Information density and phonetic structure: Explaining segmental variability

Erika Brandt

Dissertation
zur Erlangung des akademischen Grades
eines Doktors der Philosophie
an den Philosophischen Fakultäten
der Universität des Saarlandes

vorgelegt von
Erika Brandt
aus Berlin

Saarbrücken, 2019

Dekan der Philosophischen Fakultät (P): Univ.-Prof. Dr. Heinrich Schlange-Schöningen

Erstgutachter: Univ.-Prof. Dr. Bernd Möbius

Zweitgutachter: Univ.-Prof. Dr. François Pellegrino

Drittgutachter: Univ.-Prof. Dr. Ann Bradlow

Tag der letzten Prüfungsleistung: 19. Februar 2019

Acknowledgment

First of all, I would like to thank my supervisor, Prof. Bernd Möbius, for his continued guidance, patience and encouragement during my time as a Phd student. I would also like to thank Prof. François Pellegrino for agreeing so promptly to review this thesis. I am grateful that you are investing your valuable time for the review process.

During my time in Saarbrücken, I have particularly enjoyed working with my colleagues Bistra Andreeva, Iona Gessinger, Jeanin Jügler, Jürgen Trouvain, and Frank Zimmerer. Thank you, Iona, for keeping up with me while sharing an office. I am also deeply indebted to Yoonmi Oh who I have met during my first months of my Phd, and whose enthusiasm for research was infectious. I would also like to thank the student assistants in my project group, especially Zhe Wang and Kirstin Kolmorgen. Your strong work ethic was inspiring.

I would also like to thank Volker, my husband, for his continued support and encouragement. Thank your for cheering me on all this time! I am also grateful for my family who has supported me during ups and downs.

Finally, I must express my gratitude to the Deutsche Forschungsgemeinschaft (DFG) who has funded the SFB 1102 'Information Density and Linguistic Encoding' at Saarland University. I was lucky to be part of this SFB as a Phd student in Project C1 'Information Density and the Predictability of Phonetic Structure'.

To Arno

Abstract

There is growing evidence that information-theoretic principles influence linguistic structures. Regarding speech several studies have found that phonetic structures lengthen in duration and strengthen in their spectral features when they are difficult to predict from their context, whereas easily predictable phonetic structures are shortened and reduced spectrally. Most of this evidence comes from studies on American English, only some studies have shown similar tendencies in Dutch, Finnish, or Russian. In this context, the Smooth Signal Redundancy hypothesis (Aylett and Turk, 2004, 2006) emerged claiming that the effect of information-theoretic factors on the segmental structure is moderated through the prosodic structure.

In this thesis, we investigate the impact and interaction of information density and prosodic structure on segmental variability in production analyses, mainly based on German read speech, and also listeners' perception of differences in phonetic detail caused by predictability effects. Information density (ID) is defined as contextual predictability or surprisal ($S(unit_i) = -log_2 P(unit_i|context)$) and estimated from language models based on large text corpora. In addition to surprisal, we include word frequency, and prosodic factors, such as primary lexical stress, prosodic boundary, and articulation rate, as predictors of segmental variability in our statistical analysis. As acoustic-phonetic measures, we investigate segment duration and deletion, voice onset time (VOT), vowel dispersion, global spectral characteristics of vowels, dynamic formant measures and voice quality metrics. Vowel dispersion is analyzed in the context of German learners' speech and in a cross-linguistic study.

As results, we replicate previous findings of reduced segment duration (and VOT), higher likelihood to delete, and less vowel dispersion for easily predictable segments. Easily predictable German vowels have less formant change in their vowel section length (VSL), F1 slope and velocity, are less curved in their F2, and show increased breathiness values in cepstral peak prominence (smoothed) than vowels that are difficult to predict from their context. Results for word frequency show similar tendencies: German segments in high-frequency words are shorter, more likely to delete, less dispersed, and show less magnitude in formant change, less F2 curvature, as well as less harmonic richness in open quotient smoothed than German segments in low-frequency words. These effects are found even though we control for the expected and much more effective effects of stress, boundary, and speech rate. In the cross-linguistic analysis of vowel dispersion, the effect of ID is robust across almost all of the six languages and the three intended speech rates. Surprisal does not affect vowel dispersion of non-native German speakers. Surprisal and prosodic factors interact in explaining segmental variability. Especially, stress and surprisal complement each other in their positive effect on segment duration, vowel dispersion and magnitude in formant change. Regarding perception we observe that listeners are sensitive to differences in phonetic detail stemming from high and low surprisal contexts for the same lexical target.

Kurzzusammenfassung

Informationstheoretische Faktoren beeinflussen die Variabilität gesprochener Sprache. Phonetische Strukturen sind länger und zeigen erhöhte spektrale Distinktivität, wenn sie aufgrund ihres Kontextes leicht vorhersagbar sind als Strukturen, die schwer vorhersagbar sind. Die meisten Studien beruhen auf Daten aus dem amerikanischen Englisch. Nur wenige betonen die Notwendigkeit für mehr sprachliche Diversität. Als Resultat dieser Erkenntnisse haben Aylett und Turk (2004, 2006) die Smooth Signal Redundancy Hypothese aufgestellt, die besagt, dass der Effekt von Vorhersagbarkeit auf phonetische Strukturen nicht direkt, sondern nur die prosodische Struktur umgesetzt wird.

In dieser Arbeit werden der Einfluss und die Interaktion von Informationsdichte und prosodischen Strukturen auf segmentelle Variabilität im Deutschen sowie die Wahrnehmungsfähigkeit von Unterschieden im phonetischen Detail aufgrund ihrer Vorhersagbarkeit untersucht. Informationsdichte (ID) wird definiert als kontextuelle Vorhersagbarkeit oder Surprisal ($S(unit_i) = -log_2 P(unit_i|context)$). Zusätzlich zu Surprisal verwenden wir auch Wortfrequenz und prosodische Faktoren, wie primäre Wortbetonung, prosodische Grenze und Sprechgeschwindigkeit als Variablen in der statistischen Analyse. Akustisch-phonetische Maße sind Segmentlänge und -löschung, voice onset time (VOT), Vokaldispersion, globale und dynamische vokalische Eigenschaften und Stimmqualität. Vokaldispersion wird nicht nur im Deutschen, sondern auch in einer sprachübergreifenden Analyse und im Kontext von L2 untersucht.

Wir können vorherige Ergebnisse, die auf dem Amerikanischen beruhten, für das Deutsche replizieren. Reduzierte Segmentlänge und VOT, höhere Wahrscheinlichkeit der Löschung und geringere Vokaldispersion werden auch für leicht vorhersagbare Segmente im Deutschen beobachtet. Diese zeigen auch weniger Formantenbewegung, reduzierte Kurvigkeit in F2 sowie erhöhte Behauchtheitswerte als Vokale, die schwer vorhersagbar sind. Die Ergebnisse für Wortfrequenz zeigen ähnliche Tendenzen: Deutsche Segmente in hochfrequenten Wörtern sind kürzer, werden eher gelöscht, zeigen reduzierte Werte für Vokaldispersion, Formantenbewegungen und Periodizität als deutsche Segmente in Wörtern mit geringer Frequenz. Obwohl wir bekannte Effekte für Betonung, Grenze und Tempo auf segmentelle Variabilität in den Modellen beobachten, sind die Effekte von ID signifikant. Die sprachübergreifende Analyse zeigt zudem, dass diese Effekte auch robust für die meisten der untersuchten Sprachen sind und sich in allen intendierten Sprechgeschwindigkeiten zeigen. Surprisal hat allerdings keinen Einfluss auf die Vokaldispersion von Sprachlernern. Des weiteren finden wir Interaktionseffekte zwischen Surprisal und den prosodischen Faktoren. Besonders für Wortbetonung lässt sich ein stabiler positiver Interaktionseffekt mit Surprisal feststellen. In der Perzeption sind Hörer durchaus in der Lage, Unterschiede zwischen manipulierten und nicht manipulierten Stimuli zu erkennen, wenn die Manipulation lediglich im phonetischen Detail des Zielwortes aufgrund von Vorhersagbarkeit besteht.

Ausführliche Zusammenfassung

Die sprachliche Enkodierung ein und derselben Nachricht kann auf verschiedene Arten erfolgen. Diese Unterschiede zeigen sich auf unterschiedlichen linguistischen Ebenen, zum Beispiel auf der Wort-, Silben- oder auf der Phonemebene. Was genau Sprecher dazu bewegt, bestimmte Enkodierungsstrategien zu verwenden, ist eine der grundsätzlichen Forschungsfragen in der Phonetik.

Viele Faktoren beeinflussen die Variabilität in der gesprochenen Sprache. Suprasegmentelle Strukturen sind, zum Beispiel, maßgeblich an der phonetischen Ausprägung der segmentellen Ebene beteiligt (Kuzla and Ernestus, 2011; Ramus, 2002). Aber auch der phonologische Kontext spielt dabei eine entscheidende Rolle (Stevens and House, 1963; Strange and Bohn, 1998). Zudem gibt es eine wachsende Anzahl von Studien, die informationstheoretische Prinzipien verwenden, um die Variabilität von phonetischen Strukturen zu erklären. Diese Prinzipien gehen auf die Informationstheorie von Shannon (1948) zurück, die die maximale Informationsmenge quantifiziert, die während eines Kommunikationsprozesses tranferiert wird, mit besonderem Augenmerk auf Störfaktoren während dieses Prozesses. Phonetische Strukturen werden in ihrer Dauer und spektralen Eigenschaften reduziert, wenn sie aufgrund ihres Kontextes leicht vorhersagbar sind.

Es gibt viele verschiedene Maße, die verwendet werden, um die Informationsmenge während eines Kommunikationsprozesses zu quantifizieren. In dieser Dissertation wird das Maß *Surprisal* verwendet, das besonders in der Psycho- und Computerlinguistik etabliert ist. Surprisal korreliert positiv mit dem Verarbeitungsaufwand von Sprache in kognitiven Prozessen. Dieser Zusammenhang wurde für sprachliche Einheiten auf verschiedenen linguistischen Ebenen bestätigt (Demberg, Sayeed, et al., 2012; Hale, 2001; Levy, 2008, 2011). Es wird mittels folgender Gleichung 1 definiert, wobei P für Wahrscheinlichkeit und $unit_i$ für linguistische Einheit steht. *context* bezeichnet den meist vorhergehenden Kontext, da das Maß von inkrementeller Sprachproduktion und -verarbeitung ausgeht. Surprisal wird in Bits angegeben.

$$S(unit_i) = -log_2 P(unit_i|context) \tag{1}$$

Leicht vorhersagbare Kombinationen von linguistischen Einheiten haben kleinere Werte für Surprisal als Kombinationen, die schwer vorherzusagen sind (Hale, 2001; Levy, 2013). Surprisal kann als ein lokales Maß von Vorhersagbarkeit interpretiert werden, da es sich je nach lokalem Kontext ändert.

Hohe Surprisalwerte von Wörtern im Kontext werden mit erhöhtem kognitiven Aufwand assoziiert. Dieser positive Zusammenhang wurde durch Studien mittels verhaltensbezogener und neurophysiologischer Maße bestätigt. So kamen zum Beispiel Demberg and Keller (2008) zu dem Schluss, dass Surprisal ein signifikanter Prädiktor für Lesedauern im amerikanischen English ist. Wörter mit hohem Surprisal wurden langsamer gelesen als Wörter, die leichter vorhersagbar waren. Surprisal sagt

auch kognitive Anstrengung gemessen an Eye-Tracking Werten hervor (Delogu et al., 2017).

In dieser Arbeit werden Surprisalwerte basierend auf Sprachmodellen berechnet. Wir verwenden phonembasierte Sprachmodelle, da der Fokus der Arbeit auf lokalen phonetischen Strukturen liegt. Die Beziehung zwischen informationstheoretischen Prinzipien und phonetischen Strukturen wird am besten durch phonembasierte Sprachmodelle ausgedrückt, zum einen weil diese hierarchische strukturelle Informationen, zum Beispiel Silben- oder Wortgrenzen, beinhalten (Oh et al., 2015; Raymond et al., 2006). Phonembasierte Sprachmodelle werden zum Beispiel verwendet, um zu vermeiden, dass Lücken im Vokabular des Sprachmodells dazu führen, dass bestimmte bisher im Trainingskorpus ungesehene Wörter nicht im Testkorpus vorhergesagt werden können (Kneissler and Klakow, 2001).

Vorgängerstudien, die den Einfluss von Vorhersagbarkeit oder Frequenz auf die Ausprägung von phonetischen Strukturen untersucht haben, haben sich hauptsächlich auf das amerikanische Englisch fokussiert. Hochfrequente Wörter, Silben und Segmente, die leicht aufgrund ihres Kontextes vorhersagbar sind, zeigen reduzierte Dauerwerte verglichen mit Einheiten, die weniger frequent und leicht vorhersagbar sind (Aylett and Turk, 2004, 2006; Cohen Priva, 2015; Jurafsky, Bell, Gregory, et al., 2001). Auch die Dauer des Stimmeinsatzes (voice onset time (VOT)) wird durch Wortfrequenz und lokale Vorhersagbarkeit beeinflusst: Hochfrequente, leicht vorhersagbare Wörter haben kürzere VOTs als Wörter mit geringer Wortfrequenz, die aufgrund ihres Kontextes leicht vorhersagbar sind (Cohen Priva, 2017; Yao, 2009). Bezüglich der Segmentlöschung gibt es eine Reihe von Studien, die zeigen dass /t, d/ Löschung durch Wortfrequenz beeinflusst wird. Hochfrequente Wörter zeigen höhere /t, d/ Löschungsraten als Wörter mit geringer Häufigkeit (Bybee, 2002; Coetzee and Kawahara, 2013; Jurafsky, Bell, Gregory, et al., 2001). Hohe lokale Vorhersagbarkeit von Konsonanten vergrößert die Wahrscheinlichkeit einer Konsontantenlöschung (Cohen Priva, 2015). /ə/ Löschung wird auch durch informationstheoretische Faktoren beeinflusst: So wird /ə/ zum Beispiel im amerikanischen English häufiger in hochfrequenten Wörtern gelöscht als in Wörtern mit geringer Frequenz Patterson et al. (2003).

Neben diesen Studien zu Dauer und Löschung gibt es auch eine Reihe von Arbeiten, die sich mit dem Einfluss von Frequenz und Vorhersagbarkeit auf die spektralen Eigenschaften von gesprochener Sprache beschäftigt haben. Für das amerikanische Englisch wurde gezeigt, dass Vokale in hochfrequenten Wörtern (Munson and Solomon, 2004; Munson, 2007; Pierrehumbert, 2000; Scarborough, 2006; Wright, 2004) und leicht vorhersagbaren Kontexten (Aylett and Turk, 2006; Clopper and Pierrehumbert, 2008; Jurafsky, Bell, Gregory, et al., 2001) mehr zentralisiert und damit weniger distinktiv in ihren spektralen Eigenschaften sind als Vokale in Wörtern mit geringer Frequenz und in Kontexten, die schwer vorhersagbar sind. Die Distinktivität von Vokalen wird als Dispersion eines Vokales in Relation zum Mittelpunkt des Vokalraumes des jeweiligen Sprechers definiert. Vokale, die eine große Distanz

zum Mittelpunkt des Vokalraumes aufweisen, werden in Perzeptionsexperimenten als leichter verständlich wahrgenommen (Bradlow et al., 1996).

Suprasegmentelle Faktoren, wie zum Beispiel Wortbetonung, prosodische Grenze oder Sprechgeschwindigkeit, haben einen immensen Einfluss auf die phonetische Ausprägung von Segmenten und anderen phonetischen Strukturen. Die Smooth Signal Redundancy Theorie (Aylett and Turk, 2004, 2006) betont den Zusammenhang zwischen prosodischen und informationstheoretischen Faktoren in ihrem jeweiligen Einfluss auf akustische Redundanz. Sie besagt, dass Vorhersagbarkeit nur mittels prosodischer Strukturen auf die akustische Ausprägung von phonetischen Strukturen wirkt. In ihren eigenen Untersuchungen schwächen die Autoren diese Abhängigkeit bereits ab, da sie eigenständige Effekte von Vorhersagbarkeit auf Vokaldauer und Formantfrequenzen F1 und F2 beobachten, die nicht durch prosodische Faktoren gemittelt werden. In dieser Arbeit verwenden wir sowohl informationstheoretische als auch prosodische Faktoren, um segmentelle Variabilität zu untersuchen. Wir stützen uns dabei direkt auf die Smooth Signal Redundancy Theorie und untersuchen insbesondere Interaktionen zwischen Surprisal und prosodischen Faktoren.

Hörer nehmen geringe Unterschiede in phonetischem Detail besser wahr, wenn diese in Wörtern vorkommen, die schwer vorhersagbar sind verglichen mit Wörtern in vorhersagbaren Kontexten (z. B., Beaver et al., 2007; Lieberman, 1963; Manker, 2017). Vorhersagbarkeit und Wortfrequenz spielen auch eine Rolle bei der Verständlichkeit von Wörtern in einem gestörten Kommunikationskanal (Kalikow et al., 1977; Luce and Pisoni, 1998; Savin, 1963).

Zum einen repliziert diese Dissertation Ergebnisse bezüglich Segmentdauer und -löschung, VOT sowie Vokaldispersion, die zum großen Teil auf englischen Daten beruhen, für das Deutsche. Zum anderen erweitert diese Arbeit die Bandbreite der akustisch-phonetischen Untersuchungen, indem auch dynamische Formanttrajektorien, globale spektrale Eigenschaften von Vokalen sowie Stimmqualität im Kontext von Vorhersagbarkeit und Prosodie untersucht werden. Der Einfluss von Vorhersagbarkeit und Prosodie auf Vokaldispersion wird nicht nur in Deutsch, sondern auch in einer sprachübergreifenden Studie in sechs Sprachen (amerikanisches Englisch, Deutsch, Finnisch, Französisch, Polnisch und Tschechisch) untersucht.

In einer Pilotstudie wird zudem für das Deutsche untersucht, ob informationstheoretische Faktoren der Zielsprache die phonetischen Ausprägungen von Sprachlernern erklären können, auch in Bezug auf unterschiedliche Level von Sprachkompetenz. Neben diesen Produktionsstudien untersucht diese Arbeit auch in einem Perzeptionstest, die Wahrnehmung von Hörern von gebrochenen Erwartungen bezüglich phonetischer Ausprägung von Wörtern aufgrund ihrer Vorhersagbarkeit.

Methode Die meisten Produktionsanalysen basieren auf dem deutschen Siemens Synthesis Korpus (Schiel, 1997), das gelesene Zeitungstexte aus dem Frankfurter Allgemeinen Korpus von zwei professionellen Sprechern beinhaltet. Neben den Sprachaufnahmen gibt es auch elektroglottographische Signale. Für die sprachübergreifende

Analyse von Vokaldispersion wurde ein Teil des BonnTempo Korpus (Dellwo et al., 2004) verwendet. Je sechs Sprecher und Sprecherinnen des amerikanischen Englisch, Deutschen, Finnischen, Französischen, Polnischen und des Tschechischen wurden ausgewählt. Die Analyse von globalen spektralen Eigenschaften von deutschen Vokalen in unterschiedlichen Kontexten beruht auf dem PhonDat2 Korpus (PHONDAT2 – PD2, 1995). Für die Studie von Vokaldispersion bei Sprachlernen haben wir Passagen aus dem EUROM-1 corpus (Chan et al., 1995) mit Muttersprachlern, Sprachanfängern und fortgeschrittenen Lernern aufgenommen.

Für jede der untersuchten Sprachen wurde ein Sprachmodell berechnet, das auf Korpora mit geschriebener Sprache beruht. Für das Deutsche wurden zwei Modelle berechnet: ein Sprachmodell basierend auf dem Zeitungskorpus Frankfurter Rundschau (Elsnet, 1992 – 1993) und eines, das auf einem webbasierten Korpus beruht, Stuttgart German Web-as-Corpus (SDeWaC) (Baroni and Kilgarriff, 2006). Das Sprachmodell für das amerikanische English wurde basierend auf dem Corpus of Contemporary American English (COCA) (Davies, 2008-) trainiert. Für Polnisch und Tschechisch wurden webbasierte Frequenzlisten zur Sprachmodulierung verwendet (Zséder et al., 2012). Das finnische Sprachmodell beruhte auf dem Finnish Parole Korpus (Department of General Linguistics, 1996–1998), während für das französische Lexique (New et al., 2001) verwendet wurde.

Alle Sprachdaten wurden automatisch segmentiert und annotiert. Segmentgrenzen und -labels wurden von Annotatoren mit Erfahrung in phonetischer Analyse verifiziert. Die Reliabilität der Verifizierung zwischen Annotatoren wurde als hoch eingestuft. Segmentgrenzen der beiden Annotatoren des Siemens Synthesis Korpus zeigten hohe Übereinstimmungen ($\rho = 0.93$).

Die akustisch-phonetischen Analysen wurden mittels verschiedener Tools durchgeführt: Praat (Boersma and Weenink, 2017), CPPS Tool (Hillenbrand, Cleveland, et al., 1994; Hillenbrand and Houde, 1996) für die kepstrale Stimmqualitätsanalyse, das Matlab-Skript peakdet.m (Michaud, 2007) für die Analyse des elektroglottographischen Signals und SPTK (Kobayashi et al., 2017) für die Berechnung von spektraler und temporaler Distanz zwischen zwei Signalen. Die Sprachmodelle wurden mit Hilfe von SRILM (Stolcke, 2002) oder Perl-Skripten berechnet.

Eine explorative Analyse der Korrelation zwischen Vokaldispersion und Surprisal basierend auf verschiedenen Kontextgrößen eröffnete, dass kleine Kontextgrößen von einem (Biphon) oder zwei Phonemen (Triphon) die besten Korrelationswerte zwischen den Variablen zeigen. Aus diesem Grund wird Surprisal in dieser Arbeit auf Biphon- oder Triphon-Kontexten berechnet.

Die statistische Analyse beruht auf gemischten Modellen, die mit R (R Development Core Team, 2008) berechnet wurden. In der Regel beinhalten die Modelle die informationstheoretischen Variablen Wortfrequenz und Surprisal sowie die prosodischen Faktoren Wortbetonung, prosodische Grenze und Sprechgeschwindigkeit. Je nach Phänomen werden auch andere Kontrollfaktoren in die Auswertung einbezogen, zum Beispiel Wortklasse, durchschnittliche Lautdauer oder phonologischer Kontext.

Ergebnisse Im Folgenden werden die Resultate der Dissertation nach den Hauptfaktoren in den statistischen Modellen dargestellt.

Surprisal Leicht vorhersagbare deutsche Laute zeigen kürzere Dauern (auch in ihrer VOT) und eine stärkere Tendenz zur Löschung als Segmente, die schwer vorhersagbar sind. Vokale in niedrigem Surprisalkontext sind weniger distinktiv, weisen weniger dynamische Formantenveränderungen auf und zeigen erhöhte Behauchung verglichen mit deutschen Vokalen in hohem Surprisalkontext.

Wortfrequenz Deutsche Laute in hochfrequenten Worten sind kürzer und neigen eher zur Löschung als Laute in Wörtern mit niedriger Frequenz. Deutsche Vokale in hochfrequenten Worten zeigen weniger Vokaldispersion, verringerte Dynamik in ihren Formantbewegungen und weniger Periodizität als Vokale in Wörtern mit geringer Frequenz.

Primäre Wortbetonung Laute in betonten Silben wurden mit längerer Dauer produziert, auch bezüglich ihrer VOT; sie waren weniger anfällig gegenüber Segmentlöschung, zeigen erhöhte Vokaldispersion und Formantenbewegung (vector length (VL), vowel section length (VSL), F1/F2 slope, F1 velocity, F2 DCT2) sowie längere Öffnungsphasen der Stimmlippen und ausgeprägtere Amplituden der Stimmlippenöffnung im elektroglottographischen Signal als unbetonte Laute.

Prosodische Grenze Wenn ein Segment direkt vor einer prosodischen Wort- oder Phrasengrenze steht, führt dies zu einem Effekt von prosodischer Längung in der Dauer (Byrd, 2000; Wheeldon and Lahiri, 1997). Dieser wurde auch in dieser Arbeit für das Deutsche beobachtet. Zudem waren Laute an prosodischen Grenze weniger anfällig für Segmentlöschung und sie zeigen erhöhte Werte in den Formantenbewegungen in den Maßen VL, F1/F2 slope und F1 velocity. Auf der anderen Seite führte das Aufkommen einer prosodischen Grenze aber auch zu weniger Vokaldispersion, geringeren Werte in der Formantdynamik für die Maße VSL und F2 DCT2 sowie verringerter Periodizität gemessen an den kepstralen und elektroglottographischen Maßen.

Sprechgeschwindigkeit Eine erhöhte Sprechgeschwindkeit führte zu reduzierten Dauerwerten (auch für VOT), höheren Löschungsraten, verringerter Vokaldispersion und Formantbewegung in den Maßen VL, VSL, F1 slope und velocity sowie kürzeren Öffnungsphasen der Stimmlippen.

Wie bereits erwartet, waren informationstheoretische Prädiktoren meist weniger effektiv als die prosodischen Variablen in den Modellen segmenteller Variabilität (Aylett and Turk, 2006). Zudem haben wir beobachtet, dass Interaktionen zwischen Surprisal und den prosodischen Faktoren häufig die Modellperformanz verbesserten.

Die Richtung des Effekts entsprach der des Haupteffekts des prosodischen Faktors. So zeigten zum Beispiel deutsche betonte Vokale in hohem Surprisalkontext ausgeprägtere Dispersion als unbetonte Vokal in niedrigem Surprisalkontext.

Die Analyse spektralerer Distanz zwischen den gleichen Vokalphonemen in unterschiedlichen Kontexten ergab, dass sowohl Surprisal als auch korpusspezifische Silbenfrequenz signifikante Prädiktoren waren. Allerdings zeigte nur Silbenfrequenz die erwarteten Tendenzen: Vokale waren sich ähnlicher in ihren globalen spektralen Eigenschaften, wenn sie in demselben Kontext standen, während Vokale in unterschiedlichen Kontexten größere spektrale Distanzen aufwiesen.

Je höher der Surprisalwert des Biphons eines Vokals, desto höher war seine Vokaldispersion. Zu diesem Ergebnis kam diese Arbeit auch in der sprachübergreifenden Analyse von Vokaldispersion. Dieser positive Zusammenhang zwischen Surprisal und Vokaldispersion war unabhängig von Sprechgeschwindigkeit. Allerdings zeigte Finnisch keinen Zusammenhang zwischen den beiden Variablen, was auf die sehr schwach ausgeprägte spektrale Reduktion von Vokalen in dieser Sprache zurückzuführen war (Bertram et al., 2004).

Obwohl wir in der Pilotstudie zu Vokaldispersion bei deutschen Spracherlernen mit bulgarischer Muttersprache keinen signifikanten Effekt von Surprisal gefunden haben, weder für die deutschen Muttersprachler noch in den Modellen für die Sprachlerner, können wir zusammenfassen, dass fortgeschrittene Sprachlerner mehr dazu neigten zielsprachliche Muster mit Bezug auf die Faktoren Vokalgespanntheit, Wortklasse und Durchschnittsdauer in ihrer Vokaldispersion zu produzieren. Sprachlerner mit mittlerem Kompetenzlevel zeigten diese erwarteten Muster in Vokaldispersion nicht.

In dem Perzeptionsexperiment in dieser Dissertation haben wir Hörer gefragt, ob sie eine manipulierte Aufnahme oder die Originalaufnahme aus dem Siemens Synthesis Korpus als natürlicher einschätzen. Die manipulierten Aufnahmen enthielten Zielwörter, die in einem Kontext mit höherem oder niedrigerem Surprisal geäußert wurden als in der Originalaufnahme. Die Hörer waren in der Lage, die weniger natürlichen Aufnahmen korrekt zu identifizieren, aber nur wenn diese in der Reihenfolge Manipulation – Original präsentiert wurden. Diskriminationsaufgaben zeigen häufig so eine Art von Reihenfolgeeffekt (Schiefer and Batliner, 1991; Wherry, 1938; Wickelmaier and Choisel, 2006).

Zusammenfassung In dieser Arbeit konnten wir vorherige Ergebnisse bezüglich des Effektes von Vorhersagbarkeit auf Segmentdauer, Segmentlöschung und Vokaldispersion basierend auf dem amerikanischen Englisch replizieren und auf das Deutsche übertragen. Zusätzlich zu den bereits etablierten Zusammenhängen zwischen segmenteller Variabilität und Vorhersagbarkeit beinhaltete diese Arbeit auch Ergebnisse zu dynamischen Formanttrajektorien, globalen spektralen Eigenschaften und Stimmqualität. In einer Pilotstudie zu Vokaldispersion von deutschen Sprachlernen haben wir gezeigt, dass Vorhersagbarkeit neben anderen Faktoren nützlich darin sein kann, Kompetenzlevel von Sprachlernen einzuschätzen. Zudem beinhaltete diese Arbeit

auch einen sprachübergreifenden Ansatz bezüglich der Analyse von segmenteller Variabilität und Vorhersagbarkeit: Vokaldispersion ließ sich auch hier zum Teil durch die Vorhersagbarkeit der Vokale in den einzelnen Sprachen erklären. Neben den Produktionsanalysen zeigte diese Arbeit auch, dass Hörer sensibel gegenüber Unterschieden im phonetischen Detail aufgrund von verschiedenen Surprisalkontexten sind.

Contents

Part I

Introduction and Background

Chapter 1

Introduction

> *There exists an equilibrium between the magnitude or degree of complexity of a phoneme and the relative frequency of its occurrence, in the sense that magnitude or degree of complexity of a phoneme bears an inverse relationship to the relative occurrence of its frequency.*
>
> – Zipf (1935, p. 49)

1.1 Motivation and research aims

Speakers choose different strategies of linguistic encoding to convey the same message. These encoding strategies are apparent at different linguistic levels. One of the main research aims in the field of phonetics is to investigate different sources of linguistic variability in the speech signal.

In recent years, information-theoretic principles have been used to gain insight into frequency or predictability effects on linguistic variability (e. g., Jaeger, 2010; Jurafsky, Bell, Gregory, et al., 2001; Levy and Jaeger, 2007). Studies have found that linguistic units are prone to reduction or deletion when they are easily predictable from their context. These findings hold at the discourse, sentence, word and syllable level. So far, only little attention has been payed to the segmental level. This thesis aims to fill this gap. It focuses on the relationship between information-theoretic factors, such as predictability or frequency, on the phonetic structure at the segmental level.

Vocalic formant trajectories contain information about the place of articulation of preceding and following consonants (Delattre et al., 1955). Therefore, one can observe the following phenomenon: when all consonants are cut out in a stream of continuous speech, listeners are still able to decode the message in the signal. However, if all the vowels are discarded from the signal, listeners cannot retrieve the message (Trouvain, 2004). This well-known phenomenon motivated to put the focus of this thesis on

vowel characteristics. We investigated vowel dispersion, dynamic formant trajectories, spectral similarity of vowels and voice quality in the context of information-theoretic factors. We assumed that an information-theoretic account of reduction and expansion will be most pronounced in vocalic characteristic, based on the observation that they carry more information in speech perception than consonants.

Furthermore, we replicated well-established relations between temporal features of speech and information-theoretic variables and confirmed these findings for German. These analyses were performed on all segments, and not restricted to vowels. Also, the analysis of vowel dispersion in German and other languages as a function of predictability and frequency conducted here built on previous analyses in the field (e. g., Aylett and Turk, 2006; Gahl et al., 2012; Jurafsky, Bell, Gregory, et al., 2001). We widened the scope of these analyses by investigating vowel dispersion in second language (L2) speech. As vowel dispersion is based on a snap shot of the vocalic spectral characteristics at a specific time point, we broadened this limited view on vocalic features by studying dynamic formant trajectories from an information-theoretic view point. This thesis also offers a first insight into the relation between information-theoretic factors and variability in voice quality.

One of the most influential theories that has evolved in the context of information theory and linguistic variability is the Smooth Signal Redundancy (SSR) hypothesis (Aylett and Turk, 2004, 2006). This hypothesis states that there is an inverse relationship between predictability and acoustic redundancy which is moderated by prosodic structure. In this theory, prosody is attributed a key role in the implementation of predictability and frequency effects on the phonetic structure at the segment level. This thesis aims at shedding light on this relationship by including prosodic factors in the statistical modeling procedure in all production analyses with a special focus on testing interactions between predictability and prosodic factors.

Previous research has found that listeners are more sensitive to differences in phonetic detail when they occur in unpredictable contexts (e. g., Beaver et al., 2007; Lieberman, 1963; Manker, 2017), and that they prefer speech synthesis systems which implement predictability in their prosodic model (Le Maguer et al., 2016). Building on these findings and the results from our production analyses we conducted a perception experiment which tested listeners' sensitivity towards violated expectations of phonetic encoding of words due to high or low predictability. To that means, we presented listeners with words that were originally produced in an easily predictable context within a context from which the word was difficult to predict, and vice versa. We contrasted these manipulated utterances with their baselines and asked listeners for their judgment of naturalness. This perception test allowed us to make inferences about listeners' ability to perceive fine differences in phonetic detail caused by predictability when the target words are presented within context.

The main contributions of this thesis were to replicate previous findings of the impact of information-theoretic factors on American English segment duration, deletion and vowel dispersion for German. In addition, we included dynamic formant

measures, global vocalic characteristics and voice quality metrics for German which have not yet been investigated in the context of predictability. We observed that vowel formants moved less when they were easily predictable from the context. Cepstral measures of voice quality indicated decreased harmonic richness and periodicity for easily predictable vowels. On the perception side, we found that listeners were sensitive towards violated predictability expectations in phonetic detail. Additionally, we gathered some first insight into the usefulness of predictability in explaining variability in learners' speech and making inferences about their competence level.

1.2 Structure of the thesis

First, we present the scientific background (Chapter 2) for the own contributions of this thesis presenting a short introduction to the field of information theory and its applications in linguistics with a special focus on acoustic-phonetic studies of duration and deletion, spectral features of vowels, and voice quality. We also introduce few key studies that investigated the impact of predictability and frequency on speech perception at the word level. In addition, we stress the relationship and interaction between prosodic factors and predictability by summarizing key findings from the literature. The chapter concludes with the hypotheses for the entire thesis.

The methodology chapter contains materials and data analysis steps that were universally used in this thesis. Specific information about the calculation of different acoustic-phonetic measures is given in the respective section in the results (Part III). We introduce speech corpora used for acoustic-phonetic analyses and stimuli building, as well as text corpora for language modeling. In addition, we outline the overall preprocessing procedure for all speech corpora analyzed in this thesis, and give a summary of the language modeling procedure including preprocessing of the text corpora and language model (LM) calculation. For the German production analyses, we included predictability values from the same model (Section 4.2.1), while for the cross-linguistic study separate LMs for the different languages were built (Section 4.2.2). The choice of n-phone order for these models is discussed in Section 4.2.3. This chapter is rounded off by remarks about the statistical modeling procedure universally used in this thesis and an introduction of control factors used in the statistical models built for each analysis. Control factors which were specific to certain analyses are introduced in the respective section in the results. The following Part III consists of explanations of data analysis and outlines descriptive and inferential results and their discussion for the acoustic-phonetic analyses of segment duration and deletion, voice onset time (VOT), vowel dispersion, dynamic formant trajectories and spectral similarity of vowels, as well as voice quality measured on vowels. All of these analyses were conducted on German data. In addition, the Chapter 8 also contains a cross-linguistic analysis of vowel dispersion in six languages from different language families and an analysis of vowel dispersion of Bulgarian L2 speakers of German. The

last section of this chapter contains the experimental design and the results as well as their discussion of a perception experiment which investigated listeners' sensitivity to violated predictability expectations. Here, listeners rated the naturalness of cross-spliced phrases containing lexical targets which were easily predictable from the context but originally produced under low predictability, and vice versa. We finish the thesis with a general discussion of the results of the production analyses and the perception experiment in the light of the current state of the field.

1.3 Preliminary remarks

Some of the acoustic-phonetic production analyses presented in this thesis were based on abstracts, conference proceedings, and the author's contribution to a journal article (Malisz et al., 2018). These studies were revised and extended before writing them up for this dissertation.

Two analyses were presented at meetings of German-speaking phoneticians and phonologists as one-page abstracts: the study on the effect of information-theoretic factors on German segment deletion rates (Brandt et al., 2017a) and on voice quality in German male speakers (Brandt, Andreeva, et al., 2018).

The vowel dispersion analysis in six languages (Section 8.3) is based on revised and extended works presented in Schulz et al. (2016) and Malisz et al. (2018). In the proceedings paper, we did not include American English (AE) in the analysis (Schulz et al., 2016). The simple binary prosodic boundary model used in Malisz et al. (2018) was revised for this thesis to create coherence between the different analyses.

We revised and extended the analysis of dynamic formant trajectories presented in Brandt et al. (2018). This proceedings paper only introduced simple vowel inherent spectral change (VISC) measures based on onset and offset of the vowel. In Section 9, however, we included a variety of other measures of formant change and parametric measures. Also, all analyses of dynamic formant trajectories were based on monophthongs and diphthongs, while we excluded diphthongs in Brandt et al. (2018).

Section 10 is based on a revised and extended conference paper (Brandt et al., 2017b). The main revision of this study concerned the random structure of the linear mixed-effects model (LMM) built for this analysis. We added random effects for word identity, as well as following and preceding phonological context in the work presented in this thesis.

The introduction highlighted how this thesis relates to current research on linguistic variability and information-theory, which research goal and specific hypotheses have driven this thesis, and which main contributions can be derived from its results. The following chapter presents the concept of information theory (Shannon, 1948) from a linguistic point of view. The presentation of previous findings focuses on acoustic-phonetic studies.

Chapter 2

Background

The concepts of information theory (Shannon, 1948) have been adapted in many scientific fields, among those in linguistics. This thesis contributes to a diverse range of linguistic studies investigating variability as a function of predictability. In order to show how this thesis relates to and differs from previous studies we therefore firstly introduce its information-theoretic background, and prior research in the field. In addition, theoretical concepts, information density (ID) and acoustic-phonetic metrics are introduced. We conclude this section by giving our hypotheses for this thesis.

2.1 Information-theoretic background

The concept of ID expresses the ratio between information content and a unit of time or an amount of linguistic material. The information content of a linguistic unit is traditionally determined by its semantics. In recent years, the psycholinguistic and computational linguistic fields established a different notion of information content based on information theory by Shannon (1948) (e. g., Crocker et al., 2016; Demberg, Sayeed, et al., 2012; Levy, 2008).

In this thesis, we use the term ID to refer to linguistic complexity metrics, such as *surprisal* (Equation 2.1), *unigram word probability* or *phoneme probability*, and *word frequency*. In our statistical models, we subsume the complexity measures under the umbrella term ID factors. For practical reasons, we chose to use ID in this non-standard way.

The main aim of information theory is to maximize the amount of information that is transferred through a channel in communication processes, with specific focus on noisy conditions. Information theory encompasses any type of information passed through any type of communication channel. Spoken language is regarded as a random process with an irreducible complexity below which the signal cannot be compressed (Shannon, 1948).

There are several measures to quantify the amount of information conveyed in a message (Hale, 2016). In this thesis, we use the concept of *surprisal* ($S(unit_i)$) because it is relevant for human processing difficulty of linguistic units at different levels (Demberg, Sayeed, et al., 2012; Hale, 2001; Levy, 2008, 2011). It is defined by the following Equation 2.1, while P stands for probability, $unit_i$ denotes the linguistic unit under investigation, and *context* is usually preceding context of the linguistic unit as the metric presupposes incremental language production and processing. Surprisal is expressed in bits of information by the following Equation 2.1:

$$S(unit_i) = -log_2 P(unit_i|context) \tag{2.1}$$

Combinations of linguistic units which are difficult to predict are more suprising when they occur, and vice versa, the occurrence of easily predictable units is less surprising. Surprisal is used to explain the surprise of local structures. The surprisal of independent outcomes is additive. Learning of two independent linguistic events is as if learning of each separately (Hale, 2001; Levy, 2013).

Regarding the word level, surprisal correlates positively with cognitive effort in comprehension tasks using behavioral and neurophysiological measures. In a corpus study of reading times in American English, Demberg and Keller (2008) found that surprisal was a significant predictor for the syntactic processing complexity of arbitrary words. In contrast to integration cost, the effect of surprisal was not limited to certain lexical classes. Later, Smith and Levy (2013) aimed at quantifying the relationship between conditional word probability and reading times in American English. They found that reading times increased logarithmically over six orders of magnitude in estimated word probabilities. Probability even explained differences in reading times between highly unpredictable words.

In eye-tracking and event-related potentials (ERP) studies, surprisal also predicts cognitive effort. Delogu et al. (2017) attempted to tease apart surprisal-based and construction-specific accounts for differences in comprehension of coercion in American English (e. g., *began the book*) and control expressions (e. g., *read the book*). Coercion is defined as a mismatch between the semantic features of a selector (e. g., began) and the semantic properties of the selected lexical item (e. g., book) (Lauwers and Willems, 2011). In addition to the control condition with low surprisal of the complement noun, Delogu et al. (2017) used a third condition with similar surprisal of the complement noun as in the coercion condition. The authors expected to find the lowest processing effort in the low surprisal control condition, similar processing efforts in the coercion and matched-surprisal condition with additional difficulty elicited by the coercion condition due to construction-specific reasons. Processing difficulty was assessed using eye-tracking and ERP. Analysis of the eye-tracking study revealed that there was a significant effect of surprisal in the expected direction, as well as a marginal effect of coercion on late reading measures. The ERP study only supported the surprisal-based theory without showing additional effects of the construction-specific account for coercion.

Surprisal values are estimated from LMs based on large text corpora. In this thesis, we mainly used phoneme-based LMs since the focus of the production studies was on local phonetic structures on the phoneme level. The relationship between ID and phonetic structures is thought to be best reflected by n-phone LMs (Oh et al., 2015). Hierarchical structural information, such as syllable or word boundaries, which affect segmental properties are reflected in sequences of phones, especially if these sequences explicitly include word boundaries (Raymond et al., 2006).

There have been attempts to improve LM performance by dissecting words into sub-word units on the basis of morphology (e. g., Kneissler and Klakow, 2001), but only few studies have focused on phoneme-based LMs. The motivation of LM research at the sub-lexical level is to make predictions abouts out of vocabulary (OOV) words. The basic assumption is that a LM that makes prediction about upcoming events based on sub-lexical units can avoid the OOV problem provided that unseen events in the vocabulary are constructed using sub-lexical units that are part of the LM.

Ng and Zue (1997) investigated the performance of sub-word LMs for spoken document retrieval in American English using several different phoneme-based units for their modeling procedure. The following list exemplifies these units with different representations of the lexeme $< weather >$:

- Single phonemes: /w ɛ ð ɚ/
- Overlapping clusters of three phonemes: /wɛð ɛðɚ/
- Broad phonemic classes: /liquid frontvowel voicefric retroflexfrontvowel/
- Phone multigrams: /wɛðɚ/
- Phonemes organized in syllables: /wɛ ðɚ/

They found that overlapping sub-word units were the most effective in retrieving speech messages, and that their performance equaled that of the baseline text-based word-units. Introducing errors into the transcription led to a decrease in performance, especially for those LMs that were built with (non-overlapping) syllables or multigrams, while overlapping units were more robust to variation caused by transcription errors.

LMs are language-specific. Phoneme- or syllable-based LMs have been computed for several languages, such as Polish (Kłosowski, 2017; Ziółko and Gałka, 2010), American English (Ng and Zue, 1997; Schrumpf et al., 2005; Yannakoudakis and Hutton, 1992), or German (Larson and Eickeler, 2003). In a comparative study on linguistic complexity in 18 languages from 10 different language families, Oh (2015) found that the average amount of information per second was relatively stable within these languages. Variation in ID within a language was compensated by average speech rate. This means, languages with high density of information per linguistic unit had slower average speech rates than languages with less dense ID structure

which were, in return, produced with increased average speech rate. The study also observed a negative relationship in 14 of these languages between phonological and morphological complexity concluding that languages are organized with a trade-off in complexity between these linguistic levels. This typological analysis of universal tendencies across languages motivated the study introduced in Section 8.3.

LMs are not only language-specific, but they also depend immensely on the corpus that they are trained on. Depending on the text corpus LMs show a bias to overestimate or underestimate certain linguistic phenomena, and predictions made about human processing efforts. Fine et al. (2014) tested how well frequency estimates from American English Google n-gram (written) (Brants and Franz, 2006), the Switchboard corpus (Godfrey et al., 1992), CELEX (spoken and written) (Baayen, Piepenbrock, et al., 1995), and the British National Corpus (spoken and written) (BNC Consortium, 2007) predict reaction times in lexical decision tasks, word naming, and in picture naming. They found that Google n-gram had a strong bias towards the occurrence of lexical items from technology and adult topics, while Switchboard, for instance, allocated high frequency counts to colloquialisms and back-channel expressions. They concluded that the nature of a corpus (spoken vs. written) and its register influenced the lexical frequency counts that are used as predictors in psycholinguistic studies.

However, these considerations mainly apply to studies on the word level. At the sub-word level, more specifically at the phoneme-level, the vocabulary of the LM is considerably smaller than at the word level. For instance, the German LM used in this thesis that was based on Stuttgart German Web-as-Corpus (SDeWaC) contained 45 different phonemes. Domain-specific characteristics of the LM corpus are thus less likely to affect the frequency counts or surprisal values derived from this LM. In fact, we used a smaller German LM corpus in a first analysis of vowel dispersion in different languages (Schulz et al., 2016), German WebCELEX (4.5 m lexical tokens) (Max Planck Institute for Psycholinguistics, 2001), and then updated the surprisal values using a much larger German text corpus, Frankfurter Rundschau (41 m lexical tokens) (Elsnet, 1992 – 1993), in order to test the reliability of our results. Both corpora did not only differ considerably in size, but also in their domain. While Frankfurter Rundschau is a newspaper corpus, WebCELEX is a web-based text source. However, Pearson's correlation values between German vowel dispersion and biphone surprisal of the preceding context of WebCelex ($r = 0.36$) and Frankfurter Rundschau ($r = 0.30$) differed only slightly.

In order to test how well a LM predicts linguistic structures and elements it is tested on test data. *Cross-entropy* ($H_{P(T)}$) (Equation 2.2) gives a measure for LM evaluation. It estimates how many bits are needed to encode the test set (T) with a certain length (W_T) relying on the LM that was trained on a different set of linguistic data, while P stands for probability (S. F. Chen and Goodman, 1996). This means that cross-entropy gives the average surprise of a LM. Lower values for cross-entropy equal better performance in LM application. Usually, the value for cross-entropy

drops the higher the order of the model gets.

$$H_{P(T)} = -1/W_T log_2 P(T) \tag{2.2}$$

Cross-entropy is related to *perplexity* ($PP_{P(T)}$) by the following Equation 2.3:

$$PP_{P(T)} = 2^{H_{P(T)}}. \tag{2.3}$$

It gives the reciprocal of the average probability (P) that is assigned by the LM to each word in the test set. LM performance is usually reported by referring to model perplexity. Clearly, the lower the perplexity, the better the LM (Manning and Schütze, 1999).

2.2 Information density and phonetic structure

Several studies have focused on the impact of ID factors on phonetic and phonological phenomena. This line of research has resulted in information-based theories of language production that share similar ideas: the Probabilistic Reduction Hypothesis (Jurafsky, Bell, Gregory, et al., 2001), the SSR hypothesis (Aylett and Turk, 2004, 2006), the informational redundancy hypothesis (Pluymaekers et al., 2005a), and the Uniform Information Density (UID) theory (Jaeger, 2010; Levy and Jaeger, 2007).

The *Probabilistic Reduction Hypothesis* claims that words are reduced when they have high word probability. The concept of word probability goes beyond simple word frequency or predictability, and also encompasses probability based on neighboring words, syntactic and lexical structure, as well as semantic and discourse structure. Thus, word probability is estimated based on the probability of occurrence based on different linguistic levels (Jurafsky, Bell, Gregory, et al., 2001).

The *Smooth Signal Redundancy (SSR) hypothesis* posits that predictability does not have a direct effect on surface phonetics, but rather its effects are mediated and implemented through prosodic structure. Prosodic prominence structure explains acoustic redundancy to a large extent, while it coincides with unpredictable sections of speech. The SSR claims that there is an inverse complementary relationship between acoustic and language redundancy (i. e., predictability) leading to an equal recognition likelihood of each element in the speech signal (Aylett and Turk, 2004, 2006).

Similarly, the *informational redundancy hypothesis* (Pluymaekers et al., 2005a) claims that linguistic units that carry little information are produced with less articulatory effort than informative units. How informative these units are is defined on different dimensions, such as word frequency, contextual predictability, previous mention, or syntactic probability. These factors influence linguistic encoding independently and additively.

The *Uniform Information Density (UID)* hypothesis argues that speakers as rational beings structure their utterances optimal with regard to ID, i. e., the amount of

information conveyed per unit of the utterance. Speakers are thought to avoid peaks and troughs in their ID profile when producing an utterance. Information is, thus, spread uniformly across the utterance (Jaeger, 2010; Levy and Jaeger, 2007).

2.2.1 Segment duration

According to Zipf's law (1949) frequently used linguistic units are under greater pressure to be efficient than less frequent elements. For instance, this is why frequently used words get shortened over time, such as the morphological reduction of "Automobil" to "Auto", or the syntactical reduction of "in das" to "ins" in Standard German. More recent cross-linguistic studies have found that it is not frequency, but predictability which is more efficient in explaining variability in word length (Piantadosi et al., 2011). Easily predictable words and sub-lexical units, i. e., syllables or phonemes, are produced with shorter average duration than words and sub-lexical units which are difficult to predict.

Jurafsky, Bell, Gregory, et al. (2001) investigated whether function and content word durations in American English are predicted by ID factors, such as conditional probability or word frequency. They performed two separate analyses on subsets of the Switchboard corpus (Godfrey et al., 1992) showing that there was a tendency for different ID factors to explain function and content word durations respectively. Preceding and following word bigram, as well as conditional probability of both surrounding words predicted function word duration of the 10 most frequent function words in the Switchboard corpus. The duration of content words ending in /t/ or /d/ was also significantly affected by preceding and following word bigram probability, as well as by word frequency of the target.

Later, this line of research was continued by Bell et al. (2009). Lexical class, word frequency, and both preceding and following predictability were identified as significant predictors of word duration in American English conversational speech (Switchboard corpus (Godfrey et al., 1992)). However, when the data set was split based on lexical class Bell et al. (2009) observed that content and function words behaved differently with regard to ID factors, while controlling for prosodic factors (speech rate, accent, position within intonational phrase) and word form (combination of average word length, number of segments and syllables). Following conditional predictability and word frequency explained most of the variation in content word duration, while previous conditional predictability and repetition did not show a significant effect. Excluding high-frequency homonyms from the sample led to a significant result for the impact of repetition on pronounced duration of content words. Function words were analyzed treating high-frequency separately from mid/low-frequency items since high-frequency function words constituted a separate peak in the frequency distribution. Duration of high-frequency function words was mainly impacted by the significant factor previous conditional probability and moderately by word frequency, while for mid/low-frequency function words following conditional probability reached

significance level.

In addition, listeners seem to attend differently to changes in phonetic structure depending on lexical class. Listeners performed significantly better at detecting production errors made in English content words than in function words which was interpreted a being conditioned by differences in their syntactic predictability (Manker, 2017). On average, content words had lower syntactic predictability than function words.

Production studies on the relationship between ID and duration following Bell et al. (2009) usually either included the factor lexical class in their statistical analysis or focused only on one lexical class, or even only on one lexeme. For instance, Tily et al. (2009) investigated whether syntactic probabilities affect the pronounced duration of *to* in dative alternations in spontaneous speech of the Switchboard corpus (Godfrey et al., 1992). Following word bigram was a significant predictor of *to* duration, while controlling for syntactic probability conditioned by the verb, and speech rate. The authors concluded that the probability of syntactic choices in spontaneous speech was reflected in speakers' utterances. This finding for *to* was generalized to other words in NP NP constructions. However, the observed effect of syntactic probability on the duration of *to* was small, and there was large unexplained variability in the data. Low-probability data, as by its nature, was rarely found in the corpus which was why the data was skewed towards high-probability events.

In their analysis of American English word duration, Gahl et al. (2012) focused on CVC monomorphemic content words in the Buckeye corpus (Pitt et al., 2005). They included a variety of ID factors, such as word bigram probability of preceding and following context, word frequency, and previous mention, in their word duration model, while using speech rate as a prosodic control. Other control factors were average word duration, phonological neighborhood density (PND), syntactic category, and orthographic length. Results showed the expected effects of longer baseline duration for longer word duration, and increasing speech rate, bigram probabilities, and frequency for longer word durations. In addition, increased PND was associated with shorter durations.

Predictability does not only affect word duration, but also the duration of sublexical units. Syllabic duration in American English conversational and read speech was significantly influenced by language redundancy (Aylett and Turk, 2004, 2006). Here, language redundancy was defined using the categories high, mid and low based on log-transformed unigram, bigram and trigram probabilities of syllables. The authors also included prominence (primary lexical stress) in their statistical model and controlled for prosodic boundary. Both in spontaneous dialogues (Aylett and Turk, 2004) and in citation speech of professional speakers (Aylett and Turk, 2006) there was a complementary inverse relationship between language redundancy and acoustic redundancy in the durational domain which was mediated through prosodic structure. The factors prominence and language redundancy each contributed uniquely to explaining durational variability, while also interacting with each another.

ID factors were also introduced in studies focusing on phoneme durations in American English (Cohen Priva, 2015; van Son and van Santen, 2005). Intervocalic consonants were reduced in their duration when they were highly predictable based on their *information content* (van Son and van Santen, 2005). The authors defined this ID factor using Equation 2.4, where P stands for probability.

$$Information(segment) = -log_2(P(segment)) \tag{2.4}$$

Primary lexical stress and word boundary position were included in this consonant duration analysis in order to compare the findings to Aylett and Turk (2004). The effect of stress on consonant duration depended on the identity of the consonant and its position within the word (initial or final). Overall, stressed consonants were longer than unstressed ones. At word boundary, van Son and van Santen (2005) reported increased consonant durations in American English read speech compared to no boundary position which was interpreted as an increase in articulatory effort to mark prosodic constituents.

Cohen Priva (2015) confirmed the impact of information content, called segment probability in his study, on intervocalic consonant duration in American English based on the Buckeye corpus (Pitt et al., 2005). In addition, he found that low conditional probability also led to higher segment duration. High *informativity* of the consonant was also predictive of increased durations. Informativity was defined as the average contextual ($context$) predictability (P) of a linguistic unit ($unit_i$) (Equation 2.5).

$$Informativity(unit_i) = -\sum_{context} P(context|unit_i)\ log_2 P(unit_i|context). \tag{2.5}$$

2.2.2 Segment deletion

Coronal stop deletion Deletion of /t, d/ has been intensively studied in phonetic studies over the past decades, in both medial and final word position (Guy, 1980; Raymond et al., 2006; Zimmerer, 2009; Zue and Laferriere, 1979). Across studies, following phonological context had a larger impact on the deletion of coronal stops than preceding phonological context or speech rate variation. Following phonologically similar segments induced higher deletion rates of coronal stops than other consonants, and following consonants were more predictive of deletion than following vocalic segments (Guy, 1980; Tanner et al., 2017).

In models of coronal stop deletion, pause was usually defined as missing following phonological segment. This led to contradictory results of the influence of pause on coronal stop deletion (CSD). Pauses were found to lead to the lowest /t/ deletion rates of all following contexts in British English (Tagliamonte and Temple, 2005), while others observed that following pauses were more predictive of /t deletion than vocalic segments in modern Appalachia (Hazen, 2011). When pause duration was

introduced as a gradient continuous factor in the analysis of British English, deletion rates increased with decreasing pause duration (Tanner et al., 2017).

With regard to ID measures, word frequency has been identified as a predictor of CSD (Bybee, 2002; Coetzee and Kawahara, 2013; Jurafsky, Bell, Gregory, et al., 2001). Higher deletion rates were found in high-frequency words which reflected that these words were more prone to reduction than low-frequency words. Tanner et al. (2017) included word frequency and conditional probability of the bigram of the following word in their model of CSD in British English based on the Big Brother corpus (Sonderegger et al., 2017) investigating whether these factors reduced the influence of the following phonological context on deletion rates. Conditional probability combines two measures of ID: the joint probability of linguistic units, and the relative frequency of neighboring linguistic units (Jurafsky, Bell, Gregory, et al., 2001). They found that the influence of the following phonological context increased with the conditional probability of the following word, however conditional probability as a single fixed effect had no significant influence on CSD. Word frequency, on the other hand, had a significant positive effect on the deletion of final /t, d/ with higher deletion rates in high-frequency words. Tanner et al. (2017) argued that the effect of conditional probability was possibly masked by the effect of word frequency because both factors were confounded in the analysis.

Mechanisms of predictability also play a role in the comprehension of segment deletion. Bendixen et al. (2014) found that the omission of a predictable speech segment caused a larger omission response, i. e., increase in cognitive effort, than the omission of an unpredictable speech segment. They tested this for predictability measured as cloze probability at the sentence level and as repetition of lexical items (predictable) versus random presentation of lexical items (unpredictable).

In a follow-up study, Steinberg and Scharinger (2018) tested if ID factors at the sentence, word and phoneme level had different effects on this negative mismatch found in the omission response. They observed a hierarchy of these factors with cloze probability at the sentence level overriding the effect of word frequency, and word frequency overriding phoneme probability effects. This means that segment deletion in high-frequency words did not show a strong effect in the N400, even if the deleted segment had a low unigram phoneme probability. In the same vein, if segments were deleted in a low-frequency word that was produced with low cloze probability, it did not show such a strong neuronal activation as the same word being produced with segment deletion in a high cloze probability context. As outlined above for the findings of Tanner et al. (2017) predictability measures from different linguistic levels seem to be related, and that possibly in a hierarchical fashion. However, the authors in both studies compared frequency or unigram probability measures to contextual predictability which relies on the linguistic context and might therefore be considered a stronger predictor by definition.

Cohen Priva (2015) investigated the effect of different ID measures on consonant deletion in a corpus of spontaneous speech of American English (Pitt et al.,

2005). Only consonants in intervocalic and postvocalic pre-consonantal position were included in the analysis. In his logistic regression, he used unigram segment probability, segment informativity, word frequency and segment predictability as ID variables. As results, the author reported that segment probability trended in the expected direction: low segment probability decreased likelihood to delete for consonants in both phonological contexts. However, there was no strong effect of segment probability, possibly due to the low number of observations per segment identity. For logistic regressions, the authors argued, a higher number of observations per factor level is needed to lead to statistically strong effects. High segment informativity and low local predictability, on the other hand, significantly predicted consonant deletion for intervocalic consonants. For postvocalic consonants, local predictability only trended in the expected direction. The factor word frequency did not reach significance level in either of the two models for different phonological context (intervocalic and postvocalic).

As we have seen above, studies on /t/ deletion usually only include word frequency, and in some exceptions also word n-gram information in their statistical models. The only study known to the author to also include n-phone ID information in their analysis on /t, d/ deletion is Raymond et al. (2006). Here, biphone frequency of the following and preceding context for /t, d/ estimated from the Buckeye corpus for American English (Pitt et al., 2005) were used. Frequency values were calculated from the same corpus that was the basis for their /t, d/ deletion analysis. Raymond et al. (2006) argued that these local frequency estimates are informative with regard to articulatory processes, while word n-grams assess the speaker's effort in speech planning and lexical access. Following biphone frequency was predictive of /t, d/ deletion rates, while preceding biphone frequency did not reach significance level. The authors controlled for numerous extra-linguistic and linguistic variables, amongst other things phonological context and word frequency.

/ə/ deletion /ə/ is the most frequently deleted vowel in both German read and spontaneous speech, while admittedly it is also the most frequent German vowel. In their analysis of German /ə/ productions in the Kiel Corpus of spontaneous speech (IPDS, 1997) and read speech (IPDS, 1994), Kohler and Rodgers (2001) found that /ə/ was deleted in 44 % of all cases in read speech, and in 64 % of all cases in spontaneous speech. This analysis included content and function words. Out of all deleted vowels, /ə/ deletion made up 95 % in read speech, and about 83 % in spontaneous speech. Preceding phonological context was identified as being conducive to /ə/ deletion after a sonorant in German. /ə/ was deleted in 90 % of all cases in the cluster *oral plosive* + /ə/ + *nasal consonant* in function words or in unstressed syllables following stressed syllables. Nasal consonants assimilated in 40 % to the place of articulation of the preceding plosive. These reduced sequences were interpreted as lexicalized realizations in German. /ə/ realization in these clusters was seen as marked. Preceding fricatives similarly led to high deletion rates of /ə/ (about 80 % in both spontaneous

and read speech). Preceding nasals and liquids were predictive of /ə/ deletion in spontaneous speech (95 % and 90 %), but not so much in German read speech (62 % and 68 %). Lowest /ə/ deletion rates were found when /ə/ followed a vowel, such as in < *gehen* > /geːən/.

Preceding context was not the only informative factor for /ə/ deletion. Word and phrase position, morphological information as well as following segmental context also predicted /ə/ deletion. There was no proof of word-final /ə/ deletion, except in function words, such as "wäre" vs. "wär". In addition, /ə/ was always produced when standing before a non-sonorant segment. Kohler and Rodgers (2001) did not control for word frequency or other ID factors in their deletion analysis. However, they commented on high deletion rates possibly correlating with high lexical frequency of certain tokens or syllables in the corpus.

In their analysis of /ə/ deletion in spontaneous American English using the Switchboard corpus (Godfrey et al., 1992), Patterson et al. (2003) observed that lexical stress pattern was the most effective factor in predicting /ə/ deletion. This study also controlled for word frequency. Here, /ə/ was significantly more likely to be deleted in high-frequency words than in low-frequency words.

Hume (2004) argued that predictability is one of the main driving sources of vowel epenthesis and deletion. For instance, phonotactically illegal consonant patterns in American English are perceived with epenthetic /ə/ (Pitt, 1998). This phenomenon could also be explained by the lack of acoustic salience of the vowel. While the salience account may hold for /ə/ epenthesis in American English, it cannot be used explaining French vowel epenthesis. In French, vowel epenthesis is constructed using a rounded vowel (/ø/) with roundedness as an acoustic salient feature. However, this phoneme is the most predictable French vowel supporting the claim made by Hume (2004). Similarly, 37 % of the lexemes in the French lexicon contain an optional French /ø/ (Adda-Decker et al., 1999) indicating that predictability affects vowel segment deletion.

2.2.3 Voice onset time

The analysis of voice onset time (VOT) in the context of ID factors in this thesis (Section 7) is interpreted as an extension of the studies on duration and ID presented above. VOT is defined as the duration of the onset of periodic glottis vibration after a stop consonant release. Usually, positive, zero, and negative VOT are differentiated. Negative VOT denotes voice onset before the release, while positive VOT occurs in stop consonants with start of voicing after the burst. For zero VOT to happen voice onset and stop consonant release have to coincide (Lisker and Abramson, 1964). Numerous studies have investigated universal and language-specific characteristics of VOT (e. g., Chao and L. M. Chen, 2008; Kehoe et al., 2004; Kessinger and Blumstein, 1997).

Depending on the place of articulation VOT durations vary systematically. Velar

stops have longer aspiration phases than alveolars and bilabials. This observations has been made cross-linguistically and is interpreted as a universal effect of the physiological principles in the production of stop consonants. Articulatory closures which are made further back in the vocal tract are produced with higher oral pressure than closures in the front of the vocal tract because they involve a smaller pharyngeal cavity (Fischer-Jørgensen, 1954; Peterson and Lehiste, 1960). On average, female speakers have longer VOT durations than male speakers (Whiteside, Henry, et al., 2004; Whiteside and Marshall, 2001). American English function words have shorter VOTs than content words, even after controlling for word frequency (Yao, 2009). Vowel height has been shown to have an effect on VOT duration in American English (Barry and Moyle, 2011). English voiceless stop consonants have longer VOTs when followed by a high, close vowel than followed by a low, open vowel (Klatt, 1975). In addition, VOT is strongly influenced by the idiosyncratic articulatory behavior of different speakers (Allen et al., 2003).

The effect of speech rate on VOT duration, however, is language-specific. In their production study on VOT differences across different speaking rates in English and Catalan, Solé and Estebas (2000) hypothesized that VOT duration in English should vary with speech rate since syllable-initial aspiration of voiceless stop consonants is implemented as a phonological rule. Aspiration should be longer at slow tempo and decrease in length at fast speech rate. In Catalan, however, aspiration duration was expected to remain constant at different speech rates because it was considered an effect of physiological phonetic implementation of a speech sound contrast. In the English data, there was a positive relationship between speech rate and VOT duration. At slow speech rate, speakers tended to increase their aspiration phases of voiceless stop consonants. For the Catalan speakers, on the other hand, there was no effect of speech rate on aspiration duration. For both languages, differences in VOT duration induced by place of articulation were irrespective of speech rate differences.

Some studies have also included word frequency in their models of variability in VOT duration (Pierrehumbert, 2000; Yao, 2009). Pierrehumbert (2000) introduced her production-based account of exemplar theory using the historical changes in lenition in the English language as evidence for her theory. She argued that detailed phonetic structures are associated with specific word forms and therefore influenced by the frequency of these tokens. According to her theory, frequent tokens show larger variability in phonetic form than infrequent tokens. Also, recency of encountering a token determines how active the exemplar is in the perceptual memory. Speakers select an exemplar randomly from the exemplar cloud for a given token. The phonetic target may not be achieved exactly by the speaker as stored but with slight deviations in the phonetic implementation. Work on hypo- and hyper-articulation has shown that there is a systematic bias in speech production favoring lenition of stop consonants over fortition (Lindblom et al., 1990). If interpreted in the exemplar-based production framework, low-frequency words have a stronger tendency to show the bias for lenition than high-frequency words.

Yao (2009) analyzed VOT in American English word-initial voiceless stop consonants in two speakers of the Buckeye corpus (Pitt et al., 2005). The two speakers were chosen based on their maximal differences: one elderly woman with the slowest speech rate in the corpus, and one young man with the highest overall speech rate in the Buckeye corpus. Function words were excluded from the analysis. The author controlled for place of articulation (labial, alveolar, or velar), word-frequency, preceding and following context (consonant or vowel), local speech rate of a 3 word chunks with the target word in medial position, utterance position (final or non-final), and duration of following phone. Regression models were trained for each speaker individually. Regression analysis showed that VOT duration increased with more backward articulation (noticeably, only for the male speaker, though), high-frequency words had shorter VOTs than low-frequency words, and that preceding vowels shortened VOT, while following vowels increased VOT duration. Regarding utterance position there was an effect of final lengthening in the VOT of utterance-final stop consonants. Noticeably, word frequency only explained a small amount of variation in the data, whereas speech rate had a stronger effect on VOT.

Cohen Priva (2017) proposed that word-final consonant lenition is determined by the information value of affected consonants. Consonants with low language-specific informativity values (Equation 2.5) had a high likelihood to undergo word-final lenition in that language. For instance, /t/ in American English only provided 1.35 bits of information, and Spanish /s/ only 3.37 bits. These values were the lowest final consonant informativity values in the respective languages, and across the seven languages that Cohen Priva (2017) studied. Both word-final /t/ in American English and /s/ in Spanish reportedly showed the highest lenition scores among the word-final consonants in these languages. The author concluded that low informativity puts pressure on word-final consonants to lenite or even delete. He tested this claim in a deletion study of American English arguing that deletion is the most extreme case of lenition, and found that informativity was inversely correlated with deletion, while controlling for phonological context, phoneme identity, prosodic factors (speech rate, stress) and other ID factors, such as unigram word and segment probability, and residualized segment predictability.

Predictability also has an effect on VOT productions in the context of convergence. In unpredictable contexts, American English speakers accommodate their /k/ VOT productions to a higher degree to a model speaker than in predictable contexts. This effect is prevalent when speakers were not told to imitate the model, and even more pronounced when they were told to imitate (Manker, 2017). Here, predictability was calculated using the cloze probability of the preceding context. The same experimental design based on predictability based on the following context showed no significant difference in speakers' convergence in the two ID conditions.

Buz, Jaeger, and Tanenhaus (2014) conducted a production experiment focusing on VOT productions when the target was confusable with a contextual competitor, e. g., *bill* versus *pill*. The competitor items were presented on a screen alongside the

target items while the participants were instructed to produce the target in an interactive communication task. The authors found that VOTs were on average 9.1 ms longer when the competitor was present compared to trials without competitors. The total word duration was not significantly affected by this result indicating that speakers hyperarticulate a specific feature to increase contrast when contextual confusability was present, rather than the entire lexical item.

2.2.4 Spectral features of vowels

There is disagreement about the question whether spectral characteristics of vowels are best described using local formant peaks or global representations, also called static and dynamic features respectively. Formant patterns are defined as "the resonance frequencies of the oral part of the vocal tract or those resonant frequencies that show a continuity with the oral resonances of an adjacent sound" (Fant, 1960, p. 25). Global spectral features of vowels, on the other hand, are characterized by the energy of the vowel as a function of frequency in any given range (Stevens, 2002).

Static spectral features

Vowels have a long history of being described by their formant patterns (Fant, 1960; Joos, 1948). Based on the *F-pattern* of vowels one can predict the filter function, and thus decompose source and filter in Fant's model (1960). Furthermore, vowel formants and their continuities in adjacent sounds yield the possibility of inferring articulation patterns, and are important cues for speech perception (Fant, 1960). Vowel formants provide salient information for vowel identification in perception experiments, while additional properties of the spectral envelope, such as spectral tilt, relative formant amplitudes, or formant bandwidth, are perceived as speaker- or channel-specific information (Klatt, 1980).

Static or target formant measurements are used to characterize vowel quality. Phonetic height is determined by the position of F1, while phonetic frontness depends on the position of F2 (Joos, 1948; Lindblom, 1963). This relationship is usually plotted in a two-dimensional space with F1 on the x-axis and F2 on the y-axis. When both axes show reversed scaling, a multilateral shape emerges which supposedly resembles the articulatory dimension of vowels (Harrington, 2010). One of the earliest descriptions of this *formant chart* can be found in Joos (1948). Modern studies in sociophonetics use the term *vowel space* (Neumeyer et al., 2010; Simpson and Ericsdotter, 2007; Weirich and Simpson, 2014).

Static formant measurements are usually taken at assumed target positions of the vocalic segment. Monophthongs are usually described at temporal mid point which is presumably the least affected by consonantal context (Lindblom, 1963; Stevens and House, 1963). However, the notion of target positions in vowels is controversial. There is evidence that tense and lax vowels in German and American English differ

systematically in the relative timing at which the target position is reached (Lehiste and Peterson, 1961; Strange and Bohn, 1998).

This is why, there are some attempts at finding the most stable part of the vowel systematically. van Son and Pols (1990) introduce five methods for target detection of vowel formants: temporal midpoint of the vowel (method Centre), averaged formant frequency over complete duration of vocalic segment (method Average), measuring at the point of maximal energy (method Energy), or at the point of minimal or maximal F1/F2 value of the vowel, measuring at the most stable part with least variance in the log of F1 - F3. In their study of 1,178 vowel realizations from one Dutch speaker they only find small, non-significant differences between these methods, and therefore propose to use the method which is most convenient.

Vowel dispersion In this thesis, we used a static measurement of vowel distinctiveness which is widely used in sociophonetic studies (e. g., Munson, 2007; Weirich and Simpson, 2014): vowel dispersion. It is defined as the Euclidean distance between the centre of the vowel space for all targets and each speaker and formant values for each vowel measured at the temporal mid point of the vowel (Bradlow et al., 1996). Vowels with a large vowel dispersion are most distinct from vowels produced with central tongue height and frontness.

Vowel space expansion is greater under slow speech rate, compared to normal or fast speech rate in American English (Turner et al., 1995) and German (Weiss, 2007) which is also reflected in the perception of speech tempo for German (Weirich and Simpson, 2014). American English vowels are also more dispersed when they precede a stretch of slow speech (Gahl et al., 2012). Both American English and German vowel formants move to a more central position in the F1/F2 vowel space under fast speech rate when investigated in intended tempo deviations (Malisz et al., 2018; Turner et al., 1995), and when analyzed in naturally occurring differences in speech rate (Weiss, 2007).

Vowel dispersion is also influenced by phonological context, average vowel duration, sex of the speaker, and vowel identity. Gahl et al. (2012) found that American English vowels were more dispersed following back consonants, and in targets with greater vowel duration. On average, German female speakers have larger vowel spaces than male speakers (Simpson and Ericsdotter, 2007). Vowel identity also has a tremendous influence on vowel dispersion. American English peripheral vowels are by their nature more dispersed than interior vowels (Wedel et al., 2018). Furthermore, increased vowel dispersion in American English is associated with increased intelligibility (Bradlow et al., 1996).

There are conflicting results on the relationship between phonological neighborhood density (PND) and vowel dispersion. While Gahl et al. (2012) found that high PND is associated with less dispersion in American English, this was contrary to previous studies on American English vowel dispersion and PND (Munson and Solomon, 2004; Munson, 2007; Wright, 2004). Gahl et al. (2012) explained this by differences

in material: they used spontaneous speech of the Buckeye corpus (Pitt et al., 2005), whereas previous studies analyzed single-word productions or words in short carrier phrases. The authors argued that in conversations more or less extreme variations of articulatory targets are produced compared to list items. In a picture naming task, Buz and Jaeger (2016) found a tendency for a negative effect of PND on American English vowel dispersion supporting Gahl et al. (2012).

There are only few studies with a focus on ID factors and their impact on vowel dispersion. Word frequency (Munson and Solomon, 2004; Munson, 2007; Pierrehumbert, 2000; Scarborough, 2006; Wright, 2004) and language redundancy or predictability (Aylett and Turk, 2006; Clopper and Pierrehumbert, 2008; Jurafsky, Bell, Gregory, et al., 2001) have been identified as significant factors on vowel distinctiveness in American English. Munson and Solomon (2004) investigated word frequency and its effect on American English vowel expansion in isolated word productions. Vowels in high-frequency words were produced with shorter durations and less vowel space expansion than vowels in low-frequency words. More detailed investigations of these findings have shown that in an immediate-response condition both word frequency and PND influence vowel dispersion, while in long-delay conditions only PND had a positive effect on vowel expansion (Munson, 2007).

In a similar vein, Wright (2004) examined American English vowels in two different groups of 68 CVC words, classified as "easy" and "hard" targets for recognition from the Easy-hard word database (Torretta, 1995). "Easy" targets for recognition are high-frequency words which have little competition from their phonological neighbors. He found that vowels in lexical difficult words were produced with larger vowel dispersion than vowels in the "easy" targets group. This finding was supported by a multivariate analysis of variance taking vowel identity, difficulty and speaker into account. The greatest difference in vowel dispersion between "easy" and "hard" targets was found for the vowels /i, æ, ɑ, ɔ, u/, whereas the remainder of the investigated set (/aɪ, aʊ, e, ɛ, ɪ, o, ʌ/) showed slight or no expansion in vowel space. As an interpretation of his results, Wright (2004) stated that speakers make choices in linguistic encoding of vowel dispersion to increase the intelligibility of a message.

In a production task, Scarborough (2006) investigated the effect of both PND and contextual predictability measured as cloze probability on American English vowel realizations in the temporal and spectral domain. The author found an additive effect of both factors on vowel dispersion (and duration): vowels in high-frequency words with few lexical competitors, i. e., "easy" targets, were most reduced when they were also easily predictable from the preceding sentential context. "Hard" targets which were difficult to predict, on the other hand, were more dispersed in their spectral characteristics.

Jurafsky, Bell, Gregory, et al. (2001) analyzed vowel reduction in function words of American English conversational speech from the Switchboard corpus (Godfrey et al., 1992) using narrow transcription as their analysis tool. Preceding and following bigram conditional probability at the word level were both predictive of vowel reduc-

tion in function words. Highly predictable function words showed significantly more vowel reductions than words which were difficult to predict. Even though the authors controlled for the bigram of both the preceding and following word, they still found an additional, but weak, effect of the trigram predictability based on the preceding and following word for vowel reduction.

Clopper and Pierrehumbert (2008) tested the interaction between predictability measured as cloze probability and regional dialect variation in American English of female speakers from Northern, Midland and Southern dialects on vowel dispersion and duration. They used data from the Nationwide Speech Project corpus (Clopper and Pisoni, 2006) and the Indiana Speech Project corpus (Clopper, Carter, et al., 2002). The analysis focused on the vowels /i, æ, ɑ, ʌ/. Individual vowel formants F1 and F2, as well as vowel dispersion were significantly more reduced in easily predictable contexts for Southern speakers, but not for Northern or Midland speakers. Therefore, dialect and predictability complemented each other in their effect on vowel dispersion in American English. The authors also stressed differences in the overall effect of predictability on vowel dispersion depending on the vowel identity.

In their study on the influence of prosodic structure and ID on vowel characteristics in American English, Aylett and Turk (2006) investigated read speech from the Rhetorical Corpus. F1 and F2 values were measured at the temporal midpoint of the vocalic nuclei of /ɑ, æ, ɛ, i, u/. The language redundancy model was designed using high, mid and low language redundancy based on log-transformed unigram, bigram and trigram probabilities of syllables. The prosodic model consisted of prominence (none, primary lexical stress, or high probability of having a phrasal stress) and boundaries (none, word boundary, high probability of following phrase boundary). Results of the study showed that vowels were more centralized with increased language redundancy, vowel quality in prominent syllables was more distinct than in syllables that were not prominent, and spectral characteristics of vowels were also more distinct in syllables before prosodic boundaries than in syllables at word or no boundary. Aylett and Turk (2006) concluded that language redundancy and acoustic redundancy showed an inverse relationship which was mediated and implemented through prosodic structure.

The majority of studies investigating the impact of ID on vowel dispersion focuses on American English. van Son, Bolotova, et al. (2004) broadened this field by including typologically unrelated languages, such as Dutch (IFAcorpus (van Son, Binnenpoorte, et al., 2001)), Finnish and Russian (Intas 915 project (Bondarko et al., 2003)). They investigated the impact of the information content (Equation 2.4) of a vowel on its dispersion in both read and spontaneous speech in these three languages. They found that word frequency was a significant predictor of vowel dispersion in all three languages, irrespective of speech register. Dutch vowels were more dispersed in vowels with high information content in both read and spontaneous speech. However, there was no significant relation between vowel dispersion and information content for Finnish and Russian in any of the corpora investigated. This work allows the follow-

ing two conclusions: first, word frequency and predictability are not only conceptually different metrics of ID, but also seem to have different influences on segmental variability. Second, the impact of ID on segmental variability is, to some extent, dependent on the language under investigation, and not necessarily language-universal.

In their cross-linguistic analysis of vowel dispersion based on a subset of the BonnTempo corpus (Dellwo et al., 2004) including six speakers of each language (German, French, Finnish, Czech, and Polish), Schulz et al. (2016) found only a tendency for a positive effect of biphone surprisal of the preceding context on vowel distinctiveness. In their model, however, all prosodic factors (primary lexical stress, boundary, intended speech rate) showed significant effects on vowel dispersion. Vowels were more dispersed in stressed syllables, before a prosodic boundary and at normal and slow speech rate compared to fast speech.

Dynamic spectral features

Formant measurements are criticized as an estimate of the spectral characteristics of vowels because formant positions are not stable. Point-wise measurements at target position do not replicate the dynamic quality of vowels, whereas dynamic descriptions of vowels are in line with dynamic properties of speech production (Lindblom, 1963). Speakers vary considerably in the temporal spacing of the target position of the vowel, as well as in the shape of their formant contours (McDougall and Nolan, 2007). This is why assumed target positions at temporal midpoint are not necessarily fit to successfully describe vowel characteristics. In addition, dynamic features of vowels are better suited to classify vowels in statistical pattern recognition (Hillenbrand, Getty, et al., 1995; Hillenbrand and Houde, 2003; Zahorian and Jagharghi, 1993). Listeners can identify vowels successfully, even if the mid section of the vowel is substituted by silence, solely relying on cues from onset and offset of the vowel (Jenkins et al., 1999; Strange and Bohn, 1998). Steady-state information alone leads to poor results in listeners' vowel classification (Hillenbrand and Gayvert, 1993).

Nearey and Assmann (1986) proposed measures of VISC including the initial (Fn_i) and final portion (Fn_f) of the vowel which functions as way of time-normalization and enables comparisons between vowels of different duration. Both measures were interpreted by the authors as dual target measures, especially with regard to the analysis of diphthongs. ΔFn are calculated as

$$\Delta Fn = Fn_{\mathrm{f}} - Fn_{\mathrm{i}}. \tag{2.6}$$

As a measure of spectral change, Nearey and Assmann (1986) proposed F1 and F2 slope which incorporate the overall vowel duration. The difference between initial and final formant frequency is set into relation to the duration of the vowel since it is a strong predictor of the magnitude of spectral change. On average, short vowels show less spectral change than long vowels. This measure can be thought of as an

estimator of initial target plus spectral slope. Fn slope are calculated as

$$FnSlope = \frac{\Delta Fn}{VowelDur}. \tag{2.7}$$

There are several other measures of dynamic formant trajectories expressing the relationship between equidistant, time-normalized formant measurements of vowels. In the F1/F2 plane, vector length (VL) can be interpreted as an indicator of the amount of formant change. This measure is expressed as the Euclidean distance between the onset $(F1_i, F2_i)$ and offset $(F1_f, F2_f)$ of F1 and F2 values. The longer the distance between those values, the greater is the magnitude of change within the vowel. VL is calculated as

$$VL = \sqrt{((F1_i - F1_f)^2 + (F2_i - F2_f)^2)}. \tag{2.8}$$

A more fine-grained but similar measure is trajectory length (TL) which in addition uses the sampling points within onset and offset of the vowel. TL is defined as the overall sum of individual vowel section length (VSL) from one sampling point to its neighboring one. VSL is calculated with the following equation

$$VSL_n = \sqrt{((F1_n - F1_{n+1})^2 + (F2_n - F2_{n+1})^2)}, \tag{2.9}$$

while TL is the sum of all VSLs of the number of vowel sections (N) investigated. It is calculated as

$$TL = \sum_{n=1}^{N} VSL_n. \tag{2.10}$$

Spectral change, however, varies with respect to the portion of the vowel under investigation. For that reason, one can also include rate of change (roc) measurements for TL and VSL incorporating the overall vowel duration and vowel section at which the formant measurement was taken (Fox and Jacewicz, 2009). TL_roc and VSL_roc are calculated as

$$TL\_roc = \frac{TL}{0.5 * VowelDur}, \tag{2.11}$$

and

$$VSL\_roc_n = \frac{VSL_n}{0.15 * VowelDur}, \tag{2.12}$$

respectively, when 0.15 denotes the percentage distance between measurement points in the vowel.

Time-varying formant contours can be expressed by only a few coefficients by using orthogonal polynomials or by fitting a discrete cosine transformation (DCT) to the series of formant values. Orthogonal cubic polynomials of the form

$$f(x) = ax3 + bx2 + cx + d \tag{2.13}$$

reduce the formant contour to four coefficients or parameters. The constant coefficient is equivalent to the mean formant value. The linear parameter can be thought of as the slope of the formant. The quadratic coefficient is a curve with one turning point and draws an average picture of the formant curvature, while the cubic coefficient gives the curvilinear shape of the formant contour, a more detailed reflection of the overall formant movement.

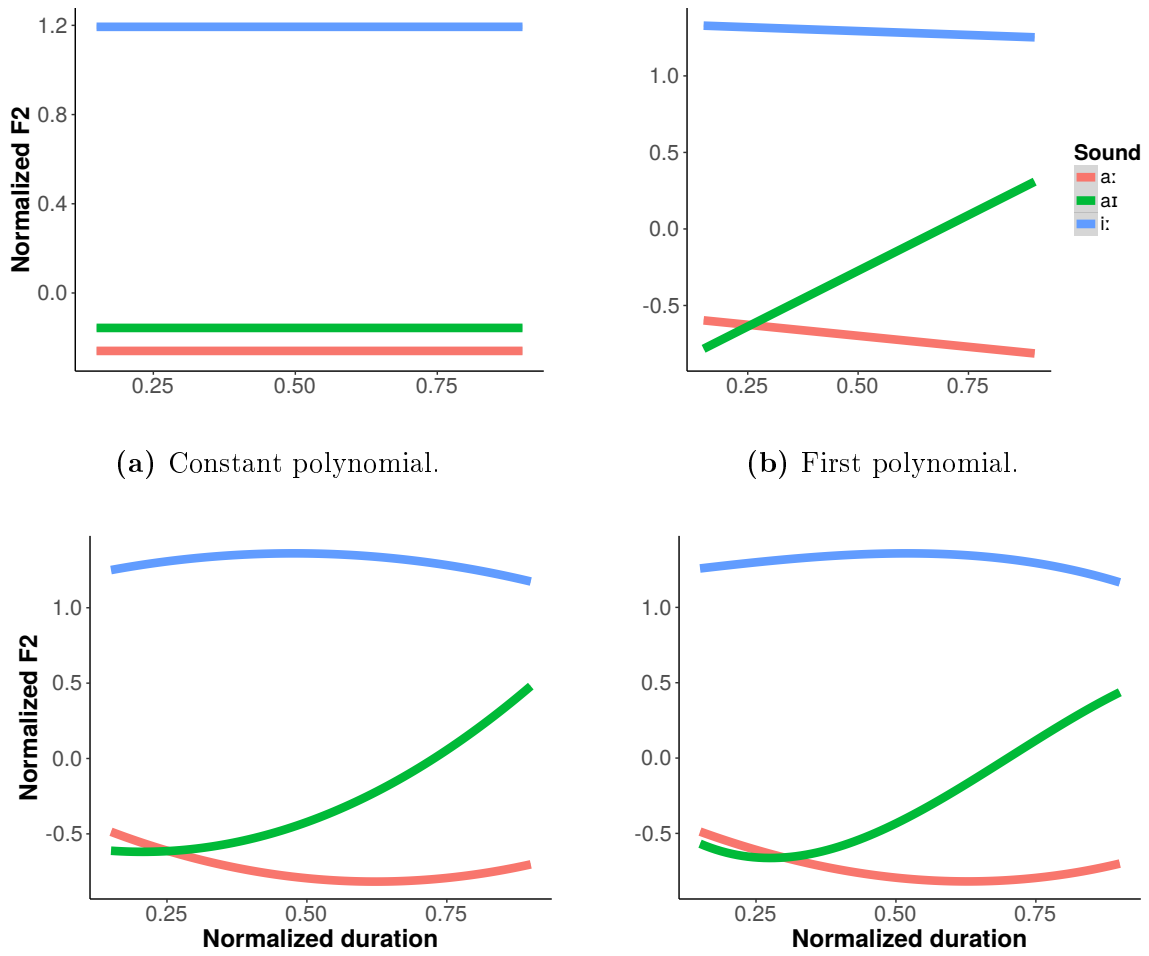
Polynomial coefficients have been used to describe pitch contours via parametric stylization (Grabe et al., 2007), pupil movements from eye-tracking data (McMurray et al., 2010), or formant trajectories (Morrison, 2009; Risdal and Kohn, 2014). The orthogonal cubic polynomial parameter of F2, for instance, has been shown to successfully differentiate African American English vowels from European-American vowels (Risdal and Kohn, 2014). Figure 2.1 visualizes the first four polynomial coefficients of F2 for the German vowels /aː, iː, aɪ/.

DCT was developed in the context of pattern recognition in digital processing (Ahmed et al., 1974). DCT is a cosine transformation of a Fast Fourier transform (FFT) decomposing the signal into half-cycles of that wave. Amplitudes measured at each half-cycle are defined as DCT coefficients of ascending order. Similarly to the polynomials, DCT coefficients break down the signal into a few single parameters. DCT0 is proportional to the signal's mean, DCT1 represents the slope, and DCT2 the curvature of the signal. Higher order DCT coefficients are used to describe higher frequencies of the spectrum (Harrington, 2010; C. I. Watson and Harrington, 1999) (Figure 2.2).

DCT coefficients have been shown to be highly useful in distinguishing different phonetic categories provided these have distinct spectra. For instance, the first two DCT coefficients played a significant role in discriminating vowel identities in a vowel classification experiment (C. I. Watson and Harrington, 1999). Combining the mean and slope value of the formant values were sufficient in characterizing vowel spectra if these did not show complex curvatures.

In a comparative analysis of parametric measures of formant trajectories (polynomial and DCT coefficients) and simple onset-offset measures of VISC, Hillenbrand, Clark, et al. (2001) did not find that the more complex parametric measures outperformed simple measures in a classification experiment of American English vowels.

An additional parametric representation of the spectral characteristics of vowels are mel-frequency cepstral (MFC) or mel-generalized cepstral (MGC) coefficients. The FFT of the log Mel-spectrum of a sound is converted via DCT to result in a MFC (Davis and Mermelstein, 1980; Muda et al., 2010). The MGC representation of



(a) Constant polynomial.

(b) First polynomial.

(c) Second polynomial.

(d) Third polynomial.

Figure 2.1: First four average polynomial coefficients of F2 for the German vowels /aː, iː, aɪ/.

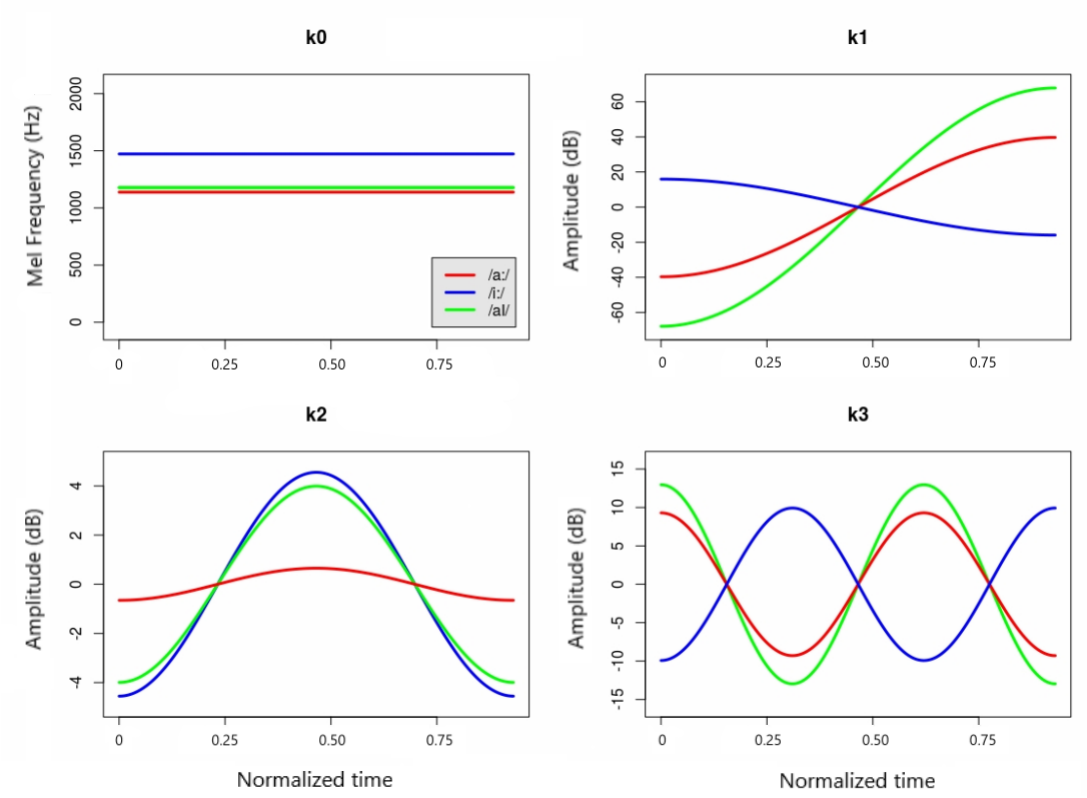


Figure 2.2: Average DCT coefficients DCT0–DCT3 (k0 - k3) of F2 for the German vowels /aː, iː, aɪ/.

speech is the same as the MFC, but calculated on a generalized log-transformed Mel-spectrum (Tokuda et al., 1994). In contrast to cepstral analysis, generalized cepstral analysis is not affected by the fine structure of the FFT spectrum. Parameters of MGC can be optimized for speech recognition, synthesis or analysis. Both acoustic vectors, MGC and MFC, can be measured at one specific time point within the vowel, or they can be used to describe dynamic changes within the vowel.

In order to estimate the spectral similarity between vowels one can use spectral distance measures, such as mel-cepstral distortion (MCD) (Equation 2.14) (Kubichek, 1993). This measure is usually applied to compare synthetic and natural speech in terms of their similarity (Lasarcyk et al., 2015; Toda et al., 2004). As input, MCD takes the acoustic vector of a mel-cepstral representation of the spectrum, e. g., MFC or MGC. If the acoustic vector is given without its deltas and double-deltas, MCD cannot take account of speech dynamics, neither short-term nor long-term. Also, it cannot deliver information about possible distortions in the pitch contour (Kominek et al., 2008).

$$MCD(v_{tar}, v_{ref}) = \alpha/T^{'} \sum_{t=0}^{T-1} \sqrt{\sum_{d=s}^{D} (v_{tar}(t) - v_{ref}(t))^2} \tag{2.14}$$

$$\alpha = 10\sqrt{2}/ln10 \tag{2.15}$$

Equation 2.14 measures the MCD between target (v_{tar}) and reference vector (v_{ref}). The term α is expressed in 2.15. T stands for the frame length, while $T^{'}$ is the number of non-silence frames in the analyzed signal. D expresses the number of analyzed dimensions which is equal to the vector size used to compute the mel-cepstral spectra. d stands for starting dimension. Its argument s can take either value 1 or 0. Starting from the first coefficient ($s = 1$) means that the power term is ignored, whereas starting from the 0th coefficient includes the overall signal power. If the result might be influenced by differences in power between the target and reference signal it is advisable to exclude the 0th coefficient from the MCD calculation and to use $s = 1$ as the starting dimension (Kominek et al., 2008).

Different vowel phonemes can be successfully distinguished by the usage of dynamic spectral features (Hillenbrand, Getty, et al., 1995; Hillenbrand and Houde, 2003; Zahorian and Jagharghi, 1993). This means that dynamic formant trajectories and parametric descriptions of the vocalic spectrum are highly influenced by the phoneme identity of the vowel. Diphthongs are known to show more spectral change than monophthongs (Harrington and Cassidy, 1994). Dynamic formant trajectories also help to distinguish tense/lax vowel pairs beyond their difference in total vowel duration (Slifka, 2003; C. I. Watson and Harrington, 1999).

F1 and F2 change their position with regard to consonantal context (Stevens and House, 1963). Specifically, the place of articulation of neighboring consonants affects vowel formant movements. The consonantal environment /hVd/ was identified

as the "null" environment leading to similar formant measurements as in isolated vowels. In this condition, differences between speakers were less pronounced than in the conditions with other consonantal contexts. Stevens and House (1963) only used symmetrical consonantal contexts differentiating between labial, alveolar and velar place of articulation. They observed that F1 was less affected by consonantal context than F2. For F2 the largest shift was seen for rounded vowels in alveolar contexts. Hillenbrand, Clark, et al. (2001) followed up on these results also including non-symmetrical consonantal contexts and investigating spectral change patterns within the vowels. All in all, they replicated earlier findings adding that formants in non-symmetrical contexts did not deviate massively from symmetrical contexts. Also, including spectral change information based on VISC measures improved vowel classification. Men showed less spectral change in their vowel formants than female speakers.

The amount of formant change also depends on the duration of the vowel. In American English, longer vowels tend to show more formant movement, even when it is calculated relative to vowel duration. Vowel duration is ultimately affected by prosodic factors, such as prosodic boundary, stress, or speech rate (Fox and Jacewicz, 2009). At fast speech rate, magnitude of formant movement is significantly smaller than at slow speech rate (Gay, 1978). The effect of speech rate on formant movement interacts with vowel and speaker identity (Weismer and Berry, 2003).

To the author's knowledge, so far there have been no studies investigating the relationship between ID factors as defined in this thesis and variability in formant change and movement. This includes studies based on dynamic formant trajectories and parametric measures. The analysis presented in Chapter 9 on dynamic formant trajectories in German male speakers as a function of ID and prosodic factors sheds a first light on this issue.

2.2.5 Voice quality

Voice quality changes as a function of the prosodic structure of an utterance, or because of dysfunctional and functional reasons. Some languages also use phonation to create linguistic contrast in their phoneme system, e. g., Gujarati (Fischer-Jørgensen, 1967). In the context of voice pathology, one usually distinguishes four different main types of voice quality (for an extensive overview see Laver (1980)):

- Modal voice
- Breathy voice
- Rough voice
- Hoarse voice

The modal voice quality is defined as the healthy voice without traces of noise in the perturbation signal due to complete vibration cycles of the entire glottal folds. Breathiness is defined by incomplete closure of the glottis during the closing phase, reduced tenseness in the glottis and slight noise in the signal. The breathy glottal waveform is more rounded and closer to a sinus shape than that of a healthy, modal voice. Because of this waveform shape the amplitude of the first harmonic (H1) is increased which is why typically H1 is measured when describing breathiness in the signal (Hillenbrand, Cleveland, et al., 1994). The rough voice quality is caused by an increased tenseness which leads to irregularities in the glottal signal (Laver, 1980). Hoarse voice quality can be interpreted as a combination of breathy and rough abnormal voice quality with individually varying degrees of breathiness and hoarseness in the signal (Ferrand, 2011).

Cepstral peak prominence

Several spectral features have been identified as correlates of voice quality, for instance spectral tilt, harmonics-to-noise ratio or spectral noise (Maryn et al., 2009). In this thesis, we used acoustic-phonetic measures which correlate well with perceived breathiness and hoarseness (Heman-Ackah et al., 2002; Hillenbrand, Cleveland, et al., 1994; Hillenbrand and Houde, 1996): cepstral peak prominence (CPP) and cepstral peak prominence smoothed (CPPS). CPP measures the difference in amplitude (in dB) between the cepstral peak and the corresponding fundamental quefrency. In contrast to the more established measures, it describes the entire cepstrum, and does not depend on a selective analysis. Additionally, as a cepstral analysis, CPP offers a better representation of the spectral envelope and its periodicity than a traditional spectral analysis. Although CPP relies on pitch tracking, the F0 measurements are not entered into the metric calculation. Other perturbation measures, such as harmonics-to-noise ratio, include pitch tracking and also possible errors into their metric calculations.

In a meta-analysis of 81 acoustic measures of voice quality used in the literature, CPP and CPPS have been identified to be among the few reliable metrics to assess overall voice quality in both sustained vowels and continuous speech (Maryn et al., 2009). However, CPP measures have been deemed unuseful to predict perceived roughness (Heman-Ackah et al., 2002; Moers et al., 2012). Low CPP values are associated with a higher degree of abnormality in the voice, while high CPP values are an indicator of harmonic richness and periodicity in the voice signal.

CPP metrics have been mainly tested in studies on laryngeal pathologies, but they were also integrated into the assessment of speech intelligibility (Haderlein et al., 2011), the detection of cognitive load (Yap et al., 2015; Yap et al., 2011), and voice attractiveness (Babel et al., 2014; Balasubramanium et al., 2012).

Haderlein et al. (2011) extended an established method for speech intelligibility evaluation from the field of automatic speech recognition to the field of voice pathology by integrating CPP measures in their algorithm. CPP was measured on single

vowels and continuous speech. A statistical model including CPP from continuous speech, prosodic features, and word accuracy correlated strongly with the voice evaluation of professional human evaluators, i. e., speech pathologists. Although CPP does not directly measure speech intelligibility, but breathiness, it is useful in a model of perceived speech intelligibility due to known correlations between these two factors (Haderlein, 2007).

Speech produced under high cognitive load, i. e., during a demanding cognitive task, contains less breathiness than speech produced under low cognitive load (Yap et al., 2011). The authors compared CPP to harmonics-to-noise ratio and H1 minus second harmonic (H2) (H1-H2) in separating tasks with different difficulty levels of cognitive load. CPP differentiated the three classes of low, medium, and high cognitive load the most successfully out of all tested acoustic measures. Under high cognitive load CPP values increased significantly.

Female and male voices which were judged sexually attractive had overall higher values of CPP than other voice samples with lower attractiveness scores (Balasubramanium et al., 2012). The authors interpreted this results with listeners' preference for well-defined harmonic structures in the voice signal indicated by higher CPP values. Babel et al. (2014) included CPP in a set of acoustic measures tested on isolated words that were recorded for a perceptual assessment of voice attractiveness. CPP accounted for a small amount of variability in attractiveness ratings, while H1-H2 and other long distance measures of H1 contributed the most to explaining variability in the attractiveness judgments.

Electroglottographic signal

Voice quality cannot only be assessed using spectral or cepstral information, but also by information collected directly at the glottis source. The electroglottographic (EGG) signal offers a precise, noninvasive method to measure the opening and closing of the vocal folds by using a current flow through electrodes which are placed on opposite sides of the neck of a speaker (Michaud, 2004).

For the purpose of analyzing the signal, usually the first derivative of the glottal waveform (DEGG) is used which offers a more accurate description of physiological events during phonation than the EGG signal. At glottal closure, the DEGG signal displays a sharp negative pulse which is more distinct than differences in slope in the EGG signal for different phases of the glottal vibration cycle.

The measure open quotient (OQ) gives the percentage of the time of the glottal cycle in which the glottis is opened. An opening phase is defined as the time span between a positive peak in the DEGG signal and its neighboring negative peak (Childers and C. K. Lee, 1991). Derivative-EGG open peak amplitude (DEOPA) measures the peak amplitude of the open phase of the DEGG signal. It has been criticized to be a less precise measure than the closing peak, because glottal opening is less abrupt than the closing gesture (Henrich et al., 2004).

Both measures are indicators of harmonic richness: the lower the value for OQ or DEOPA, the more well-defined is the harmonic structure of the speech signal. For instance, perceived age and OQ values are strongly positively correlated in male speakers (Winkler and Sendlmeier, 2006), i. e., male speakers with lower harmonic richness in their vocal quality are perceived as older than other male speakers with a more well-defined harmonic structure based on OQ measurements. This relationship, however, is not found for female speakers.

To the author's knowledge, so far there have been no studies investigating the relationship between ID factors as defined in this thesis and variability in voice quality. This includes studies based on the spectral or cepstral characteristics of speech and studies based on the EGG signal. The analysis presented in Chapter 11 on voice quality in German male speakers and ID factors sheds a first light on this issue.

2.2.6 Speech perception

There is some evidence that listeners are more sensitive to differences in phonetic detail when these occur in unpredictable compared to predictable contexts (e. g., Beaver et al., 2007; Lieberman, 1963; Manker, 2017), although naturally produced or manipulated differences between both conditions are relatively small. Predictability and word frequency both also affect speech intelligibility in noisy channel conditions (Kalikow et al., 1977; Luce and Pisoni, 1998; Savin, 1963).

One of the earliest studies on speech perception and predictability was conducted by Lieberman (1963). In this study, the author recorded American English sentences with target words with different sentence cloze probability and n-gram probability (based on a cloze test) at fast intended speech rate. Easily predictable targets were produced with less duration and amplitude than targets which were difficult to predict from the context. The target words were extracted from the sentences and then presented to listeners in isolated form. Results of the perception test showed that listeners had lower recognition rates of targets produced at high predictability (26 % correct) than for targets taken from low predictability context (61 % correctly identified). Overall, the author concluded that the degree of intelligibility of a target word was inversely proportional to its redundancy.

In a discrimination experiment, Manker (2017) presented listeners with American English words in predictable and unpredictable contexts, measured as cloze probability, repeated the target word after a second of silence and following static noise (0.5 sec), and asked whether the target was produced exactly the same as heard in the previous sentence. Half of the target items were manipulated with doubled VOT of the word-initial consonant and a 20 Hz pitch raise to increase the prominence of the word. Subjects were better at discriminating experimental trials with manipulated words in an unpredictable context than in a predictable context. This effect only held when predictability was measured based on the cloze probability of the preceding context. The same experiment using cloze probability of the following context

failed to show a significant difference in listeners' discriminations.

Listeners were also able to detect probabilistic reduction in duration in a perception experiment of second occurrence focus (Beaver et al., 2007). This linguistic phenomenon occurs when a focus-sensitive operator is focused in the utterance, but is a repeat of an earlier focus. Second occurrence focus operators were produced with longer durations (6 ms) and more energy than their baselines.

In low speech-to-noise ratio conditions, low-frequency English words were perceived as phonologically similar high-frequency words (Savin, 1963). Despite these channel conditions, speakers were able to perceive some phonetic detail about the uttered word and used these as cues for word perception. Similarly, Kalikow et al. (1977) found that listeners were more successful at identifying easily predictable words than words which were difficult to predict when they were presented in babble-type noise. Predictability was measured as sentence cloze probability. Luce and Pisoni (1998) used evidence from word recognition in noise to support their *neighborhood activation model*. In addition to word frequency, PND and number of phonetically similar words were observed as crucial factors for word identification in noisy conditions.

Le Maguer et al. (2016) used information-theoretic metrics to enrich the descriptive feature set of the prosodic model in an English text-to-speech (TTS) synthesis system. They proposed to integrate word and syllable trigram predictability measures into the prosody related decision tree of the TTS system. For F0, word trigram information imposed a reorganization of the decision tree, while syllable trigram predictability only had a small effect on F0 implementation. In an AB preference perception test, listeners were asked if they preferred the baseline TTS system or the system which included both word and syllable trigram predictability information. 72.6 % of the listeners preferred utterances synthesized with the system including the ID information. However, according to spectral distance measures between both systems, such as MCD or root mean square error (RMSE) for duration and F0, there were no significant differences.

2.3 Interaction between prosodic factors and information density

According to the SSR hypothesis effects of language redundancy are moderated by the prosodic structure of an utterance (Aylett and Turk, 2004, 2006). This means that the effect of language redundancy is not additive to that of prosodic structure. However, the authors found that there was an unexpected unique contribution of language redundancy to explaining variability in their models of syllabic duration and F1/F2 characteristics of American English vowels. The authors also stated that their prosodic model consisting of primary lexical stress and prosodic boundary explained most of the variability in the data, followed by the shared contribution of prosody and language redundancy.

Closely related studies investigated the impact of ID factors on realizations of prominence marking and intonation structure and found that they are indeed influenced by predictability (Turnbull, 2017; Turnbull et al., 2015; D. G. Watson et al., 2008).

D. G. Watson et al. (2008) tried to tease apart effects of importance and predictability on acoustic prominence measured as F0 maximum and minimum, duration, and intensity. They set up a verbal version of the game "Tic Tac Toe" in which game moves were either predictable or unpredictable. Pitch movement and duration increased for non-predictable words compared to predictable words in this game scenario. Important game moves, however, were expressed by overall higher intensity by the participants.

Later, Turnbull et al. (2015) built on this work, but included a cross-linguistic angle into their investigation of the impact of predictability on the prosodic marking of focus. Native speakers of American English and Paraguayan Guaraní were asked to play an interactive game which contained more or less predictable objects based on their visual context. Predictable focus was on the adjective, noun or on the noun phrase. In the unpredictable trials, it was not predictable from the visual context whether adjective, noun or whole noun phrase would be in focus. In American English, acoustic cues for focus marking were enhanced in the unpredictable context compared to the predictable context. In Guaraní, contextual predictability was marked differently in the prosodic structure: accented adjectives had steeper F0 slopes in the unpredictable condition compared to the predictable condition, irrespective of focus. The authors concluded, that unpredictability is marked in the prosodic structure, but this is language-specific, and also may be influenced by word order and headedness.

Turnbull (2017) built on these findings in his investigation of the impact of predictability on F0 peak in experimental corpora of American English. Predictability was defined as discourse mention, utterance probability, and semantic focus. The study controlled for phonological effects of pitch accent types, and part-of-speech (adjectives vs. nouns). Effects of focus condition on F0 peak were inconsistently shown for nouns, but not found on adjectives. Second mention did not affect F0 peak in neither adjectives nor nouns after controlling for pitch accent type. Increased utterance probability led to a significant decrease of F0 maximum and minimum for adjectives and F0 maximum of nouns.

Another cue of prosodic salience is boundary signaling. Turk (2010) argued that word boundary marking is inversely related to language redundancy: unpredictable phrases are produced with more salient word boundary markers than predictable phrases. In order to investigate the impact of ID on prosodic boundaries the authors suggested to use both predictability of the preceding and following context, as well as ID measured based on word n-grams. However, Turk (2010) did not perform a data analysis in her theoretical paper.

There is also evidence that ID factors affect overall speech rate in a language. In a

cross-linguistic study, Pellegrino et al. (2011) investigated the relationship between ID and speech rate hypothesizing that there is a balance between both factors which is robust across languages. They included data from English, French, German, Italian, Javanese, Mandarin Chinese, and Spanish in their study. As their unit of ID, they used syllables, and also calculated speech rate as syllables per second. Results showed that the amount of information conveyed per unit of time, i.e., information rate, was significantly different for some language pairings (involving Javanese and English), but the majority of the languages investigated clustered together in that they did not differ significantly in their information rate.

2.4 Hypotheses

This thesis investigates the relationship between ID and prosodic factors at the segmental level. It is motivated by previous findings and resulting theories of an inverse relationship between phonetic reduction and predictability or language redundancy (Aylett and Turk, 2004; Jaeger, 2010; Jurafsky, Bell, Gregory, et al., 2001; Levy and Jaeger, 2007; Pluymaekers et al., 2005a). The overarching hypothesis for this thesis is that segments which are difficult to predict from the context are expanded in their spectral distinctiveness and increased in their durational features, while easily predictable segments are reduced spectrally and temporally. We test this hypothesis conducting several production analyses of segment duration and deletion, voice onset time, vowel dispersion, dynamic formant trajectories, vocalic spectral distance, and voice quality. Specifically, we expect to observe the following findings in these studies:

1. Easily predictable segments in and segments in high-frequency words are shorter in their overall duration and VOT (for plosives), as well as more likely to delete than segments in low-frequency words and segments which are difficult to predict.

2. Easily predictable vowels and vowels in high-frequency words show decreased vowel dispersion, magnitude of formant change in VISC and parametric measures, as well as decreased periodicity and harmonic richness compared to vowels that are difficult to predict and vowels in low-frequency words.

3. Vowels in the same ID context are more similar to each other in their spectral characteristics than vowels in different ID contexts.

For the segment duration, deletion and VOT analysis, we directly build on findings from previous studies, and aim to replicate these findings in another Germanic language, German. To the author's knowledge, formant dynamics, global spectral characteristics and voice quality have not yet been analyzed in the context of ID. We therefore extend the spectrum of acoustic-phonetic measures in this field by including VISC and parametric measures, MCD between vowels, as well as the perturbation

measurements CPP and CPPS, and EGG parameters which describe the regularity of the glottal fold vibration. These analyses can offer a first insight into the relationship between ID and dynamic formant trajectories, global spectral characteristics as well as voice quality.

Prosodic factors The prosodic factors integrated into all statistical models for the production analyses were primary lexical stress, prosodic boundary, and articulation rate extending the model in Aylett and Turk (2006) (Section 4.4). With regard to these variables we formulate the following general hypotheses:

1. Stressed segments show increased duration (also in VOT), lower deletion rates, increased vowel dispersion and magnitude of formant change, and presumably more harmonic richness than unstressed segments.

2. Segments immediately preceding a prosodic boundary also show increased duration, vowel dispersion, magnitude of formant change and are less likely to delete. We expect to find these effects to be more pronounced at a higher-level boundary (phrase level) than at a low-level boundary (word level) in the prosodic hierarchy of boundaries. Vowels at phrase boundary position are assumed to have decreased periodicity and harmonic richness.

3. Speech rate acceleration leads to decreased segment durations, vowel dispersion, magnitude of formant change, and harmonic richness, as well as higher likelihood of segment deletion. These effects are seen when speech rate is measured at the global sentence level and at the local word level. In case of a mismatch between those rates, we expect the local speech rate to have a stronger effect on the segment because of its immediate local influence compared to the global speech rate.

In accordance with Aylett and Turk (2006) we expect to find small, but robust effects of ID on the acoustic-phonetic measures investigated in the production analyses. The prosodic factors, especially primary lexical stress, are assumed to have a stronger effect on these measures compared to the ID effect. Also, we expect to find interactions between prosodic factors and ID supporting the idea that language redundancy is, to some extent, moderated by prosodic structure (Aylett and Turk, 2004).

In the analysis of spectral distance, we compare same vowel identities in the same or different stress and speech rate categories. Here, the hypotheses are the following for the prosodic factors:

- Vowels at both slow speech rate and vowels in both stressed conditions have the smallest spectral distances.

- Vowels in unstressed position and vowels produced at fast speech rate are assumed to show the largest amount of variability, and thus spectral distance to other vowels in contrasting and non-contrasting condition.

Vowel dispersion in six languages Regarding vowel dispersion, we also conducted a cross-linguistic analysis based on previous research by Oh (2015), Pellegrino et al. (2011), and van Son, Bolotova, et al. (2004) which stressed the importance of including typologically different languages when analyzing the effect of ID factors on acoustic-phonetic measures. Most of the research in this field focused on (American) English (Jaeger and Buz, 2017) and to a lesser extent on Dutch (Pluymaekers et al., 2005a,b). Our cross-linguistic analysis of vowel dispersion in six languages includes data from three different subfamilies of Indo-European (Germanic, Slavic and Romance), as well as a Finno-Ugric language, Finnish. In addition, we can make inferences about the relation between vowel dispersion and ID at different intended speech rates because of the nature of the corpus investigated here (Dellwo et al., 2004) (Section 3.1.2). For this analysis we follow these hypotheses:

- Across all languages, vowel dispersion is reduced in easily predictable contexts.
- The relation between vowel dispersion and ID holds across all speech rates because of the inverse relationship between those factors which is presumably not affected by intended acceleration or deceleration.

Vowel dispersion of Bulgarian L2 speakers of German Vowel dispersion is also analyzed in the context of language learning. We assume that native speakers of a language share the same LM with individual variability due to idiolectal, sociolinguistic or regional factors. They share certain knowledge about the predictability of upcoming and preceding linguistic events and units based on their context.

Language learners are exposed to a L2 and presumably also built mental models about the predictability of linguistic units in this language. These models presumably differ regarding the speaker's competence level and their level of exposure. We investigate if ID factors of the target language can explain phonetic variability, i. e., vowel dispersion, of L2 speakers at different competence levels. This analysis introduces one possible approach at language learning from an information-theoretic point of view. We expect to observe the following finding:

- Observed patterns for vowel dispersion and ID factors in L1 speakers are also apparent in advanced competence level (C2) language learners, but less pronounced or even non-existent in intermediate competence level (B2) speakers.

Perception experiment The impact of ID factors has been reported as small in explaining F1/F2 characteristics (Aylett and Turk, 2006) or durational variability

(Cohen Priva, 2015). We also know that ID variables have an effect on listeners' sensitivity to differences in phonetic detail, even if these are very small, and their ability to recognize words in noisy conditions (Section 2.2.6). Therefore, we conducted a perception experiment using cross-splicing of naturally produced material to create stimuli. Same word identities taken from a high ID context were presented in a low ID context, and vice versa. Listeners were asked if the manipulated or baseline recording sounded more natural. We had the following hypotheses:

- Listeners are sensitive to differences in the acoustic-phonetic realization of words because of their predictability, and judge the baseline to be more natural than the cross-spliced stimuli.
- Listeners are equally sensitive to these differences in either of the cross-splicing directions.

This chapter gave an overview of the state of the field of acoustic-phonetic studies investigating ID and prosodic factors. We introduced theoretical concepts and metrics which are later used in this thesis. Also, we outlined our general hypotheses regarding ID and prosody and their impact on segmental variability, in addition to hypotheses regarding the perception of violated ID expectations. The following Part II gives the methodology (materials and data analysis) utilized here.

Part II

Methodology

Part II Methodology: Structure

Part II of this thesis contains the methodology and is divided into two main chapters: materials (Chapter 3), and data analysis (Chapter 4). In the materials chapter we first introduce the speech corpora (Section 3.1) that were used in this thesis to conduct acoustic-phonetic analyses, and to design experimental stimuli for our perception experiment. The following speech corpora were used in this thesis:

- Siemens Synthesis corpus (Schiel, 1997)
- BonnTempo corpus (Dellwo et al., 2004)
- EUROM-1 corpus (Chan et al., 1995)
- PhonDat2 (PHONDAT2 – PD2, 1995)

All of these corpora contain read speech. For each corpus, we briefly summarize the materials used for recording, comment on the recording setting, technical devices used and the speakers who recorded the data sets. Additional material contained within the corpora are also presented, e. g., transcriptions or automatic segmentations.

Second, the chapter materials introduces the text corpora which were used for language modeling purposes (Section 3.2). Languages are given with their respective corpora for LM building:

- American English: Corpus of Contemporary American English (COCA) (Davies, 2008-)
- Czech: Frequency dictionary (Zséder et al., 2012)
- Finnish: Finnish Parole corpus (Department of General Linguistics, 1996–1998)
- French: Lexique (New et al., 2001)
- German: Stuttgart German Web-as-Corpus (SDeWaC) (Baroni and Kilgarriff, 2006)
- German: Frankfurter Rundschau corpus (Elsnet, 1992 – 1993)
- Polish: Frequency dictionary (Zséder et al., 2012)

In the second chapter of Part II on methodology, we present general procedures that were used to preprocess the speech material including an inter-rater reliability test of the manual verification process of segment boundaries (Section 4.1), as well as general processes regarding the language modeling procedure for the German LM based on SDeWaC (Baroni and Kilgarriff, 2006) (Section 4.2.1) which was the basis for all surprisal and word frequency estimates used in the acoustic-phonetic studies on

German, and the LMs built for the cross-linguistic analysis of vowel dispersion (Section 4.2.2). We aimed to keep the statistical modeling procedure regarding collinearity analysis, choice of control factors, model selection, and effect size calculation relatively stable across all production analyses conducted within the scope of this thesis (Section 4.3). We introduce the following control factors in more detail because they were used in almost all LMMs built here (Section 4.4):

- Word class
- Word frequency
- Average duration
- Phonological context
- Primary lexical stress
- Prosodic boundary
- Speech rate

Chapter 3

Materials

This chapter describes the speech and text corpora that were used in this thesis. Speech corpora were used for acoustic-phonetic analyses. The text corpora were utilized to train LMs for several languages.

3.1 Speech corpora

In this thesis, only corpora containing read speech are used. This clearly bears the drawback that we cannot make inferences about the impact of ID on acoustic-phonetic measures in different speech registers. However, we argue that we would find the same results, if not more pronounced in spontaneous speech that we found in our analyses of read speech because spontaneous speech contains a considerably larger amount of segmental reduction than read speech, even yielding in massive reductions (Johnson, 2004). Also, there is some evidence that for American English and Dutch positive relationships between ID and phonetic variables, such as duration and vowel dispersion, hold irrespective of the register of the speech data (Aylett and Turk, 2004, 2006; van Son, Bolotova, et al., 2004). Since German is closely related to both languages we expect to find this pattern in German as well.

3.1.1 Siemens Synthesis corpus

Most of the production analyses, i. e., segment duration and deletion, VOT, vowel dispersion, dynamic formant trajectories, and voice quality, were based on the German Siemens Synthesis corpus (SI1000P). In addition, this corpus was utilized to build stimuli for the perception test introduced in Chapter 12.

The Siemens Synthesis corpus (SI1000P) contains audio recordings from two professional male speakers with exceeding broadcasting experience (Schiel, 1997). Each speaker read 1,000 sentences from the Frankfurter Allgemeinen corpus (SI1000).

There are audio and EGG recordings for each speaker. The speakers are abbreviated "ai" and "wo" in the corpus. They were asked to read as if in a setting of broadcast announcing, i. e., fluent speech with a Standard German variety. The speakers were recorded in a total echo-canceling studio at the Institute of Phonetics at the University of Munich. The audio signal was recorded using a Sennheiser MKH20 omnidirectional with a controlled distance of 30 cm from the mouth. For the EGG signals, the laryngograph LxProc of Laryngograph Ltd. London was utilized. The recordings were done at 48 kHz, 16 bit, and then filtered and down-sampled to 16 kHz.

Canonical transcriptions, as well as automatic word and phoneme segmentations are available in the corpus. In addition, it contains manually labeled symbolic prosodic segmentations including three boundary and three accent markers which can be compared to German Tones and Break Indices (GToBI) parametrization (Grice and Baumann, 2002). Accent markers refer to the word where the accent is marked, while boundary markers are provided for left and right neighbors of the boundary. Syllable information is not included in the accent model of BAS. The accent and boundary markers were defined as follows:

- B3: full intonational boundary with strong marking
- B2: intermediate phrase boundary
- B9: boundary due to hesitations or repairs
- PA: phrase accent
- NA: secondary accent
- EK: emphatic accent

3.1.2 BonnTempo corpus

For the cross-linguistic analysis of vowel dispersion (Section 8.3), we analyzed a subset of the BonnTempo corpus (Dellwo et al., 2004) with three female and three male speakers of American English (AE), Czech (CES), German (DEU), Finnish (FIN), French (FRA) and Polish (POL). FIN, POL, and three AE speakers (two females, one male) were added to the BonnTempo corpus using the original instructions. The speakers were digitally recorded at a sampling rate of 48 kHz, 16 Bit in a sound-attenuated booth using a stationary DPA 2011A cardioid microphone. The recordings were down-sampled to 16 kHz sampling rate. Speakers were given an excerpt of a novel in their native language, and were asked to familiarize themselves with the text. Next, speakers were recorded at what they considered to be reading at normal pace. Then, subjects were asked to slow down, and to slow down even more. In a third step, fast speech rate was recorded asking speakers to speak fast, and speed up their speech rate until they considered they could not speed up any more. From these acceleration

Table 3.1: BonnTempo corpus: vowel qualities and token frequencies per language for vowel dispersion.

Language	Frequency	Quality
AE	560	/i, ɪ, ɑ, ɔ, ʊ/
CES	1,156	/iː, ɪ, ɛː, ɛ, aː, a, u/
DEU	825	/iː, ɪ, eː, ɛː, ɛ, aː, a, uː, ʊ/
FIN	1,178	/iː, i, eː, e, æ, æː, ɑː, ɑ, uː, u/
FRA	689	/i, e, a, u/
POL	790	/i, ɛ, a, u/

steps, normal speech rate, as well as the first steps of slow and fast speech rate were used for the analysis.

For the analysis of vowel dispersion we chose vowel phonemes to facilitate a comparative analysis between the different languages in the corpus. If available in the data, tense and lax vowels in closed front, closed back, open and front mid position were used for the analysis (Table 3.1). Based on these positions in the vowel space we allocated the vowel tokens to four categories which were used as factor levels in the statistical analysis. The total number of analyzed tokens was 5,198.

3.1.3 Production data for L1/L2 analysis

For the analysis of vowel dispersion in language learners (Section 8.4) we recorded Bulgarian speakers of German and German native speakers while reading five text passages from the EUROM-1 corpus (Chan et al., 1995) (Section B). The recordings took place in a quiet environment with a head mounted microphone (AKG C520), digitized with an Audiobox (M-Audio Fast Track). We recorded the speakers using the software Praat (Boersma and Weenink, 2017) and its default settings for audio recordings (41 kHz, 16 Bit). The speakers were asked to read fluently and as if they were engaged in telephone conversations in professional settings.

Six L2 speakers at intermediate competence level (B2) between the ages of 19 and 24 ($M = 20.17, SD = 1.77$), and six L2 speakers at advanced competence level (C2) between the ages of 36 and 54 ($M = 42.84, SD = 5.79$) were recorded in addition to six native speakers of German (L1) between the ages of 28 and 52 ($M = 36.67, SD = 9.09$). All speakers were females. The B2 Bulgarian speakers originated from Pernik and Sofia, while the C2 Bulgarian speakers came from Sevlievo, Sliven and Sofia. The German natives came from Wiesbaden, Kiel, Saarbrücken, Stuttgart and Berlin.

Only vowels in accented position were used for analysis in order to exclude effects stemming from that factor. Tense and lax vowels in the corner positions of the German vowel space were chosen for analysis: /aː, a, iː, ɪ, oː, ɔ/. In total, we analyzed 2,393 vowel tokens (L1 = 796, C2 = 797, B2 = 798).

Table 3.2: Data for MCD analysis: total number of MCD values per vowel identity.

Vowel	Number of items		
/ə/	180,155	/iː/	2,751
/ɐ/	255,957	/ɪ/	97,902
/øː/	489	/oː/	4,254
/œ/	12,921	/ɔ/	30,976
/aː/	103,096	/uː/	205
/a/	91,384	/ʊ/	105,720
/eː/	10,636	/yː/	518
/ɛ/	11,548	/ʏ/	11,817
Total	940,676		

3.1.4 PhonDat2

Read speech material from PhonDat2 Version 2.8 (PHONDAT2 – PD2, 1995) of 16 German natives (m = 10, f = 6) was used for the analysis of spectral similarity of vowels (Chapter 10). In PhonDat2, each speaker read 200 different screen-prompted sentences from a train inquiry task. They were asked to read carefully but fluently as if they were engaging with an automatic dialog system. Speakers were recorded in sound-canceling studio environments with various Sennheiser microphones (e. g., MKH 20 P48) at 48 kHz, 12 Bit. The data was digitally filtered to 8 kHz and down-sampled to 16 kHz sampling rate. The recordings took place at three different sites in Germany (University of Kiel, University of Bonn, University of Munich). This procedure was followed to ensure that the data contained regional variability of the German language. The corpus was designed to help improve automatic speech recognition and automatic segmentation procedures of regional varieties of German.

Automatic alignment, word segmentation and phonological segmentation of the speech data is provided in the corpus. Canonical transcription of the corpus including primary lexical stress information is given in the documentation. We decided to check the agreement between the manual phonological segmentation of the speech corpus with our segmentation guidelines. Because of reasons of practicability we manually checked a subset of 120 sentences per speaker, adding up to a total of 1,920 sentences. Only vowels in content words were analyzed. Table 3.2 lists the number of MCD values per vowel.

3.2 Language modeling corpora

Since LMs are language-specific we calculated individual models for all languages investigated in this thesis. In this section, we present the text corpora used for LM building for eliciting surprisal and word frequency values for German and for the

cross-linguistic analysis (Section 8.3).

3.2.1 German language model

The German LM was based on the SDeWaC corpus which was derived from the DeWaC corpus (Baroni and Kilgarriff, 2006). The web-crawled corpus contains 846,159,403 running words and about 1,094,902 millions lexical types with a diverse range of genres. The SDeWaC corpus was generated at the Institute for Natural Language Processing in Stuttgart choosing sentences from German Web-as-Corpus (DeWaC) that were parsable with a standard dependency parser for German. This cleaning procedure included removal of web-specific structures, such as HTML structures or long lists, removal of duplicate sentences and sentences that were grammatically ill-formed (Faaß and Eckart, 2013). SDeWaC is formatted with one sentence per line.

3.2.2 Cross-linguistic study

For each of the six languages, individual LMs were built using six different corpora. The FRA corpus used, Lexique 3.80 (New et al., 2001), already provides phonetic transcription and syllabification which were utilized for our language model building. For DEU, the Frankfurter Rundschau corpus (Elsnet, 1992 – 1993) was transcribed and syllabified using the WebMaus grapheme-to-phoneme (g2p) tool (Kisler et al., 2017). For AE, the same procedure was applied to process the COCA (Davies, 2008-). For FIN, the Finnish Parole corpus (Department of General Linguistics, 1996–1998) was acquired online. For both CES and POL, frequency dictionaries derived from a large-scale web corpus were used (Zséder et al., 2012). The FIN and CES text corpora were automatically converted into phonemes by the eSpeak speech synthesizer (Duddington, 2015) and automatically syllabified using a custom bash script. The POL frequency dictionary was converted into phonemic symbols and syllabified by an automatic tool for transcription and syllabification (Zeldes, 2008–2014). Corpus sizes are given in Table 3.3.

Materials that were used in this thesis comprised speech corpora for acoustic-phonetic analyses, and the design of experimental items, in addition to large text corpora for the purpose of language modeling. These were introduced in this chapter. The following chapter presents the data analysis procedures that were used to preprocess and analyze the data presented here.

Table 3.3: Cross-linguistic study: corpora and corpus sizes (in million tokens) for language modeling.

Language	Corpus	Size
CES	Zséder et al. (2012)	398
POL	Zséder et al. (2012)	901
AE	COCA	410
DEU	Frankfurter Rundschau	41
FRA	Lexique 3.80	9
FIN	Finnish Parole	180

Chapter 4

Data analysis

The following chapter gives details on the preprocessing of speech corpora that were used in the acoustic-phonetic analyses and are described in Section 3.1. Concrete descriptions of the methods used to extract spectral and durational characteristics, or to identify segment deletion are described in the respective sections on these analyses. Language modeling and statistical modeling followed uniform procedures which are described in this chapter. Admittedly, the language modeling procedure used for the cross-linguistic study on vowel dispersion differed to some extent from the German LM because they were not produced by the first author. They will be outlined in this chapter but are described in more detail in Oh (2015). Regarding statistical modeling we also introduce control factors that were universally used in the statistical models in Part III of this thesis. Control factors that were only used for specific analyses are described in the modeling sections of the respective analyses.

4.1 Preprocessing of speech corpora

The German speech data used in this thesis was annotated using the automatic forced alignment offered by WebMaus (Kisler et al., 2017). For the cross-linguistic study we used SPPAS (Bigi, 2013) for FRA because it was not yet implemented in WebMaus, and WebMaus for all other languages. Since there was no automatic segmentation tool available for CES, WebMaus implementations for other languages were tested. Hungarian WebMaus proved to be the most effective for CES because both languages have a largely similar consonant inventory, and vowel length is phonemic in both CES and Hungarian.

The manual verification process of the automatic segmentation of CES data was completed by an expert on Slavic languages. For all the other languages, the automatic segmentation was also manually verified by phonetic experts using general criteria in order to facilitate a comparative analysis between the different languages

in the corpus. POL, DEU and AE TextGrids were manually verified by native or near-native phonetic experts. FIN and FRA TextGrids were manually verified by non-native phonetic experts with experience in annotating foreign languages.

All TextGrids produced by WebMaus have the same basic structure. The first tier contains the orthographic transcription of the words in the speech material (= ORT) with word boundaries. The second tier holds information about the canonical phonemic transcription of the words with known variants in the transcription (= KAN). Here, word boundaries are segmented. The third tier consists of segmentations and annotations based on the canonical transcription in tier 2 (= MAU). Segment boundaries and pauses are segmented. The MAU tier was copied and annotators included their manually verifications in this fourth tier.

Manual verification followed a segmentation manual which included general guidelines on spectrogram settings and segmentation procedures. For instance, phonemic boundaries were verified using information from both the spectrogram and the waveform, as well as perceptual judgment. The largest portion of the manual describes detailed guidelines on how to segment and annotate specific cases, such as neighboring homorganic plosives, nasal releases, or double bursts in plosives. Another important guideline concerns the duration analysis and the spectral analysis of vocalic segments in the data. Following the segmentation manual the beginning of vowels was marked when F1 was clearly visible, and endings of vowels were marked using the end of a visible F2 structure.

Segment deletion, insertion or substitution were marked in the TextGrids by inserting a minimal interval of below 0.001 sec. Annotating these phenomena in our annotations allowed for an analysis of segment deletion (Chapter 6). We used the following annotations to mark the three phenomena:

- Deletion: "SOUND –"
- Insertion: "– SOUND"
- Substitution: "SOUND – otherSOUND"

Annotations of the Siemens Synthesis corpus were manually verified by two phonetically trained annotators. One was a native speaker of German, while the second annotator was a Russian native with C2 German competence level. The native speaker verified the majority of the TextGrids ($n = 1,157$). The near-native speaker completed a total of 827 manual verifications. We performed an inter-rater reliability test between the two annotators. The German native annotator verified 5 % ($n = 41$) of the same TextGrids that the near-native annotator completed. We selected a random sample from the TextGrids for the inter-rater test. We correlated segment durations between the annotations of the German native and the Russian native annotator using Spearman's rank correlation. We used this correlation method because the durations were not normally distributed. The correlation was very strong

($\rho = 0.93, S = 1427500000, p < 0.001$) indicating that there was a large amount of agreement between the two annotators. In light of the segment deletion analysis (Chapter 6) we were also interested in the inter-rater reliability for deleted segments. The German native annotator tended to mark slightly more segment deletions ($n = 146$) than the non-native annotator ($n = 132$). There was very low disagreement about 0.006 % of the segments of either being produced or deleted between the two annotators.

4.2 Language modeling

In order to obtain surprisal values we calculated LMs for all languages investigated in this thesis. The following section introduces the training procedure of these models and discusses different n-phone orders.

4.2.1 German language model

The German LM was based on the SDeWaC corpus (Section 3.2.1). Data preprocessing of the corpus included lower-casing and punctuation removal. The corpus was transcribed using the g2p tool implemented in German-Festival (Möhler et al., 2000). The transcriptions of the most frequent 1,000 words in the corpus were manually verified by the author of this thesis. Systematic errors were identified and corrected for all lexical items in the corpus. Next, we split the data into training (80 %) and test corpus (20 %) in order to test the model's performance (Section 4.2.3).

The training corpus was used to train n-phone LMs using the SRILM toolkit (Stolcke, 2002). All LMs included sentence markers. We included both function and content words in the LM, and trained different LMs including and excluding word boundaries. The default LM in SRILM calculates conditional probability of a linguistic unit occurring with preceding context. In order to calculate conditional probabilities based on the following context we used the inbuilt function of SRILM *reverse-text*. Each line of the corpus is read as a sentence, and the order of the linguistic units is reversed. Neither the words themselves nor their order within the sentence order are changed. This procedure resulted in four different LMs:

- Phoneme model including word boundaries
- Phoneme model without word boundaries
- Phoneme model reversed including word boundaries
- Phoneme model reversed without word boundaries

The smoothing technique applied for all models was Witten-Bell. This technique was used because of the limited lexicon of the LM. Count-of-counts statistics used

by more frequently applied discounting methods, such as Kneser-Ney, produced erroneous output. Surprisal values for the preceding and following context were obtained from the output of the LM. Surprisal was log-transformed due to positive skewness.

The following list shows the phonemes that were included in the phoneme inventory of the German LM based on SDeWaC. /ʒ/ and the nasalized vowels /ɛ̃, ã, õ/ were included in the phonemic transcription although they qualified as non-native phonemes. Allophones of German /ʁ/ were expressed by one symbol (/ʀ/).

- **Vowels**
  - /øː, ɐ, œ, ə, ɛ, ɛː, ɛ̃, ɪ, ɔ, ɔʏ, ʊ, ʏ, a, aː, aɪ, aʊ, ã, eː, aː, oː, õ, uː, yː/
- **Consonants**
  - /ʔ, ç, ŋ, ʀ, ʃ, ʒ, b, d, f, g, h, j, k, l, m, n, p, s, t, v, x, z/

In addition to the German *n*-phone LMs described above, we also trained a LM with SDeWaC on the word level. The text corpus was used in its graphemic version. Here, we used Kneser-Ney smoothing and included sentence and word boundaries. This model was trained to calculate surprisal values based on the word level for the perception test described in Chapter 12.

4.2.2 Cross-linguistic study

For each of the six languages investigated in the cross-linguistic study on vowel dispersion (Section 8.3), an individual LM was trained. Preprocessing of the data included cleaning of erroneous entries with non-alphabetic characters, lower-casing and punctuation removal. The data was phonemically transcribed and if not already provided in the data automatically syllabified. The FRA corpus contained phonemic transcriptions and syllabification. The text corpora for FIN, POL, and CES were automatically converted into phonemes by the speech synthesizer eSpeak (Duddington, 2015), and then syllabified by a bash shell script. The text corpora for AE and DEU were automatically transcribed using the WebMaus g2p tool (Kisler et al., 2017). The transcriptions of the most frequent 1,000 lexical items in the corpora were checked by native speakers or phonetic experts of the languages, and systematic errors were fixed. The LMs were calculated using custom-built perl-scripts. LMs included word and syllable boundary, as well as lexical stress information.

4.2.3 *n*-phone size

In order to establish a baseline for modeling the relationship between phonetic encoding at the segmental level and ID we tested *n*-phone dependencies of different sizes, as well as predictability values based on *n*-phone LMs with and without word boundaries. As a test case, we used vowel dispersion as the dependent phonetic variable

Table 4.1: Perplexity of the n-phone LMs. Test conducted on different orders of n-phone with word boundary (wb), and without word boundary (nb). Reversed corpora (r) were used for LMs based on the following context.

LM	n-phone order		
	6	4	3
wb	3.99657	5.50475	7.64124
wb, r	3.99655	5.50474	7.64124
nb	5.44327	8.39234	11.6828
nb, r	5.44379	8.39236	11.6828

because the positive relationship between vowel dispersion and ID is well-established by many studies that have found vowels to be more dispersed when they were difficult to predict from the context, and less dispersed when they were easily predictable (Section 2.2.4).

We tested n-phone dependencies up to the order of 6 based on LMs trained on the phonemically transcribed training corpus of SDeWaC with and without word boundary. The analysis was performed for ID values based on the preceding and following context. We reversed the order of phonemes within sentence boundaries in order to train and test LMs for the following phonemic context with SRILM. Perplexity (Equation 2.3) decreased with LM order, and was better for models with word boundary than for models without word boundary. Judging from these results, phoneme models with word boundaries of order 6 performed the best out of all the LMs tested (Table 4.1).

The results of the LM testing, however, do not necessarily have to be indicative of the most informative n-phone order for calculating ID measures for segmental variability. Perplexity measures the probability of the appearance of a test sentence based on a LM (Manning and Schütze, 1999). It presupposes that the test material is not part of the training data, and therefore qualifies as unseen material. This measure is therefore purely based on written text material, and might not be effective in determining the most informative n-phone order for modeling linguistic features of spoken data. For that reason, we did not use the model perplexity results to derive a starting point for n-phone size for this thesis, but identified it experimentally.

We correlated vowel dispersion calculated in German read speech (Section 8.2) with surprisal values calculated on different n-phones (order 2 to 6) from LMs with (wb) and without word boundaries (nb) using only vowels in content words (CON) and vowels in all words (ALL: content and function words). Figure 4.1 shows that across all four conditions larger n-phone order did not increase the Pearson's r value for the correlation between vowel dispersion and surprisal. We found that triphone surprisal had the highest correlation values in all four conditions. Generally, correlation values decreased in strength for higher n-phone order than 3. Using vowels in all words for

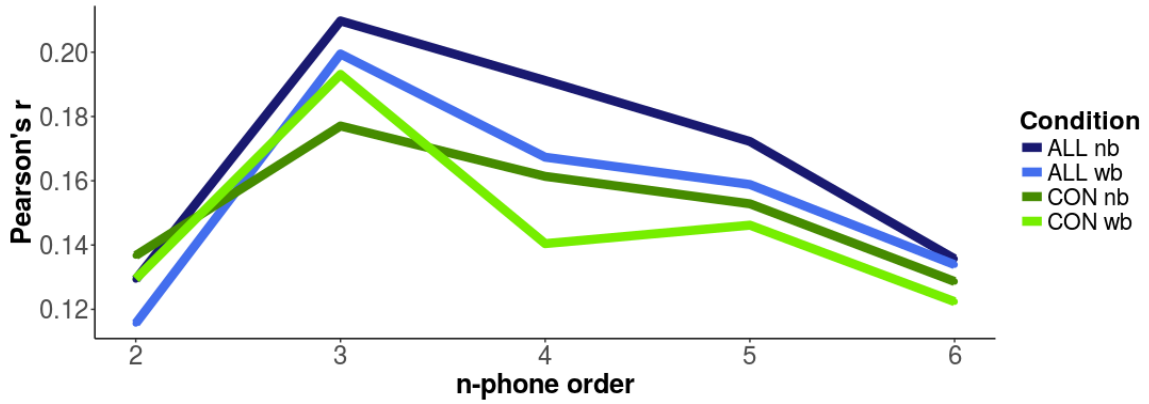


Figure 4.1: Correlation between vowel dispersion and surprisal based on LMs with (wb) and without (nb) word boundary using only vowels in content words (CON), and in all words (ALL).

the analysis resulted in overall higher correlation values for all n-phone sizes. When limiting the analysis to vowels in content words the Pearson's r was slightly higher at order 3 for the LM including word boundaries. As a result of this analysis we decided to test small n-phone sizes based on LMs including word boundary for the statistical models of ID and phonetic variability calculated in this thesis.

Regarding ID values for the following phonological context we found that there were only very low, and mostly negative, correlations between surprisal values based on different n-phone sizes and vowel dispersion. For that reason, we did not include surprisal of the following context in this analysis.

We found stronger similarities for surprisal between similar n-phone sizes than for surprisal with larger differences in n-phone size, irrespective of whether word boundaries were included in the LM or not. Figure 4.2 shows Pearson's r correlation values for surprisal values with different n-phone sizes. Biphone and sixphone surprisal, for instance, had the lowest correlation ($r = 0.1/0.15$), while the highest correlation was found for fivephone and sixphone surprisal values ($r = 0.90$). This finding, again, confirmed that larger n-phone context was not more informative than smaller n-phone dependencies for an analysis of the relation between phonetic encoding and ID.

In addition, data sparsity increased with increasing n-phone size, i. e., we found a considerably larger number of missing values in the higher order n-phone LMs when matching the surprisal values using n-phone keys. Missing values were due to discrepancies in transcriptions between the g2p tool of German Festival (Möhler et al., 2000) and the g2p tool implemented in WebMaus (Kisler et al., 2017) which was used for forced alignment. Non-native speech sounds were mostly not included in the LM, but were prominent in the analyzed audio data, for instance in foreign names. When these phonemes appeared in the phonemic transcription of the analyzed data respective n-phones could not be matched with the n-phones used in the LM. In addition, data sparsity occurred when the order of preceding or following context was

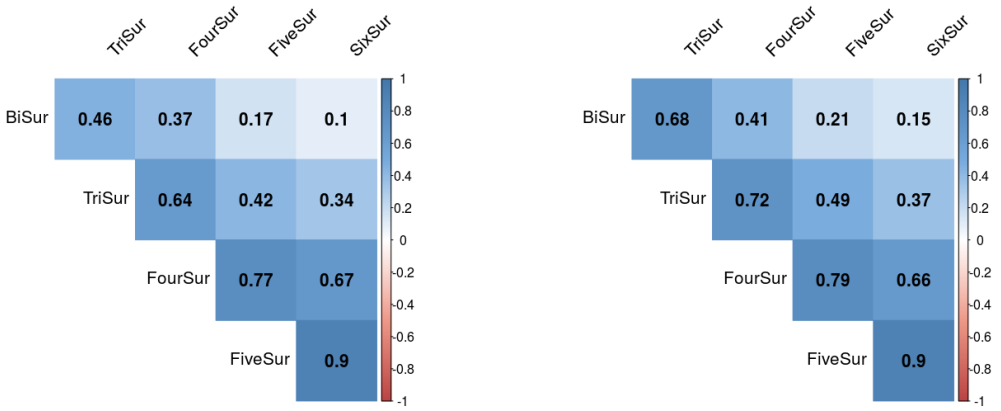


Figure 4.2: Correlation matrix (Pearson's r) for surprisal values based on LMs with word boundary (left), and without word boundary (right). Surprisal of the preceding context with n-phone order biphone (BiSur), triphone (TriSur), fourphone (FourSur), fivephone (FiveSur), and sixphone (SixSur) are given.

larger than the number of preceding or following phones present in the audio files. This concerned segments in words at the beginning or end of a sentence, or segments in words at phrase break. Another reason for larger n-phones having a larger number of missing values was that they included all missing values of their respective n-phone size below, i. e., all missing values in biphone surprisal were included in triphone surprisal and so on.

The finding that surprisal based on small n-phone orders showed stronger positive correlations with phonetic correlates than surprisal based on larger n-phone orders was not an artifact of the increasing missing values for surprisal of higher n-phone orders. We also tested the correlation between biphone, triphone, fourphone and fivephone surprisal with vowel dispersion in a smaller corpus in the analysis on vowel dispersion in Bulgarian L2 speakers of German (Section 8.4). We analyzed six different vowel phonemes in 91 words, adding up to about 132 data points per speaker. The total number of data points was 796. The variability in n-phone grew with size, as expected. There were 48 different biphones for these data points, 64 triphones, 98 fourphones, and 111 fivephones. We replicated our results on the n-phone size analysis of a larger corpus of German with triphone surprisal of the preceding context showing the strongest correlation with vowel dispersion ($r = 0.23, t(795) = 6.58, p < 0.001$) compared to biphone ($r = 0.14, t(794) = 3.84, p < 0.001$), fourphone ($r = 0.18, t(776) = 5.13, p < 0.001$), and fivephone surprisal of the preceding context ($r = -0.002, t(698) = -0.05, p = 0.96$). Judging from the degrees of freedom of the t-statistics from the Pearson's correlation test one can see that there were no missing values in surprisal for biphone surprisal, and that the number of missing values remained quite low up until fourphone surprisal, and then increased noticeably for fivephone surprisal.

4.3 Statistical modeling procedure

We calculated linear mixed-effects models (LMMs) or generalized linear mixed-effects models (GLMMs) for all analyses in this thesis with the R (R Development Core Team, 2008) packages "lme4" (Bates et al., 2015) and "lmerTest" (Kuznetsova et al., 2017) which gives additional p-values calculated based on Satterthwaite's approximations. These models incorporate both fixed and random effects in their model structure to evaluate the conditional mean of the dependent variable. The main difference between random and fixed effects is that fixed effects are repeatable, while random effects are not repeatable. The model does not generalize to the populations that are defined in the random structure. However, fixed effects estimate contrasts in the mean of the dependent variable which can be generalized to the defined factor levels (Baayen, Davidson, et al., 2008).

The random structure of the models built in this thesis included random intercepts and random slopes. Random intercepts describe the difference between the average predicted response for the fixed effects and the responses predicted by different levels of the random structure. For instance, it estimates the difference between the average population level and effects due to individual differences between subjects. Random slopes estimate the variance introduced by fixed effects per random effects. This means that in addition to differences introduced by the individual-level effects of random intercepts, these effects may vary with respect to specific fixed effects. In order to capture this effect, random slopes for fixed effects per random effects were used in the models. We used the notation of correlated random intercept and slope $(x + (x|g))$ which is the default in lme4. It assumes that the coefficients of same random effects terms are correlated (Bates et al., 2015). This strategy allowed us to discard random slopes from the model structure if they were perfectly correlated with the random intercept.

Prior to model building, we performed a collinearity analysis for the fixed effects. This analysis identified dependencies between fixed effects, and led to the exclusion of a fixed effect if it was correlated with one or more other fixed effects. For instance, unigram phoneme probability was tested in some models in this thesis as a possible predictor for segmental variability, but it was moderately correlated with surprisal values estimated on both context directions and was therefore excluded from the analyses.

For all analyses, we first built baseline models, and in a second step interaction models. We included interaction terms for the prosodic factors and surprisal, if surprisal was a significant predictor in the model. Model comparisons of the baseline model and the model including the interaction term were performed via ANOVA tests. This test compares the deviance of a model containing the interaction term to the baseline model which is otherwise identical in random effects structure.

For the LMMs in this thesis, we used a backward model selection procedure with maximal random structure to identify the largest converging model. We included

random intercepts for the random effects, and random slopes for all fixed effects. Due to convergence errors we reduced the maximal model structure stepwise. First, random slopes were removed from the model structure, and then, if necessary, random intercepts. We also removed the random slope when perfect correlations were found for the random slopes and random intercepts. These random slopes were degenerate and did not contribute to the model.

Because of the categorical nature of the dependent variable in the segment deletion analysis (produced vs. deleted) we calculated GLMMs in Chapter 6. We ran the *glmer* function and selected the binomial distribution of the data and the logistic link function by using "binomial" as the model family. In this analysis, forward model selection procedure was applied to identify the largest converging model because GLMMs are less robust against many predictor variables than LMMs. We enlarged the fixed structure by adding fixed effects in a stepwise fashion. The random structure for the segment deletion model did not converge including random slopes.

After calculating the statistical models we checked for homoscedasticity by visual inspection of the residuals. Reported models in this thesis did not reveal any obvious deviations from homoscedasticity or normality.

In order to estimate the effect size of the models we used functions within the R package "MuMIn" (Bartoń, 2017). This function gives both the marginal and conditional R^2 for a model. Marginal R^2 calculates the effect size of all fixed effects, while conditional R^2 informs about the overall model effect size including both fixed and random effects. The effect size of individual fixed effects was estimated subtracting the marginal R^2 from the model including the fixed effect from a second model which were the same in model structure except for the inclusion of the fixed effect. Only effect sizes of significant fixed effects were reported.

4.4 Control factors in the regression models

Segmental variability is clearly influenced by supra-segmental structures, and vice versa (e. g., van Bergem, 1993). For that reason, each of the models investigating the influence of information-theoretic factors on segmental variability included a prosodic model with three different factors: primary lexical stress, boundary, and speech rate.

One of the main objectives of this dissertation is to shed light on the question whether segmental variability is mainly influenced by prosodic factors or information-theoretic factors, and how these two potentially interact in explaining variance in spoken language. Regarding the question whether ID or prosody has a larger influence on segment variability one could argue that the better the model that is constructed for each of the variables, the more variance this model will eventually explain, and will therefore be identified as the better predictor for spoken language variance. This is why, we aimed to build optimized models for each of the two factors, ID and prosody, for the respective research question.

In addition to the prosodic factors, we controlled for word class, average segment duration, and phonological context. They were included because of their well known impact on segmental variability. Other control variables for individual analyses are defined in the respective results sections (Part III).

4.4.1 Word class

The lexical category distinction between function words and content words has proven to be informative with regard to linguistic variability. Function words are closed-class words that have grammatical roles. Prepositions, pronouns, conjunctions, auxiliary verbs, and grammatical articles were counted as function words, while all other lexemes are defined as content words.

In their durational analysis, Bell et al. (2009) found that content and function words differ in their sensitivity towards different ID factors. Word frequency and conditional probability have different effects on both lexical classes (Section 2.2.1). Since this thesis also attempts to tease apart predictability from frequency effects we decided to control for word class in our analyses. As a general rule, we decided to exclude function words from the analysis, and only report results on content words. This procedure assured that we excluded a confounding variable from our analysis. If the data set was relatively small, as described in sections on VOT (Chapter 7), and vowel dispersion of Bulgarian L2 speakers of German (Section 8.4), word class was included as a control factor in the analysis. Then, the analysis was performed on the entire data set including both lexical classes. We did not include word class as a control factor in the analysis of vowel dispersion in six languages (Section 8.3).

4.4.2 Word frequency

Contextual predictability can be calculated on a variety of linguistic levels. In this thesis we decided to train LMs on phonemes to explain segmental variability. Predictability measured on the basis of other linguistic levels certainly also plays a role in explaining local variability in speech segments. Previous work has mostly focused on calculating predictability based on preceding syllables or words (Aylett and Turk, 2006; Demberg and Keller, 2008; Tanner et al., 2017). For that reason, we have not only tried to include random intercept for word in our statistical models whenever possible, but also used word frequency as an additional control factor.

Word frequency is well established as a predictor of segmental variability (e. g., Gahl, 2008; Jurafsky, Bell, Gregory, et al., 2001; Pluymaekers et al., 2005b; Zhao and Jurafsky, 2009). It is not a predictability measure and correlates highly with word class (Bell et al., 2009). Word frequency and n-phone surprisal are largely independent of each other. We only found very weak negative correlation between the two factors indicating that high-frequency words had low n-phone surprisal values, in contrast to the more pronounced relationship between phoneme probability and

surprisal which was tested in some models. We estimated German word frequency from the training corpus of SDeWaC, the same corpus that was used for the German LM. Due to its positive skewness word frequency was log-transformed.

4.4.3 Average duration

Average duration is known to be a strong predictor of segmental variability. Average duration depends on the phonemic identity of the segment. Vowels and consonants are known to behave differently in their durational characteristic under stress (Burdin and Clopper, 2015; Klatt, 1976) or variations in speech rate (Miller and Volaitis, 1989). Systematic durational differences between vocalic qualities due to tenseness, phonemic status (monophthong vs. diphthong), or vowel height are well known (Möbius and von Santen, 1996). Tense vowels tend to be longer than lax vowels. Monophthongs are generally shorter in duration than diphthongs. Due to physiological reasons open vowels are longer than closed vowels.

Consonantal duration, on the other hand, is mainly influenced by the manner of production (continuant vs. abrupt), the place of articulation, and voicing. Most continuant consonants are shorter than abrupt consonants involving a closure phase in the articulatory process. As a general rule, consonants produced further in the back of the oral cavity tend to have a longer duration than those produced with a more anterior place of articulation. Voiceless plosives are longer than voiced plosives including all phases of their production (Möbius and von Santen, 1996).

We used average segment duration as a control factor in our analyses of segmental duration, vocalic spectral characteristics and voice quality. Average segment duration was identified as a strong predictor of segment duration in previous studies (Gahl, 2008, 2012). Assumingly, this factor will explain a lot of the variability in the data, and is therefore an indispensable control factor when investigating the effects of information-theoretic factors on duration. In the models for vocalic spectral characteristics we included average vowel duration as a factor to control for formant change due to durational differences between vowel phonemes. Following the "H and H" theory (Lindblom et al., 1990) we expect vowel distinctiveness and magnitude of formant movement to increase with vowel duration.

4.4.4 Phonological context

We controlled for variance introduced by the segmental environment by including the identity of the following and preceding segment into the random structure of our statistical models. Due to coarticulatory effects phonological context predicts segmental variability in the spectral and temporal domain. Preceding and following phonological context influence the articulation of segments. For instance, formant trajectories move towards a target frequency as an obstruction is produced. The place of articulation of the stop consonants defines the target frequency of the formants.

In perception experiments, listeners use cues of formant trajectories to identify the consonant identity when the consonant is not present in the stimuli (Delattre et al., 1955). In addition, phonological voicing of neighboring segment influences segmental duration (Lisker, 1978; Stevens and House, 1961).

Another reason for including preceding and following context into our model structure was that information-theoretic variables on the segment level mirror language-specific phonotactic structures to a certain extent. Therefore, phonological context was part of the random model structure to control for potential collinearities between surprisal and phonological context.

Factor levels of phonological context for the analyses of segment duration and deletion, VOT, and voice quality were defined using the phonological categories based on manner of articulation: obstruent, sibilant, and sonorant. Pause was included as a fourth factor level. In the models for vowel dispersion and dynamic formant trajectories, phonological context was defined based on the place of articulation to control for differences in formant movement because of the known impact of place of neighboring consonants. Three levels of place of articulation were used for consonants: coronal, dorsal, and labial. If context consisted of a pause or vocalic context, this was also added to the factor levels. Including pause in the factor levels of context caused an overlap between the factors prosodic boundary and phonological context. Information on preceding and following context did not include word boundary.

4.4.5 Primary lexical stress

Prominence was a binary factor using primary lexical stress (stressed vs. unstressed) based on the corpus text. For monosyllabic words, function words were counted as unstressed, whereas content words were identified as stressed.

We decided against using the provided accent and boundary labels in the Siemens Synthesis corpus because the label information was marked on the word and was therefore not optimal for analyses on the sub-word level. For instance, if a word carried an accent, irrespective of the type of accent, all syllables were marked as carrying that accent, although typically only the syllable with primary lexical stress can also be accented in a phrase. Also, we aimed at keeping the factor definition constant across all analyses in this thesis. Since neither BonnTempo corpus, nor PhonDat2, or EUROM-1 contain comparable accent labels we decided against using them to model prominence.

Secondary stress was not included in the factor definition of prominence because not all languages investigated here have secondary stress, i.e., Polish, Czech and French. With regard to the BonnTempo corpus the number of words in the corpus carrying secondary stress would be sparse. For the sake of cohesiveness of this thesis, we kept the binary definition of the factor levels for stress constant across all analyses. Admittedly, this approach might be limiting. Including secondary stress might reveal

more fine-grained differences between segmental characteristics with different degrees of prominence.

4.4.6 Prosodic boundary

As part of the prosodic factors included in this thesis we used prosodic boundary with three factor levels following Aylett and Turk (2006):

- No following prosodic boundary
- Following word boundary
- Following phrase boundary

Segments were marked as preceding a word or phrasal boundary if theses boundaries were in their immediate neighboring context. All other segments were marked as not standing in a boundary position. While Aylett and Turk (2006) marked all pauses larger than 100 ms as high probability of qualifying as a phrase boundary, we only defined annotated pauses as phrase boundary markers in our model of prosodic boundary. Since all segment labels and boundaries were manually checked in our corpora we used this approach over automatically extracting pauses that meet the criteria of $> 100\,ms$. In the prosodic hierarchy, the phrase boundaries used here are more likely to mark full intonational phrases (IPs) or intermediate IPs than in the approach described in Aylett and Turk (2006). Our model of prosodic hierarchy is a simplified model and can by no means capture the complexity of this hierarchy.

4.4.7 Speech rate

Speech rate can be interpreted as an integral part of the prosodic structure of speech. There is disagreement about which measure to use when calculating speech rate (Morgan and Fosler-Lussier, 1998; Ramus, 2002). We decided to measure speech rate as articulation rate excluding pauses. Although articulation rate is the more precise technical term, the global term speech rate is used in the thesis. We calculated phonemes per second on the sentence level for global speech rate, and phonemes per second on the word level for local speech rate. As a linguistic unit of speech, we decided to use phonemes because the corpora were manually verified on the phoneme level. This was the most accurate information about speech rate available for the current analyses. Global and local speech rate were mean-centered, separately for each speaker.

In the cross-linguistic analysis of vowel dispersion (Section 8.3), we did not use laboratory speech rate, but intended speech rate with three different levels: normal, slow, and fast. The BonnTempo corpus lent itself to being utilized in that way because it contains intended speech rate deviations.

We introduced the data analysis procedures regarding the preprocessing of speech material, language modeling, as well as statistical modeling in this chapter. Before we present the results of this thesis in Part III, we summarize all analyses conducted, their materials, measures, and their control factors in the following Table 4.2.

Table 4.2: Overview of analyses, their materials and methods. DurAverage = average duration of segments, PreVoicing = phonological voicing of preceding segment, Wordfreq = word frequency, Syllfreq = syllable frequency, Wordprob = word probability, SentencePos = sentence position, PhStatus = phonemic status.

Corpus analysis	Corpus	Measure	Control factors
Segment duration	Siemens Synthesis corpus	• Duration	• Surprisal • Wordfreq • Stress • Boundary • Global tempo • DurAverage • PreVoicing
Segment deletion	see above	• Deleted vs. produced	• Surprisal • Wordfreq • Stress • Boundary • Global tempo • Sound class
VOT	see above	• positive VOT	• Surprisal • Stress • Global tempo • Local tempo • Voicing
Vowel dispersion in German	see above	• Vowel dispersion	• Surprisal • Wordfreq

Continued on next page.

Table 4.2 Continued from previous page.

Corpus analysis	Corpus	Measure	Control factors
			• Stress • Boundary • Global tempo • Local tempo • DurAverage
Vowel dispersion in 6 languages	BonnTempo corpus	see above	• Surprisal • Stress • Boundary • Intended tempo • Vowel identity
Vowel dispersion of L2 speech	EUROM-1	see above	• Surprisal • Word class • Tenseness • DurAverage
Dynamic formant trajectories	Siemens Synthesis corpus	• Formant slopes • VL • VSL • Formant velocity • DCT coefficients	• Surprisal • Wordfreq • Stress • Boundary • Global tempo • Local tempo • DurAverage

Continued on next page.

Table 4.2 Continued from previous page.

Corpus analysis	Corpus	Measure	Control factors
			• PhStatus
Spectral similarity	PhonDat 2	• Mel-cepstral distortion	• Surprisal • Wordprob • PD2 Wordfreq • PD2 Syllfreq • Stress • Global tempo
Voice quality	Siemens Synthesis corpus	• CPP • CPPS • OQs • DEOPA	• Surprisal • Wordfreq • Stress • Boundary • Global tempo • Local tempo • DurAverage • SentencePos
Perception experiment	see above	• d'	• Surprisal • Order

Part III

Results

Part III Results: Structure

Part III of this thesis contains the results of the production analyses of segmental variability and its relation to ID and prosodic factors, as well as the results of the perception experiment in which we violated ID expectations in the implementation of phonetic detail via cross-splicing. This part consists of eight chapters:

1. Segment duration
2. Segment deletion
3. Voice onset time
4. Vowel dispersion
5. Dynamic formant trajectories
6. Spectral similarity of vowels
7. Voice quality
8. Perceptual sensitivity of violated ID expectations

Each chapter outlines the method of the data analysis, the results based on descriptive and inferential statistics and their interim discussion. Most of the chapters are dedicated to one acoustic-phonetic measure analyzed on one data set.

There are some chapters, however, that contain multiple analyses. For segment deletion (Chapter 6), we performed separate investigations for /t/ (Section 6.2.2) and /ə/ deletion in German (Section 6.2.3). Vowel dispersion (Chapter 8) was analyzed as a function of ID and prosody in German (Section 8.2), in a cross-linguistic analysis (Section 8.3), and in L2 German of Bulgarian speakers (Section 8.4). Multiple metrics of magnitude and curvature of formant change were used in Chapter 9. After performing a collinearity analysis we identified the following dynamic formant metrics for further analysis:

- Formant change measures (Section 9.2.2)
  - vector length (VL)
  - vowel section length (VSL)
  - F1/F2 slope
  - F1 velocity
- Parametric measure (Section 9.2.3)
  - F2 DCT2

In chapter 11, we subsume two analyses of the impact of ID and prosody on voice quality. First, the results of a cepstral analysis of German vocalic voice quality features are presented (Section 11.1). Second, we outline how ID and prosody affect the EGG signal analyzing open quotient smoothed (OQs) and DEOPA values (Section 11.2).

The last chapter of this part contains the experiment design, the results and their discussion of a perception test (Chapter 12). We present how experiment items were manipulated, and the experiment was set up (Section 12.1.1), alongside information about the participants who took part in the experiment (Section 12.1.2). This chapter also contains a durational and spectral post-hoc analysis of the target items in comparison to their respective baselines (Section 12.2.3).

Chapter 5

Segment duration

Duration of different linguistic units has been widely investigated in relation to a diverse range of concepts of ID (e. g., Bell et al., 2009; Tily et al., 2009). The following chapter outlines an analysis of segment duration in German and its relationship with ID and prosodic factors. We aimed at replicating findings of previous studies as a baseline for other analyses in this thesis.

5.1 Method

Segment duration was measured using the segment boundary information in the annotations for the Siemens Synthesis corpus. Forced-aligned segmentation using WebMaus (Kisler et al., 2017) was manually verified by phonetic experts who showed strong inter-rater reliability (Section 4.1). Only content words were included in the analysis because function and content words have been shown to be affected by different ID factors (Bell et al., 2009). All segments were included in the analysis except for inserted segments because they were not part of the underlying phonological structure. Plosives following a pause were also excluded from the duration analysis because their closure phase cannot be recognized in the visualizations of the audio signal. Furthermore, non-native phonemes were not considered in the analysis, such as /w/ or /ð/, because they did not have corresponding n-phones in the LM. In total, 175,652 segments from two speakers were analyzed.

5.2 Results

5.2.1 Descriptive statistics

Figure 5.1 shows segment duration per phonemic category in the corpus. The diphthongs /aɪ/ ($M = 129\,ms, SD = 28\,ms$), /ɔʏ/ ($M = 139\,ms, SD = 27\,ms$) and /aʊ/

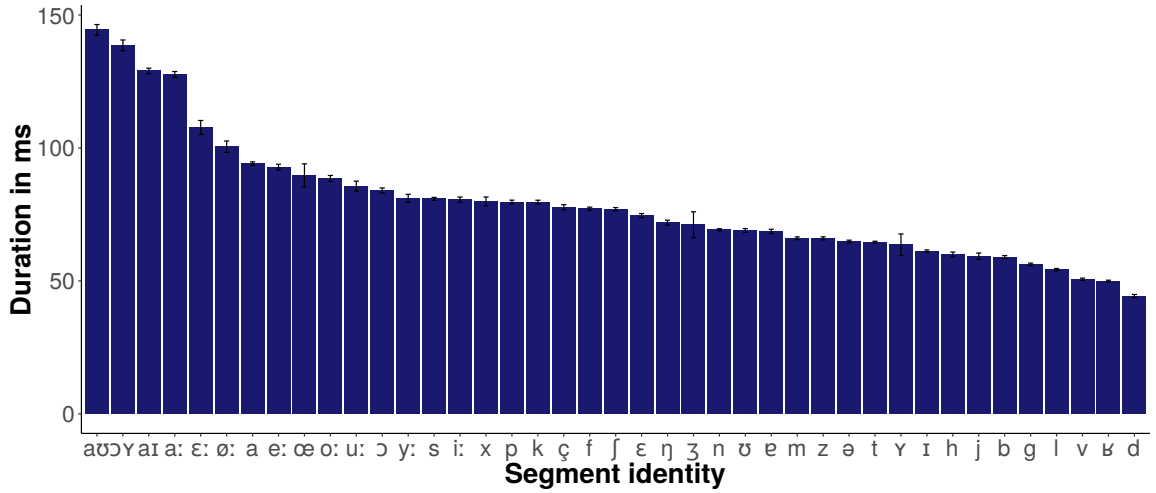


Figure 5.1: Segment duration (ms) in descending order, differentiated by phonemic category.

($M = 144\,ms, SD = 36\,ms$) were the longest segments, while /v/ ($M = 51\,ms, SD = 13\,ms$), /ʁ/ ($M = 50\,ms, SD = 15\,ms$) and /d/ ($M = 44\,ms, SD = 20\,ms$) were the shortest segments in the corpus. Typically, segment duration was positively skewed and therefore log-transformed for the following analysis.

The data was annotated for primary voicing of the preceding phonological segment, primary lexical stress, boundary information, and articulation rate of the sentence. Most segments were preceded by a voiced segment ($n = 131,334$). Voicing of the preceding context only seemed to have a subtle effect on segment duration with segments preceded by voicing ($M = 67\,ms, SD = 31\,ms$) being slightly longer than segments without preceding voicing ($M = 66\,ms, SD = 31\,ms$). The majority of the analyzed segments were found in syllables without primary lexical stress ($n = 102,823$). On average, segments in unstressed syllables ($M = 69\,ms, SD = 30\,ms$) were shorter than segments in stressed syllables ($M = 77\,ms, SD = 33\,ms$).

Three levels of prosodic boundary were defined as none, word boundary and phrase boundary (Section 4.4.6). As expected, most segments did not precede a prosodic boundary ($n = 141,773$), a considerable amount of segments was followed by a word boundary ($n = 29,768$), and the rest by a phrasal boundary ($n = 4,742$). Segment duration was the longest for segments preceding a phrasal boundary ($M = 111\,ms, SD = 42\,ms$), shorter for segments before a word boundary ($M = 73\,ms, SD = 33\,ms$), and even shorter for segments without immediate prosodic boundary following ($M = 71\,ms, SD = 30\,ms$).

5.2.2 Linear mixed-effects modeling

The information-theoretic variables included in the analysis were word frequency, unigram phoneme probability, as well as biphone and triphone surprisal of the preceding and following context.

Prior to model building we performed a collinearity analysis for the fixed effects. There were only weak dependencies between the fixed effects of the models, except for a moderate negative relationship between biphone surprisal of the preceding context and phoneme probability ($r = -0.54, t(176200) = -268.10, p < 0.001$). Surprisal values of the same context direction but with different n-phone sizes were positively correlated. This relationship was stronger for following context ($r = 0.65, t(174110) = 356.86, p < 0.001$) than for preceding context ($r = 0.51, t(175660) = 250.80, p < 0.001$). Therefore surprisal values of the same direction of context were included in separate models. This strategy ensured that correlated surprisal values were not included in the same model.

Biphone surprisal correlated positively with segment duration. This significant correlation was stronger for surprisal of the following ($r = 0.11, t(176060) = 47.96, p < 0.001$) than for surprisal of the preceding context ($r = 0.02, t(176200) = 6.73, p < 0.001$). The same trend was also found in the Pearson's correlations between segment duration and triphone surprisal. ID values of triphone of the following context ($r = 0.11, t(174230) = 45.42, p < 0.001$) showed a slightly stronger positive correlation than for the preceding context ($r = 0.10, t(175700) = 42.05, p < 0.001$).

Two different baseline models for segment duration with biphone and triphone surprisal were run. Since the positive correlation between segment duration and biphone surprisal of the preceding context was so weak, this information-theoretic factor was not included in the LMM. When surprisal had a significant effect on segment duration interaction models were calculated investigating potential interaction effects between surprisal and prosodic factors. The three prosodic factors used here, speech rate, primary lexical stress and boundary, were entered separately as interaction terms in the models.

As fixed effects, ID factors, such as SURPRISAL and WORD FREQUENCY, prosodic factors, i. e., STRESS, SPEECH RATE, and BOUNDARY, as well as VOICING, and AVERAGE DURATION were used. Random factors were SPEAKER, SOUND, WORD, PRECEDING CONTEXT and FOLLOWING CONTEXT. The final LMM converged with an additional random slope for SURPRISAL per WORD, and STRESS per PRECEDING CONTEXT (Model structure 5.1). Interaction terms were not used in this baseline model. All categorical variables were treatment-coded. All continuous variables were log-transformed due to positive skewness, except for SPEECH RATE which was mean-centered.

$$\begin{aligned} Duration \sim\ & BiFolSur + Wordfreq + \\ & Stress + Boundary + GlobalTempo + \\ & DurAverage + Voicing + \\ & (1|Speaker) + (1 + BiFolSur|Word) + (1|Sound) + \\ & (1 + Stress|Preceding) + (1|Following) \end{aligned} \tag{5.1}$$

Biphone SURPRISAL of the following context was not a significant predictor of segment duration. However, WORD FREQUENCY was predictive of segment durations.

Table 5.1: Segment duration in German: regression coefficients, standard error (SE) and statistical output of LMM analysis including biphone surprisal of the following context.

	Terms	Coeff.	SE	t-value	p-value
ID model	Surprisal	-0.004	0.004	-1.09	= .28
	Word frequency	-0.004	0.0007	-6.29	< .001
Prosodic model	Global tempo	-0.03	0.001	-30.25	< .001
	Boundary (phrase – none)	0.48	0.01	41.39	< .001
	Boundary (word – none)	0.12	0.002	49.69	< .001
	Stress (y – n)	0.09	0.002	44.19	< .001
Other control	Average duration	1.04	0.02	48.98	< .001
	Voicing (y – n)	0.05	0.003	14.13	< .001

On average, high-frequency words contained segments with shorter durations than low-frequency words. The prosodic factors also explained variability in segment duration. Segments were longer when they appeared in a syllable with primary lexical STRESS compared to unstressed position. Segment duration increased when segments preceded a word BOUNDARY, and even further when they preceded a phrasal BOUNDARY. Regarding SPEECH RATE we found a significant negative effect on segment duration indicating that segments were reduced in duration as the SPEECH RATE increased. Expected effects were also seen for the control factors AVERAGE DURATION and VOICING. When the preceding phonological segment had VOICING, segment duration increased significantly. AVERAGE DURATION strongly predicted individual segment durations (Table 5.1).

About half of the segment duration variance was captured by the final model structure based on the conditional pseudo-R^2 ($Var = 48.71\,\%$). The strongest predictor of segment duration was AVERAGE DURATION ($Var = 30.80\,\%$), followed by the prosodic model explaining $4.85\,\%$ of model variance (BOUNDARY: $4.00\,\%$, STRESS: $0.37\,\%$, SPEECH RATE: $0.48\,\%$). WORD FREQUENCY was a less effective predictor of segment duration ($Var = 0.04\,\%$), while VOICING had an even smaller effect of predicting segment duration in the model ($Var = 0.03\,\%$). The fixed effects explained a total of $35.72\,\%$ of data variance in segment duration.

In a second LMM, the effect of triphone SURPRISAL on segment duration was tested, while keeping the model structure constant. Fixed effects and random structure were the same as for the model presented above (Model structure 5.2). We therefore expected to find the same effects for the prosodic factors, controls and the ID factor WORD FREQUENCY as observed in the LMM for segment duration including biphone surprisal. Instead of biphone surprisal of the following context both triphone

Table 5.2: Segment duration in German: regression coefficients, standard error (SE) and statistical output of LMM analysis including triphone surprisal.

	Terms	Coeff.	SE	t-value	p-value
ID model	Surprisal preceding	0.04	0.002	20.94	$< .001$
	Surprisal following	0.004	0.002	2.06	$= .04$
	Word frequency	-0.003	0.0007	-4.76	$< .001$
Prosodic model	Global tempo	-0.03	0.001	-29.86	$< .001$
	Boundary (phrase – none)	0.48	0.01	48.11	$< .001$
	Boundary (word – none)	0.13	0.003	49.91	$< .001$
	Stress (y – n)	0.08	0.002	42.67	$< .001$
Other control	Average duration	1.01	0.02	47.30	$< .001$
	Voicing (y – n)	0.05	0.003	16.29	$< .001$

surprisal of the preceding and following context were used in this model.

$$\begin{aligned} Duration \sim\ & TriSur + TriFolSur + Wordfreq + \\ & Stress + Boundary + GlobalTempo + \\ & DurAverage + Voicing + \\ & (1|Speaker) + (1 + BiFolSur|Word) + (1|Sound) + \\ & (1 + Stress|Preceding) + (1|Following) \end{aligned} \tag{5.2}$$

We found the same significant effects of the fixed effects that were found in the baseline LMM for segment duration with biphone surprisal (Table 5.2). In addition, there were significant effects of triphone SURPRISAL of both the preceding and the following context. Segment duration increased with increasing triphone SURPRISAL of both context directions.

Regarding the effect size of the model, marginal pseudo-R^2 indicating how much data variance was explained by the fixed effects added up to 36.14 %, while the entire model explained 50.94 % variance in segment duration. AVERAGE DURATION was by far the most powerful predictor ($Var = 30.64\,\%$), followed by prosodic factors with 4.98 % effect size (BOUNDARY: 4.15 %, SPEECH RATE: 0.67 %, and STRESS: 0.16 %). SURPRISAL added 0.16 % explained variance to the model. The additional ID factor WORD FREQUENCY had a stronger effect than SURPRISAL ($Var = 0.19\,\%$). VOICING of the previous context also had a small effect in explaining segment duration ($Var = 0.15\,\%$).

All interaction models performed significantly better than the baseline model, except for the model containing the interaction term between SURPRISAL of the following context and SPEECH RATE ($\chi^2(1) = 3.39, p = 0.06$) (Table 5.3). Both triphone SURPRISAL of the preceding and the following context interacted positively with primary lexical STRESS in explaining variability in segment duration. Segments under

Table 5.3: Segment duration model: interaction of triphone surprisal with prosodic factors.

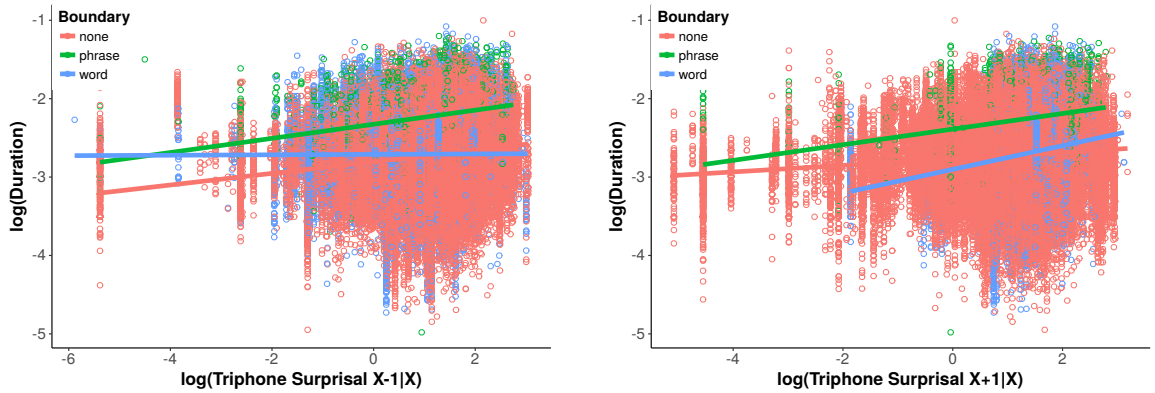
Context	Terms	Coeff.	SE	t-value	p-value
Preceding	Surprisal * Speech rate	0.007	0.001	6.33	$< .001$
	Surprisal * Stress	0.03	0.003	11.51	$< .001$
	Surprisal * Boundary (Phrase)	-0.03	0.005	-5.90	$< .001$
	Surprisal * Boundary (Word)	-0.04	0.002	-18.49	$< .001$
Following	Surprisal * Stress	0.02	0.002	8.92	$< .001$
	Surprisal * Boundary (Phrase)	0.03	0.005	6.76	$< .001$
	Surprisal * Boundary (Word)	-0.02	0.004	-4.72	$< .001$

stress and high surprisal had longer duration than unstressed segments in low surprisal triphone contexts. The interaction terms added only little to the overall explained variance (preceding context: $Var = 0.02\,\%$, following context: $Var = 0.01\,\%$).

High SURPRISAL values of the preceding context and an increased SPEECH RATE complemented each other in predicting longer segment duration. This finding was somewhat unexpected since we assumed to find shorter durations under high SPEECH RATE, even if the triphone SURPRISAL of that segment was high. The interaction term, however, showed a moderate negative correlation with the main effect for SPEECH RATE ($r = -0.67$) which required cautious interpretation of the effect direction of the interaction term.

With regard to the LMM including triphone SURPRISAL of the preceding context we observed that the regression estimates for both interaction terms for SURPRISAL and BOUNDARY were negative in the interaction LMM. The interaction plot, however, revealed that SURPRISAL and phrasal BOUNDARY interacted positively with each other when predicting segment durations (Figure 5.2a). The LMM result was possibly due to the weak negative correlation between the interaction term and the fixed effect of triphone SURPRISAL of the preceding context ($r = -0.22$). For the interaction term of word BOUNDARY and triphone SURPRISAL, we found, indeed, a negative relationship in predicting segment durations, both visible in the interaction plot and in the LMM regression estimates. This finding was possibly caused by moderate to strong negative correlations between the fixed effect word BOUNDARY and the interaction term ($r = -0.61$), as well as between the fixed effect triphone SURPRISAL of the preceding context and the interaction variable ($r = -0.46$). The interaction term added $0.11\,\%$ to the overall explained variance of the model.

The interaction term between triphone SURPRISAL of the following context and word BOUNDARY was negative in the LMM, although when investigated in an interaction plot it became visible that there was a positive relationship between the interacting variables and segment duration. This positive relationship was most pronounced for segments preceding phrasal or word BOUNDARY compared to segments



(a) Preceding context. **(b)** Following context.

Figure 5.2: Interaction of triphone surprisal of different context directions with the factor boundary on segment duration in German.

at no BOUNDARY position (Figure 5.2b). This interaction term added 0.02 % to the overall explained variance of the model.

The analysis of model effect size showed that the LMM including an interaction term of triphone SURPRISAL of the preceding context and BOUNDARY explained the largest amount of variance in segment duration of all models tested in this analysis ($Var = 51.05\,\%$). It followed that the best model for predicting segment duration in the analyzed corpus of German read speech was based on triphone surprisal and an additional interaction term for SURPRISAL of the preceding context and BOUNDARY information, while the rest of the fixed and random structures were equal to other models tested in the analysis.

5.3 Discussion

Word frequency and triphone surprisal of both context direction were predictive of segment duration. Segments were shorter in high-frequency words and predictable contexts than in low-frequency words and in unpredictable contexts. The model including triphone surprisal had a higher performance than the model with non-significant biphone surprisal. As word frequency and n-phone surprisal did not show dependencies in the collinearity analysis both factors added to the model performance. Triphone surprisal of the following context was a better predictor than triphone surprisal of the preceding context. Word frequency was slightly more effective than triphone surprisal of the following context in explaining duration variance.

These findings were in accordance with previous studies on duration and predictability or frequency (Aylett and Turk, 2004; Bell et al., 2009; Gahl et al., 2012; Jurafsky, Bell, Gregory, et al., 2001; Tily et al., 2009), and they were a valuable addition to the studies on segment durations in the context of information theory because they were based on German which was not investigated yet in previous research. In

previous research, predictability based on the following context was also identified as a stronger predictor of duration than ID based on the preceding context (Bell et al., 2009). Unfortunately, studies on the duration of sub-lexical units usually use ID variables calculated on the preceding context (Aylett and Turk, 2004; Cohen Priva, 2015), or information content based on the unigram probability of a segment without taking contextual information into account (van Son and van Santen, 2005).

Primary lexical stress had the expected expanding effect on segment duration (Aylett and Turk, 2004). Segments were longer when they immediately preceded a word or phrase boundary compared to no boundary position. This effect was more pronounced for phrase boundaries than for word boundaries which is in line with observations on the effect of boundaries with different prosodic strength, i. e., level in the prosodic hierarchy (Turk, 2010). As expected, speech rate acceleration led to overall reduction in segment duration (Bell et al., 2009; Gahl et al., 2012). Changes in speech rate seemed to affect acoustic redundancy on the global level, while still leaving room for local modulations due to the ID profile of the utterance (Turk, 2010). Boundary was by far the strongest predictor of segment duration of all the prosodic factors. The effect of final segment lengthening when preceding a prosodic boundary (Wheeldon and Lahiri, 1997) was strongly pronounced in the data.

Prosody and ID factors had independent effects on segment duration. In both models, the prosodic factors had a higher overall effect size than the ID factors. This finding was expected considering that redundancy factors also only contributed to a small degree to the duration model performance in Aylett and Turk (2004, 2006). We also found interactions between prosodic variables and surprisal indicating that these factors complemented each other in explaining segment duration variability. Thus, our findings were in line with a weak version of the SSR hypothesis (Aylett and Turk, 2004, 2006).

The control factors average duration and phonological voicing of the preceding segment had the expected effects on segment duration (Klatt, 1976; Lisker, 1978; Stevens and House, 1961). Average segment duration was the strongest predictor of individual segment durations. This finding was expected because the average values were based on the production data itself, and both measures were moderately correlated ($r = 0.55, t(176280) = 276.19, p < 0.001$). Phonological voicing, on the other hand, was a weak predictor of segment duration. This was probably due to a considerate amount of overlap between the random intercept for preceding phonological context and the fixed effect voicing.

This chapter presented the German segment duration analysis using the Siemens Synthesis corpus. The following chapter outlines the results of the segment deletion analysis performed on the same corpus. We performed separate statistical analyses for /t/ and /ə/ deletion in German.

Chapter 6

Segment deletion

While the impact of ID on segment duration has been investigated thoroughly, there are only a few studies focusing on how ID factors, especially contextual predictability, influence segment deletion (e.g., Cohen Priva, 2015; Hume, 2004; Jurafsky, Bell, Gregory, et al., 2001; Tanner et al., 2017). This chapter presents an analysis of segment deletion in German, with a focus on /ə/ and /t/ deletion, and its relationship with ID and prosodic factors. This deletion analysis was presented in the form of a one-page abstract in a conference proceedings in Brandt et al. (2017a), and was extended and revised for this thesis.

6.1 Method

The process of segment deletion is gradient. Often one can find underlying articulatory gestures when there is no trace of the articulation in the fine phonetic detail of the acoustics. For the purpose of this analysis segment deletion was coded as a binary process: a segment was either realized or deleted. Any instance of surface realization was interpreted as a case of non-deletion. Deletions were defined based on the canonical transcription of the utterances: if there was a discrepancy between the expected phoneme string based on the canonical transcription and the uttered phoneme string, this was marked by annotators. The label "SOUND –" was used to mark deletion (Section 4.1). In the case of /t/ preceding a homorganic plosive there was often only one closure with accompanying burst visible in the speech signal (Figure 6.1). In these cases, /t/ was marked as being deleted and the following plosive was annotated as being produced by the speaker.

/ə/ can only be analyzed as being deleted if the underlying phonological representation of a word is interpreted as including /ə/. For that reason we included an additional factor for the analysis of only /ə/ deletion (Section 6.2.3), namely the pronunciation rules for German /ə/ from the *Duden Aussprachwörterbuch* (Kleiner

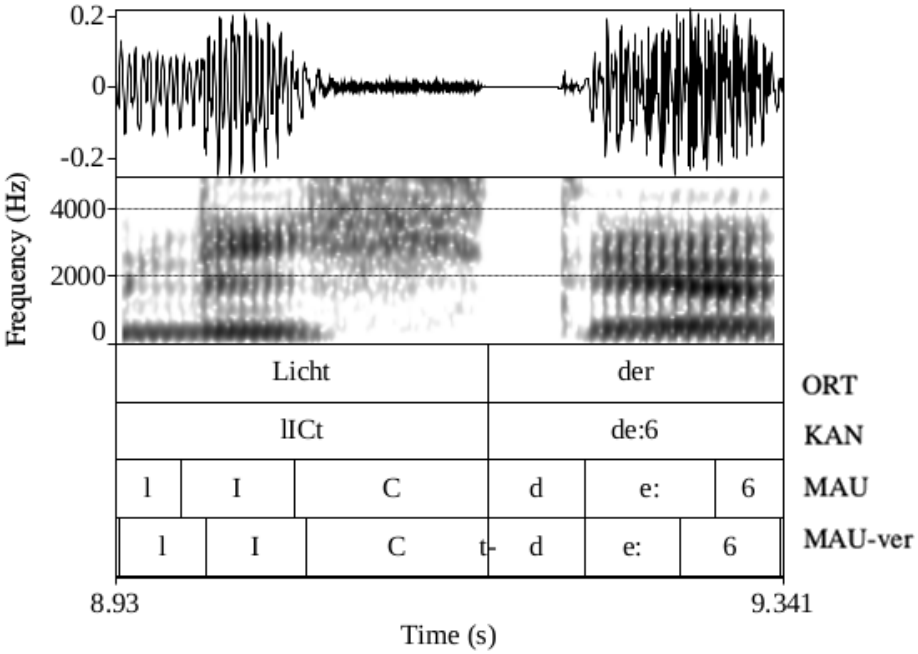


Figure 6.1: Example of /t/ deletion in context with homorganic plosive taken from the Siemens Synthesis corpus.

et al., 2015). The binary factor based on the Duden rules had two levels: deleted and produced. Duden has an exhaustive rule set of /ə/ production in German, but they do not describe /ə/ realization in word-final position. For "normal" pronunciation, one can find the following rules in the dictionary:

1. Pronunciation of /əm/
   - /ə/ is deleted after all German fricatives
   - /ə/ is produced after /p, b, t, d, k, g/ and /m, n, ŋ, l, r/
2. Pronunciation of /ən/
   - /ə/ is deleted after all German obstruents, except in diminuitive /çən/
   - /ə/ is produced in all other contexts
3. Pronunciation of /əl/
   - /ə/ is deleted after all German obstruents and nasals
   - /ə/ is produced before and after vowels, and after /ʁ/

The prosodic factors included as controls in this study were speech rate, primary lexical stress, and boundary, as described in Section 4.4. As ID factors, we included surprisal of the following and preceding context and word frequency. In addition, speech sound class with the factor levels consonant and vowel was introduced as

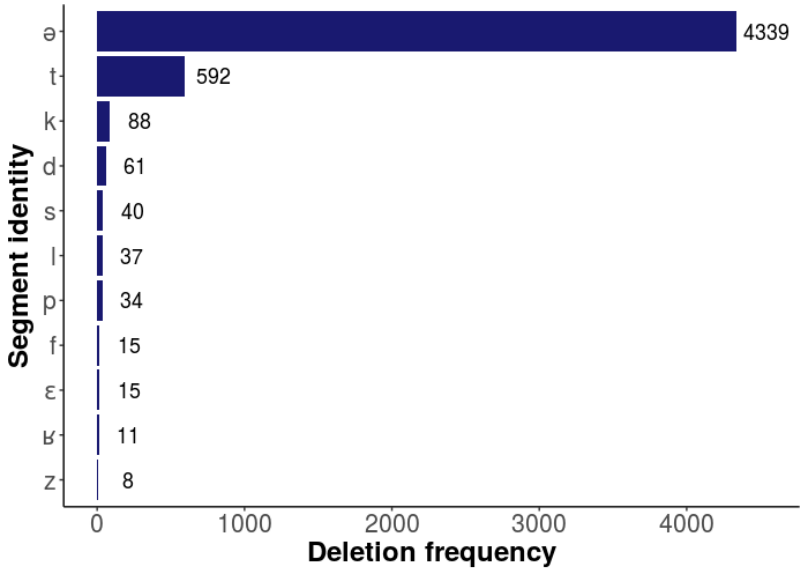


Figure 6.2: Total number of deletion per segment identity. Only deletion frequencies larger than 5 are shown.

a control factor. Preceding and following phone were defined as either obstruent, sibilant or sonorant. Pause information was also included in the factor preceding and following context (Section 4.4.4).

6.2 Results

6.2.1 Segment deletion

Descriptive statistics

The total number of analyzed segments was 182,120 with 5,269 (2.89 %) deleted segments. The majority of the deleted segments were /ə/ ($n = 4,339$) and /t/ ($n = 592$) (Figure 6.2). Vowels made up the majority of deleted segments ($n = 4,367,83\,\%$), while only 902 consonants were deleted (17.12 %). 0.82 % of all consonants in the speech corpus were deleted, and 6.07 % of all vowels in the corpus were deleted. This included /ə/ deletion as being assumed to be part of the underlying phonological structure based on the canonical transcription of the Siemens Synthesis corpus.

4.43 % of the segments in unstressed position got deleted ($n = 4,772$), while only 0.67 % of stressed segments got deleted ($n = 497$). Most of the deletions appeared at no boundary position ($n = 4,947,93\,\%$). At word boundary, segments got deleted in 1.04 % of all cases ($n = 314$), while it was very rare that segments were not produced when a phrasal boundary followed ($n = 8, 0.16\,\%$). Speaker "wo" had higher deletion rates ($n = 2,841$) than speaker "ai" ($n = 2,428$). With regard to phonological context segments were always marked as being produced when a pause preceded, and also in

most cases when a pause followed ($n = 9, 0.20\,\%$). Deletion rates for preceding context obstruents and sibilants were equally high at around $6.50\,\%$, while segments got rarely deleted when a sonorant preceded ($n = 876, 0.80\,\%$). Regarding following context, we found highest deletion rates for following sonorants ($n = 4,470, 4\,\%$), and similar deletion rates for following obstruents and sibilants ($n = 476, 1\,\%; n = 314, 1\,\%$).

Since the phonemes /ə/ and /t/ were deleted most frequently in the corpus and both speech sounds are known to delete under specific circumstances (Section 2.2.2), separate analyses were conducted for both speech sounds in Sections 6.2.3 and 6.2.2.

Generalized linear mixed model

Running a correlation analysis with *pairs.panels* from the R package *psych* (Revelle, 2017), we found the strongest relationship between biphone surprisal of the following context and deletion ($r = -0.19$). With increasing n-phone size (triphone) this relationship decreased in strength ($r = -0.11$). This finding was replicated for surprisal of the preceding context. Biphone surprisal ($r = -0.11$) showed a little stronger negative correlation with deletion than triphone surprisal ($r = -0.10$). Therefore, we used biphone surprisal as an ID measure in the following analyses.

For statistical analysis, GLMMs were calculated using "lme4" (Bates et al., 2015) and "lmerTest" (Kuznetsova et al., 2017). Categorical factors were treatment-coded. The continuous variables surprisal and word frequency were log-transformed due to positive skewness. Surprisal of the following and preceding biphone context showed a low positive correlation ($r = 0.26$). Weak correlations were found for word frequency and stressed segments ($r = -0.16$), word boundary and segment class vowel ($r = 0.25$), word boundary and carrying primary lexical stress ($r = 0.11$).

Forward model selection method was applied resulting in a final model with biphone SURPRISAL of the following and the preceding context, STRESS, SPEECH RATE, BOUNDARY, WORD FREQUENCY and SOUND CLASS as fixed effects and random intercepts for SPEAKER, PRECEDING CONTEXT and FOLLOWING CONTEXT. Including WORD as additional random intercept led to convergence errors of the model (Model structure 6.1).

$$\begin{aligned} Deletion \sim\ & BiSur + BiFolSur + Wordfreq + \\ & Stress + Boundary + GlobalTempo + \\ & SoundClass + \\ & (1|Speaker) + (1|Preceding) + (1|Following) \end{aligned} \tag{6.1}$$

All factors were significant in explaining variability of segment deletion. Segments which were easily predictable were more likely to be deleted than segments which were more difficult to predict. This finding held for both preceding and following context. If a segment appeared in a word with high WORD FREQUENCY it was more

Table 6.1: Segment deletion in German: regression coefficients, standard error (SE) and statistical output of GLMM analysis including biphone surprisal.

	Terms	Coeff.	SE	z-value	p-value
ID model	Surprisal following	-1.36	0.02	-56.59	$< .001$
	Surprisal preceding	-1.42	0.03	-42.55	$< .001$
	Word frequency	0.07	0.006	11.02	$< .001$
Prosodic model	Global tempo	0.05	0.02	2.63	$< .01$
	Boundary (phrase – none)	-3.19	0.94	-3.37	$< .001$
	Boundary (word – none)	-0.82	0.06	-12.76	$< .001$
	Stress (y – n)	-1.54	0.05	-28.55	$< .001$
Other control	Sound class (V – C)	1.35	0.04	30.76	$< .001$

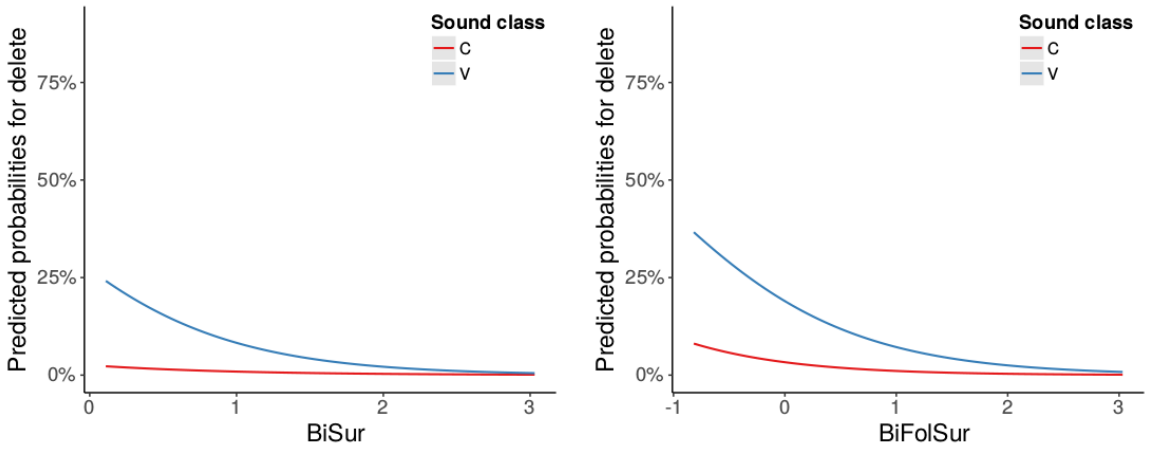
likely to be deleted than in a low-frequency word. Regarding the prosodic factors, at slower SPEECH RATE segment deletion rate was significantly lower than at fast tempo. Segments in syllables that did not carry STRESS were more likely to be deleted than segments in syllables with primary lexical STRESS. At word and phrasal BOUNDARY position, segments had lower deletion rates than at no BOUNDARY position. Vowels were more likely to be deleted than consonants (Table 6.1).

The regression estimates for SURPRISAL of the preceding and following context were similar. Investigating the marginal effects of the model predictors, however, showed that at low SURPRISAL of the following context there was a higher predictability to delete than at low SURPRISAL for the preceding context which was also mirrored in the previous correlation analysis (Figure 6.3).

The random intercept for SPEAKER explained 0.02 % ($SD = 0.12\,\%$) of the variance in the deletion rates, while random intercept for PRECEDING CONTEXT explained 3.76 % ($SD = 1.94\,\%$) and random intercept for FOLLOWING CONTEXT explained 0.73 % ($SD = 0.85\,\%$) of the data variance.

The individual level effects of the random intercept for SPEAKER showed that, on average, speaker "wo" produced more deletions than speaker "ai" ($\beta = 0.09; \beta = -0.09$). Segments had the highest deletion rates if their preceding phone was a sibilant or an obstruent. Segments were less likely to delete if their preceding context was sonorant or a pause. Regarding following phonological context the effects of factor levels were exactly opposite: segments following pauses or sonorants showed higher deletion rates, while following obstruents and sibilants led to lower likelihood to delete.

Further investigation of the predicted probabilities for sound class to delete showed that both consonants and vowels were more prone to deletion when they stood in low surprisal conditions of the preceding or following context. This means that we found the same tendencies for both sound classes to delete in low ID conditions (Figure 6.3b).



(a) Preceding context. **(b)** Following context.

Figure 6.3: Predicted probabilities to delete by surprisal of different context directions and sound classes.

Updated GLMM with reduced data set

If all /ə/ deletions that were predicted by the DUDEN RULES were discarded from the data set, only 2.07 % of all /ə/ were interpreted as being deleted. This was in contrast to the high deletion rates of /ə/ (27.39 %) if DUDEN RULES were not taken into account, and /ə/ was assumed to be underlying in all phonological structures of /əm/, /ən/ and /əl/. The number of /t/ deletion was reduced from 592 to 317 when contexts with homorganic plosives were excluded from the analysis. This reduced the percentage of /t/ being deleted from 3.42 % to 1.89 % of all cases. Excluding /t/ deletions in contexts with homorganic plosives and /ə/ deletions predicted by the DUDEN RULES reduced the number of deleted segments to 868 (0.48 %). 27.76 % of the deleted segments were vowels, and the majority of the segment deletions were consonants (72.24 %). The reduced data set had a total number of 176,084 data points.

Rerunning the GLMM above for this reduced data set replicated the results presented above, except for the factor SOUND CLASS. Now, vowels were significantly less often deleted than consonants (Table 6.2).

The random intercept for PRECEDING CONTEXT explained 0.51 % ($SD = 0.72$ %) of the variance in the data, while SPEAKER explained 0.03 % ($SD = 0.17$ %). Investigation of the Best Linear Unbiased Prediction (BLUP) of the random intercept for PRECEDING CONTEXT revealed that segments were more likely to delete when preceded by a sonorant or obstruent, and had lower deletion rates when preceded by a sibilant or pause.

In a second analysis step, we tested the performance of interaction models including interaction terms of surprisal and prosodic factors compared to the baseline model. We used ANOVA tests to conduct model comparisons. Interaction models with

Table 6.2: Segment deletion in German: regression coefficients, standard error (SE) and statistical output of updated GLMM analysis including biphone surprisal.

	Terms	Coeff.	SE	z-value	p-value
ID model	Surprisal following	-0.68	0.06	-12.22	$< .001$
	Surprisal preceding	-0.47	0.07	-6.91	$< .001$
	Word frequency	-0.03	0.01	-1.61	< 0.01
Prosodic model	Global tempo	0.21	0.04	4.89	$< .001$
	Boundary (phrase – none)	-1.77	0.38	-4.61	$< .001$
	Boundary (word – none)	-0.21	0.04	-4.69	$< .001$
	Stress (y – n)	-0.38	0.08	-4.68	$< .001$
Other control	Sound class (V – C)	-0.20	0.09	-2.33	$= .02$

SPEECH RATE and SURPRISAL of the following or preceding context did not perform better than the baseline deletion model. The GLMM with an interaction between SURPRISAL of the following context and STRESS failed to converge. The only model that outperformed the baseline GLMM was the interaction model with SURPRISAL of the preceding context and STRESS ($\beta = -0.69, SE = 0.16, z = -4.42, p < 0.001$). STRESS and SURPRISAL complemented each other in predicting segment deletion in German. Segments in stressed syllables and under high surprisal were less likely to undergo deletion than unstressed segments in low suprisal contexts. This interaction term was significant, while both main effects stayed significant in the model output.

6.2.2 /t/ deletion

Descriptive statistics

/t/ was the most frequently deleted consonant in the corpus ($n = 592$). More than half of all /t/ deletions marked in the corpus were in contexts of following homorganic plosives ($n = 313, 53\,\%$). However, /t/ was not necessarily deleted when a homorganic plosive followed. In about half of all cases of following /d/, /t/ was not deleted (48.84 %). When /t/ was being followed by another /t/, it was deleted in 63.21 % of all cases.

Around 46 % of the /t/ deletions appeared when there was another consonant following ($n = 275$). And only in 4 cases, /t/ was deleted when a vowel followed (0.68 %). When /t/ occurred before alveolar consonants it showed a tendency to delete in contexts with /s/ ($n = 159$) and /z/ ($n = 22$), but not in contexts with following nasal /n/. When a non-homorganic plosive followed, /t/ showed highest deletion rates before /b/ ($n = 33$), followed by /p/ ($n = 24$).

Regarding preceding phonological context we found that most /t/ deleted when they were preceded by a sonorant ($n = 444$), or sibilant ($n = 117$), and only few when

Table 6.3: Number of produced and deleted /t/ within context of preceding (OB = obstruent, P = pause, SI = sibilant, SO = sonorant) and following phoneme (N = neutralizing, C = other consonant, V = vowel).

Context	Produced	Deleted	Context	Produced	Deleted
OB /t/ C	26	19	SI /t/ N	1,296	33
OB /t/ N	360	10	SI /t/ V	3,162	1
OB /t/ V	431	2	SO /t/ C	291	173
P /t/ N	18	0	SO /t/ N	5,940	270
P /t/ V	12	0	SO /t/ V	4,782	1
SI /t/ C	71	83	*Total*	16,389	592

obstruents preceded ($n = 31$). /t/ did not delete in the context of a preceding pause (Table 6.3).

Generalized linear mixed model

In order to test if SURPRISAL still had a significant influence in explaining deletion rates of /t/ when phonemic context was controlled for in the model we ran a separate GLMM for the subcorpus of the phoneme /t/ with biphone SURPRISAL of the following and the preceding context, WORD FREQUENCY, STRESS, SPEECH RATE, and BOUNDARY as prosodic factors, as well as random intercepts for PRECEDING CONTEXT, FOLLOWING CONTEXT, WORD and SPEAKER. Preceding phonemic context was defined as in the models above with factor levels obstruent (OB), sibilant (SI), sonorant (SO) and pause (P) ($n = 848, 4646, 11457, 30$). Following context was coded using the three levels: neutralizing (i. e., homorganic stops) (N), other consonants (C), and vowels (V) ($n = 663, 7927, 8391$) following Tanner et al. (2017). All categorical factors were treatment-coded, and all continuous variables were log-transformed. Due to convergence errors the random intercept for WORD, and then step-wise fixed effects WORD FREQUENCY and BOUNDARY were removed from the model structure.

Because the random intercept for FOLLOWING CONTEXT explained a large quantity of the variance in the data ($Var = 9.43\,\%, SD = 3.07\,\%$) we updated the GLMM using this factor as a fixed effect in the model (Model structure 6.2). FOLLOWING CONTEXT was helmert-coded resulting in the following contrasts: neutralizing segments compared to other consonants, all consonants compared to vowels. This improved model performance significantly ($\chi^2(1) = 18.54, p < 0.001$).

$$\begin{aligned} /t/ - Deletion \sim\ & BiSur + BiFolSur \\ & Stress + GlobalTempo + \\ & FollowingContext + \\ & (1|Speaker) + (1|Preceding) \end{aligned} \tag{6.2}$$

Table 6.4: /t/ deletion in German: regression coefficients, standard error (SE) and statistical output of GLMM analysis including biphone surprisal.

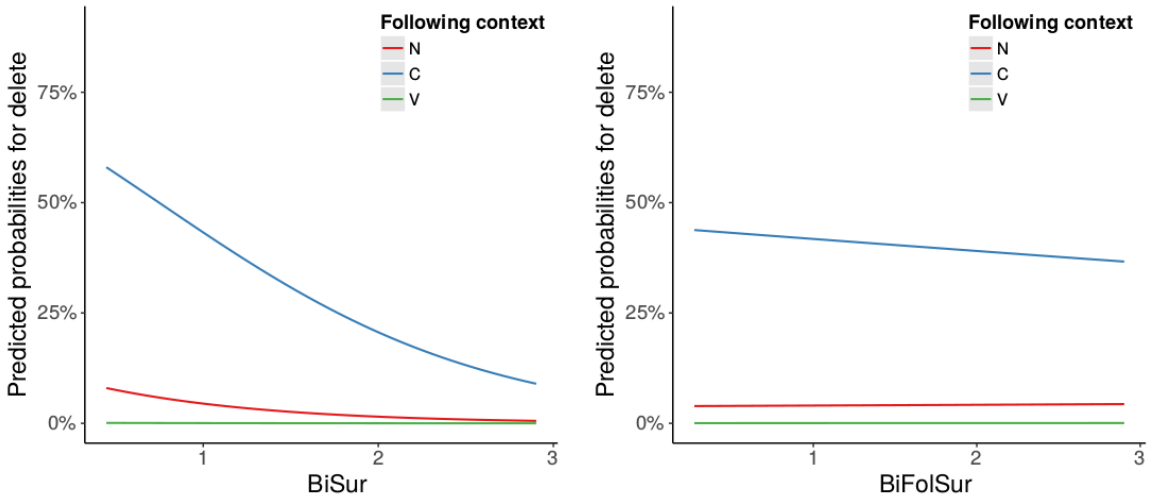
	Terms	Coeff.	SE	z-value	p-value
ID model	Surprisal following	-0.13	0.09	-1.42	$= 0.15$
	Surprisal preceding	-1.38	0.18	-7.47	$< .001$
Prosodic model	Global tempo	0.34	0.06	5.71	$< .001$
	Stress (y – n)	0.04	0.10	0.41	$= .68$
Following context	N – C	1.46	0.05	28.29	$< .001$
	C – V	-1.98	0.17	-11.78	$< .001$

This updated GLMM for the /t/ deletion analysis showed that biphone SURPRISAL of the following context was no longer significant in explaining variance of /t/ deletion. However, SURPRISAL of the preceding context reached significance level in explaining variance in the data. /t/ with low SURPRISAL of the preceding context was more likely to be deleted than with high SURPRISAL, even though preceding phonemic context was included in the model as random effect. We found the same effect for SPEECH RATE on deletion rates as for the main model: /t/ was more likely to delete at faster speech rate than at slow tempo. /t/ in neutralizing following context (N) had significantly higher deletion rates than /t/ in context of other consonants (C). /t/ with consonantal following context had significantly higher deletion rates than /t/ with following vowel (V). The predictor STRESS did not reach significance level (Table 6.4).

The marginal pseudo-R^2 showed that the majority of the variance in /t/ deletion was explained by FOLLOWING CONTEXT alone (59 %). SURPRISAL of the preceding context added only about 3 % to the explained variance. The fixed effect SPEECH RATE led to 0.5 % more variance explained. The BLUP for random intercept for PRECEDING PHONEME showed that /t/ was most frequently deleted following a sonorant, compared to all other preceding phonological contexts. The random intercept for PRECEDING PHONEME accounted for 0.04 % ($SD = 0.20$ %) of the variance in the data. Differences between the speakers in /t/ deletion explained 0.07 % ($SD = 0.26$ %) of the variance. Both fixed and random effects of the GLMM explained 64 % of the variance in /t/ deletion rates in the data.

Investigation of the partial effect plots for SURPRISAL showed that the effect of ID on /t/ deletion depended on the following context. For SURPRISAL of the following context we found no relationship between ID and deletion rates when neutralizing consonants or vowels followed /t/ (Figure 6.4b). For SURPRISAL of the preceding context, on the other hand, neutralizing and consonantal context both showed the observed effect of ID on deletion rates of /t/, while this relationship did not apply when vowels followed (Figure 6.4b).

We also built interaction models introducing interaction terms between SURPRISAL



(a) Preceding context. **(b)** Following context.

Figure 6.4: Predicted probabilities of /t/ deletion by surprisal different context directions and for different following phonological contexts (N = neutralizing, C = other consonants, V = vowels).

and prosodic factors in the baseline model for /t/ deletion. Based on the output of model comparison ANOVA tests between baseline and interaction model none of the interaction terms between SURPRISAL of the preceding context and STRESS ($\chi^2(1) = 1.61, p = 0.20$) or SPEECH RATE (($\chi^2(1) = 0.32, p = 0.57$) increased model performance significantly.

6.2.3 /ə/ deletion

Descriptive statistics

/ə/ showed a strong tendency to delete in contexts with preceding obstruent and following alveolar nasal ($n = 2081, 48\,\%$), and preceding sibilant and following alveolar nasal ($n = 1773, 41\,\%$). The majority of /ə/ deletions occurred with /l, m, n/ following ($n = 4136, 95\,\%$). In other consonantal contexts /ə/ was rarely deleted ($n = 195, 4\,\%$), and even less with following vocalic context ($n = 8, 0.2\,\%$). /ə/ was never deleted when a pause followed. In more than half of all cases of preceding sibilant, /ə/ was deleted ($n = 1841, 56\,\%$). Every fourth /ə/ was not produced when an obstruent preceded ($n = 2340, 26\,\%$), and sonorants were the least predictive of /ə/ deletion ($n = 158, 0.04\,\%$) (Table 6.5).

Generalized linear mixed effects model

For the /ə/ deletion model, an additional predictor based on the DUDEN RULES of /ə/ pronunciation was introduced (Section 6.1). It was a binary predictor with factor levels "deleted" and "produced" ($n = 5544, 10294$). For the /ə/ GLMM, the predictor

Table 6.5: Number of produced and deleted /ə/ within context of preceding (OB = obstruent, SI = sibilant, SO = sonorant) and following phoneme (C = other consonant, P = pause, V = vowel).

Context	Produced	Deleted	Context	Produced	Deleted
OB /ə/ C	4,401	174	SI /ə/ /n/	243	1,773
OB /ə/ /l/	293	44	SI /ə/ P	120	0
OB /ə/ /m/	249	37	SI /ə/ V	182	2
OB /ə/ /n/	801	2,081	SO /ə/ C	928	3
OB /ə/ P	312	0	SO /ə/ /l/	22	0
OB /ə/ V	439	4	SO /ə/ /m/	124	0
SI /ə/ C	786	18	SO /ə/ /n/	2,232	153
SI /ə/ /l/	54	36	SO /ə/ P	90	0
SI /ə/ /m/	30	12	SO /ə/ V	193	2
			Total	11,499	4,339

SEGMENT CLASS was not included because only vowels were investigated. STRESS was not included either because it was not meaningful for /ə/ which can only be unstressed in German. Fixed effects of the GLMM were biphone SURPRISAL of the following and the preceding context, SPEECH RATE, DUDEN RULES and random intercepts for SPEAKER and for PRECEDING CONTEXT coded as in the models above with the levels obstruent, sibilant, sonorant and pause ($n = 8835, 3256, 3747, 0$). Including a random intercept for WORD or FOLLOWING CONTEXT led to convergence errors, as well as including BOUNDARY or WORD FREQUENCY as fixed effects (Model structure 6.3). All categorical variables were treatment-coded, and all continuous variables log-transformed, except for SPEECH RATE which was mean-centered.

$$\begin{aligned} /@/ - Deletion \sim\ & BiSur + BiFolSur \\ & GlobalTempo + DudenRules + \\ & (1|Speaker) + (1|Preceding) \end{aligned} \tag{6.3}$$

Only SURPRISAL of the following context as well as the DUDEN RULES were significant in explaining variability of /ə/ deletion. If the DUDEN RULES predicted /ə/ deletion it was more likely to be deleted. At low biphone SURPRISAL of the following context /ə/ was more likely to be deleted than at high SURPRISAL. Neither SURPRISAL of the preceding context nor SPEECH RATE reached significance level in the model (Table 6.6).

/ə/ was most likely to delete following a sibilant or sonorant, while it was more often realized following an obstruent. The random intercept for preceding context explained 0.17 % (0.42 %) of the model variance, while speaker added 0.06 % (SD = 0.24 %) of explained variance. The marginal pseudo-R^2, indicating how much variance

Table 6.6: /ə/ deletion in German: regression coefficients, standard error (SE) and statistical output of GLMM analysis including biphone surprisal.

	Terms	Coeff.	SE	z-value	p-value
ID model	Surprisal following	-1.62	0.08	-20.74	$< .001$
	Surprisal preceding	0.04	0.06	0.61	$= .54$
Prosodic model	Global tempo	-0.03	0.04	-0.94	$= .35$
Other control	Duden rule (y – n)	4.76	0.13	36.93	$< .001$

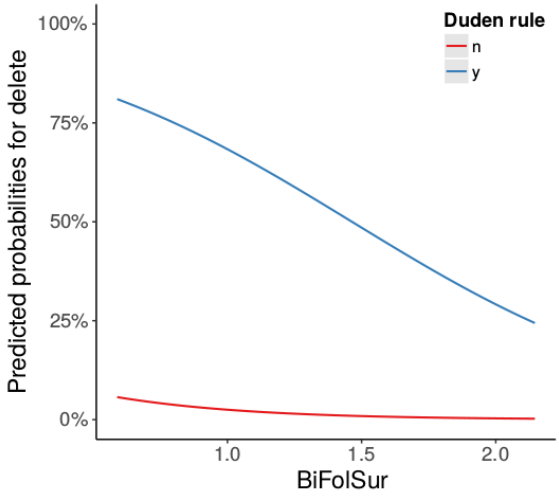


Figure 6.5: Predicted probabilities of /ə/ deletion by surprisal of the following context for different Duden rules prediction.

is explained by the fixed factors, showed that the DUDEN RULES explained 35 % of the deletion variance alone. SURPRISAL of the following context only added 2 % to the explained variance of the /ə/ deletion model. The conditional pseudo-R^2 for the variance explained by both fixed and random effects equaled 70% in the final model.

The partial effects for /ə/ deletion by SURPRISAL of the following context for different DUDEN RULES showed that the relationship between ID and deletion rates held across segments that were predicted by DUDEN RULES to be deleted and to be realized. The relationship was stronger, of course, for segments that were predicted to be deleted by the DUDEN RULES (Figure 6.5).

We also tested whether SURPRISAL interacted with SPEECH RATE in its effect on /ə/ deletion rates. We entered an interaction term between SURPRISAL of the following context and SPEECH RATE. ANOVA model comparison between the baseline model and the interaction model showed that there was no significant difference in both model performances ($\chi^2(1) = 1.47, p = 0.23$).

6.3 Discussion

In summary, low predictability of segments was predictive of reduced likelihood to delete, even after controlling for the prosodic factors primary lexical stress, boundary, and speech rate, as well as for speech sound class and the information-theoretic variable word frequency. Biphone surprisal of the following context was stronger in predicting deletion than biphone surprisal of the preceding context. This was probably due to the process of /ə/ deletion dominating the data set, and this phenomena usually appears in predictable biphone combinations, such as /əm/, /ən/ or /əl/ which are also reflected in the Duden pronunciation rules (Section 6.1).

As expected, high speech rate, absence of primary lexical stress, and no boundary position increased the likelihood of segment deletion (Cohen Priva, 2015). Segments were less likely to be deleted at boundary position then when they appeared at no boundary position. This finding can be explained by the large quantity of /ə/ deletions in the data set which usually appeared in the last syllable of the word, but not immediately at word boundary. Interestingly, there was a large variability in deletion at phrasal boundary position. Tanner et al. (2017) found that /t/ deletion rates were impacted by the duration of the following pause. Higher predictability to delete was associated with shorter pause duration, and longer pause duration decreased /t/ deletion rates. The large variability in segments deletion rates at phrasal boundary position found in this study can possibly be explained by differences in pause duration. However, these were not controlled in the study.

Vowels were significantly more often deleted than consonants when all /ə/ deletions were counted based on the canonical transcription of the Siemens Synthesis corpus, and including all /t/ deletions in contexts with homorganic plosives. If these /t/ and /ə/ occurrences were not included in the data set we found that consonants were significantly more often deleted than vowels. This is in line with previous findings of deletion rates in German conversational speech (Zimmerer, 2009). Here, underlying phonological structure was also controlled. /ə/ was the most frequently deleted vowel in our data set which agreed with studies on the deletion rates of vowels in German read and spontaneous speech (Kohler and Rodgers, 2001).

All in all, results of the main deletion GLMM were replicated when cases of /ə/ deletion predicted by the *Duden Aussprachwörterbuch* as well as /t/ preceding homorganic plosives were not included in the analysis. Low surprisal and high word frequency were predictive of segment deletion, and segments were more likely to be deleted at increased speech rate. Unstressed segments had higher deletion rates than stressed segments. Previous results of higher deletion rates for segments at no boundary position were also replicated in the updated GLMM.

Although the investigated data set only contained two speakers we found that the difference in tendency to delete between both speakers explained a small part of the likelihood of all models built in the deletion analysis. Individual speakers showed the same patterns of deletion, but differed in the frequency with which they used these

patterns.

Preceding context was used as a random effect in all GLMMs. In the main deletion model detailed investigation of the BLUP of preceding context showed that there were higher deletion rates for segments if their preceding phone was a sibilant or an obstruent. When running the GLMM with the reduced data set, excluding /ə/ deletion predicted by *Duden Aussprachwörterbuch* as well as /t/ preceding homorganic plosives, this result changed to preceding sonorants or obstruents being predictive of high segment deletion rates.

The segments /ə/ and /t/ were the most frequently deleted segments in the analysis of German read speech. Deletion rates of these segment identities both increased with low surprisal of the preceding and following context based on the main deletion model.

A separate analysis for /t/ showed that biphone surprisal of the following context was not predictive of higher deletion rates when following phonological context was included as a fixed effect in the statistical model. Following phonological context was a strong predictor of /t/ deletion explaining most of the variance in the data (59 %), while surprisal of the preceding context only added about 3 % to the explained variance. In following neutralizing context, /t/ was most likely to delete compared to other consonantal context, while following vocalic context led to lower deletion rates compared to following consonantal context. This findings was expected based from previous findings on CSD in British English (Tanner et al., 2017).

To the author's knowledge, there is only one other study that has included n-phone information-theoretic variables in a segment deletion analysis. In said study (Raymond et al., 2006), following phonological context and biphone frequency of the following context were included in the statistical model. However, the authors found a significant effect of biphone frequency which was not replicated in the current analysis on German data. This discrepancy may be attributable to differences in coding of the factor following phonological context, and the usage of mere frequency counts versus surprisal values.

In the /t/ deletion model, effects of speech rate found in the models based on the entire data set were replicated. As expected, accelerated speech rate was associated with higher deletion rates of /t/ (Raymond et al., 2006; Tanner et al., 2017). Whether a segment appeared in a stressed syllable did not influence the deletion rates of /t/. This finding stood in contrast to previous accounts which have found that /t/ in unstressed syllables had higher deletion rates than in stressed syllables (Labov, 1972). In a more detailed analysis of this effect, Raymond et al. (2006) found that primary lexical stress was only marginally predictive of /t/ deletion in codas, while it was a significant predictor for /t/ deletion in onsets. It might well be that most of the /t/ deletions in our data set were found in syllable codas which is why we saw a null effect of stress. Unfortunately, we did not control for this factor in this analysis.

In a separate analysis for /ə/, the pronunciation rules of the *Duden Aussprachwörterbuch* were a strong predictor of /ə/ deletion (35 % explained variance). Even

though this control factor was included in the statistics, biphone surprisal of the following context was significant in explaining /ə/ deletion rates confirming the results of the main deletion model. However, this factor had a small effect size contributing only 2 % to the explained variance. As noted above, the Duden rules are based on both the preceding and following context. The fact that we still found an effect of surprisal of the following context proved that surprisal effects go above and beyond the effect of phonological context. In addition, the wording of the Duden rules attached more importance to the preceding context than to the following. For that reason, we still saw an effect of biphone surprisal of the following context on /ə/ deletion. Interestingly, /ə/ deletion was independent of speech rate deviations.

While this chapter outlined the results of the segment deletion analysis, the following chapter gives an overview of the VOT analysis performed on German read speech from the Siemens Synthesis corpus.

Chapter 7

Voice onset time

In addition to analyses on segment duration and deletion, we also conducted a voice onset time (VOT) analysis on the same German corpus. We interpret this analysis as an extension of the previous investigations zooming in on a specific duration measure for one speech sound class. The following chapter introduces the method, results and their discussion of this analysis.

7.1 Method

VOT was measured in fortis (/p, t, k/) and lenis plosives of German (/b, d, g/). Since the Siemens Synthesis corpus is a considerably large corpus an automatic VOT tagger was used to speed up the segmentation procedure. The software package used for automatic measurement of VOT is called AutoVOT (Keshet et al., 2014). The user needs to provide pre-segmented audio data with plosive segment labels in the form of Praat (Boersma and Weenink, 2017) TextGrids. Then, AutoVOT uses a classifier to localize the VOT for the pre-segmented plosives. The software package already contains classifiers for English stop consonants. However, the makers of the software recommend to train a new classifier based on a small sub corpus of about 100 manually labeled VOTs from the data.

We decided to train AutoVOT on manually labeled VOTs of our corpus. VOT was manually labeled starting at the beginning of the release and putting an end boundary when there was a clear voice bar visible in the following vocalic segment. This means we only analyzed positive VOT values, also for voiced plosives in German. The trained annotator used the Praat pitch tracker as a cue for voice activity, but mainly relied on auditory and visual cues from the speech signal. Four classifiers were trained with each 100 manually labeled VOTs. We trained per speaker and per phonological voicing, i.e., each speaker had their individually trained classifier for both classes of phonological voicing. Then, mode 1 of feature extraction and training

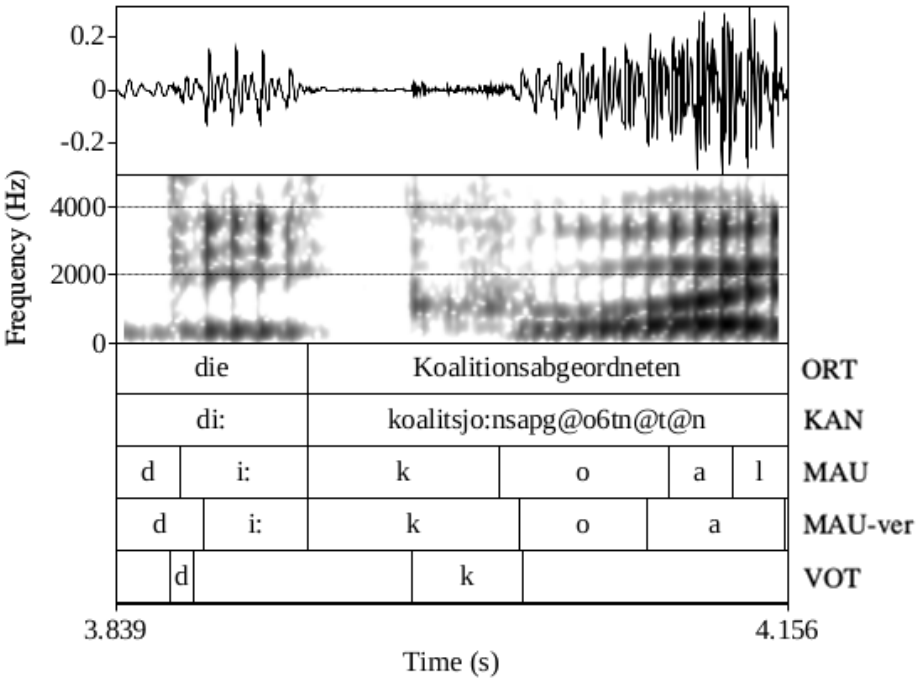
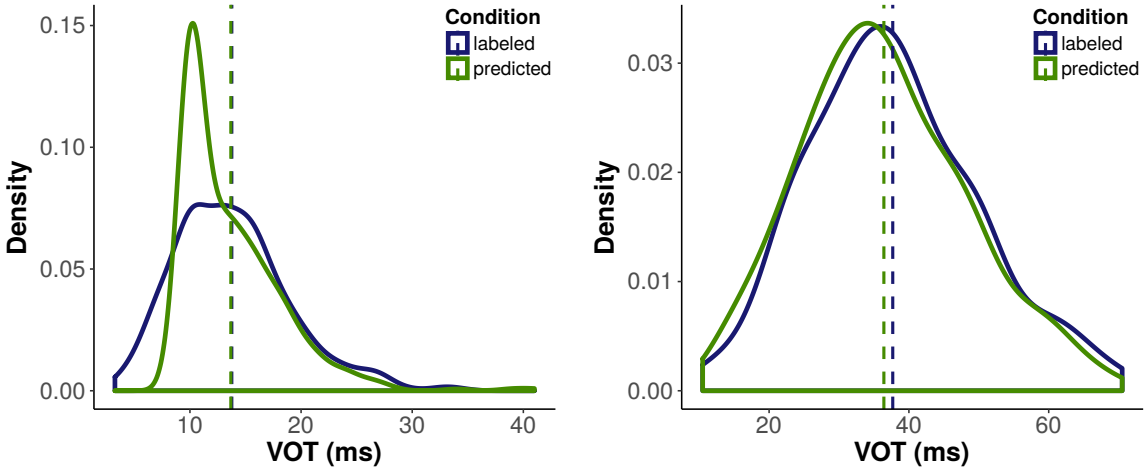


Figure 7.1: Example of automatically labeled VOT by AutoVOT of stop consonants /d/ and /k/ in German read speech. ORT: orthographic transcription, KAN: canonical transcription, MAU: automatic segmentation, MAU-ver: verified segmentation, VOT: automatic VOT segmentation.

from the software package AutoVOT was run with default values. In a second step, mode 1 of VOT decoding of the tool was performed using the trained classifier on a set of TextGrids and corresponding audio files. Here, the minimum and maximum VOT length was left at default values for voiceless plosives (15 and 250 ms, respectively). For voiced plosives, however, a range from 10 to 200 ms was empirically determined based on a comparative analysis of the tagger performance with default values, and with 5 ms minimum and 100 ms maximum. Measurement of the tagged VOTs led to the minimum and maximum VOTs used for the entire corpus for labeling VOT of voiced stop consonants (Figure 7.1).

There is an inbuilt performance check for AutoVOT which relies on a set of labeled VOTs by a trained annotator and the predicted VOTs from the automatic tagger for the same stop consonants. In order to check performance a random sample of 5 % of the automatically tagged data files of voiced plosives ($n = 97$), and of 5 % of the data files of voiceless plosives ($n = 57$) with predicted VOTs were manually verified by a trained annotator. Then, the performance check for AutoVOT was run comparing the verified to the automatically predicted labels. Two checks were performed, one for the data set with voiced plosives, and one for the data with voiceless plosives. Duration values for the verified and predicted VOT labels were obtained. Spearman's rank correlation of VOT durations between verified and predicted labels was strong for voiceless plosives ($\rho = 0.74, S = 54415, p < 0.001$) and moderate for voiced data



(a) Voiced plosives. **(b)** Voiceless plosives.

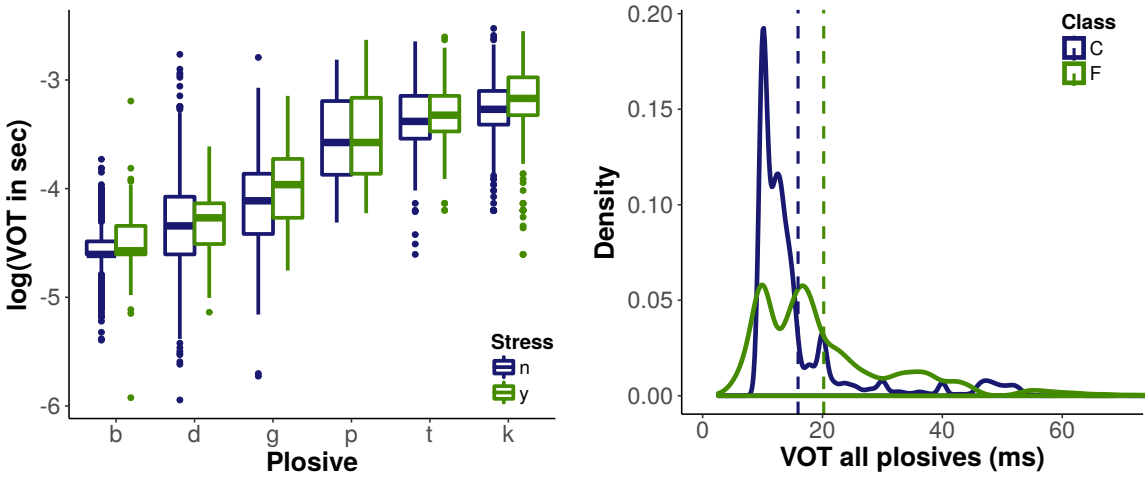
Figure 7.2: Density plot of German VOT durations (ms) for labeled and predicted boundaries. Mean durations are indicated by vertical lines.

Table 7.1: VOT in German: number of observations per stop consonant and word class.

Word class	/b/	/d/	/g/	/p/	/t/	/k/
Content	1,500	972	1,503	454	270	818
Function	180	5,049	126	0	16	260

($\rho = 0.40, S = 12853000, p < 0.001$). The average difference between labeled and predicted VOT for voiced stops ($M = 14.05\,ms, SD = 37.25\,ms$) was higher than for voiceless stops ($M = 8.78\,ms, SD = 35.67\,ms$). Figures 7.2a and 7.2b illustrate that the tagger performance was more accurate for voiceless than for voiced stops. On average, VOTs of voiceless and voiced stops were increased in their duration during the verification of the annotator. This process widened the narrow peak of the density distribution of the voiced VOT.

Only word-initial German stop consonants followed by a vowel were used for the VOT analysis. Stop consonants preceded by a pause were excluded from the analysis because closure phase and VOT cannot be labeled precisely. Based on visual inspection of boxplots for VOT duration per sound identity 9 data points were excluded from the data set because they were identified as outliers. In total, 11,148 stop consonants were analyzed. The majority of the data was made up by voiced consonants ($n = 9329$). Word class was included as a factor in the statistical analysis because about half of the data set contained function words ($n = 5633$). Most of the analyzed stop consonants were /d/ ($n = 6020$) with 84 % of function words (Table 7.1).



(a) Per stop consonant and primary lexical stress (y: stressed, n: unstressed).

(b) Per word class (C: content word, F: function word).

Figure 7.3: VOT durations.

7.2 Results

7.2.1 Descriptive statistics

Voiceless plosives ($M = 20.48\,ms, SD = 9.31\,ms$) had longer VOTs than voiced plosives ($M = 17.61\,ms, SD = 11.25\,ms$). Velar voiceless stop consonant had the longest VOTs, followed by /t/ and /p/. We found the same descending order of VOT duration for place of articulation for the voiced stop consonants: /g/ had longer VOTs than /d/ and /b/ (Figure 7.3a).

With regard to primary lexical stress and its impact on VOT duration we found that VOT in stressed stop consonants was longer ($M = 18.75\,ms, SD = 12.21\,ms$) than in unstressed stop consonants ($M = 17.93\,ms, SD = 10.71\,ms$). It should be noted that 82 % of the stop consonants were the onset of syllables without primary lexical stress ($n = 9,116$), while the rest stood in stressed position. All stop consonants were found in both unstressed and stressed position in the data set. This pattern was most evident in /d/, and only subtle in velar and bilabial stop consonants (Figure 7.3a).

Prior to investigating the relationship between articulation rate and VOT duration, we log-transformed the duration values because of positive skewness, and also mean-centered local and global speech rate per speaker. In the correlation analysis, we found a significant negative relationship between log-transformed VOT durations and local and global tempo. This means that VOT duration decreased with increasing speech rate. The negative correlation was weak for local tempo and VOT ($r = -0.29, t(11104) = -31.34, p < 0.001$), and very low for global speech rate ($r = -0.06, t(11146) = -6.21, p < 0.001$).

More than two thirds of the stop consonants were preceded by a sonorant ($n = 7,728$), while the frequency counts for preceding sibilants and obstruents were similar ($n = 1,785, n = 1,635$). VOT was longer when an obstruent preceded the stop consonant ($M = 18.71\,ms, SD = 11.42\,ms$) than when preceded by a sonorant ($M = 18.11\,ms, SD = 10.81\,ms$) or sibilant ($M = 17.35\,ms, SD = 11.41\,ms$). All stop consonants in the data set were observed with all three preceding contexts.

With regard to average VOT durations for different word classes we observed that VOT was longer for function words ($M = 20.19\,ms, SD = 11.86\,ms$) than for content words ($M = 15.92\,ms, SD = 9.58\,ms$) when comparing raw duration values, and longer for content words ($M = -4.04, SD = 0.56$) than function words ($M = -4.26, SD = 0.39$) when using the log-transformed values. This finding was possibly due to the larger positive skewness in VOT duration values for content words ($v = 2.52$) compared to skewness in VOT duration for function words ($v = 1.56$) (Figure 7.3b).

7.2.2 Linear mixed-effects model

The LMs built to estimate ID included word boundary markers. Since VOT was only measured in word-initial stop consonants, biphone ID measures were not investigated because of their low variability in the n-phone context. This is why triphone and fourphone ID measures were used. As in all other analyses in this thesis, surprisal was used as an ID measure. Surprisal values were log-transformed due to positive skewness.

VOT showed significant positive correlations with all tested surprisal measures. Both surprisal measures for the preceding context had higher correlation values than for the following context. For triphone of the preceding context ($r = 0.36, t(11128) = 40.62, p < 0.001$) and fourphone of the preceding context ($r = 0.31, t(11112) = 33.85, p < 0.001$) the Pearson's product-moment correlations were moderate. For following context there was a very low positive correlation between triphone surprisal and VOT duration ($r = 0.14, t(10995) = 15.06, p < 0.001$), and an even smaller correlation between fourphone surprisal and VOT duration ($r = 0.02, t(11089) = 2.23, p < 0.001$). Also, there was a significant weak negative correlation between log-transformed word frequency and log-transformed VOT duration ($r = -0.29, t(11025) = -32.02, p < 0.001$), and a low negative correlation between phoneme probability and log-transformed VOT ($r = -0.12, t(11149) = -13.67, p < 0.001$).

In the following modeling procedure, we included triphone and fourphone surprisal of the following and preceding context, phoneme unigram probability, and word frequency as ID factors. Prosodic controls for VOT duration were global and local speech rates, as well as primary lexical stress. Boundary was not included as a prosodic factor because all observations were taken from word-initial position, and all data points preceded by a pause were excluded from the analysis. Since the statistical analysis was performed with a model on the entire data set phonological voicing was included

to control for VOT duration differences because of that factor. The binary factor word class was introduced to investigate whether VOT duration differences were predicted by the target word being a content or function word. As mentioned before, surprisal and VOT duration were log-transformed because of their positive skewness. This also applied to word frequency. Speech rate was mean-centered. All categorical variables were treatment-coded.

In order to avoid collinearity in the LMMs we tested dependencies between the fixed effects prior to model training. Triphone and fourphone surprisal of the same context direction were related to each other, preceding context ($r = 0.80$) more so than following context ($r = 0.53$). This led to separate models for triphone and fourphone surprisal. Word frequency and phoneme probability were moderately correlated ($r = 0.59$). Both ID measures were also negatively correlated with triphone surprisal of the preceding context ($r = -0.70$) and fourphone surprisal of the preceding context ($r = -0.62$). Since these ID measures apparently had the same collinearities with other factors in the analysis only one was entered into the final model structures. Phoneme probability was chosen over word frequency because of the strong positive correlation between word class and word frequency ($r = 0.81$). Interestingly, global and local speech rate were only weakly related ($r = 0.18$).

Backward model selection was applied to train two different models, one with triphone surprisal and one with fourphone SURPRISAL. Both models had the same final model structure. SURPRISAL of the preceding and following context (triphone or fourphone, respectively), and PHONEME PROBABILITY were entered as ID measures in the model. Both SPEECH RATES (global and local), as well as primary lexical STRESS were used as prosodic controls. Additionally, phonological VOICING of the target stop consonant and WORD CLASS were used as predictors. In both models, WORD CLASS did not show a significant effect on VOT duration, neither in the model with triphone SURPRISAL ($\beta = 0.03, SE = 0.02, t(97) = 1.31, p = 0.19$) nor in the model with fourphone SURPRISAL ($\beta = 0.02, SE = 0.02, t(41) = 0.97, p = 0.34$) which was why this factor was disregarded in the following analysis. The random structure contained random intercepts for SPEAKER, WORD, STOP CONSONANT, and PRECEDING CONTEXT with the factor levels sibilant, sonorant, and obstruent. Following context was not included because all segments following the stop consonant were vowels. In addition, the models converged with a random slope for both SURPRISAL values per WORD (Model structure 7.1).

$$\begin{aligned} VOT \sim\ & TriSur/FourSur + TriFolSur/FourFolSur \\ & Stress + GlobalTempo + LocalTempo + \\ & Voicing + (1|Speaker) + \\ & (1 + TriSur/FourSur + TriFolSur/FourFolSur|Word) + \\ & (1|Stop) + (1|Preceding) \end{aligned} \tag{7.1}$$

In the triphone model, all fixed effects but PHONEME PROBABILITY reached sig-

Table 7.2: VOT in German: regression coefficients, standard error (SE) and statistical output of LMM analysis including triphone surprisal.

	Terms	Coeff.	SE	t-value	p-value
ID model	Surprisal following	0.03	0.007	4.89	$< .001$
	Surprisal preceding	0.10	0.05	1.99	$= .04$
	Phoneme probability	0.20	0.30	0.73	$= .49$
Prosodic model	Global tempo	-0.01	0.003	-2.80	$< .01$
	Local tempo	-0.009	0.001	-8.30	$< .01$
	Stress (y – n)	0.04	0.01	2.86	$< .01$
Other control	Voicing (n – y)	0.87	0.15	5.95	$< .01$

nificance level. Both SURPRISAL factors indicated that VOT durations increased with increasing surprisal values. Regarding the prosodic factors included in this analysis we found that both SPEECH RATES were significant in explaining VOT duration variance. VOT durations decreased with increasing local and global tempo. When stop consonants were the onset of syllables carrying primary lexical STRESS, they had longer VOT durations than unstressed stop consonants. Voiceless stop consonants had significantly longer VOT durations than voiced stop consonants (Table 7.2).

In the LMM for VOT duration with triphone surprisal, marginal pseudo-R^2 of the fixed effects explained 44.94 % of the model variance. Total explained variance of the entire model given by conditional pseudo-R^2 added up to 73.62 %. The largest effect size of all fixed effects for this model had VOICING with 42.47 % explained variance in the data. Both SPEECH RATES explained about 1.01 % of data variance, and STRESS 0.14 %. The ID measure SURPRISAL of the following context added 0.43 % to the explained variance, while SURPRISAL of the preceding context was a stronger predictor ($Var = 0.89\,\%$). Based on the effect size of the model the ID factors had a slightly larger impact on VOT duration than the prosodic factors used here.

Since the LMM with fourphone SURPRISAL of the preceding and following context as fixed effects was run on the same data as the previously described model we found the same main significant effects for local and global SPEECH RATE, primary lexical STRESS, and phonological VOICING. Also, there was no significant effect of PHONEME PROBABILITY on VOT duration. SURPRISAL of the following context, however, did reach significance level, while higher SURPRISAL of the preceding context showed a tendency to predict longer VOT durations (Table 7.3).

Marginal pseudo-R^2 of the LMM with fourphone surprisal indicated that the fixed effects explained a total of 44.05 % of the variance in the VOT duration data. Conditional pseudo-R^2 giving the entire model effect size was 73.46 %. Again, VOICING had the largest effect size with 42.82 % explained variance of VOT duration. Both SPEECH RATES explained 0.98 % of the model variance, while SURPRISAL of the following context had an effect size of 0.14 %. Primary lexical STRESS added 0.11 % explained

Table 7.3: VOT in German: regression coefficients, standard error (SE) and statistical output of LMM analysis including fourphone surprisal.

	Terms	Coeff.	SE	t-value	p-value
ID model	Surprisal following	0.01	0.003	3.90	$< .001$
	Surprisal preceding	0.05	0.03	1.89	$= .06$
	Phoneme probability	0.15	0.26	0.57	$= .59$
Prosodic model	Global tempo	-0.01	0.003	-3.02	$< .01$
	Local tempo	-0.009	0.001	-8.31	$< .01$
	Stress (y – n)	0.05	0.01	3.99	$< .001$
Other control	Voicing (n – y)	0.88	0.14	6.13	$< .01$

Table 7.4: VOT duration model: explained data variance in % of ID and prosodic model for triphone and fourphone LMM.

n-phone	ID model		Prosodic model	
	Preceding	Following	Speech rate	Stress
triphone	0.89	0.43	1.01	0.14
Total	1.32		1.15	
fourphone		0.14	0.98	0.11
Total	0.14		1.09	

variance to the model. Table 7.4 compares the effect sizes of ID and prosodic factors for both LMMs.

Investigation of the marginal effects of the continuous factors surprisal and speech rate in both LMMs showed that global speech rate was less effective in predicting German VOT duration than speech rate measured at word level. This was visible in less steeper regression lines for global rate compared to local rate.

For both LMMs, including triphone or fourphone surprisal, the random effects explained about a fourth of the variance in VOT duration. For the triphone model the percentage was 28.68 %, and only a little lower for the fourphone model (23.37 %). On average, the bilabial stop consonants were shorter in VOT durations than the average predicted VOT for the set of fixed-effect values used in the LMM, while all other stop consonants were longer than this predicted average. The random intercept for preceding context showed that VOT durations with preceding sonorant were shorter than the predicted average VOT duration, whereas VOT of stop consonants with preceding sibilants and obstruents were slightly longer than all VOTs on average. There were also inter-speaker differences in VOT durations which added to the explained variance. Speaker "wo" produced shorter VOTs than speaker "ai".

One main objective of this thesis was to investigate possible interactions between information-theoretic variables and prosodic factors. The final baseline LMMs with

Table 7.5: VOT in German: interaction of triphone surprisal of the preceding and following context with prosodic factors.

Context	Terms	Coeff.	SE	t-value	p-value
Preceding	Surprisal * Global tempo	-0.03	0.01	-2.41	= .02
	Surprisal * Local tempo	-0.009	0.003	-2.53	= .01
Following	Surprisal * Local tempo	-0.003	0.001	-3.34	< .001

maximal model structure were used to build interaction models testing the interaction terms in a step-wise procedure. Model comparison of the baseline LMM for VOT duration and the model including the interaction term were performed via ANOVA tests.

The interaction models for the triphone baseline models were trained with interaction terms between triphone SURPRISAL of the preceding and the following context because both were significant predictors of VOT durations in the LMM. According to the likelihood ratio tests, only the interaction models with local SPEECH RATE, and the interaction model with triphone SURPRISAL of the preceding context and global SPEECH RATE performed significantly better than the respective baseline models. High surprisal and accelerated speech rate complemented each other in predicting a decrease in VOT duration (Table 7.5).

For the LMMs with fourphone surprisal, only interactions with SURPRISAL of the following context and prosodic factors were tested because the main effect of surprisal of the preceding context was non-significant. According to the likelihood ratio tests, none of the three interaction models performed significantly better than the baseline model.

7.3 Discussion

About half of the data points for the VOT analysis contained plosives in function words. We therefore included the factor word class into our statistical analysis in favor of word frequency, both are known to correlate strongly. Word class did not have a significant effect on VOT durations which justified running the analysis on both function and content words. This result stood in contrast to Yao (2009) who identified word class as a significant predictor of VOT durations in American English. Her analysis, however, was conducted on spontaneous speech. In this register, reduction processes due to word class or word frequency may be more pronounced than in read speech (Johnson, 2004).

As expected, surprisal was a significant predictor of VOT in German stop consonants. In the model with triphone surprisal, high surprisal of both preceding and following context predicted longer VOT durations, while only surprisal of the following context was a significant predictor when fourphone surprisal was used in the

LMM. We concluded that larger n-phone size in the preceding context for estimating surprisal did not have an effect on subtle differences in VOT durations for stop consonants when controlling for other factors at the same time.

Fourphone surprisal of the following context had a smaller effect on VOT than triphone surprisal of the following context. In the triphone model, surprisal of the preceding context explained twice as much variance as surprisal of the following context. This was expected from the prior Pearson's r correlation analysis. Apparently, the amount of context on which surprisal was calculated and the direction of context (preceding vs. following) played a key role in deciding how effective it was in explaining the dependent variable VOT duration. To sum up, the effect of surprisal was subtle and depended on context and direction, as expected from previous studies on the relationship between duration and ID variables (Yao, 2009).

In both LMMs, primary lexical stress and both speech rates were significant predictors of VOT durations in German stop consonants. Stressed stop consonants had longer VOTs than unstressed plosives. Surprisingly, primary lexical stress only had a very subtle effect on VOT duration. In both models this effect was around 0.1 % explained variance. Apparently, stress did not play a huge role in predicting subtle differences in VOT durations which were primarily determined by phonological categories, such as voicing and place of articulation. However, primary lexical stress has been identified as a reliable predictor of phoneme and word duration (Aylett and Turk, 2006; Bell et al., 2009). Our findings showed that a strong effect of stress on duration does not apply generally to all duration measures in speech production, when controlling for other factors.

Regarding speech tempo, we found that at fast speech rate VOT was shortened compared to slow speech rate. Similarly to English, VOT durations of German stop consonants were implemented as a phonological rule since they varied with speech rate (Solé and Estebas, 2000). In both LMMs, variation in speech rate had a stronger effect on VOT duration than primary lexical stress. Analysis of the marginal effect size of the different speech rate measures revealed that local speech rate was much more informative in predicting VOT duration than speech rate based on the entire sentence. This was also mirrored in the correlation analysis prior to statistical modeling. Here, we found a moderate negative correlation for local tempo and VOT, but only a very low negative correlation between global tempo and VOT.

None of the tested interaction models containing an interaction term between surprisal and primary lexical stress had a better fit for predicting VOT durations in German read speech than the respective baseline model. This is in contrast to our findings for segment duration (Chapter 5). Here, surprisal interacted positively with stress in explaining durational variability. In addition, this finding contradicts a prior account of an interaction between stress and predictability on VOT durations in a shadowing experiment (Manker, 2017). We assume that we did not find an interaction between the two variables because in our data stress only had a very subtle effect on VOT duration.

However, the interaction model with local speech rate and triphone surprisal of both context directions performed significantly better than the baseline model. The same was true for the interaction between global speech rate and triphone surprisal of the preceding context. The conditional effect of speech rate by surprisal on VOT was that VOT decreased with faster speech rate and higher surprisal. Investigating the collinearities between the model's terms revealed that there was a strong correlation between triphone surprisal of the preceding context and the interaction term between (local and global) tempo and surprisal. Therefore, interpreting single effects of the predictors was difficult because they were not independent. In contrast, the interaction between global tempo and triphone surprisal of the following context did not correlate with fixed effects. Despite that, it did not increase marginal pseudo-R^2 of the entire model. Summing up, the interaction terms either introduced collinearities with main effects or were independent terms but had only vanishingly small effect sizes.

One of the main objectives of this thesis was to investigate whether ID variables or prosodic factors are more informative in predicting variation in phonetic structures. For the analysis of VOT duration of German stop consonants in read speech, we found that in a model with triphone surprisal of both context directions the ID factors explained more variance than the prosodic factors (primary lexical stress, local and global speech rate). If the same VOT duration values were modeled in a LMM with fourphone surprisal of both context directions and the same prosodic model, the prosodic model turned out to be much more effective in predicting VOT duration. This finding demonstrated that it is not only crucial how well model predictors of prosody or ID are defined and calculated but how well the specific definition is appropriate for the predicted variable. That is to say, for VOT duration, surprisal based on triphones proved to be more informative than surprisal based on fourphones. Local speech rate was much more informative for modeling VOT duration than globally measured tempo on the sentence level. Considering that VOT is a measure that is below the phone-level this finding was to be expected.

Place of articulation was implicitly included in the model as random intercept of phoneme identity. Investigation of the BLUP of the random intercept for phoneme identity showed that bilabial stop consonants were shorter in VOT duration than the average. The descriptive statistics also confirmed that bilabials had shorter VOTs than alveolar or velar stop consonants. We therefore confirmed the known universal effect of place of articulation on VOT duration (Lisker and Abramson, 1964; Stevens, 1993).

Regarding the BLUP of the random intercept of preceding phonological context we found that VOT was shorter with preceding sonorants than the average values for VOT in the model. This finding was in line with Yao (2009) who observed shortening in VOT with preceding vowels. However, Yao (2009) defined phonological context with a binary factor of consonants versus vowels which is why results were only comparable to a certain extent.

Although only two different speakers were investigated the BLUP of the random intercept for speaker showed that the speakers differed in their VOT productions. This result added to the research that found strong effects of idiosyncratic articulatory behavior in VOT productions (Allen et al., 2003) which in some cases ultimately led to individual statistical models per speaker (Yao, 2009).

The previous three chapters focused on durational acoustic-phonetic measures and their relationship to ID and prosodic factors. The following production analyses outline the spectral acoustic-phonetic measures used in this thesis, starting with vowel dispersion.

Chapter 8

Vowel dispersion

The next chapter introduces three analyses of vowel dispersion: in German, in six different languages from a cross-linguistic perspective, and in L2 Bulgarian speakers of German. ID factors have been shown to influence vocalic spectral characteristics. Vowels that are difficult to predict from the context are not only more distinct in their quantity, but also in their quality (Aylett and Turk, 2004). Most of the research, however, focuses on English. We therefore intended to extend this line of research to an investigation of vowels in other languages (Sections 8.2 and 8.3). In addition, we studied if vowel dispersion of L2 speakers at different proficiency levels can be explained by ID factors of the target language (Section 8.4).

8.1 Method

F1 and F2 were measured at the temporal midpoint in vocalic nuclei. Formant analysis was conducted with the Burg algorithm in Praat (Boersma and Weenink, 2017) with a maximum of five formants, window size of 25 ms, pre-emphasis from 50 Hz, and a maximum formant threshold of 5000 Hz (male speakers) and 5500 Hz (female speakers). Formant values were cleaned and manually checked before speaker-dependent normalization was applied to control for differences in formant values due to sex or speaker (Adank et al., 2004). As a measure for vowel distinctiveness, the Euclidean distance between midpoint of the vowel space and formant values for every vowel were calculated for each speaker (Bradlow et al., 1996). The larger the distance between the vowel space midpoint and individual vowels gets, the more distinct is the vowel quality. This measure is independent of differences in vowel inventory between the languages because it assumes that vowel distinctiveness is defined by vowel space expansion.

Recently, the interpretation of vowel dispersion has been broadened with respect to vowel specific movements within the vowel space with regard to competitor vowels.

Wedel et al. (2018) argued that vowels are under competition from neighboring vowels depending on their position in the vowel space. Peripheral vowels, such as /i, e/, for instance, are under competition from interior vowels /ɪ, ɛ/. Wedel et al. (2018) showed that in cases of lexical competition peripheral vowels move further away from the vowel space center to the periphery, while interior vowels move closer to the center. In the light of these findings, one could argue that vowel dispersion is not an ideal measure of vowel space expansion.

From the languages investigated here only American English and German, and Czech to some extent, have contrastive interior vowels in their phoneme systems. In addition, Wedel et al. (2018) limited their study to vowels in stressed position. Surely, in unstressed position, German interior vowels, and /ɪ/ in American English, face competition from the mid central vowel /ə/. Admittedly, the F1/F2 Euclidean distance from the vowel space center cannot possibly capture all vowel movements within the vowel space, and vowels might behave differently with regard to the amount of dispersion from the center. We have included vowel identity as a random or fixed factor in the statistical models to account for these differences.

8.2 Vowel dispersion in German

The following section contains results and discussion of a vowel dispersion analysis of German read speech in relation to ID and prosodic factors. It is based on data from the Siemens Synthesis corpus (Section 3.1.1).

8.2.1 Descriptive statistics

The total number of vowel tokens in content and function words was 79,395. Diphthongs were not included in this analysis. This data set was reduced to 57,743 data points from only content words. Outlier cleaning of vowel dispersion led to a removal of 15 data points.

On average, stressed German vowels ($M = 1.59, SD = 0.67$) were more dispersed than unstressed vowels ($M = 1.14, SD = 0.58$). Vowels were more distinct at no boundary position ($M = 1.34, SD = 0.64$) than at phrase ($M = 0.95, SD = 0.67$) or word boundary ($M = 0.84, SD = 0.52$). Investigating vowel dispersion per vowel identity showed that there were clear differences in the amount of dispersion depending on the phoneme. Peripheral vowels in the acoustic space, such as /aː/, /oː/ /uː/, and /iː/ were amongst the vowels with the highest vowel dispersion values. Interior vowels, on the other hand, showed less dispersion because of their position in the acoustic space (Figure 8.1).

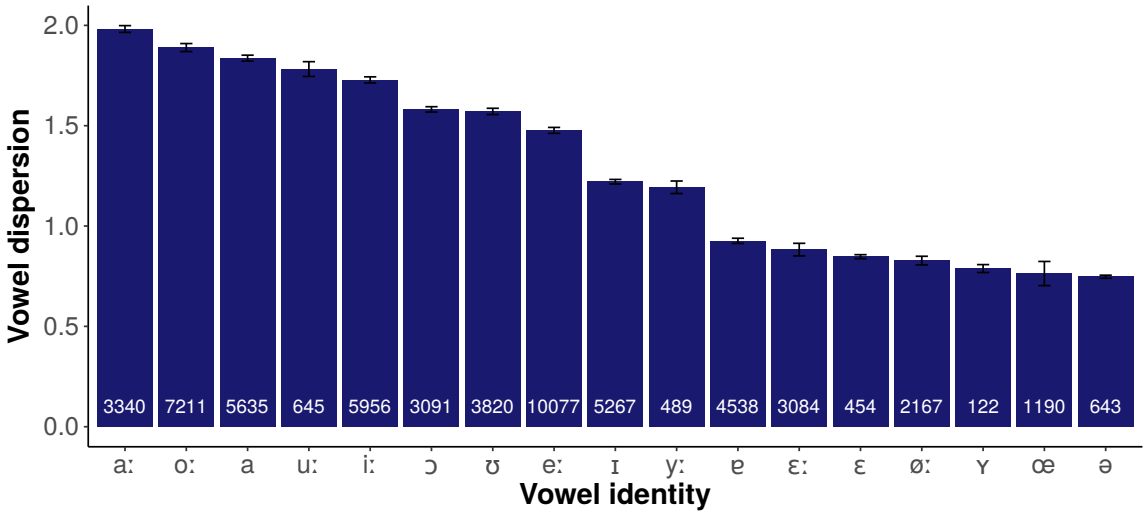


Figure 8.1: Vowel dispersion in German per vowel identity. Number of tokens per vowel are given at the bottom of individual bars.

8.2.2 Linear mixed-effects model

Prior to model building, we performed a collinearity analysis testing possible relations between the ID factors, surprisal of the preceding and following context, word frequency, and phoneme probability, as well as prosodic factors, such as local and global speech rate, boundary, stress, and the additional control factor average vowel duration (Section 4.4). Local and global speech rate were only weakly related ($r = 0.15$). Primary lexical stress was positively correlated with average vowel duration ($r = 0.24$). Phoneme probability was moderately correlated with surprisal of the preceding context, biphone ($r = -0.46$) and triphone ($r = -0.31$), but not with word frequency ($r = 0.05$). Word frequency, however, was only weakly related to triphone surprisal of the preceding context ($r = -0.11$), and showed even lower correlations with the other ID factors. In congruence with other LMM structures in this study, we decided to exclude phoneme probability from the model because of its relations to surprisal. As seen in Figure 4.1, surprisal values of the preceding (biphone and triphone), and the following context (biphone and triphone) were related moderately.

The LMM for vowel dispersion was run with the ID factors biphone SURPRISAL of the preceding context, and WORD FREQUENCY, prosodic factors, such as primary lexical STRESS, prosodic BOUNDARY, local and global SPEECH RATE, as well as average vowel DURATION. Surprisal of the following context was not included in the model because correlation values were not significant for neither biphone ($r = 0.004, t(57689) = 1.01, p = 0.36$) nor triphone context ($r = -0.004, t(57499) = -0.84, p = 0.27$). Random structure included random intercepts for VOWEL, SPEAKER, as well as preceding and following PHONOLOGICAL CONTEXT, and random slopes and intercept for SURPRISAL and STRESS per WORD. PHONOLOGICAL CONTEXT was defined as having five factor levels based on consonantal place of speech (labial, coronal, dorsal), vo-

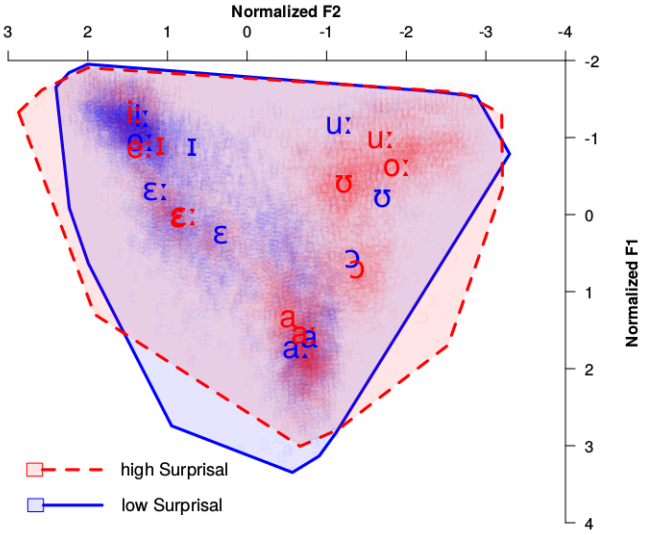


Figure 8.2: Vowel dispersion in German under high and low biphone surprisal of the preceding context averaged across both speakers. Binning of surprisal was based on 20 % of the highest and lowest values in the data set.

calic context and following or preceding pause in the speech signal (Section 4.4.4). The categorical factors STRESS and BOUNDARY were treatment coded before entering into the LMM (Model structure 8.1).

$$\begin{aligned} VowelDispersion \sim\ & BiSur + Wordfreq + \\ & Stress + Boundary + GlobalTempo + LocalTempo + \\ & DurAverage \\ & (1|Speaker) + (1 + BiSur + Stress|Word) + \\ & (1|Preceding) + (1|Following) + (1|Vowel) \end{aligned} \tag{8.1}$$

German vowel dispersion was significantly affected by both ID factors, SURPRISAL and WORD FREQUENCY. Vowels in high surprisal context were more dispersed than in low surprisal context (Figure 8.2). In addition, we found that high-frequency words contained less dispersed vowels than low-frequency words. Regarding the prosodic factors integrated in the model, we observed that both SPEECH RATES had a significant negative effect on vowel dispersion. At higher SPEECH RATE, at word and sentence level, vowel dispersion decreased. Preceding a BOUNDARY, at word and phrase level, vowels showed a decrease in dispersion compared to no BOUNDARY position. Both primary lexical STRESS and average vowel DURATION had a positive effect on vowel dispersion (Table 8.1).

Regarding the LMM for vowel dispersion with biphone surprisal of the preceding context marginal pseudo-R^2 of the fixed effects explained 6.19 % of the model variance. Total explained variance of the entire model given by conditional pseudo-R^2 added up to 73.93 %. The ID factors both explained a total of 0.32 % of the data variance in

Table 8.1: Vowel dispersion in German: regression coefficients, standard error (SE) and statistical output of LMM analysis including biphone surprisal of the preceding context.

	Terms	Coeff.	SE	t-value	p-value
ID model	Surprisal	0.03	0.009	4.01	$< .001$
	Word frequency	-0.003	0.001	-2.57	$= .01$
Prosodic model	Global tempo	-0.007	0.002	-2.87	$= .004$
	Local tempo	-0.007	0.0007	-9.50	$< .001$
	Boundary (phrase – none)	-0.12	0.04	-3.28	$= .001$
	Boundary (word – none)	-0.06	0.01	-5.48	$< .001$
	Stress (y – n)	0.25	0.008	32.88	$< .001$
Other control	Average vowel duration	0.18	0.006	30.81	$< .001$

vowel dispersion with WORD FREQUENCY ($Var = 0.20\,\%$) being a stronger predictor than SURPRISAL ($Var = 0.12\,\%$). The largest effect size of all fixed effects for this model had STRESS ($Var = 3.60\,\%$). Other significant prosodic factors explained much less variability in vowel dispersion, BOUNDARY 0.31 % and SPEECH RATE 0.30 %. Vowel DURATION was a relatively strong predictor of vowel dispersion with 1.66 % explained variance.

Most of the variance in the model for vowel dispersion was explained by the random intercepts for WORD ($Var = 21.63\,\%$) and for SOUND ($Var = 20.70\,\%$) (Figure 8.3). Closed vowels (/iː, ɪ, uː, ʊ, eː/), back vowels (/oː, ɔ/), and open vowels (/aː, a/) were more dispersed than all vowels on average while all other vowels were less dispersed than the average. Vowels that were less dispersed than the average were vowels at mid position (/ɛː, ɛ, øː, œ/), and more centralized vowel phonemes (/ə, ɐ, yː, ʏ/). While the random intercept for FOLLOWING PHONOLOGICAL CONTEXT did not explain a large quantity of variance in vowel dispersion ($Var = 0.44\,\%$), the random intercept for PRECEDING PHONOLOGICAL CONTEXT did ($Var = 3.42\,\%$). Vowel dispersion was higher than the average for vowels with preceding pauses and other vowels, while preceding consonants led to lower values in vowel dispersion. The random intercept for speaker did not add much to the model performance ($Var = 0.03\,\%$).

Running the same LMM with triphone instead of biphone surprisal of the preceding context we found no significant effect of surprisal ($\beta = 0.004, SE = 0.006, t(2781) = 0.65, p = 0.52$), but consistent significant effects which were previously observed for all other fixed effects in the model. Based on this result and our correlation analysis for surprisal of larger n-phone size and vowel dispersion, we did not expect to find significant effects of n-phone surprisal larger than triphone in a more complex LMM. This was why these were not tested.

In a second step, we investigated interaction effects of surprisal and prosodic factors on vowel dispersion. Interaction terms between individual prosodic factors and

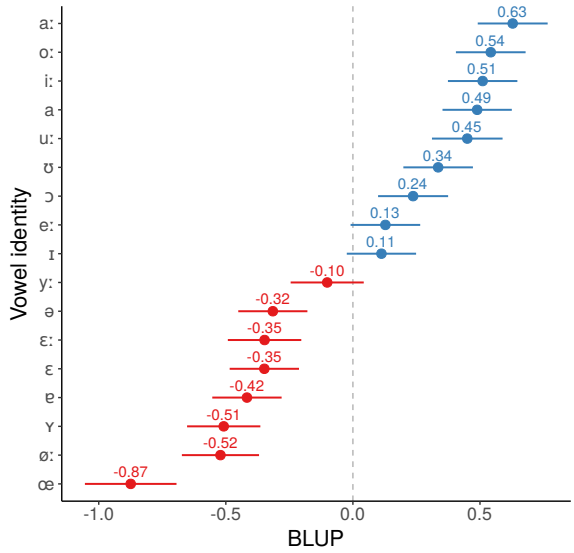


Figure 8.3: Effect of the random intercept for vowel identity (BLUP) in German vowel dispersion model including biphone surprisal of the preceding context.

surprisal were entered into separate interaction models. Using ANOVA model comparison we tested whether the interaction model had a significantly better performance than the baseline model. If this was the case, effects of the interaction term were reported. All interaction models except for one had better performance than the baseline model. The interaction model with global tempo and surprisal did not yield a significantly better performance ($log(L) = -23816$) than the baseline LMM for vowel dispersion ($log(L) = -23815; \chi^2(1) = 0.34, p = 0.56$).

SURPRISAL interacted positively with STRESS in its effect on vowel dispersion. At high surprisal and in a syllable carrying primary lexical stress vowels were more dispersed than at low surprisal in unstressed position. The interaction term added 0.06 % of explained variance to the model. Phrase BOUNDARY interacted significantly with surprisal, while word boundary did not. There was a positive effect on vowel dispersion of the interaction term between phrase boundary and surprisal which added 0.06 % of explained variance to the model. Surprisal also interacted significantly with local SPEECH RATE. At higher speech rate and higher surprisal we found a significant reduction in vowel dispersion. This effect added 0.09 % to the overall explained variance of the fixed effects (Table 8.2).

8.2.3 Discussion

Word frequency and biphone surprisal of the preceding context had the expected effects on vowel dispersion: high-frequency words contained less dispersed vowels than low-frequency words. Vowels in high surprisal context were more dispersed than in low surprisal context. The analysis of the effect size of single fixed effects revealed

Table 8.2: Vowel dispersion in German: interaction of biphone surprisal of the preceding context with prosodic factors.

Terms	Coeff.	SE	t-value	p-value
Stress * Surprisal	0.16	0.02	10.18	$< .001$
Boundary (phrase) * Surprisal	0.20	0.04	4.99	$< .001$
Boundary (word) * Surprisal	-0.005	0.03	-0.19	$= .85$
Local tempo * Surprisal	-0.009	0.001	-6.95	$< .001$

that word frequency had a stronger effect on vowel dispersion than surprisal. To the knowledge of the author, only word frequency and cloze probability (Scarborough, 2006), or word frequency and information content (van Son, Bolotova, et al., 2004) were used simultaneously in studies on ID and vowel dispersion. These previous works also found additive effects of ID factors on vowel dispersion. Although, it should be mentioned that van Son, Bolotova, et al. (2004) stressed the language-dependency of this relation in their cross-linguistic study on Dutch, Finnish and Russian vowels.

As expected, vowel dispersion was higher in stressed than in unstressed vowels. The factor stress also had the strongest effect size of all fixed effects in the model. Its effect was larger than both of the ID factors' effects combined which was assumed considering the findings in Aylett and Turk (2006). In addition, we found a positive interaction between surprisal and stress on vowel dispersion, i. e., vowels were more distinct in high surprisal and stressed position. This observation further strengthened the hypothesis that ID is contingent upon prosodic structure (Aylett and Turk, 2004).

We found that at both word and phrase boundary vowels were less dispersed than at no boundary position. This finding was not in line with Turk's (2010) assumption of a positive relation between prosodic boundary and acoustic salience. At the level of word boundaries, and even stronger at higher levels of prosodic boundaries in the prosodic hierarchy, Turk (2010) claimed that boundary markers correlated inversely with language redundancy. This was why they controlled for this factor by discarding vowels at boundary position from their analysis of vowel duration and F1/F2 properties (Aylett and Turk, 2006). In this thesis, we found that acoustic salience at boundary position was not expressed by vowel distinctiveness but by expansion in segment duration (Chapter 5).

However, the interaction of phrase boundary and surprisal led to an increase in vowel dispersion. The interaction between surprisal and word boundary for vowel dispersion was not significant. This finding showed that in German acoustic salience is expressed at boundary position only when this prosodic boundary is at a high level of the prosodic hierarchy (phrase boundary vs. word boundary), and the vowel stands in a high surprisal context. In the interaction analysis, we therefore confirmed that higher prosodic boundaries have a stronger influence on the acoustic salience of vowels than lower acoustic boundaries (Turk, 2010). In addition, there is a high degree of

optionality and variability in the effect of prosodic boundaries on acoustic salience. Processes of acoustic salience may also be weaker at higher prosodic boundaries (Kuzla and Ernestus, 2011). Considering our results of vowel dispersion at different boundary positions, and the interaction between boundary and surprisal, these observations can be explained by different properties in the ID profiles across boundaries.

We controlled for the known effect of speech rate deviation on vowel dispersion (Turner et al., 1995; Weiss, 2007). In contrast to Section 8.3, speech rate was not intentionally varied by the speakers. We therefore included speech rate as a continuous factor in our analysis. Both accelerated global and local speech rate led to a decrease in vowel dispersion. This finding was in line with similar studies investigating the relationship between vowel distinctiveness and naturally occurring differences in speech rate (Weiss, 2007). The interaction analysis showed that even at high surprisal vowels were less distinct when they were produced with a fast local speech rate. Apparently, local speech rate overruled surprisal in its effect on vowel distinctiveness. Global sentence rate did not interact significantly with surprisal in explaining vowel dispersion variability.

Vowel dispersion depends on average vowel duration of the phoneme. Longer vowels tended to show more dispersion than short vowels ($r = 0.34, t(57741) = 87.87, p < 0.001$). Average vowel duration was a relatively strong predictor compared to the other fixed effects ($Var = 1.66\,\%$). This relationship is well known for German, but also for other languages (Gendrot and Adda, 2005). Previous research has found that this relationship is not pronounced in the same way for all vowel phonemes. Open vowels that are inherently longer than closed vowels also showed a more pronounced correlation for vowel dispersion and duration (Aylett and Turk, 2006). In our statistical analysis vowel openness was not included, but vowel identity as a random effect, because effects of vowel dispersion were assumed to show the same tendencies and only vary randomly across vowel phonemes.

The main proportion of data variance in vowel dispersion was explained by the random structure. We found expected tendencies for peripheral vowels to be more dispersed than interior vowels (Figure 8.3). These systematic differences in vowel dispersion as a function of vowel identity explained about 20 % of the data variance. Interestingly, preceding phonological context still explained some of the data variance (about 3 %), even though we included biphone surprisal of the preceding context as a fixed effect. This finding suggested that n-phone surprisal and phonological context do not express the exact same predictor, although they surely express similar relations based on phonotactics. Following phonological context, on the other hand, was much less informative in the model structure ($Var = 0.44\,\%$). This can explain why surprisal of the following context did not show a positive relationship with vowel dispersion either. Speaker identity was even less informative in the vowel dispersion LMM which was probably due to speaker normalization of formant values used in the analysis.

8.3 Vowel dispersion in six languages

This section outlines results and discussion of vowel dispersion as a function of ID and prosodic factors from a cross-linguistic perspective. A subset of the BonnTempo corpus (Section 3.1.2) containing AE, CES, DEU, FIN, FRA, and POL data was analyzed. The content of this chapter is based on the author's contribution to a journal paper (Malisz et al. (2018)), but was extended and revised for this thesis.

8.3.1 Descriptive statistics

There was not a large range of average vowel dispersion among the six languages that were analyzed ($R = 1.21, 1.28$). DEU had the largest vowel dispersion ($M = 1.28, SD = 0.59$), and Polish vowels the least amount of vowel distance from the center ($M = 1.22, SD = 0.70$). Vowels in CES ($M = 1.26, SD = 0.62$), FIN ($M = 1.26, SD = 0.63$), and FRA ($M = 1.25, SD = 0.63$) were almost equally dispersed on average, while vowels in AE ($M = 1.24, SD = 0.65$) were a little less dispersed.

Averaged over all languages, vowel dispersion was higher in stressed ($M = 1.28, SD = 0.65$) than in unstressed vowels ($M = 1.24, SD = 0.63$). In vowels preceding a word boundary, vowel dispersion was the lowest ($M = 1.19, SD = 0.61$), followed by no boundary ($M = 1.30, SD = 0.66$), and phrase boundary ($M = 1.33, SD = 0.57$). As expected, vowels at fast speech rate showed the least dispersion ($M = 1.20, SD = 0.60$), and dispersion increased with speech rate normal ($M = 1.24, SD = 0.64$), and slow ($M = 1.32, SD = 0.65$). With regard to vowel identity, the largest distance from the vowel space midpoint was found for closed back vowels ($M = 1.89, SD = 0.55$), followed by closed front vowels ($M = 1.44, SD = 0.62$). The mid open vowels ($M = 1.19, SD = 0.56$) were more dispersed than the mid front vowels ($M = 0.90, SD = 0.50$).

8.3.2 Linear mixed-effects model

Across languages, biphone surprisal of the preceding context showed the strongest relationship with vowel dispersion (Table 8.3). We therefore used this measure in the following analysis.

Overall, we saw the same tendency for a positive correlation between vowel dispersion and surprisal across all languages, except for FIN ($r = -0.01, t(1176) = -0.42, p = 0.68$). This relationship was also consistent across intended speech rates (Figure 8.4). In the following LMM analysis, we therefore used LANGUAGE as a random effect, and not as a fixed effect. The assumption behind this was that vowel dispersion varied randomly across languages with (almost) all languages showing the same tendencies following Pellegrino et al. (2011).

We performed a collinearity analysis to assure that all fixed effects were independent factors. There was a weak positive correlation between stress and surprisal

Table 8.3: Vowel dispersion in six languages: Pearson's correlation coefficients and tests ($\alpha = 0.05$) between vowel dispersion and biphone surprisal of the following and preceding context, and triphone surprisal from neighboring context.

	Biphone following	Biphone preceding	Triphone
AE	0.10*	0.26***	0.16***
CES	0.06*	0.24***	n.s.
DEU	-0.09**	0.30***	0.14***
FIN	0.12***	n.s.	n.s.
FRA	0.25***	0.18***	0.26***
POL	0.20***	0.12***	n.s.

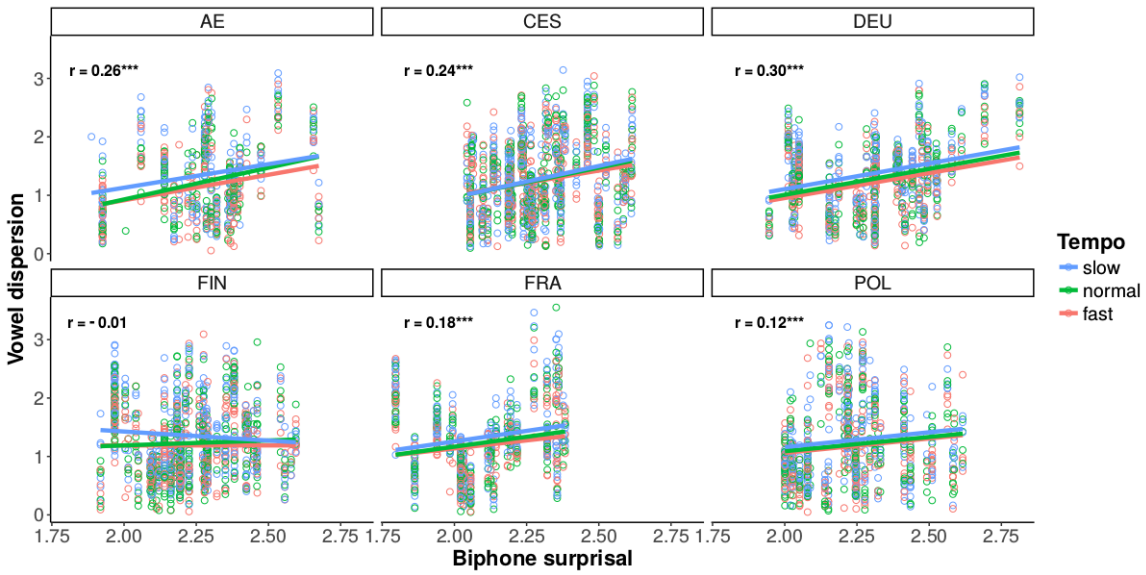


Figure 8.4: Correlation between vowel dispersion and biphone surprisal of the preceding context at different intended speech rates in AE, CES, DEU, FIN, FRA, POL.

($r = 0.24$) indicating that vowels in high surprisal contexts showed a tendency to stand in syllables carrying primary lexical stress. Other potential dependencies between the fixed effects were not found.

For the baseline vowel dispersion model, the fixed effects SURPRISAL of the preceding biphone, VOWEL IDENTITY, STRESS, SPEECH RATE, and BOUNDARY, as well as GENDER were entered. The final model converged with random intercepts for FOLLOWING SEGMENT and PRECEDING SEGMENT, LANGUAGE and WORD, and random slopes for SURPRISAL per WORD. The fixed effect GENDER and the random intercept SPEAKER did not explain any variance in the data and were removed from the model structure. SURPRISAL values were log-transformed because of positive skewness. VOWEL IDENTITY was sum-coded, whereas all other categorical factors were treatment-coded (Model structure 8.2).

$$\begin{aligned} VowelDispersion \sim\ & BiSur \\ & Stress + Boundary + IntendedTempo + \\ & VowelIdentity + \\ & (1 + BiSur|Word) + (1|Language) + \\ & (1|Preceding) + (1|Following) \end{aligned} \tag{8.2}$$

In the baseline model, all fixed effects, but BOUNDARY and STRESS, reached significance level (Table 8.4). As expected, vowel dispersion was positively affected by SPEECH RATE: as the speech rate got slower, the vowel dispersion measure increased. There were significant differences between vowel dispersion at normal and fast, and between slow and fast SPEECH RATE. Additional post-hoc analysis (Tukey Contrasts) revealed that vowels at slow SPEECH RATE were more dispersed than vowels at normal SPEECH RATE ($\beta = 0.08, z = 6.95, p < 0.001$). We found a tendency for an effect for vowels in stressed syllables to be more dispersed than vowels in unstressed syllables. Regarding the ID measure, vowels with high biphone SURPRISAL values were significantly more dispersed than vowels with lower SURPRISAL values. Vowel dispersion also depended on the vowel identity. On average, /i/ was significantly more dispersed than the grand mean, while vowels /a/ and /e/ were less dispersed than the grand mean.

The marginal pseudo-R^2 indicating how much variance is explained by the fixed factors showed that the baseline prosodic factors explained 0.66 % of the vowel dispersion variance. The significant effect SPEECH RATE contributed 0.24 % to that explained variance. The effect size increased by 2.24 % when SURPRISAL was included in the additive model. A large amount of variance was explained when VOWEL IDENTITY was added to the model (17.52 % increase). The conditional pseudo-R^2 for the variance explained by both fixed and random effects equaled 86.66 % in the final model.

In a second step, interactions were entered into the baseline model of vowel dispersion. Interactions between all prosodic factors and SURPRISAL were tested comparing

Table 8.4: Vowel dispersion in six languages: regression coefficients, standard error (SE) and statistical output of LMM analysis including biphone surprisal of the preceding context.

Terms	Coeff.	SE	t-value	p-value
Surprisal	0.70	0.26	2.65	= .009
Stress (y – n)	0.02	0.04	0.39	= .69
Boundary (word – none)	-0.08	0.05	-1.72	= .08
Boundary (phrase – none)	-0.05	0.04	-1.35	= .18
Tempo (normal – fast)	0.04	0.01	3.16	= .01
Tempo (slow – fast)	0.12	0.01	10.03	< .001
Vowel identity (/a/ – Mean)	-0.11	0.03	-3.25	= .001
Vowel identity (/e/ – Mean)	-0.59	0.04	-16.84	< .001
Vowel identity (/i/ – Mean)	0.35	0.04	8.60	< .001

the interaction model to the baseline model. None of the interaction models had a better fit than the baseline model.

8.3.3 Discussion

We followed the hypothesis that encoding strategies of vowel dispersion in different surprisal contexts are consistent across intended speech rates and different languages. This assumption was based on Pellegrino et al. (2011) who found a systematic relationship between information transmission and speech rate as part of phonetic encoding. By and large, this hypothesis was confirmed for all languages investigated here, except for Finnish. Here, we found a non-significant relationship between vowel dispersion and biphone surprisal of the preceding context ($r = -0.01, t(1176) = -0.42, p = 0.68$). This finding was due to two possible reasons. Finnish has a weak expression of vowel quality differences in different stress conditions which was why we did not find a positive relationship between surprisal and vowel dispersion, similar to van Son, Bolotova, et al. (2004) who studied stressed word-initial Finnish vowels. Additionally, vowel quality in Finnish is morphophonemic. Front vowels (/æ, y, ø/) never appear with back vowels (/ɑ, u, o/) in the same lexeme (Bertram et al., 2004) which is why vowel reduction would be detrimental to identifying lexeme boundaries in this language.

We found a significant positive effect of surprisal on vowel dispersion. Vowels in high surprisal contexts were more dispersed in their spectral characteristics than in low surprisal contexts, as expected (Aylett and Turk, 2006; Jurafsky, Bell, and Girand, 2002; Jurafsky, Bell, Gregory, et al., 2001). Based on marginal pseudo-R^2 values, surprisal explained a larger quantity of the vowel dispersion variance than the prosodic factors used here. This result was contrary to findings in Aylett and

Turk (2006) who reported an overall smaller effect of language redundancy on vowel formants F1 and F2 than for the prosodic model. These contrasting findings regarding effect size should be interpreted cautiously because stress and surprisal were weakly positively correlated (see below).

In contrast to Schulz et al. (2016), we only found a tendency of a positive effect of stress on vowel dispersion, and not a significant effect. This difference might be due to the weak positive correlation between surprisal and stress ($r = 0.23$). Effects for both variables cannot be fully separated in a statistical model. In addition, Schulz et al. (2016) analyzed only five languages of the BonnTempo corpus, DEU, CES, POL, FIN, and FRA. The present study also included American English.

The current study replicated results regarding differences between vowel dispersion as a function of speech rate. Vowel dispersion increased with decreasing speech rate (Turner et al., 1995; Weirich and Simpson, 2014; Weiss, 2007). Vowel formants moved to a more central position in the F1/F2 vowel space under fast speech rate when investigated in intended tempo deviations (Turner et al., 1995) and in naturally occurring differences in speech rate (Weiss, 2007). We also saw the same effect in our previous analysis for German vowel dispersion as a function of continuous naturally occurring speech rate deviations (Section 8.2).

The factor prosodic boundary was not significant in the LMM, in contrast to the analysis above on vowel dispersion in German which showed that vowels were less distinct at boundary position, and that phrase boundary interacted positively with surprisal resulting in higher vowel dispersion (Section 8.2). This discrepancy might be due to the diversity of languages included in one statistical model. Prosodic boundaries are possibly not expressed in the same way in all of these languages.

Aylett and Turk (2006) emphasized the large degree of variability of unique or shared contributions of their redundancy and prosodic model in explaining variance in F1 and F2 of AE vowels among different vowel phonemes. The current study also showed that the impact of ID and prosody largely depended on the investigated vowel identity, although a different measure of vowel dispersion was used. The factor vowel identity explained 17.5 % of the variance in the vowel dispersion measurements. In addition, vowel identities differed in their magnitude of dispersion compared to the mean. Closed front vowels were significantly more dispersed than the grand mean, while open mid vowels and front mid vowels were less dispersed than the grand mean.

8.4 Vowel dispersion of Bulgarian L2 speakers of German

The following section presents a vowel dispersion analysis and its relation to ID in the context of L2 speech. We analyzed vowels extracted from read production data based on passages from the EUROM-1 corpus (Section 3.1.3). The text passages were

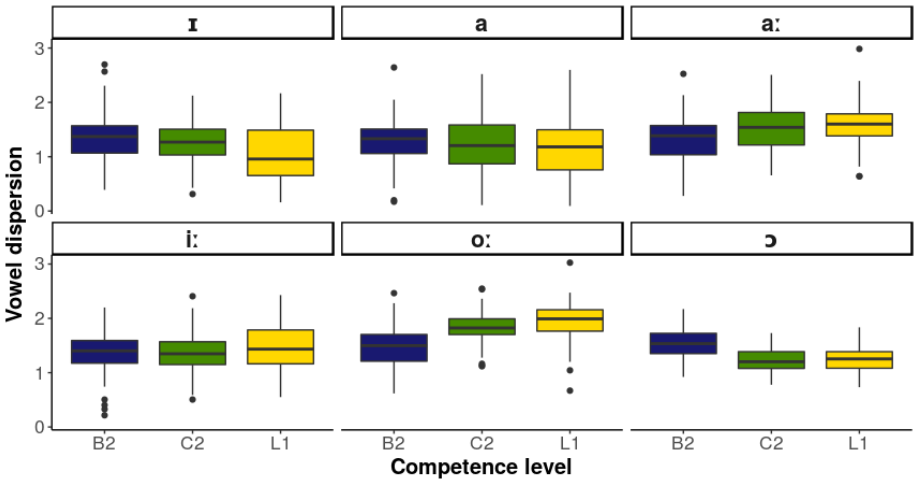


Figure 8.5: Vowel dispersion of German natives (L1) and Bulgarian speakers of German at intermediate (B2) and advanced proficiency level (C2) per analyzed vowel phoneme.

read by Bulgarian L2 speakers at intermediate competence level (B2) and advanced competence level (C2) level, as well as by a German native control group (L1).

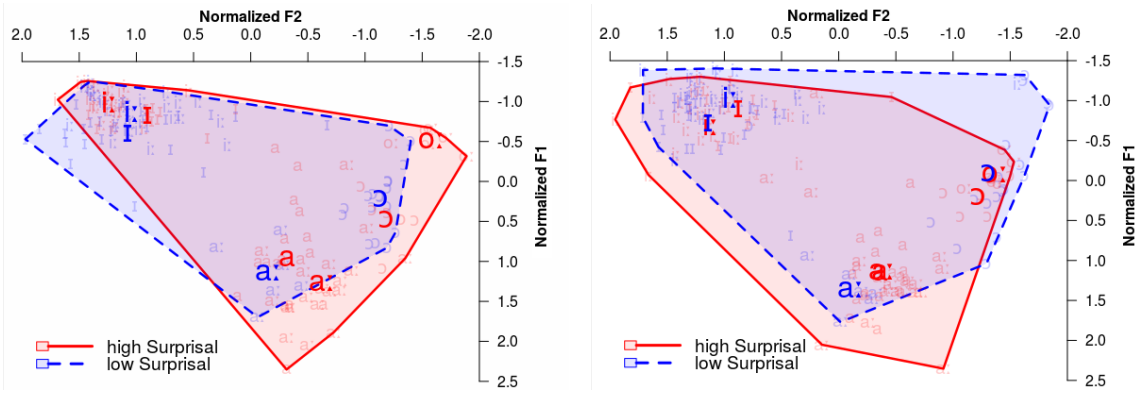
8.4.1 Descriptive statistics

On average, Bulgarian L2 speakers showed larger vowel dispersion calculated over all vowels than German native speakers ($M = 1.32, SD = 0.50$). B2 Bulgarian speakers ($M = 1.36, SD = 0.37$) were slightly more dispersed in their German vowel production than C2 speakers ($M = 1.34, SD = 0.43$).

L2 speakers at both proficiency levels approached native German production of /aː, a/ successfully. The closed back vowels /oː, ɔ/ were similarly dispersed in C2 and L1 speakers, whereas B2 speakers showed less dispersion for /oː/, and more for /ɔ/ compared to the other two speaker groups. With regard to the closed front vowels /iː, ɪ/ we found that neither the C2 nor the B2 learners reached the same level of dispersion as the German native speakers. While the target /iː/ was approached with a little less vowel dispersion than in the native speech, the L2 speakers showed much higher dispersion values for the lax vowel /ɪ/ (Figure 8.5).

Advanced Bulgarian speakers of German approached the pattern for vowel space expansion in different ID contexts that was observed in Section 8.2 for German natives (Figure 8.2). For advanced Bulgarian L2 speakers, German back and mid vowels were less dispersed under low surprisal than under high surprisal. Regarding the front vowels, we observed that /iː/ and /ɪ/ approached a similar position in the vowel space under low surprisal, while there were clearly separated under high surprisal condition (Figure 8.6a).

In contrast, intermediate speakers did not show the expected pattern of vowel space reduction under low compared to high surprisal. Instead, we found that low



(a) Advanced proficiency level (C2). **(b)** Intermediate proficiency level (B2).

Figure 8.6: Vowel space of Bulgarian L2 speakers under high and low surprisal at different proficiency levels. Binning of surprisal was based on 10 % of the highest and lowest values in the data set.

surprisal vowels were raised in relation to high surprisal vowels. This observation reflected the native Bulgarian vowel reduction pattern (Andreeva et al., 2013) (Figure 8.6b).

8.4.2 Linear mixed-effects model

We tested the relation between vowel dispersion and surprisal per speaker group in correlation tests. Replicating our results presented in Section 4.2.3, we found the strongest positive correlation between vowel dispersion and triphone of the preceding context for the L1 speakers ($r = 0.23, t(794) = 6.58, p < 0.001$). The correlation decreased in strength with increasing n-phone size that was used to calculate surprisal values. There was also a significant positive correlation between these measures for the L2 speakers at C2 level ($r = 0.14, t(795) = 3.86, p < 0.001$), but no significant relationship for the B2 speakers ($r = -0.02, t(796) = -0.69, p < 0.49$). Following the results of the correlation analysis we calculated three different LMMs for each speaker group using triphone surprisal of the preceding context as an ID measure.

Due to the phonemic status of the vowels included in the analysis we decided to control for vowel tenseness (tense vs. lax) in the statistical model. Also, we added average vowel duration based on the production data as a control variable.

A collinearity analysis was performed to identify potential dependencies of the individual fixed effects. Due to the small number of data points per group we included content and function words in the statistical analysis, while introducing the factor word class to the model. Word frequency and surprisal were moderately related ($r = -0.62$). Vowel phonemes in words with high frequency showed low surprisal values. Word frequency and word class were strongly correlated ($r = 0.80$) with function words showing higher frequency values than content words. Word class and surprisal, on the other hand, showed a weaker correlation ($r = -0.50$) than

word frequency and surprisal. Average vowel duration and surprisal were only weakly related ($r = 0.26$). Higher surprisal values were correlated with longer vowel duration. A similar relationship was observed for vowel dispersion and average vowel duration ($r = 0.23$). Vowel tenseness was correlated with average vowel duration as well ($r = 0.50$) indicating that, on average, tense vowels were longer than lax vowels.

As a result of the collinearity analysis SURPRISAL, WORD CLASS, average vowel DURATION, and vowel TENSENESS were included as fixed effects. WORD FREQUENCY was excluded as a predictor in this model because it showed strong correlations with WORD CLASS, and a moderate correlation with SURPRISAL. The random structure of the model consisted of random intercepts for SPEAKER and WORD. LMMs with larger random structure did not converge considering the small amount of data points per model. Vowel TENSENESS was sum-coded, WORD CLASS was treatment-coded, and both continuous predictors were log-transformed (Model structure 8.3).

$$\begin{aligned} VowelDispersion \sim\ & TriSur + WordClass + \\ & Tenseness + DurAverage \\ & (1|Speaker) + (1|Word) \end{aligned} \tag{8.3}$$

In the LMM for German natives, we found expected significant effects for vowel TENSENESS, WORD CLASS, and DURATION. Long, tense vowels were more dispersed than short, lax vowels. Vowels in function words were less dispersed than vowels in content words. However, we only found a tendency for a positive effect of SURPRISAL on vowel dispersion, possibly due to data sparsity, and the strong effects of the control factors.

In the model for the C2 speakers, there were significant effects of TENSENESS and DURATION in the expected directions, and a tendency for a negative effect of WORD CLASS. Since we did not find a significant effect for SURPRISAL on vowel dispersion in the model for German natives, we did not expect to observe a significant effect in the models for L2 speakers since correlation values between SURPRISAL and vowel dispersion were lower or non-significant for these two groups. In the LMM for Bulgarian B2 speakers of German, there were no significant effects of any of the fixed effects to explain vowel dispersion (Table 8.5).

We calculated effect sizes for all three models and their significant effects separately. The largest overall effect size of the entire model indicated by conditional pseudo-R^2 was found for the German natives ($Var = 63.39\,\%$). The same model structure explained only 37.71 % of the variance in the data of L2 advanced speakers, and even less variance in the vowel dispersion of L2 intermediate speakers ($Var = 24.11\,\%$). For both German native speakers ($Var = 18.72\,\%$) and L2 advanced speakers ($Var = 10.17\,\%$) average vowel DURATION was the strongest predictor of vowel dispersion. Vowel TENSENESS added 6.42 % explained variance for the German native data, and 1.48 % in the model for C2 vowel dispersion. While WORD CLASS was not

Table 8.5: Vowel dispersion of L1 and Bulgarian L2 speakers of German: regression coefficients, standard error (SE) and statistical output of LMM analyses including triphone surprisal of the preceding context. The LMM was run on content (C) and function (F) words.

	Terms	Coeff.	SE	t-value	p-value
L1 speakers	Surprisal	0.02	0.08	0.25	= .80
	Tenseness (lax – tense)	-0.21	0.04	-5.27	< .001
	Word class (F – C)	-0.19	0.07	-2.53	= .01
	Average vowel duration	0.42	0.11	3.81	< .001
L2 speakers (C2)	Surprisal	-0.03	0.07	-0.50	= .62
	Tenseness (lax – tense)	-0.13	0.03	-4.13	< .001
	Word class (F – C)	-0.11	0.06	-1.98	= .05
	Average vowel duration	0.31	0.12	2.58	= .01
L2 speakers (B2)	Surprisal	-0.05	0.06	-0.79	= .43
	Tenseness (lax – tense)	-0.04	0.03	-1.36	= .18
	Word class (F – C)	-0.001	0.05	-0.03	= .98
	Average vowel duration	0.11	0.11	0.99	= .32

significant in the L2 models, it explained 2.03 % of data variance in German native vowel dispersion.

8.4.3 Discussion

This study investigated whether Bulgarian L2 speakers of German behave similar to German native speakers in their vowel dispersion in different surprisal contexts, and whether their vowel productions depended on their proficiency level of German. Vowel dispersion was measured for the tense vowels /iː, oː, aː/ and their lax counterpart /ɪ, ɔ, a/ in read speech from 6 German natives, 6 advanced Bulgarian L2 speakers (C2), and 6 intermediate Bulgarian L2 speakers of German (B2).

Replicating our previous results on the relationship between vowel dispersion and surprisal (Section 8.2), we found a significant positive correlation between triphone surprisal of the preceding context and vowel dispersion in German native speakers. In addition, advanced L2 speakers showed a significant positive relationship between those two measures, while this relation was not visible in intermediate L2 vowel productions. These findings showed that advanced L2 speakers of German show a tendency to modulate their vowel productions in the same way as German natives with regard to ID factors, whereas intermediate L2 speakers were not able to differentiate in their vowel productions according to measures of German native predictability. This result indicated that proficiency level of L2 speakers can be expressed as the degree of familiarity with German native language structures and their predictabilities on a

sub-word level.

Although we found significant positive correlations between surprisal and vowel dispersion for German natives and C2 speakers, this effect was not significant in a more complex LMM analysis with control factors word class, average vowel duration, and vowel tenseness. This was probably due to the small amount of data points per LMM, the restricted number of surprisal contexts because of the short length of the text passage, and the specific nature of vowel phonemes and their behavior under high and low surprisal. The overall range of triphone surprisal of preceding context in the data ($R = 0.52, 2.43$) was less than half as large as the surprisal range in the big data set for German vowel dispersion ($R = -1.94, 3.05$) (Section 8.2). Detailed analysis of the relationship between surprisal and vowel dispersion for different vowel identities showed that for /ɪ/ ($r = -0.12, t(177) = -1.64, p = 0.10$) and /ɔ/ ($r = 0.01, t(64) = 0.07, p = 0.94$) the regression was either negative or flat, while the positive relation between the two measures held across all other vowel phonemes in the data set. Regarding the triphone context we found that /ɔ/ ($n = 198$) and /oː/ ($n = 198$) had the smallest diversity with 7 and 9 different contexts respectively. The largest number of different triphone contexts was observed for /a/ with 15 contexts for 666 data points per sub group of the data set.

Advanced L2 speakers of German were able to differentiate their vowel productions with regard to differences in tenseness and vowel duration. They also showed a tendency to produce native-like differences between vowel tokens depending on whether they stood in function or content words. These effects were not found in B2 speakers of German. We can therefore clearly separate the two proficiency levels using vowel dispersion as an acoustic measure. Interestingly, the amount of German proficiency of the three groups was also mirrored in the effect sizes of the corresponding LMMs. The effect size of the model decreased with decreasing proficiency level of German. Average vowel duration was the strongest predictor for vowel dispersion for both German natives and advanced L2 speakers. However, this effect size should be interpreted with caution because vowel duration and tenseness were positively correlated ($r = 0.50$).

The descriptive analysis of vowel dispersion in L1 and Bulgarian L2 speakers showed that proficiency level depended on the vowel phoneme under investigation. While both L2 groups approached native-like mid-open vowel /a/ productions, they did not successfully discriminate between /iː/ and /ɪ/ in the way that German natives did. Also, B2 speakers were not able to distinguish /oː/ and /ɔ/ as German natives did. Both had similar vowel dispersion values and did not show native-like dispersion patterns (Figure 8.5).

We found that Bulgarian L2 speakers, in particular at B2 level, showed more vowel dispersion than German natives. This finding was not surprising considering that under low surprisal Bulgarian B2 speakers of German raised their vowel space mirroring native Bulgarian vowel raising in unstressed condition (Andreeva et al., 2013) (Figure 8.6b). Although none of the analyzed vowels carried an accent we found this characteristic reduction pattern for low surprisal vowels in the B2 speakers.

This can be interpreted as a certain degree of awareness in Bulgarian B2 speakers of the German native phonological structures and their predictabilities. But they were not able to produce the native-like reduction pattern for the respective vowels in low surprisal context. In contrast, they relied on their native reduction pattern.

In this chapter, we investigated the static spectral characteristics of vowels as vowel dispersion and their relation to ID and prosody in German read speech, in a cross-linguistic analysis, and a study on L2 speakers of German. The following chapter is concerned with dynamic spectral characteristics of vowels and shows how dynamic formant trajectories relate to ID and prosody.

Chapter 9

Dynamic formant trajectories

Vowel dispersion is based on formant measurements at the assumed steady state of the vowel. It takes a snap shot of vowel spectral characteristics as opposed to a description of the dynamic formant trajectories over the entire duration of the vowel. We use dynamic formant metrics based on the onset and offset of formant movements, so called VISC measurements, as well as metrics based on several data points within the vowel, such as VSL or TL, in addition to parametric measurements, i. e., DCT and polynomial coefficients. The following chapter first outlines how these metrics are related. Based on a correlation analysis we limited the number of dynamic formant measurements, and investigated to what extent formant movement is predicted by ID and prosodic factors.

This chapter is a revised and extended version of an analysis of VISC measures and ID presented in Brandt et al. (2018). In this paper, we ran the data analysis only on the monophthongs in the corpus and not on all vowels which is presented here.

9.1 Method

Formant measurements were taken using the tool FormantPro (UCL, 2013) for large-scale analyses of vowel formants and formant movements written for Praat (Boersma and Weenink, 2017). From the output of FormantPro the files containing information about time-normalized formant values, mean formant values, as well as time-normalized formant velocity were used for the following analysis. FormantPro estimates formant values using the Praat command "To Formant (burg)" with the default values of time step 0, a maximum of five formants, the maximum formant value set at 5,000 Hz for male speakers, an analysis window of 25 ms, and pre-emphasis from 50 Hz. Additionally, we left the default parameter of FormantPro for the number of normalized times per analyzed interval at 20. Continuous trajectories of F1, F2, F3

(in Hz) and F2_3 (in Hz/s) which is calculated as

$$F2\_3 = 0.5 * (F2 + F3), \tag{9.1}$$

and formant velocity values of F1, F2, F3, and F2_3 calculated as

$$V = (F\_n + 1 - F\_n)/time\_step^{1}) \tag{9.2}$$

were merged with formant means values extracted by FormantPro containing information about vowel duration, as well as mean formant values. F2_3 is interpreted as a joint indicator of tongue frontness.

In order to analyze formant trajectories one needs at least two sampling points within the vowel duration. Typically, this minimum set of sampling points contains formant measurements taken at the onset and offset of the vowel (Nearey and Assmann, 1986). Usually, formants are also estimated at target position or assumed steady state (Hillenbrand, Getty, et al., 1995). Target values can be defined as the maximum or minimum value reached for a particular formant (Weismer and Berry, 2003), as the formant frequency average over the complete vowel duration, or as the formant frequency at the point of maximal energy within the vowel (van Son and Pols, 1990). Most studies on vowel formants use measurements at temporal midpoint of the vowel which is assumed to come closest to a steady state position (Broad and Wakita, 1977; Schouten and Pols, 1979). Alternatively, the steady state is defined as the time point in the vowel with the least overall formant movement estimated over neighboring frames (Miller and Volaitis, 1989).

In order to calculate more fine-grained measures of formant movement, such as spectral rate of change (roc), more than three sampling points within one vowel token are necessary. Combinations of these procedures depending on the frontness and height of vowels, as well as on the amount of formant movements have to elicit reliable formant measurements (C. I. Watson and Harrington, 1999). At least two additional sampling points are used adding up to a total of five data points per vowel token (Fox and Jacewicz, 2009). Alternatively, one can use data points at each tenth time-normalized point within the vowel (at 10 %, 20 %, 30 % ...etc.). From the 20 time-normalized points in each vowel interval given by FormantPro six equally spaced points were chosen for further analysis. Vowel measurements were taken at 15 %, 30 %, 45 %, 60 %, 75 % and 90 % of the normalized duration. We did not use a denser sampling set of formant measurements because each of these sampling points had to be cleaned identifying tracking errors and leading to a potential loss in data points. If we had chosen a higher number of data points per vowel, more extensive cleaning would have been necessary.

The cleaning procedure was performed on the entire data set including vowels in function and in content words. Data cleaning involved plotting the F1 and F2 values

[1] FormantPro uses Praat's default time step of 0.25 sec and multiplying with window length of 0.025 sec which equals 0.00625 sec.

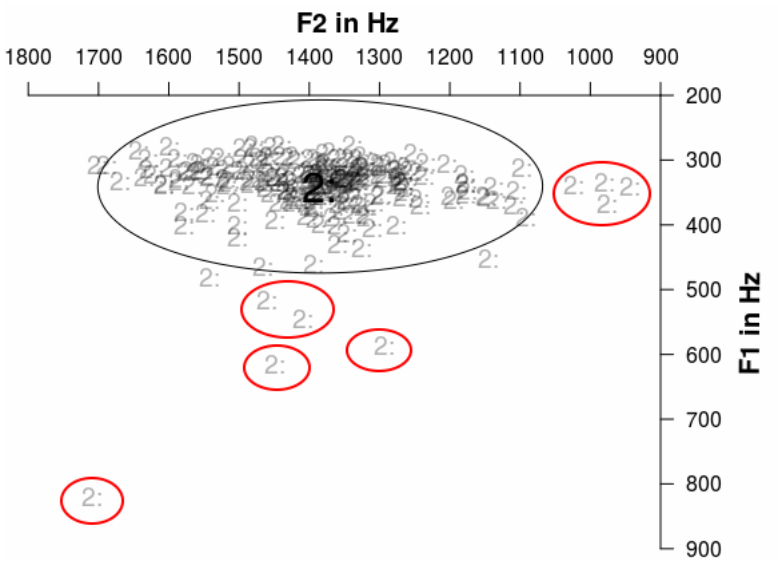


Figure 9.1: Example of identification of potential formant tracking errors using vowel /øː/ (SAMPA) of speaker "ai" with ellipsis at 95 % confidence interval. Red circled formant measurements indicate potential candidates of formant tracking errors that were manually checked.

of the vowel phonemes per speaker with their respective ellipse at 95 % confidence interval to identify spurious values, as described in Harrington (2010) (for example Figure 9.1). These data points were manually checked and excluded from the data set if the formant values were tracking errors. Starting from the first time stamp in the vowel and continuing chronically to the last time stamp 312 vowels were excluded for speaker "wo", and 321 for speaker "ai". This means that from originally 86,706 only 0.73 % were discarded. Manual correction of the formant measurements would not have been feasible considering the values needed to be comparable to the time-normalized data taken by FormantPro.

Formant values were normalized per speaker using the Lobanov normalization which has outperformed other normalization procedures in comparative studies (Adank et al., 2004). All measures of formant movement were calculated based on this normalization procedure, except for the DCT coefficients. For this measure, we followed the analysis procedure recommended by Harrington (2010) and used mel-scaled formant values. DCT coefficients based on formant values from an auditory scale have proven to be more efficient in distinguishing phonetic categories than DCT measures based on raw Hz values. A small number of DCT coefficients on a mel-scale are more efficient than the same number of coefficients on a Hz scale (Tychtl and Psutka, 1999). This normalization method for formant values for DCT analysis contrasted with studies that used raw (C. I. Watson and Harrington, 1999) and log-transformed Hz value (Morrison, 2009), or non-linear frequency scaling of Hz values (Zahorian and Jagharghi, 1993). For the orthogonal polynomial fit of the formant data we followed Risdal and Kohn (2014) in using formant values normalized via Lobanov

transformation.

We used velocity of formants F1, F2, F3, and F2_3 calculated by FormantPro as indicators of the magnitude of formant change throughout the duration of the vowel. In addition, other established measures to estimate the amount of formant change were included in the analysis. On the one hand, VSL and TL, as well as rate of change (roc) of these two measures were calculated over all sampling points within a vowel phoneme. Formant delta and slope, as well as VL were taken for formants at the onset and offset of the vowel. Parametric measures, such as DCT and polynomial coefficients, were taken to describe the overall formant trajectory. The first four DCT coefficients and polynomial coefficients were calculated. All of these measures were based on F1 and F2 values.

We included ID factors surprisal and word frequency, as well as prosodic factors, such as prosodic boundary, primary lexical stress, and speech rate (global and local) in the analyses. Additional controls in the models were average vowel duration and vowel category of the vowel. All factors are explained in more detail in Section 4.4. Since the statistical analysis was performed with a model on the entire data set vowel category (monophthong vs. diphthong) was included to control for differences in formant change because of that factor.

9.2 Results

Excluding vowels in function words from the analysis there were 62,590 data points including all German vowels, monophthongs and diphthongs. Around 62 % of the data points stood in syllables in unstressed position ($n = 38,727$), while all others were marked as standing in stressed syllables ($n = 23,863$). Most vowels were identified as standing in no boundary position ($n = 56,790$), some at direct word boundary ($n = 4,826$), and only a few vowels were found to stand at phrase boundary position ($n = 974$). There were 4,862 diphthongs in the data set and 57,728 monophthongs (Figure 9.2).

9.2.1 Correlation analysis between dynamic formant metrics

Prior to a more detailed descriptive and inferential statistical analysis of the measures taken in this study it was crucial to perform a correlation analysis between them. The aim of this analysis was to exclude measures which highly correlated from further analysis. Since we calculated several measures that describe the amount of formant change, and did both DCT and polynomials for spectral curve fitting, a correlation analysis identified redundant measurements for further statistical analysis. We used absolute values for both DCT and polynomial coefficients, as well as for formant deltas, slopes and velocity in order to describe magnitude of formant change and movement and not the direction of these measurements. If not indicated otherwise, all reported Pearson's r correlations were at significance level $\alpha < 0.001$.

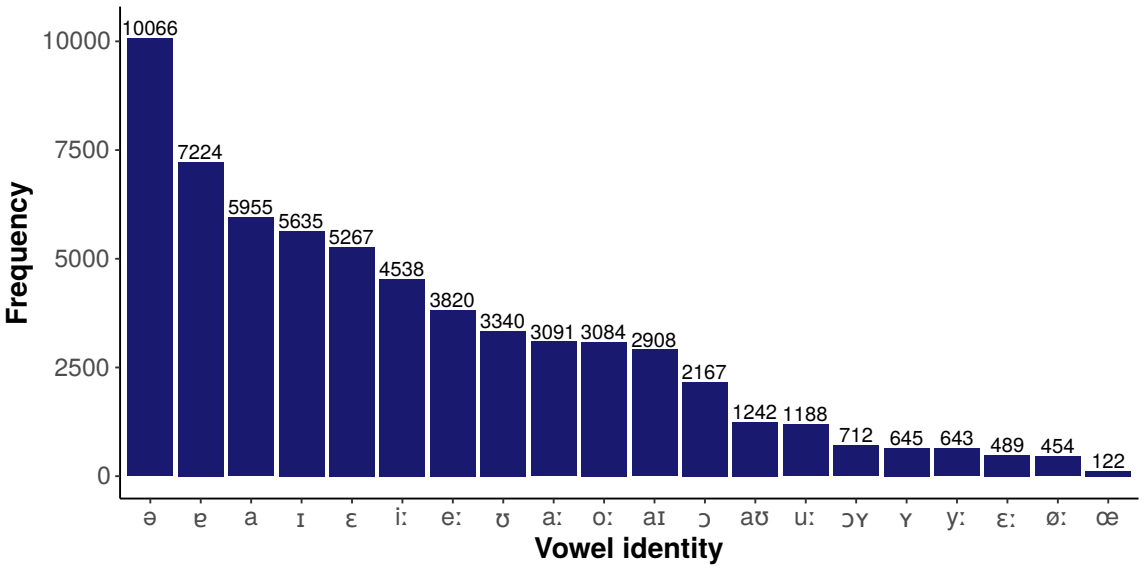


Figure 9.2: Frequency plot of number of tokens per vowel identity in content words of the Siemens Synthesis corpus.

DCT and polynomial coefficients DCT and polynomial coefficients are related since both transformations describe the entire spectral curve. The zeroth DCT coefficient and the constant polynomial coefficient both express the mean value of a formant trajectory averaged over all sampling points. F1 DCT0 strongly correlated with F1 Constant ($r = 0.96$), as well as F2 DCT0 with F2 Constant ($r = 0.98$). Since DCT1 and the linear polynomial both represent the slope of the signal they were also highly correlated for both F1 ($r = 0.85$) and F2 ($r = 0.80$). The formant measurements are fitted to the quadratic polynomial in a single curve with one turning point, while DCT2 describes the overall curvature of the signal. Both measures were strongly correlated for F1 ($r = 0.88$) and F2 ($r = 0.83$). DCT3 and the cubic polynomial fit are less straightforwardly related to each other than the other coefficients explained above. While DCT3 describes higher frequencies of the spectrum, the cubic polynomial fit as a function with two turning points and one inflection point allows a more detailed description of formant movement than lower polynomial coefficients. The measures were less strongly related for F1 ($r = -0.75$), and only moderately for F2 ($r = -0.50$).

There were no significant correlations within the group of DCT coefficients above moderate level ($r => 0.50$). For the polynomials, on the other hand, relationships between coefficients for F1 were above moderate correlation level. Since DCT and polynomial coefficients were correlated strongly and DCT coefficients did not show any intra-correlations above moderate level, we decided to use DCT coefficients above polynomials in the following analysis.

VSL and TL VSL is measured per sampling point in the vowel, whereas TL is the sum of all these measurements. Therefore the strong correlation found for both measures was expected ($r = 0.84$). Also, roc for both VSL and TL are related measures. VSL_roc is calculated at each sampling point, and TL_roc is the roc of TL at vowel midpoint. As expected, roc of TL and VSL were strongly correlated ($r = 0.87$). The roc measures were not highly informative either because they showed moderate to strong positive relationships with the measures VSL and TL. This was true for roc and the respective measures from which it was derived: VSL_roc and VSL ($r = 0.69$), TL_roc and TL ($r = 0.76$). Less strong relationships were observed for the correlations between VSL_roc and TL ($r = 0.62$), and TL_roc and VSL ($r = 0.58$).

Since VSL and TL were strongly related and their derived roc measures were not very informative, we decided to only use one of these measures in the following analysis. We decided in favor of VSL over TL because it is measured per sampling point in the vowel and is therefore assumed to give a more fine-grained profile of formant movement.

VISC measures According to our correlation analyses absolute formant slope and delta were not independent measures. Very strong correlations were found for F1 slope and Δ F1 ($r = 0.87$), as well as for F2 slope and Δ F2 ($r = 0.95$). Formants with steep slopes also showed larger differences between onset and offset values. VL was also related to absolute formant slopes and deltas: there were moderate correlations for F2 slope ($r = 0.54$), and Δ F2 ($r = 0.60$), as well as strong correlations for F1 slope and Δ F1 (both $r = 0.75$).

Due to these strong correlations between the VISC measures we decided to use formant slopes as measures of formant change over formant deltas because slopes are calculated relative to the vowel duration. We also included VL in the following analysis, despite its correlations with formant slopes, because it is a distance measure combining information on movement in both formants.

Formant velocity The velocity measures of different formants were not independent from each other. Based on the correlation analysis there was a low to moderate tendency for stronger formant change in one formant if it was also found in other formants. Strong correlations were observed between V F2_3 and V F2 ($r = 0.78$) and V F3 ($r = 0.83$). We did not find correlations between formant velocity and VSL, nor between velocity and VISC measures above a very low threshold. Therefore, they were interpreted as being independent measures.

Formant change and formant curve We also investigated the relationship between measures of formant change and formant curve. We could only find low correlations between VSL and DCT coefficients. Curvature of the signal (DCT2) for both formants F1 and F2 and these measures of formant change based on multiple

sampling points in the vowel showed only very low to low positive correlations. The more curved the profile of F2 in DCT2, the more formant change one can expect in VSL ($r = 0.24$). The same was true for F2 DCT3 and VSL ($r = 0.10$). However, we did not find these relationships for F1 DCT coefficients and VSL. Interestingly, there was a low negative relationship between mean values of F1 and F2 and measures of formant change. F1 DCT0 had less formant change in VSL ($r = -0.31$) as its value increased. Again, the same observation, although less pronounced, was made for F2 DCT0 and VSL ($r = -0.14$). None of the DCT coefficients were related to measures of (absolute) VISC nor to measures of formant velocity above a low level of correlation.

To summarize, we reduced the measures of dynamic formant trajectories to:

- Formant slopes
- VL
- VSL
- Formant velocity
- DCT coefficients

We discarded formant deltas because of their strong relationship to formant slopes, and TL because it was collinear with VSL, as well as their derived roc measures because they were strongly related with VSL and TL respectively. Polynomial and DCT coefficients had a large degree of overlap, especially in their first three coefficients. However polynomials also showed intra-correlations which was why DCT coefficients were chosen for further analysis. Formant change measures and curve-fitting DCT coefficients were only weakly related.

9.2.2 Formant change measures

Descriptive statistics

Formant slopes Formant slopes are formant deltas relative to the duration of the vowel. Formant deltas ($R = 13.25$) therefore have a much larger average range than formant slopes ($R = 0.71$). When comparing average values of formant slopes absolute values were reported because they reflect the amount of change rather than the direction of the formant movement. Analyzing spectral change separately in F1 and F2 slope based on absolute values showed that F1 values on average had more formant change ($M = 0.02, SD = 0.03$) than F2 values ($M = 0.01, SD = 0.01$).

In unstressed vowels ($M = 0.02, SD = 0.03$) there was more change in F1 slope than in stressed vowels ($M = 0.01, SD = 0.02$). However, there were slightly higher

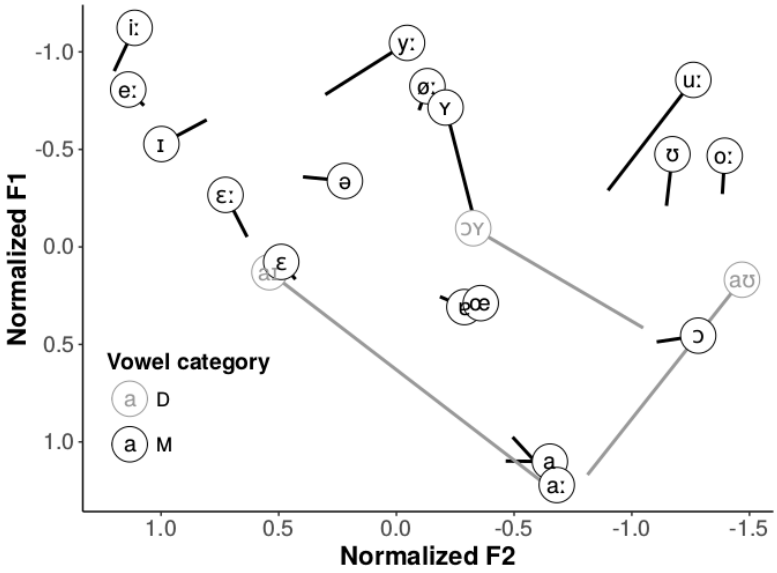


Figure 9.3: Spectral change patterns of Δ F1 and Δ F2 for all monophthongs (black), and diphthongs (grey) averaged across all phonetic environments. Only mean values are plotted.

values in F2 slope in stressed vowels ($M = 0.014, SD = 0.01$) than in unstressed vowels ($M = 0.012, SD = 0.01$). With regard to differences in formant slopes at different boundary position we found that vowels at word boundary ($M = 0.01, SD = 0.01$) showed less formant change in F1 than at no ($M = 0.02, SD = 0.03$) or phrase boundary position ($M = 0.02, SD = 0.02$), while F2 slope was constant across boundaries with the same level of variation ($M = 0.01, SD = 0.01$).

Vector length Figure 9.3 visualizes the amount of spectral change within the vowel. The length of the vector spanned by the F1 and F2 values of the vowels depicted in these figures is expressed by the measurement VL. Rounded closed vowels /ʏ, yː, uː/ showed the most spectral change in their F1 and F2 values compared to all other monophthongs. Diphthongs ($M = 1.57, SD = 0.74$) clearly had more formant value change than monophthongs ($M = 0.86, SD = 0.71$).

On average, vowels in stressed position ($M = 0.98, SD = 0.76$) showed higher values for VL than unstressed vowels ($M = 0.87, SD = 0.72$). At word boundary ($M = 0.69, SD = 0.56$), vowels had shorter VL than at phrasal boundary position ($M = 0.91, SD = 0.69$) or no boundary position ($M = 0.92, SD = 0.74$).

Vowel section length Average VSL ($M = 1.78, SD = 0.97$) showed a high standard deviation indicating that there was a lot of variability in the data in this measure. Interestingly, VSL was highest in closed vowels and diphthongs which was also observed in the data for the measure VL, although both measures only showed a weak positive correlation ($r = 0.19$). On average, VSL was slightly longer in stressed

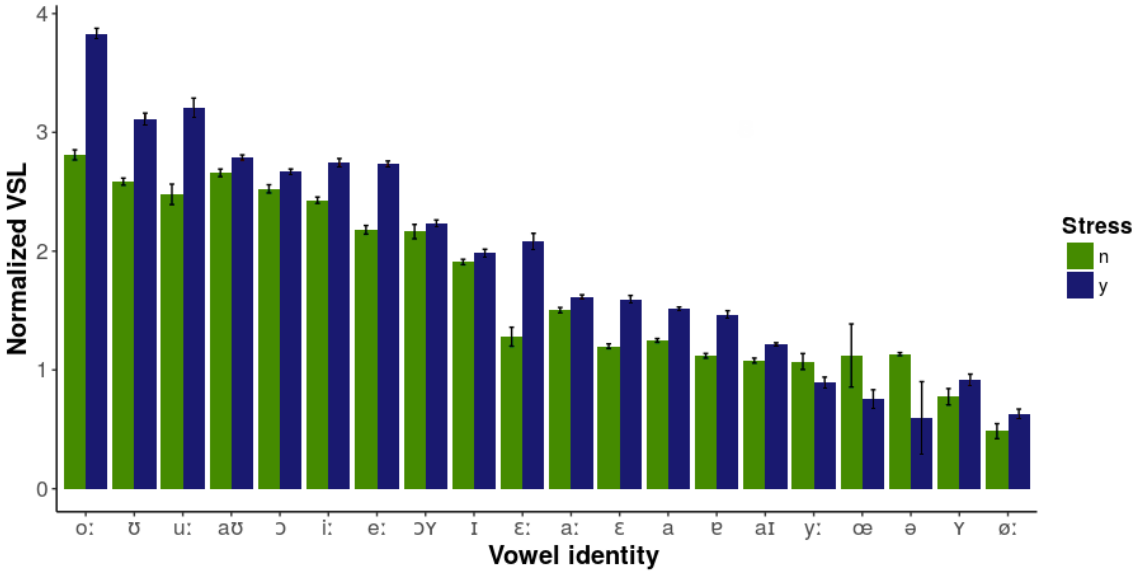


Figure 9.4: VSL per vowel identity and primary lexical stress condition.

vowels ($M = 2.06, SD = 1.00$) than in unstressed vowels ($M = 1.61, SD = 0.90$), and this effect also depended on vowel identity (Figure 9.4). Regarding the factor boundary we found highest formant change values in VSL for vowels at no boundary position ($M = 1.83, SD = 0.96$), less formant change in vowels at word boundary ($M = 1.37, SD = 0.99$), and vowels at phrase boundary ($M = 1.21, SD = 0.75$) with the least formant change.

Formant velocity Formant velocity values taken at six points within the vowel were averaged per vowel data point. These values inform about the overall direction of the formant trajectory and the amount of formant change. We also used absolute values for formant velocity. On average, F2 ($M = 2718, SD = 2612$) and F3 ($M = 2705, SD = 3045$) had a higher amount of velocity than F2_3 ($M = 2254, SD = 2438$) or F1 ($M = 1781, SD = 2279$). In stressed vowels, there was less velocity than in unstressed vowels in all formants measured here. Word and no boundary position showed increased velocity values compared to phrase position for all velocity measures (Table 9.1).

Linear mixed-effects model

VISC measures In order to analyze the relationship between measures of VISC and measures of ID we performed a correlation analysis correlating VL, F1 and F2 slope with surprisal measures of the preceding and the following context, as well as with word frequency, and phoneme probability. Word frequency and surprisal values were log-transformed due to positive skewness.

The correlation matrix in Figure 9.5 visualized previously discussed correlations between the VISC measures, in particular between absolute formant slopes and

Table 9.1: Mean absolute values (SD) of formant velocity (Hz/timestep) for boundary and stress.

	V F1		V F2		V F3		V F2_3	
none	1797	(2297)	2749	(2632)	2737	(3062)	2281	(2456)
word	1799	(2224)	2643	(2481)	2619	(3031)	2200	(2346)
phrase	766	(848)	1290	(1357)	1300	(1360)	1003	(1077)
stressed	1546	(1947)	2694	(2621)	2567	(2860)	2118	(2359)
unstressed	1926	(2450)	2733	(2605)	2791	(3151)	2339	(2481)

VL. In addition, positive correlations between the surprisal measures of the previous context ($r = 0.41, t(62407) = 112.28, p < 0.001$), and of the following context ($r = 0.59, t(62324) = 182.15, p < 0.001$) became apparent. Surprisal values of different n-phone length were related because smaller n-phone are inherently contained within longer n-phones . There were only very low negative correlations between log-transformed word frequency and surprisal measures, while phoneme probability and surprisal measures showed comparatively stronger negative correlations. Although phoneme probability and word frequency were independent measures, the negative correlations for phoneme probability and surprisal disqualified phoneme probability as a control factor in the LMM.

Following our main hypothesis (Section 2.4) we expected to observe positive correlations between ID and VISC measures. The only tendencies for very low positive correlations were found for biphone surprisal of the preceding context (BiSur) and all VISC measures, as well as for triphone surprisal of the preceding context (TriSur) and VL (Figure 9.5). As expected, word frequency correlated negatively with VISC measures: VL and F1 slope had very low negative correlations with word frequency ($r = -0.01, t(59782) = -2.33, p = 0.02$). High-frequency words, therefore, had a slight tendency to show less formant change in their vowels than low-frequency words.

Since we only found very low positive correlations for biphone surprisal of the preceding context and VISC measures, as well as for triphone surprisal of the preceding context and VL, only LMMs for the preceding context were tested. As noted above, surprisal values of the preceding context for bi- and triphone were positively correlated which was why these two factors were not integrated into one LMM. If surprisal had a significant effect on measures of VISC, interaction models were calculated investigating potential interaction effects between surprisal and prosodic factors.

In order to avoid collinearity in the LMMs we performed a correlation analysis between the fixed effects prior to model training. There was a low positive correlation between global speech rate and local speech rate ($r = 0.19$). Average vowel duration and local ($r = -0.34$) and global speech rate correlated negatively ($r = -0.13$). The factor primary lexical stress showed low positive correlations with word frequency ($r = 0.18$) and vowel duration ($r = 0.28$).

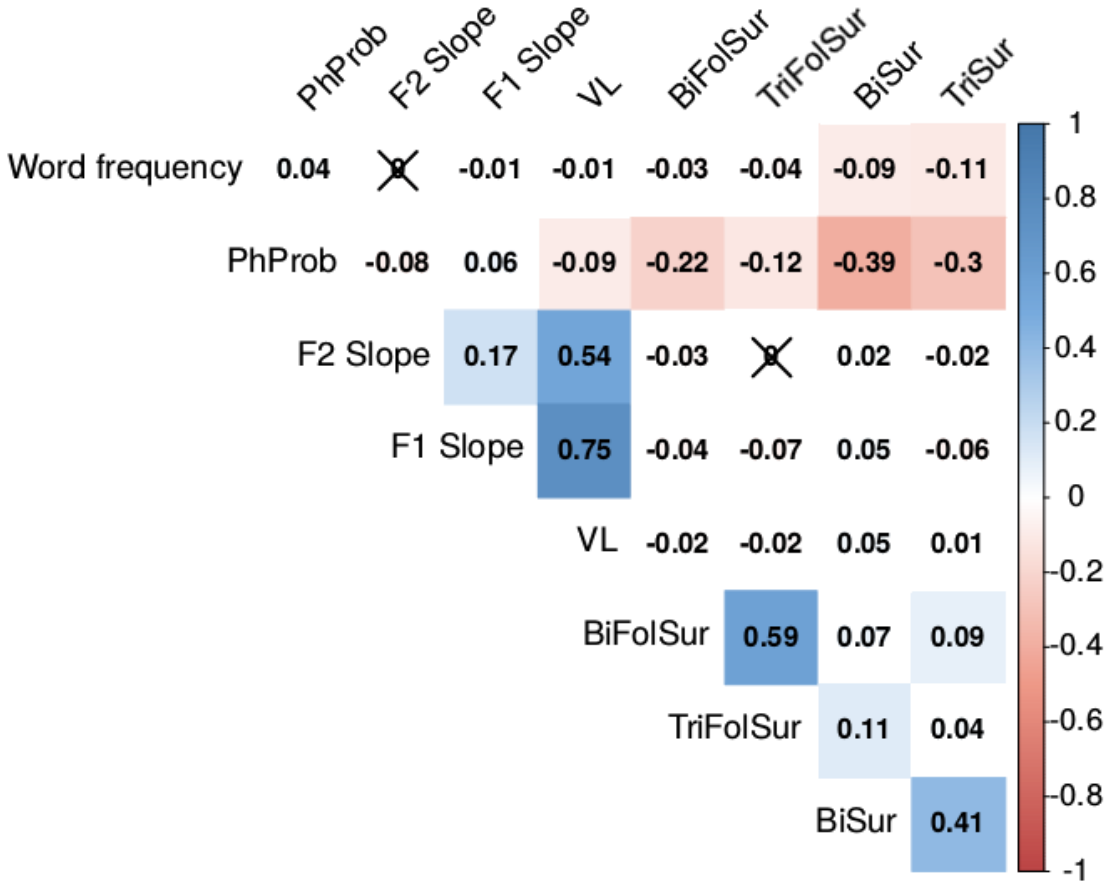


Figure 9.5: Correlation matrix (Pearson's r) for VISC measures (VL, formant slopes) and ID measures (surprisal, word frequency, and phoneme probability (PhProb)). Correlations for biphone surprisal of preceding (BiSur) and following context (BiFolSur), and triphone surprisal of preceding (TriSur) and following context (TriFolSur) are given. Non-significant correlations at significance level 0.05 are crossed out. Red shades indicate negative correlations, while blue shades show positive ones.

As fixed effects, ID factors, such as SURPRISAL, and WORD FREQUENCY, prosodic factors, i. e., STRESS, BOUNDARY, and SPEECH RATE, as well as VOWEL CATEGORY, and average vowel DURATION were used. Random factors were SPEAKER, WORD, PRECEDING CONTEXT and FOLLOWING CONTEXT. Here, context was defined based on place of articulation to control for differences in formant movement because of the impact of neighboring consonants. Three levels of place of articulation were used for consonants (coronal, dorsal, and labial). If context consisted of pause or vocalic context, this was also added to the context levels. The final model converged with all fixed effects, random intercepts for all random effects listed above, and additional random slopes for SURPRISAL and STRESS per WORD. This model structure was kept constant for all VISC measures and for F1 velocity. The models for VSL and F2 DCT2 converged with an additional random slope for BOUNDARY per WORD.

$$\begin{aligned} VL/F1slope/F2slope/F1velocity \sim\ & BiSur + Wordfreq + \\ & Stress + Boundary + \\ & GlobalTempo + LocalTempo + \\ & DurAverage + VowelCategory + \\ & (1|Speaker) + (1 + BiSur + Stress|Word) + \\ & (1|Preceding) + (1|Following) \end{aligned} \tag{9.3}$$

In the model for VL including biphone surprisal of the preceding context, all fixed effects but SURPRISAL and local SPEECH RATE reached significance level. WORD FREQUENCY showed a significant effect on VL: vowels in high-frequency words had less spectral change with regard to VL than in low-frequency words. Stressed, and longer vowels had significantly longer values for VL than unstressed, and short vowels. Vowels at word BOUNDARY position had higher values for VL than vowels at no BOUNDARY position. We observed a significant effect of global SPEECH RATE on VL variability: VL increased as global SPEECH RATE decreased. As expected, VOWEL CATEGORY was significant in the model. Overall, monophthongs (M) showed less formant change than diphthongs (D) (Table 9.2).

Regarding the random effects we found smaller VL for vowels with preceding coronal, dorsal, or labial consonants than the average VL, while preceding pauses and other vowels had a lengthening effect on VL in vowels. If the vowel was followed by a pause, labial or coronal consonant VL was shorter than the average, while following dorsal consonants and vowels had a lengthening effect on VL in vowels.

The effect size of the fixed effects of this model added up to 7.39 % of the VL model variance. Total explained variance of the entire model given by conditional pseudo-R^2 added up to 31.45 %. The largest effect size of all fixed effects for this model had VOWEL CATEGORY with 6.83 % explained variance in the data. In comparison, DURATION had a rather small effect size with 0.02 % of explained data variance. Both

Table 9.2: Vector length (VL) of German vowels: regression coefficients, standard error (SE) and statistical output of LMM analysis including biphone surprisal of the preceding context.

Measure	Terms	Coeff.	SE	t-value	p-value
ID model	Surprisal	-0.005	0.01	-0.50	= .62
	Word frequency	-0.005	0.002	-3.07	< .01
Prosodic model	Global tempo	-0.01	0.004	-3.89	< .001
	Local tempo	0.0003	0.001	0.27	= .78
	Boundary (phrase – none)	-0.03	0.06	-0.43	= .67
	Boundary (word – none)	0.06	0.01	3.89	< .001
	Stress (y – n)	0.05	0.009	5.36	< .001
Other controls	Average vowel duration	0.07	0.01	7.26	< .001
	vowel category (M – D)	-0.69	0.02	-43.12	< .001

SPEECH RATES explained about 0.19 %, STRESS 0.01 %, and BOUNDARY 0.09 % of data variance. WORD FREQUENCY had an overall effect size of 0.17 %, and SURPRISAL, although not significant, 0.08 %. Based on the effect size of the model the prosodic factors had a larger impact on VL than the ID factors used here.

In our correlation analysis, we only found a very weak relationship between VL and triphone surprisal of the preceding context ($r = 0.01$). Nevertheless, we tested the effect of triphone surprisal of the preceding context on VL in a separate model which replicated the results found in the model with biphone surprisal of the same context direction. Triphone surprisal was also not significant in predicting VL ($\beta = 0.0005, SE = 0.004, t(45910) = 0.12, p = 0.90$). Therefore, the results were not reported in more detail.

Regarding the interaction models for VL we entered interaction terms for all prosodic factors with surprisal separately in different models. Although SURPRISAL did not show a significant positive effect on VL in the baseline model, we found that in stressed vowels under high surprisal VL was longer than in unstressed vowels in low surprisal context. The interaction between SURPRISAL and STRESS added 0.03 % of explained variance to the model. The interaction model including an interaction term for BOUNDARY and SURPRISAL also performed significantly better in explaining VL than the baseline model ($\chi(2) = 6.65, p = 0.04$). Vowels at phrase boundary under high surprisal had significantly longer VL than vowels at no boundary position in low surprisal contexts (Table 9.7). This interaction added 0.02 % explained variance to the effect size of the model. Neither the interaction between SURPRISAL and local ($\chi^2(1) = 0.004, p = 0.95$) nor global SPEECH RATE ($\chi^2(1) = 0.78, p = 0.38$) added significantly to the model performance if tested against the baseline model in ANOVA model comparisons.

LMM structure was held constant across VISC measures to ensure comparisons

Table 9.3: F1 slope of German vowels: regression coefficients, standard error (SE) and statistical output of LMM analysis including biphone surprisal of the preceding context.

Measure	Terms	Coeff.	SE	t-value	p-value
ID model	Surprisal	0.002	0.0004	4.91	$< .001$
	Word frequency	-0.0001	0.00005	-2.74	$< .01$
Prosodic model	Global tempo	-0.0003	0.0001	-2.74	$< .01$
	Local tempo	-0.0001	0.00004	-1.96	$= .05$
	Boundary (phrase – none)	0.002	0.001	1.14	$= .27$
	Boundary (word – none)	0.002	0.0005	3.89	$< .001$
	Stress (y – n)	0.001	0.0003	3.90	$< .001$
Other controls	Average vowel duration	-0.02	0.0003	-63.99	$< .001$
	vowel category (M – D)	-0.01	0.0005	-19.66	$< .001$

between the models. This meant that we used the same model structure for the dependent measure absolute formant slopes as for VL. We used the model structure with fixed effects SURPRISAL, WORD FREQUENCY, STRESS, BOUNDARY, SPEECH RATE, BOUNDARY, VOWEL CATEGORY, and average vowel DURATION. The random structure was comprised of random intercepts for SPEAKER, WORD, FOLLOWING CONTEXT and PRECEDING CONTEXT, and random slopes for SURPRISAL per WORD, and STRESS per WORD (Model structure 9.3).

Absolute F1 slope showed a similar relationship to biphone SURPRISAL of the preceding context than VL to surprisal ($r = 0.05, t(62551) = 11.61, p < 0.001$). In the LMM for F1 slope including biphone SURPRISAL of the preceding context, we found a significant effect for both ID measures. Vowels in high SURPRISAL contexts showed more formant change in F1 than vowels in low SURPRISAL contexts. We also found the same effect for the ID factor WORD FREQUENCY that was observed in the VL model: vowels in high-frequency words had less spectral change in absolute F1 slope than in low-frequency words. With regard to the prosodic factors we found similar results as in the VL LMM. Vowels at word BOUNDARY showed more formant change in F1 than vowels in no BOUNDARY position. As global SPEECH RATE increased, absolute F1 slope decreased. A non-significant tendency for the same effect was found in local SPEECH RATE. In contrast to the VL model, we did not find a significant effect of primary lexical STRESS on F1 formant change. The previously observed effect in the VL model of VOWEL CATEGORY of the vowels on formant change was also observed in the F1 slope model. Overall, monophthongs showed less formant change than diphthongs. The effect of average vowel DURATION on absolute F1 slope was unexpected since the model calculated a negative effect which meant that vowels with longer duration showed less formant change in F1 than shorter vowels (Table 9.3).

Regarding the random effects we found that F1 slope was higher in vowels with

Table 9.4: F2 slope of German vowels: regression coefficients, standard error (SE) and statistical output of LMM analysis including biphone surprisal of the preceding context.

Measure	Terms	Coeff.	SE	t-value	p-value
ID model	Surprisal	-0.0005	0.0001	-2.98	$< .01$
	Word frequency	-0.0001	0.00002	-3.11	$< .01$
Prosodic model	Global tempo	-0.00004	0.00007	-0.55	$= .58$
	Local tempo	0.00003	0.00002	1.47	$= .14$
	Boundary (phrase – none)	0.001	0.001	1.21	$= .23$
	Boundary (word – none)	0.001	0.0002	2.85	$< .01$
	Stress (y – n)	0.001	0.0001	6.62	$< .01$
Other controls	Average vowel duration	-0.003	0.0002	-18.92	$< .01$
	vowel category (M – D)	-0.01	0.0002	-38.30	$< .01$

preceding coronals, pauses, and vowels, while preceding dorsal and labial consonants had a decreasing effect on F1 slope. When vowels were followed by a pause or a labial consonant there was less movement in their F1 than the average movement in F1 slope. Following coronals, dorsals or vowels had an increasing effect on F1 slope.

The fixed effects in the LMM for F1 slope with biphone surprisal of the preceding context explained a total of 9.57 % of data variance. Total explained variance of the entire model given by conditional pseudo-R^2 added up to 28.68 %. The largest effect size of all fixed effects for this model had vowel DURATION with 5.43 % explained variance in the data. VOWEL CATEGORY added 0.78 % to the explained variance in F1 slope. The ID factors both contributed 0.50 % to the explained variance, while the prosodic factors had an effect size of 2.07 % in total. SURPRISAL was a stronger predictor with 0.47 % explained variance than WORD FREQUENCY ($Var = 0.03\,\%$). The strongest prosodic predictor in the model was STRESS ($Var = 0.86\,\%$), followed by global SPEECH RATE ($Var = 0.40\,\%$), and BOUNDARY ($Var = 0.12\,\%$).

Absolute F1 slope did not correlate positively with any of the other surprisal measures tested here. Therefore, we did not calculate additional LMMs for absolute F1 slope. None of the interaction models for F1 slope performed significantly better than the baseline model. We tested interactions between SURPRISAL and STRESS ($\chi^2(1) = 0.12, p = 0.73$), SURPRISAL and BOUNDARY ($\chi^2(2) = 0.59, p = 0.75$), as well as between SURPRISAL and global ($\chi^2(1) = 1.81, p = 0.18$) and local SPEECH RATE ($\chi^2(1) = 2.72, p = 0.10$).

We calculated the LMM for absolute F2 slope with the same fixed and random structure as for the other VISC measures (Model structure 9.3). F2 slope was modeled with fixed effects SURPRISAL, WORD FREQUENCY, STRESS, BOUNDARY, SPEECH RATE, VOWEL CATEGORY, and average vowel DURATION. The random structure

contained intercepts for SPEAKER, WORD, FOLLOWING CONTEXT and PRECEDING CONTEXT, and additional random slopes for SURPRISAL per WORD, and STRESS per WORD.

In the LMM for F2 slope, we found the same effects for WORD FREQUENCY, BOUNDARY, average vowel DURATION and VOWEL CATEGORY as in the LMM for F1 slope. High-frequency words had less formant change in their F2 than low-frequency words. Vowels at word BOUNDARY displayed more formant change in F2 than vowels at no BOUNDARY position. We found a tendency for the same effect for vowels in phrase BOUNDARY position. The longer the duration of the vowel, there was less change in absolute F2 slope. Monophthongs displayed less formant change in F2 than diphthongs. In contrast to the F1 slope model, we observed that STRESS had a significant impact on absolute F2 slope. Stressed vowels had more formant change than unstressed vowels. While SPEECH RATE had a significant effect on F1 slope, we could not replicate this finding for F1 slope. Somewhat surprisingly, we found a significant negative effect of SURPRISAL on F2 slope, although the relationship was positive in simple Pearson's r correlation analysis ($r = 0.02, t(62551) = 3.93, p < 0.001$). Admittedly, the correlation was very low which potentially caused the change in sign in the more complex LMM (Table 9.4).

Regarding the random effects we found that F2 slope was higher in vowels with preceding dorsals and vowels, while all other contexts had a decreasing effect on F2 slope. When vowels were followed by a pause or a coronal consonant there was less movement in their F2 than the average movement in F2 slope. Following labials, dorsals or vowels had an increasing effect on F2 slope.

The entire LMM explained 31.52 % in data variance of F2 slope. The fixed effects, however, only contributed 4.89 % of explained data variance. The strongest predictor by far was VOWEL CATEGORY ($Var = 4.44\,\%$). STRESS ($Var = 0.34\,\%$) and vowel DURATION ($Var = 0.10\,\%$) added only little to the explained data. Even smaller effect sizes were observed for BOUNDARY ($Var = 0.005\,\%$) and WORD FREQUENCY ($Var = 0.002\,\%$).

We did not run the F2 slope model including any other surprisal measure because we did not find any other positive correlations between surprisal and F2 slope, except for biphone surprisal of the preceding context, in our correlation analysis.

Regarding the interaction models for F2 slope we entered interaction terms for all prosodic factors with SURPRISAL separately following the same procedure as in all other interaction models. The interaction model including an interaction term between SURPRISAL and STRESS ($log(L) = -311624$) performed significantly worse than the baseline model ($log(L) = -311619; \chi^2(1) = 15.50, p < 0.001$). The output of the model, however showed a positive effect of the interaction of SURPRISAL and STRESS on F2 slope indicating that both factors complemented each other in explaining higher F2 slope values. The interaction between SURPRISAL and BOUNDARY improved model performance significantly ($\chi^2(2) = 11.02, p < 0.01$). There was a positive effect of the interaction of surprisal with phrase boundary compared to no boundary position on

F2 slope (Table 9.7). This interaction term added 0.03 % to the explained variance of the fixed effects in the model. Neither the interaction between SURPRISAL and local ($\chi^2(1) = 2.41, p = 0.12$) nor global SPEECH RATE ($\chi^2(1) = 0.85, p = 0.36$) added significantly to the model performance if tested against the baseline model in ANOVA model comparison.

Vowel section length In our previous correlation analysis, we found that VSL and TL, as well as their respective roc measures showed strong correlations. Therefore, one of the two measures was used. We opted for VSL because it is measured per sampling point in the vowel and was therefore assumed to give a more fine-grained profile of formant movement. Both roc measures only had very weak correlations with surprisal values which is why they were disregarded in the analysis (Section 9.2.1).

Model selection procedure was performed as described in Section 4.3. The final model for VSL included the ID factors biphone SURPRISAL of the preceding context, and WORD FREQUENCY, prosodic factors, such as primary lexical STRESS, prosodic BOUNDARY, local and global SPEECH RATE, as well as average vowel DURATION and VOWEL CATEGORY. Random structure included random intercepts for WORD, SPEAKER, as well as preceding and following PHONOLOGICAL CONTEXT. Random slopes for the factors SURPRISAL, STRESS, and BOUNDARY per WORD yielded no convergence errors and were therefore used in the model (Model structure 9.4). The same model structure was used for F2 DTC2 (Section 9.2.3).

$$\begin{aligned} VSL/F2DCT2 \sim\ & BiSur + Wordfreq + \\ & Stress + Boundary + GlobalTempo + LocalTempo + \\ & DurAverage + VowelCategory + \\ & (1|Speaker) + (1 + BiSur + Stress + Boundary|Word) + \\ & (1|Preceding) + (1|Following) \end{aligned} \tag{9.4}$$

SURPRISAL and WORD FREQUENCY were both predictive of increased formant change in VSL. Easily predictable vowels showed less formant change than vowels which were difficult to predict. Vowels in high-frequency words had smaller values in VSL than vowels in low-frequency words. In primary lexical STRESS position, vowels displayed more formant movement than in unstressed position. Phrase and word BOUNDARY position, however, led to smaller values in VSL. At accelerated local SPEECH RATE, there was less formant change in VSL than at decelerated rate. Global SPEECH RATE showed the same tendency. The effect, however, was not significant. We also found the expected effect for vowel DURATION: longer vowels had higher values in VSL. Contrary to our expectation, monophthongs showed significantly more formant movement in VSL than diphthongs. Especially the back vowels /oː, ɔ, uː, ʊ/ and tense vowels /iː/ and /eː/ showed high values in VSL compared to the diphthongs /ɔʏ/ and /aɪ/ (Figure 9.4). Table 9.5 summarizes the model output.

With regard to the random effects we found that preceding coronal consonants and vowels had a lengthening effect on VSL in German vowels, whereas vowels with preceding dorsal and labial consonants, and preceding pauses showed shorter VSL lengths than the average. Following pauses also had a shortening effect on VSL, in addition to following coronal consonants. VSL in vowels with following dorsal or labial consonants, or other vowels were longer than the average.

Based on marginal pseudo-R^2 the fixed effects in the model explained 10.29 % of the data variance in VSL. The entire model had a quite strong overall performance of 75.52 % indicated by conditional pseudo-R^2. The ID factor BOUNDARY was the strongest predictor of VSL in German vowels ($Var = 3.12\,\%$), closely followed by STRESS ($Var = 2.68\,\%$) and average vowel DURATION ($Var = 1.94\,\%$). All other factors were less effective in explaining VSL variability. SURPRISAL ($Var = 1.75\,\%$) and WORD FREQUENCY ($Var = 0.14\,\%$) were not as effective as the prosodic factors in explaining VSL variability. VOWEL CATEGORY ($Var = 0.01\,\%$) and local SPEECH RATE ($Var = 0.32\,\%$) only had small effects on VSL.

Interaction terms were entered separately into the baseline VSL LMM. Vowels under high SURPRISAL and STRESS showed longer VSL values than under low surprisal and in unstressed vowels. These two factors complemented each other in their positive effect on VSL adding 0.41 % to the effect size of the model. The interaction model including an interaction term between BOUNDARY and SURPRISAL also performed better than the baseline model. At word boundary compared to no boundary position, vowels showed shorter VSL when they stood in a high surprisal context compared to a low surprisal context. For phrase boundaries, there was a tendency for the same effect, but it was non-significant. We also found a significant negative interaction between local SPEECH RATE and SURPRISAL on VSL (Table 9.7). The interaction model with an interaction term between global SPEECH RATE and SURPRISAL, however, did not perform significantly better than the baseline model ($\chi^2(1) = 2.51, p = 0.11$).

F1 velocity F1 velocity was the only velocity measure that showed a weak positive relationship with biphone surprisal of the preceding context ($r = 0.03, t(62551) = 7.89, p < 0.001$) which was why it was chosen for the following analysis. For the LMM for F1 velocity, the same model structure as for the VISC measures was used (Model structure 9.3).

Again, both ID factors were significant in explaining variability in the formant change measure. Vowels had higher velocity values when they were difficult to predict from the context, and vowels in high-frequency words showed less F1 velocity than in low-frequency words. The significant positive effect of STRESS on F1 velocity was expected. At word BOUNDARY, F1 velocity was significantly higher compared to vowels at no BOUNDARY position, this effect was not shown for vowels at phrase boundaries. At fast global SPEECH RATE, F1 velocity decreased. We also found significant effects of DURATION and VOWEL CATEGORY on F1 velocity. Monophthongs showed less formant velocity than diphthongs. As vowel duration increased

Table 9.5: Vowel section length (VSL) of German vowels: regression coefficients, standard error (SE) and statistical output of LMM analysis including biphone surprisal of the preceding context.

Measure	Terms	Coeff.	SE	t-value	p-value
ID model	Surprisal	0.27	0.02	14.09	$< .001$
	Word frequency	-0.02	0.003	-6.21	$< .001$
Prosodic model	Global tempo	-0.005	0.004	-1.40	$= .16$
	Local tempo	-0.003	0.001	-2.55	$= .01$
	Boundary (phrase – none)	-0.58	0.06	-8.86	$< .001$
	Boundary (word – none)	-0.49	0.02	-20.28	$< .001$
	Stress (y – n)	0.39	0.02	21.52	$< .001$
Other controls	Average vowel duration	0.30	0.01	31.50	$< .001$
	vowel category (M – D)	0.43	0.02	19.02	$< .001$

we observed a decrease in F1 velocity[2] (Table 9.6).

Regarding the random effects we found that F1 velocity was lower than the average when the vowel was preceded by dorsal and labial consonants or vowels. Preceding pauses and coronals had an increasing effect on F1 velocity. When vowels were followed by coronals or vowels their F1 velocity was higher than the average, while following pauses, dorsal and labial consonants had a decreasing effect.

The entire LMM for F1 velocity explained 30.94 % in the data variance. The fixed effects, however, contributed 8.28 % to the explained data variance. The strongest predictor was DURATION ($Var = 6.70\,\%$), followed by VOWEL CATEGORY ($Var = 0.53\,\%$), and STRESS ($Var = 0.42\,\%$). WORD FREQUENCY explained only 0.01 % of the data variance of F1 velocity, and SURPRISAL even less ($Var = 0.001\,\%$). BOUNDARY and global SPEECH RATE each contributed 0.07 % of explained variance to the model.

Interaction models were built following the procedure explained in Section 4.3. None of the interaction models performed significantly better than the baseline model described above.

9.2.3 Parametric measures

Descriptive statistics

Since DCT coefficients and polynomial parameters were strongly correlated and there were no collinearities found within the group of DCT coefficients above moderate level we opted for DCT coefficients in the following descriptive and inferential statistical

[2]This phenomenon was also seen for F1 and F2 slope and is due to the calculation of these measures which is discussed further in Section 9.3.

Table 9.6: F1 velocity of German vowels: regression coefficients, standard error (SE) and statistical output of LMM analysis including biphone surprisal of the preceding context.

Measure	Terms	Coeff.	SE	t-value	p-value
ID model	Surprisal	160	37	4.00	$< .001$
	Word frequency	-12	5	-2.77	$= .02$
Prosodic model	Global tempo	-42	12	-3.50	$< .001$
	Local tempo	-1	4	-0.35	$= .73$
	Boundary (phrase – none)	93	126	0.74	$= .47$
	Boundary (word – none)	115	46	-2.90	$= .01$
	Stress (y – n)	100	30	3.00	$< .001$
Other controls	Average vowel duration	-1661	29	-56.50	$< .001$
	vowel category (M – D)	-257	47	-5.51	$< .001$

Table 9.7: Dynamic formant trajectories of German vowels: interaction of biphone surprisal of the preceding context with prosodic factors.

Measure	Terms	Coeff.	SE	t-value	p-value
VL	Stress * Surprisal	0.06	0.02	2.70	$< .01$
	Boundary (phrase) * Surprisal	0.15	0.06	2.40	$= .02$
F2 slope	Stress * Surprisal	0.001	0.0004	3.95	$< .001$
	Boundary (phrase) * Surprisal	0.003	0.001	3.22	$< .01$
VSL	Stress * Surprisal	0.21	0.03	6.48	$< .001$
	Boundary (phrase) * Surprisal	-0.09	0.08	-1.18	$= .24$
	Boundary (word) * Surprisal	-0.38	0.06	-6.57	$< .001$
	Local tempo * Surprisal	-0.007	0.002	-3.12	$< .01$
F2 DCT2	Stress * Surprisal	9.19	0.73	12.56	$< .001$
	Local tempo * Surprisal	-0.37	0.06	-6.25	$< .001$

Table 9.8: Absolute mean values (SD) of F1 and F2 DCT coefficients (DCT1-DCT3) for boundary and stress.

DCTC	**Boundary**			**Stress**		
F1	none	word	phrase	stressed	unstressed	
DCT1	33.00	(31.82) 33.02	(31.22) 25.99	(25.24) 33.80	(32.63) 32.33	(31.09)
DCT2	23.13	(23.07) 20.66	(20.66) 18.40	(16.64) 24.95	(23.46) 21.57	(22.31)
DCT3	11.77	(14.28) 11.36	(13.74) 9.96	(11.40) 11.07	(13.79) 12.10	(14.44)
F2	none	word	phrase	stressed	unstressed	
DCT1	43.67	(41.09) 40.60	(38.10) 33.95	(38.11) 47.49	(45.08) 40.69	(39.06)
DCT2	20.57	(24.34) 14.55	(17.75) 15.80	(20.03) 24.40	(27.65) 17.34	(20.81)
DCT3	10.51	(15.64) 9.02	(12.34) 9.75	(14.42) 10.44	(15.53) 10.34	(15.32)

analysis (cf. Section 9.2.1). Similarly to the velocity data, mean values of DCT were calculated over the six sampling points per individual vowel in the data. Only DCT coefficients for F1 and F2 were calculated. We used absolute values of DCT coefficients in order to describe the amount of movement in formant trajectories, and not the direction of movement. Mean values of the zeroth DCT coefficient are equivalent to the mean values of formants. They did not add any information with regard to the shape of the formant curve and were therefore not reported.

At phrase boundary, we found the lowest values in F1 DCT coefficients compared to other boundary levels. Word and phrase boundary lower values in F2 DCT coefficients compared to no boundary position. Regarding stress we found that stressed vowels showed more change in DCT1 and DCT2 for both formants, but only a small difference in F2 DCT3 with the expected direction of the effect, and the reversed effect for F1 DCT3 (Table 9.8).

Linear mixed-effects model

Pearson's correlation tests showed that there was only a weak positive relationship between absolute F2 DCT2 and surprisal of the preceding context. For both biphone ($r = 0.12, t(62551) = 32.56, p < 0.001$) and triphone surprisal ($r = 0.13, t(62407) = 33.39, p < 0.001$) this correlation was low, but significant. Correlation analysis did not reveal any additional relationships between surprisal and the other DCT coefficients which was why the following analysis was restricted to F2 DCT2.

The LMM for DCT2 of the second formant was constructed using the same structure as for the VSL LMM (Model structure 9.4). As results, we found significant effects of SURPRISAL and WORD FREQUENCY in the expected directions. The effect of SURPRISAL previously found in a simple correlation analysis also held, while controlling for other predictors of formant movement. In addition, there were significant

Table 9.9: F2 DCT2 of German vowels: regression coefficients, standard error (SE) and statistical output of LMM analysis including biphone surprisal of the preceding context.

	Terms	Coeff.	SE	t-value	p-value
ID model	Surprisal	3.98	0.33	12.08	$< .001$
	Word frequency	-0.36	0.06	-6.11	$< .001$
Prosodic model	Global tempo	-0.22	0.11	-2.01	$= .04$
	Local tempo	0.09	0.03	2.64	$< .01$
	Boundary (phrase – none)	-8.50	1.21	-7.04	$< .001$
	Boundary (word – none)	-3.03	0.47	-6.44	$< .001$
	Stress (y – n)	2.83	0.40	7.12	$< .001$
Other controls	Average vowel duration	13.04	0.27	48.24	$< .001$
	vowel category (M – D)	-7.38	0.57	-12.93	$< .001$

effects of STRESS, average vowel DURATION, and VOWEL CATEGORY with the expected outcomes of presence of primary lexical stress and longer durations increasing formant movement, and diphthongs showing more formant movement than monophthongs. At both BOUNDARY positions, vowels had less absolute formant movement than within words. Interestingly, for SPEECH RATE we found contradicting results. At a local level, there was an increase of formant movement when the word was produced faster. At a global level, sentences read with higher SPEECH RATE included vowels with less formant movement than at slower tempo (Table 9.9).

Vowels with preceding or following pauses showed less movement in F2 DCT2 than all vowels on average. Preceding coronals, as well as following labials and vowels had a decreasing effect on F2 DCT2. In all other contexts, we found higher values for the measure compared to its average.

The entire LMM explained 53.37 % in the data variance of F2 DCT2. The fixed effects, however, contributed 10.44 % to the explained data variance. The strongest predictor by far was vowel DURATION ($Var = 3.56\,\%$), followed by VOWEL CATEGORY ($Var = 2.30\,\%$). Biphone SURPRISAL ($Var = 1.99\,\%$) of the preceding context and WORD FREQUENCY ($Var = 0.20\,\%$) both explained 2.19 % of the data variance of F2 DCT2. Equally strongly predictive as both ID factors was STRESS ($Var = 2.25\,\%$). BOUNDARY was much less effective in explaining F2 DCT2 data variance ($Var = 0.002\,\%$). Both SPEECH RATES added around 0.14 % explained variance.

STRESS and SURPRISAL also interacted positively on F2 DCT2. This interaction added a considerable amount of 1.46 % of explained data variance to the baseline model. Another interaction model that performed significantly better than the baseline model included an interaction term for local SPEECH RATE and SURPRISAL. Vowels in words with high local speech rate in a high surprisal context showed less curvature in their F2 than vowels in slowly spoken words and low surprisal contexts

(Table 9.7). This interaction term added 0.12 % explained data variance. The interaction model with an interaction term for BOUNDARY and SURPRISAL did not perform better than the baseline model ($\chi^2(2) = 5.26, p = 0.07$).

9.3 Discussion

The analyses presented in this chapter investigated the impact of prosodic structure and ID on dynamic formant trajectories in German read speech. We confirmed our hypothesis regarding a negative relationship between word frequency and vocalic formant change (Section 2.4). Vowels in high-frequency words showed less formant change than in low-frequency words in all metrics investigated here. The effect of contextual predictability on formant change was less clear. For some measures, we only found a positive effect of surprisal on formant change in interaction with the factor stress, e. g., for F2 slope and VL, while for all other measures (F1 slope, VSL, F2 DCT2, F1 velocity) we found a positive impact of the main effect surprisal on the magnitude of formant change. These findings showed that vowel formant change was significantly influenced by ID factors on both phoneme and word level.

Regarding the prosodic factors we observed a consistent effect of primary lexical stress on all formant change measures: vowels in syllables carrying primary lexical stress showed more formant change than in unstressed syllables. This was expected considering that vowels increase in their duration and vowel dispersion when they stand in stressed lexical position (Gahl et al., 2012; Malisz et al., 2018). Stress also interacted positively with surprisal on all formant change metrics, except for F1 velocity. In addition, effects of surprisal and stress overlapped to a small degree which was indicated by weak positive correlations in the collinearity analyses. Because of their weakness these dependencies were disregarded for the current analyses.

We found conflicting results for the effect of prosodic boundary on formant movement. On the one hand, F1 velocity and the VISC measures all showed the same significant effect of higher formant change values for vowels at word boundary position compared to no boundary. This finding mirrors expansion of articulatory gestures in segments that undergo final-lengthening before a prosodic boundary (Byrd, 2000; Kohler, 1988). It is also supported by our findings of an increasing effect of word boundary on segment duration (Chapter 5). Surprisingly, this finding was not replicated at phrase boundary compared to no boundary for these measures. VSL and F2 DCT2, on the other hand, were lower in vowels at word and phrase boundary compared to no boundary position. This finding goes hand in hand with the decreasing effect of boundary on German vowel dispersion (Section 8.2). Contrary to the claim made by Turk (2010), boundary position does not reduce coarticulatory effects in the spectral characteristics of segments, i. e., increased values of formant change and distinctiveness, but leads to less pronounced formant change and vowel dispersion.

In general, we found a negative effect of speech rate on dynamic formant change.

With increasing speech rate formants showed less movement. This effect was observed at a local level (VSL) and at a global level of speech rate (F1 slope, F1 velocity, VL). For F1 slope and velocity, we also found the tendency for the same effect at a local level of speech rate, and VSL showed a tendency to decrease with increasing global speech rate. These findings were in line with studies showing reduced articulatory gestures at fast speech rate compared to slow speech rate (Turner et al., 1995), and were additionally supported by our analyses of German vowel dispersion and segment duration in previous Chapter 5 and Section 8.2. Contrary to our expectation, F2 DCT2 increased significantly with increasing global speech rate.

F1 and F2 slope, as well as F1 velocity are measures of formant change relative to the duration of the vowel. Since these measures already include vowel duration we found a negative relationship between absolute formant change and average vowel duration. For instance, keeping the amount of change constant but doubling the duration of the vowel token leads to the relative amount of change being halved. VL, VSL and F2 DCT2, on the other hand, are not measured relative to the duration of the vowel, which is why we found a positive effect of vowel duration on these metrics, which in turn is in line with previous findings on formant change and duration variability (Hillenbrand, Getty, et al., 1995).

We decided to include the vowel category of the vowel in the LMMs rather than vowel identity because of data sparsity and convergence issues of the models. However, we did not find consistent results for vowel category across the metrics of dynamic formant trajectories. We expected to generally find more formant change in diphthongs than in monophthongs because they inherently involve a change in vowel quality. We confirmed this hypothesis for all measures investigated here, except for VSL. The binary coding of vowel identities helped to identify coarse differences between monophthongs and diphthongs in most formant change measures, but was not informative for all of them. Vowel height was decisive in explaining variability in VSL, and not necessarily vowel category of the vowel which also includes vowel tenseness (Figure 9.4). Dynamic cues also seem to be very specific to vowel identity, and help listeners differentiate between tense and lax vowels in addition to information about vowel-inherent duration (Strange and Bohn, 1998).

The preceding two chapters focused on the variability of spectral characteristics of vowels and its relation to ID and prosodic variables. Static and dynamic formants were analyzed. The following chapter introduces a comparative analysis of the global spectral characteristics of German vowels.

Chapter 10

Spectral similarity of vowels

This chapter investigates whether German vowels differ significantly from each other in mel-cepstral distortion (MCD) when they stand in different ID contexts. It is an expansion of the vowel dispersion and dynamic formant movement analysis of German since MCD values were based on a metric which describes the global spectral features of vowels (MGC) as opposed to a description of formant measurements only. The following analysis is based on a revised and extended conference paper (Brandt et al., 2017b). In particular, we have revised the LMM structure leading to contrasting results to those presented in the proceedings paper. The analysis was performed on a subset of the PhonDat2 corpus (Section 3.1.4).

10.1 Method

As an alternative to vowel dispersion, one can use a distance metric which is established in the evaluation of speech synthesis: mel-cepstral distortion (MCD). It is defined as the Euclidean distance between two vectors which describe the global spectral characteristics. Here, we used the mel-generalized cepstral (MGC) transform to describe the speech signal which is defined as the inverse Fourier transform of the generalized logarithmic spectrum calculated on a warped frequency scale (Tokuda et al., 1994). The smaller the MCD values, the smaller the spectral distance between two speech signals. MCD only require a simple outlier cleaning procedure, and they are less prone to errors in calculating than formants. Therefore, MCD is a suitable distance metric for the analysis of large speech corpora.

MGC representations and MCD were calculated using SPTK 3.7 (Kobayashi et al., 2017) at the temporal midpoint of each vowel over a window of 25 ms. Before the distance metric was estimated, the optimal feature vector size for the MGCs was calculated since speech sound classes differ in this feature (Tokuda et al., 1994). The optimal feature vector size for a respective data set can be estimated by using the

diagonal of covariance matrices. The variance of features at size 5, 12, 19, 24, 30 and 39 were compared. For the vowels in the current data set, vector size 30 had the lowest variance ($Var(m30) = 6.43$ e-18). Further parameters for MGC extraction were $\alpha = 0.42$, $\gamma = 2$ and frame length of 512. In a second step, the Euclidean distance between vowel vectors for the same vowel identity in different ID conditions were extracted. All MCD values larger than 10 were identified as outliers and cleaned from the data.

10.2 Results

In order to define different ID contexts for German vowels, we binned the continuous biphone surprisal values of the preceding context into three equally large groups for low, mid, and high surprisal. In total, there were six categories of surprisal context: high – high (h - h), mid – mid (m - m), low – low (l - l), high – mid (h - m), high – low (h - l), and mid – low (m - l). Unigram word probability was obtained similarly based on a word LM of SDeWaC (Section 4.2.1), and binned in the same way as surprisal values. This again led to six categories of unigram word probability, parallel to those for biphone surprisal.

In the same fashion, we binned the continuous variable global speech rate into three equally large groups of the data. This led to three factor levels of speech rate: normal, fast, and slow. This means we compared vowels in words that were spoken at different speech rates in the categories normal – normal (n - n), fast – fast (f - f), slow – slow (s - s), normal – fast (n - f), slow – fast (s - f), and slow – normal (s - n).

10.2.1 Descriptive statistics

MCD values were log-transformed because of positive skewness. For unigram word probability we found the expected hierarchy of non-contrasting conditions (l - l, h - h, m - m) having lower values than the contrasting comparisons (h - m, h - l, m - l). This hierarchy was not replicated for MCD between vowels in different surprisal conditions. Here, vowels in m - m surprisal were the most similar, followed by vowels in m - l ID condition, h - m, l - l, and h - l condition, while same vowel identities in the non-contrasting condition h - h surprisal were the most distant from each other (Table 10.1).

Vowels that were both unstressed were the most distant from each other compared to vowels that were both stressed, and the comparison between stressed and unstressed vowels. When stressed vowels were compared, we found the lowest MCD values from all three stress conditions. Also, vowels in both unstressed syllables had the largest standard deviation in MCD values. If both vowels were produced in words at fast speech rate they were more distant than vowels that were both produced at normal or slow speech rate. Regarding the contrasting comparative conditions of speech

Table 10.1: Descriptive statistics of log-transformed MCD of German vowels in different ID conditions based on surprisal and unigram word probability (WP).

Terms	Level	Mean	SD
Surprisal	m – m	-1.88	0.54
	m – l	-1.80	0.51
	h – m	-1.79	0.55
	l – l	-1.75	0.53
	h – l	-1.75	0.53
	h – h	-1.74	0.58
Unigram WP	l – l	-1.83	0.53
	m – m	-1.79	0.56
	h – h	-1.78	0.54
	h – l	-1.77	0.54
	h – m	-1.77	0.56
	m – l	-1.77	0.55

rate same vowel identities at normal and fast speech rate were almost as distant as vowels in the f - f condition, followed closely by vowels in s - f speech rate comparisons. Same vowel phonemes showed smaller spectral distances when they stood in sentences spoken at slow and normal speech rate (Table 10.2).

10.2.2 Linear mixed-effects model

PhonDat2 contains read speech from a train inquiry task (Section 3.1.4). Because of the specific domain of the speech data, word and syllable frequencies of the PhonDat2 corpus were included as control factors. In that way, effects on the spectral vowel characteristics which were due to corpus-specific frequency distributions were identified. PhonDat2 was syllabified using the g2p tool from WebMaus (Kisler et al., 2017). Frequency values were binned into three categories (low, mid and high frequency), and put into six comparative factor levels, similarly to the ID or the speech rate levels.

There were only weak correlations between the predictor values. Word and syllable frequency of PhonDat2 were positively correlated ($r = 0.18$) since both were extracted from the same data set. Word frequency of PhonDat2 and unigram word probability based on SDeWaC, however, were negatively correlated ($r = -0.11$) indicating the domain-specific word frequency distribution of the speech material.

For the baseline LMM, the fixed effects SURPRISAL of the preceding biphone, STRESS, SPEECH RATE, UNIGRAM WORD PROBABILITY, PhonDat2 WORD FREQUENCY, and PhonDat2 SYLLABLE FREQUENCY, as well as GENDER were entered.

Table 10.2: Descriptive statistics of log-transformed MCD of German vowels in different prosodic conditions.

Terms	Level	Mean	SD
Stress	n – n	-1.76	0.56
	y – n	-1.77	0.53
	y – y	-1.87	0.54
Speech rate	f – f	-1.71	0.55
	n – f	-1.72	0.55
	s – f	-1.76	0.54
	n – n	-1.78	0.55
	s – n	-1.83	0.54
	s – s	-1.88	0.53

The final random structure included random intercepts for SPEAKER, VOWEL IDENTITY, WORD, PRECEDING CONTEXT, and FOLLOWING CONTEXT. The predictor GENDER did not explain variance in the data and was therefore removed. Reference level for all ID and corpus frequency factors was the comparative condition h - h. Reference level for the predictor value SPEECH RATE was the comparison between two vowels in sentences which were both produced at fast speech rate. For STRESS, reference level was the comparison between two vowels in syllables with primary lexical stress. The dependent variable MCD was log-transformed due to positive skewness. All categorical variables were treatment coded (Model structure 10.1).

$$\begin{aligned} MCD \sim\ & Surprisal + UnigramWP + PD2Wordfreq + PD2Syllablefreq + \\ & Stress + GlobalTempo \\ & (1|Speaker) + (1|Word) + (1|Vowel) \\ & (1|Preceding) + (1|Following) \end{aligned} \tag{10.1}$$

Results of the baseline LMM showed that SURPRISAL, STRESS, global SPEECH RATE, and SYLLABLE FREQUENCY were all significant in explaining variability in MCD. None of the word ID factors reached significance level. All comparisons to the reference level h - h SURPRISAL context were significant, except for h - l. Vowels in both stressed positions were more similar to each other than vowels which both stood in unstressed positions. The spectral distance between vowels in unstressed and stressed position was larger than in vowels which both stand in unstressed position. All other SPEECH RATE conditions showed significantly larger MCD values than vowels in f - f condition, except for vowels in s - s condition (Table 10.3).

Regarding our hypotheses, post-hoc analysis using Tukey-tests were performed to identify differences between contrasting and non-contrasting conditions. Contrary to our hypothesis, non-contrasting comparative surprisal conditions were significantly

Table 10.3: Spectral similarity of vowels: regression coefficients, standard error (SE) and statistical output of LMM analysis including biphone surprisal of the preceding context.

Terms	Level	Coeff.	SE	t-value	p-value
Surprisal	h – l	-0.02	0.004	-4.42	$< .001$
	h – m	-0.02	0.003	-5.62	$< .001$
	l – l	-0.06	0.005	-12.10	$< .001$
	m – l	-0.04	0.004	-9.96	$< .001$
	m – m	-0.03	0.004	-7.91	$< .001$
Unigram WP	h – l	-0.005	0.01	-0.40	$= .69$
	h – m	-0.01	0.01	-0.79	$= .43$
	l – l	-0.03	0.03	-1.07	$= .29$
	m – l	-0.008	0.02	-0.36	$= .72$
	m – m	-0.02	0.03	-0.89	$= .38$
Stress	y – n	0.02	0.002	7.69	$< .001$
	y – y	-0.03	0.004	-6.30	$< .001$
Speech rate	n – f	0.006	0.002	3.55	$< .001$
	n – n	0.009	0.002	4.39	$< .001$
	s – f	0.02	0.002	5.90	$< .001$
	s – n	0.02	0.002	4.49	$< .001$
	s – s	0.004	0.003	1.53	$= .13$
Word frequency	h – l	0.04	0.02	1.95	$= .05$
	h – m	0.03	0.02	1.49	$= .14$
	l – l	0.008	0.04	0.21	$= .83$
	m – l	0.01	0.04	0.27	$= .79$
	m – m	-0.002	0.04	-0.06	$= .96$
Syllable frequency	h – l	0.08	0.003	25.03	$< .001$
	h – m	0.03	0.003	8.47	$< .001$
	l – l	0.01	0.005	2.63	$= .008$
	m – l	0.02	0.005	3.55	$< .001$
	m – m	-0.03	0.005	-6.42	$< .001$

different from each other in their MCD, except for the difference between h - l and h - m ($Coeff. = 0.001, SE = 0.003, z = 0.18, p = 1.00$). Vowels in both high surprisal condition had the largest spectral distance from each other compared to the other conditions. These differences were all significant. Vowels in h - l surprisal context were significantly more distant from each other than vowels in all other contexts, except for vowels in h - h surprisal condition. Vowels in h - m surprisal were more distant from each other than vowels in non-contrasting conditions m - m and l - l, as well as in the condition m - l. The smallest distance based on the post-hoc analysis was found between vowels in both low surprisal condition.

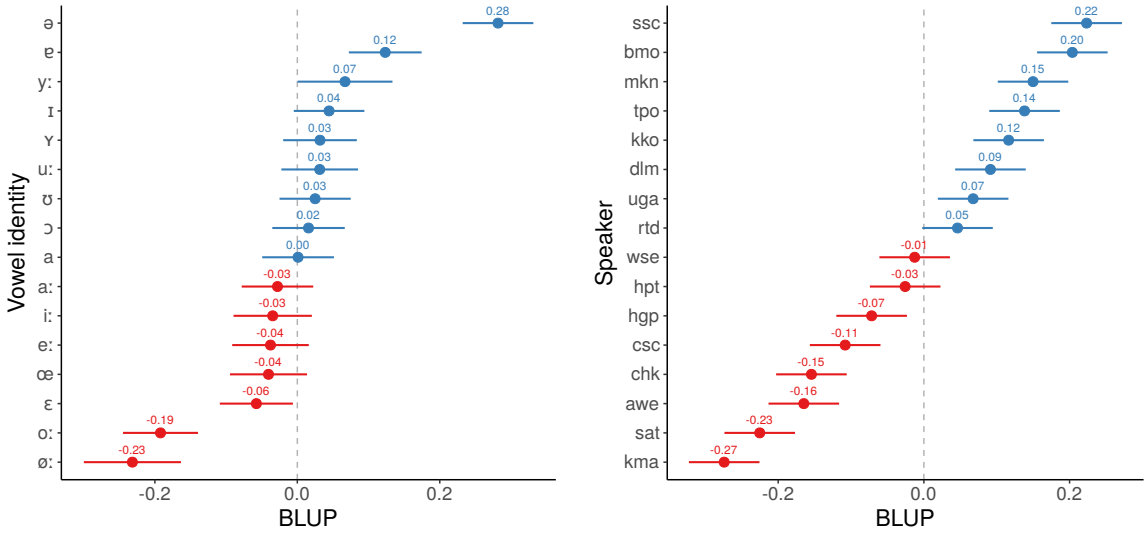
For the prosodic factors, all stress conditions differed significantly from each other in the post-hoc analysis. In addition to the statistical output from the LMM, we found that MCD values in stressed – stressed condition were significantly smaller than in stressed – unstressed ($Coeff. = -0.04, SE = 0.002, z = -17.25, p < 0.001$). In addition to the observations regarding speech rate based on the output of the LMM, we only found that s - f vowels were significantly more distant from each other than vowels in n - f condition ($Coeff. = 0.006, SE = 0.002, z = 3.25, p = 0.01$). Vowels at both slow speech rate were significantly more similar than vowels produced at s - f ($Coeff. = -0.009, SE = 0.003, z = -3.88, p = 0.001$), or s - n speech rate conditions ($Coeff. = -0.006, SE = 0.002, z = -3.39, p = 0.009$). Other comparisons for speech rate were not significant.

As can be seen in Table 10.3, all other syllable frequency conditions but m - m showed larger spectral distance than h - h. Additional post-hoc analysis revealed that the largest MCD values were found in the contrasting conditions for syllable frequency, and the smallest distance values in the non-contrasting conditions. All level comparisons were significant except for vowel spectral distances in m - l syllable frequency condition compared to l - l condition. The largest MCD values were observed between same vowel identities in h - l syllable frequency, followed by h - m, and m - l comparisons.

The marginal pseudo-R^2 indicating how much variance is explained by the fixed factors showed that SURPRISAL explained 0.24 % of the MCD variance alone, the control factor for domain-specific SYLLABLE FREQUENCY distribution added 0.20 % to the explained variance. When STRESS was added, explained variance increased by 0.10 %. The effect of speech rate was vanishingly small ($Var = 0.01\,\%$). The conditional pseudo-R^2 for the variance explained by both fixed and random effects equaled 24.84 % in the final model.

Vowel comparison differed in their average MCD per vowel identity. Vowels /ə/ ($M = -1.67, SD = 0.54$) and /ɐ/ ($M = -1.64, SD = 0.57$) had the highest MCD values, while /œ/ stood out as the vowel which showed the lowest MCD ($M = -2.10, SD = 0.47$). This segment-specific variability in MCD values was mirrored in the estimates for the random intercept of VOWEL IDENTITY (Figure 10.1a).

Similarly, speakers differed in the overall amount of spectral distance with which they produced their vowels. On average, some speakers produced vowels with large



(a) Vowel identity. **(b)** Speaker identity.

Figure 10.1: MCD for German vowels: BLUP of random intercepts.

spectral distances, while others produced vowels that were on average less distant from same vowel identities. This finding was, again, mirrored in the estimates of the random effect for SPEAKER (Figure 10.1b).

10.3 Discussion

This chapter aimed to investigate the global spectral characteristics of German vowels in different ID conditions. MCD values between same vowel identities in the same ID context and in contrasting ID conditions were compared. We hypothesized that vowels in the same ID context are more similar to each in their spectral characteristics than vowels in different ID contexts. We confirmed this hypothesis, but only for the factor surprisal and syllable frequency, and not for unigram word probability. Vowels in contrasting and non-contrasting ID conditions were significantly different from each other. This observation, however, did not hold for all comparisons. Vowels in high – mid surprisal condition were not less distant from each other than vowels in contrasting conditions high – low surprisal.

We did not find any significant differences between the factor levels of unigram word probability nor between those of PhonDat2 word frequency because we included word identity of the first and second word of the comparisons in our statistical model. Since we accounted for the identity of the word, ID factors related to word identity did not add any information to the model. This was in contrast to the other spectral analyses in this thesis. The LMM for the vowel dispersion analysis in German (Section 8.2), for instance, included random intercept for word and the fixed effect word frequency. Here, word frequency was a significant predictor of vowel disper-

sion although we controlled for random differences in vowel dispersion because of word identity. The main difference between these analyses was, however, that in the current analysis spectral distances between vowel tokens in different words were analyzed, while the vowel dispersion measure does not compare two vowel tokens. We therefore conclude that in a comparative analysis of spectral characteristics of vowels word identity overrides effects of word frequency and unigram word probability.

Additionally, we hypothesized that there is a hierarchy in the distance metric modeled as a function of ID context: smaller distances are supposedly found in the same ID conditions, while larger distances are apparent between vowels in contrasting ID conditions. Surprisingly, this hypothesis was confirmed for MCD in different syllable frequency conditions, but not for different surprisal conditions. Therefore, it seemed that syllable frequency was the better ID measure to predict differences in MCD values for German vowels. However, non-contrasting and contrasting syllable frequency conditions did not form homogeneous groups. As explained earlier, there were also significant differences between members of both categories.

As expected, the prosodic model explained variance of MCD of German vowels. Vowels in stressed syllables were less distant from each other than vowels in unstressed syllables, and when vowels in unstressed – stressed condition were compared. This finding possibly relates to larger variability in unstressed German vowels because unstressed syllables are produced with a higher degree of coarticulation (Mooshammer and Geng, 2008). Confirming our expectations, non-contrasting SPEECH RATE conditions showed a clear hierarchy of MCD with lowest values for slow – slow, followed by normal – normal and than largest differences between vowels in sentences which were both produced at fast speech rate. Again, this result can be explained by larger variability in vowels at fast compared to slow speech rate (Lindblom, 1963).

Detailed analysis of the random effects for speaker and vowel identity showed that there was variability in how much vowels were, on average, distant from each other in their spectral characteristics, and how speakers produced tokens of same vowel phonemes. High frequent vowels, such as /ə/ and /ɐ/, had larger MCD values than vowels which are less frequent, such as /œ/. Also, vowels with large formant variability (Pätzold and Simpson, 1997) showed overall large MCD values for their respective category.

Analysis of marginal pseudo-R^2 revealed that the ID conditions explained a larger part of the variance in the dependent variable than the prosodic model which was used here, contrary to previous studies (Aylett and Turk, 2006). Still, the spectral characteristics of vowels were only subtly influenced by ID, as expected (Jaeger and Buz, 2017). The amount of explained variance highly depends on the fit of the predictor values for the dependent variable. It is likely that the prosodic model increases in its strength if additional factors were added, for instance realized prominence, phrasal accent or boundary strength.

Most variance in the MCD values was explained by random intercepts for speaker and for vowel identity. This finding can be explained by vowel-inherent variability

and markedness of vowels. For instance, /ə/ was not found in the high surprisal condition, while vowels /øː, œ, yː, ʏ/ only stood in high surprisal biphones. Also, investigation of the conditional modes of the random intercepts showed that the large vowel-inherent variability within /ə/ and /ɐ/ was reflected in overall larger MCD values in pairwise comparisons than for all other German vowels.

This analysis showed that German vowels differ significantly from each other when they stand in different ID contexts. However, ID was not a strong predictor of MCD values of German vowels. The prosodic model explained even less variance than ID which was possibly due to a weak model which was based on canonical stress and a global sentence-based speech rate measure. Corpus-specific syllable frequency showed expected tendencies in MCD: smaller distances were found between vowels in the same conditions compared to vowels in contrasting conditions.

We investigated durational, as well as static and dynamic spectral characteristics of segmental variability. This chapter included a comparative analysis of global spectral characteristics of vowels in different ID and prosodic contexts. The following chapter focuses on voice quality variation as a function of ID and prosody. Voice quality was estimated on the cepstral and the EGG signal.

Chapter 11

Voice quality

The following chapter presents an analysis of voice quality in German and its relationship with ID and prosody. We chose to analyze voice quality based on cepstral features of the speech signal (Section 11.1), and on established metrics based on the EGG signal (Section 11.2). Both analyses quantify the harmonic richness of the signal. The content of the following chapter was summarized in a one-page abstract for a conference proceedings (Brandt, Andreeva, et al., 2018).

11.1 Cepstral peak prominence

This section gives the method, results and discussion of the analysis of cepstral voice parameter and their relation to ID and prosody.

11.1.1 Method

Cepstral peak prominence (CPP) measures the difference in amplitude (in dB) between the cepstral peak and the corresponding fundamental quefrency (Section 2.2.5). We have calculated CPP and cepstral peak prominence smoothed (CPPS) using the CPPS tool designed by Hillenbrand and Houde (1996). Vowels were extracted from running speech of the Siemens Synthesis corpus (Section 3.1.1) and fed into the tool using the default analysis settings for sustained vowels. CPP is calculated every 10 ms of the signal, and then averaged for every signal. CPPS is measured every 2 ms and includes additional smoothing of the cepstra before it is calculated. The smoothing procedure contains two steps: first, the cepstra are averaged over time. Second, the average of cepstral magnitude is calculated across quefrency bins (Hillenbrand and Houde, 1996).

In total, we extracted measurements for 55,093 vowels. For about 35 % of the original vowel data set, we could not retrieve CPP or CPPS measures because of the

short duration of these segments ($< 60\,ms$). Because these short vowels were not processed by the CPPS tool we found that /a/ was the most frequent vowel in the data set ($n = 7,246$), followed by /ə/ ($n = 5,643$) and /ɐ/ ($n = 4,808$). For the analysis of voice quality, we also included the German diphthongs /aɪ/ ($n = 4,498$) /aʊ/ ($n = 1,710$), and /ɔʏ/ ($n = 765$). Figure 11.1 includes vowel token frequencies per vowel identity at the bottom of the bar plots.

11.1.2 Results

For the analysis of voice quality and its relationship with ID and prosodic factors, we included the factor sentence position of the word (final vs. not final). Considering that F0 decreases, and often shows irregularities in glottal fold vibration at sentence final position (Ferrer et al., 2002; Henton and Bladon, 1988), this factor was interpreted as a useful addition to the modeling procedure. It is not only used in the analysis of CPP and CPPS in this Section, but also in Section 11.2 on EGG analysis.

Prior to the statistical analysis, we tested whether lexical class plays a role in CPP or CPPS measurements. In simple ANOVA models, lexical class had a significant influence on both CPP ($F(1) = 248.6, p < 0.001$), and CPPS ($F(1) = 386.4, p < 0.001$). On average, vowels in content words showed higher values in both metrics than in function words. In addition, lexical class was strongly related to word frequency ($r = 0.70$). We therefore decided to exclude function words from the following analysis. This reduced the data set to 40,203 vowels.

Descriptive statistics

Cepstral peak prominence Averaged over all vowel tokens in content words CPP was 20.79 dB with a low standard deviation of 3.26 dB. Open tense vowels, such as /ɛː/ ($M = 22.04, SD = 2.88$) and /aː/ ($M = 22.07, SD = 2.55$) had the largest CPP values (dB), while closed vowels showed a tendency to have smaller values in CPP (Figure 11.1).

Vowels under stress ($M = 20.84, SD = 3.18$) were very similar in their CPP values to unstressed vowels ($M = 20.74, SD = 3.33$). Vowels that immediately preceded a phrase boundary ($M = 19.38, SD = 2.80$) showed smaller CPP than values within word boundaries ($M = 20.82, SD = 3.28$) and at word boundary position ($M = 20.79, SD = 3.06$). If a vowel stood in the last word of a sentence CPP on average was lower ($M = 20.04, SD = 3.46$) than for vowels in words which did not stand at the last position in the sentence ($M = 20.83, SD = 3.24$).

Cepstral peak prominence smoothed CPPS measurements ($M = 9.74, SD = 2.60$) showed a moderately strong correlation with CPP values ($r = 0.68, t(40201) = 188.45, p < 0.001$). We found the same tendency in CPPS for open vowels to have higher values than closed vowels. /ɛː/ ($M = 10.55, SD = 2.00$) and /aː/ ($M =$

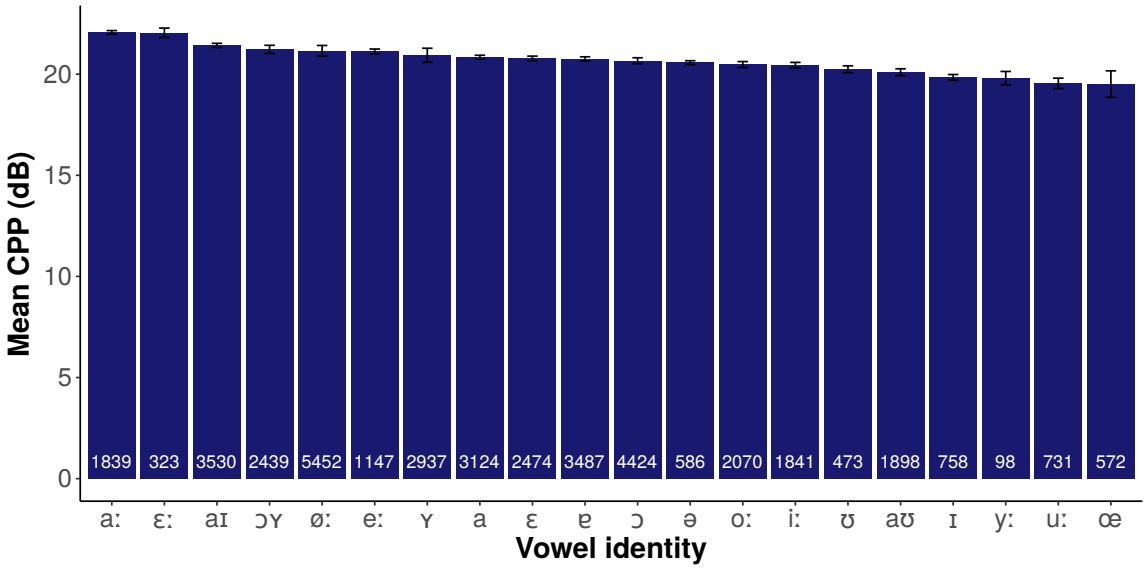


Figure 11.1: Mean CPP (dB) in German per vowel identity. Number of tokens per vowel are given at the bottom of individual bars.

$10.47, SD = 1.99$) also had the highest values in CPPS. CPP and CPPS shared some similarities when analyzed as a function of vowel identity, but they did not show the exact same pattern (Figure 11.2).

Stressed vowels ($M = 9.81, SD = 2.55$) showed larger CPPS values than unstressed vowels ($M = 9.68, SD = 2.65$). For boundary and sentence position, we observed the same patterns as for CPP: vowels immediately preceding a phrase boundary ($M = 8.41, SD = 2.39$) showed lower values in CPPS than vowels at word ($M = 9.91, SD = 2.14$) or no boundary position ($M = 9.76, SD = 2.64$). Also, vowels in the last word of a sentence ($M = 8.71, SD = 2.77$) were produced with lower CPPS values than in all other words ($M = 9.81, SD = 2.58$).

Linear mixed-effects modeling

Cepstral peak prominence Both CPP and CPPS showed the strongest positive correlations with biphone surprisal of the preceding context and triphone surprisal of the following context (Table 11.1). As these factors were only weakly related ($r = 0.08, t(38033) = 16.53, p < 0.001$), we decided to use them both in our LMMs for the perturbation metrics.

Prior to model building we performed a collinearity analysis of the fixed effects. As ID factors, we tested surprisal, word frequency, and phoneme probability. We also included prosodic factors, such as primary lexical stress, boundary, and local and global speech rate, as well as the control factors average vowel duration and final sentence position.

Out of all predictors, primary lexical stress was the most confounded with most of the other factors, except for speech rate. There were weak positive relations to

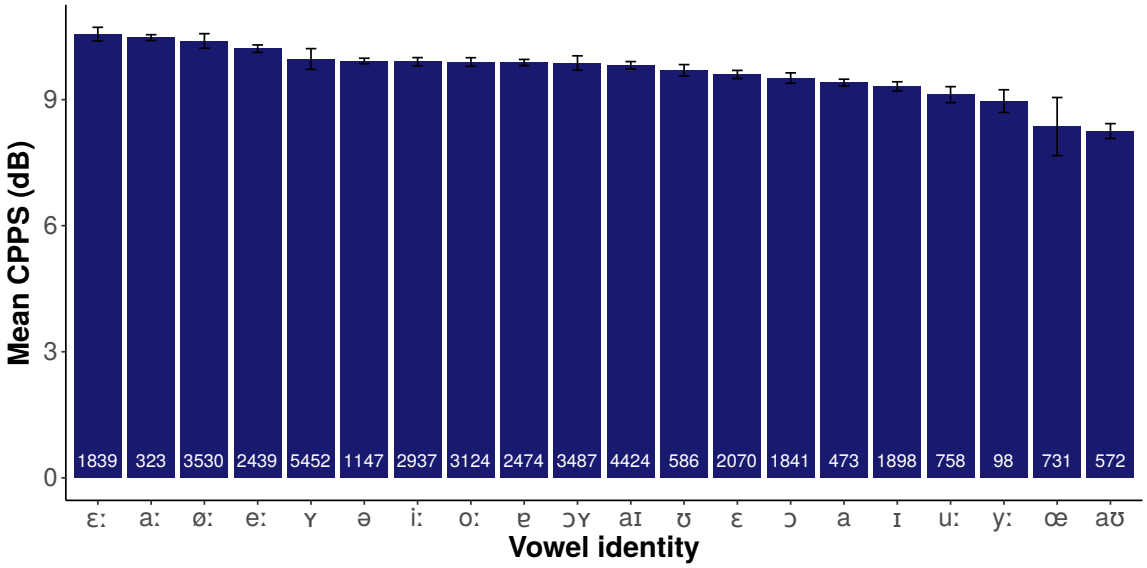


Figure 11.2: Mean CPPS (dB) in German per vowel identity. Number of tokens per vowel are given at the bottom of individual bars.

Table 11.1: Cepstral peak prominence (smoothed): Pearson's correlation coefficients and tests ($\alpha = 0.05$) between CPP measurements and biphone of the preceding (BiSur) and following context (BiFolSur), as well as triphone of the preceding (TriSur) and following context (TriFolSur).

	BiFolSur	BiSur	TriFolSur	TriSur
CPP	0.03***	0.07***	0.06***	0.01*
CPPS	-0.01*	0.07***	0.05***	-0.01*

biphone surprisal of the preceding ($r = 0.14$), and triphone surprisal of the following context ($r = 0.14$), as well as with average vowel duration ($r = 0.28$) and word frequency ($r = 0.19$). Average vowel duration was also related to triphone surprisal of the following context ($r = 0.27$). Since these correlations were weak they can be disregarded in the modeling procedure. For phoneme probability, however, we found moderately weak correlations between biphone surprisal of the preceding context ($r = -0.48$), and average vowel duration ($r = -0.42$). We therefore excluded phoneme probability from the statistical models.

The LMMs for the dependent variables CPP and CPPS were built following the procedure explained in Section 4.3. The final model for CPP contained fixed effects biphone SURPRISAL of the preceding context, triphone SURPRISAL of the following context, WORD FREQUENCY, primary lexical STRESS, BOUNDARY, local and global SPEECH RATE, average vowel DURATION, and SENTENCE POSITION. We included random intercepts for SPEAKER, VOWEL and FOLLOWING CONTEXT, as well as random intercepts and slopes for STRESS per PRECEDING CONTEXT, and STRESS, biphone and triphone SURPRISAL per WORD (Model structure 11.1). Continuous variables SURPRISAL and WORD FREQUENCY were log-transformed because of positive skewness. SPEECH RATE was mean-centered, and categorical variables were treatment-coded.

$$\begin{aligned} CPP/CPPS \sim\ & BiSur + TriFolSur + Wordfreq + \\ & Stress + Boundary + GlobalTempo + LocalTempo + \\ & DurAverage + SentencePosition + \\ & (1|Speaker) + (1 + BiSur + TriFolSur + Stress|Word) + \\ & (1 + Stress|Preceding) + (1|Following) + (1|Vowel) \end{aligned} \tag{11.1}$$

Biphone and triphone SURPRISAL were both predictive of CPP. We found a positive relationship between the dependent variable and SURPRISAL. The other ID factor WORD FREQUENCY, on the other hand, did not reach significance level in the model. Vowels were produced with reduced CPP when they stood at phrase or word BOUNDARY compared to no BOUNDARY position. As global SPEECH RATE increased, CPP values decreased. The opposite effect was found for local SPEECH RATE. Primary lexical STRESS was not significant in predicting CPP. We found, however, expected effects for average vowel DURATION and SENTENCE POSITION. Longer vowels had a more well-defined harmonic structure, while vowels in the last word of a sentence had decreased CPP values (Table 11.2).

Regarding the random structure we found that the phonological context of sibilants, sonorants, and pause had the same influence of CPP in vowels, regardless of their context direction. Following and preceding pause and sibilant both had a decreasing effect on CPP in vowels. Following or preceding sonorants, however, showed an increasing effect on CPP. If there was an obstruent following the vowels CPP was lower than average. Preceding obstruents led to an increase in CPP.

Table 11.2: Cepstral peak prominence (CPP): regression coefficients, standard error (SE) and statistical output of LMM analysis including biphone surprisal of the preceding and triphone surprisal of the following context.

	Terms	Coeff.	SE	t-value	p-value
ID model	Surprisal biphone	0.85	0.07	11.82	$< .001$
	Surprisal triphone	0.07	0.03	2.03	$= .04$
	Word frequency	-0.003	0.008	-0.40	$= .70$
Prosodic model	Global tempo	-0.21	0.02	-9.93	$< .001$
	Local tempo	0.02	0.003	7.52	$< .001$
	Boundary (phrase – none)	-3.52	1.48	-2.38	$= .02$
	Boundary (word – none)	-3.71	0.08	-4.37	$< .001$
	Stress (y – n)	-0.17	0.20	-0.81	$= .46$
Other control	Average vowel duration	1.90	0.80	2.36	$= .03$
	Sentence position (f – n)	-0.66	0.08	-8.64	$< .001$

Conditional pseudo-R^2 reached a total of 30.83 % for both fixed and random structure. The fixed effects only explained 3.08 % of the data variance in CPP. The strongest fixed effect was average vowel DURATION ($Var = 1.19\,\%$), followed by biphone SURPRISAL of the preceding context ($Var = 0.95\,\%$), and global SPEECH RATE ($Var = 0.43\,\%$). SENTENCE POSITION and BOUNDARY ($Var = 0.21\,\%$) explained the same amount of data variance. Local SPEECH RATE ($Var = 0.04\,\%$) and triphone SURPRISAL ($Var = 0.05\,\%$) were the least effective significant fixed effects in the model.

We built the interaction models as explained in Section 4.3. Including the interaction term for biphone SURPRISAL and BOUNDARY led to a significantly better model than the baseline ($\chi^2(1) = 13.10, p = 0.001$). The interaction between word BOUNDARY and high SURPRISAL predicted decreased CPP values ($\beta = -0.71, SE = 0.20$). This interaction term added 0.01 % to marginal pseudo-R^2. The interaction between triphone SURPRISAL and STRESS performed significantly better than the baseline model ($\chi^2(1) = 17.45, p < 0.001$). Both factors complemented each other positively in their effect on CPP. Vowels under STRESS and high SURPRISAL showed higher values in CPP ($\beta = 0.25, SE = 0.06$). The interaction added 0.09 % explained variance to the model.

Cepstral peak prominence smoothed For reasons of comparability, we kept the model structure constant for both measures of voice quality CPP and CPPS. The LMM for CPPS was run with the structure given in Model structure 11.1.

As results, we found the same effect of biphone SURPRISAL on CPPS as in the model for CPP. Vowels in high biphone surprisal contexts were produced with higher CPPS. The effect of triphone SURPRISAL, however, was not significant. The model

Table 11.3: Cepstral peak prominence smoothed (CPPS): regression coefficients, standard error (SE) and statistical output of LMM analysis including biphone surprisal of the preceding and triphone surprisal of the following context.

	Terms	Coeff.	SE	t-value	p-value
ID model	Surprisal biphone	1.01	0.06	16.96	$< .001$
	Surprisal triphone	0.02	0.03	0.63	$= .53$
	Word frequency	-0.005	0.006	-0.75	$= .45$
Prosodic model	Global tempo	-0.09	0.01	-6.60	$< .001$
	Local tempo	0.01	0.002	5.23	$< .001$
	Boundary (phrase - none)	-2.86	1.00	-2..88	$< .01$
	Boundary (word - none)	-0.51	0.06	-8.11	$< .001$
	Stress (y - n)	0.20	0.25	0.81	$= .47$
Other control	Average vowel duration	1.02	0.77	1.32	$= .20$
	Sentence position (f - n)	-1.01	0.05	-18.68	$< .001$

output of CPPS also showed non-significant effects for STRESS and WORD FREQUENCY, as well as the same effects for SPEECH RATE, SENTENCE POSITION, and BOUNDARY that were observed in the CPP model. While average vowel DURATION was the strongest predictor in the CPP model, it did not reach significance level in the CPPS LMM (Table 11.3).

Vowels with preceding or following sibilants showed lower values in CPPS than the overall average. With preceding or following obstruents vowels had higher values in CPPS. For the other two phonological contexts, we observed that preceding sonorants had a decreasing effect on CPPS, while following sonorants had the opposite effect. Expectedly, following pauses had a decreasing effect on CPPS, whereas it increased with preceding pauses.

Regarding the effect size of the CPPS model we found a considerably higher conditional pseudo-R^2 ($Var = 56.65\,\%$) than in the CPP model, although both shared the same model structure and were strongly correlated. The fixed effect only explained a total of $4.94\,\%$ of data variance. Biphone SURPRISAL was the strongest predictor with an effect size of $3.00\,\%$, followed by SENTENCE POSITION ($Var = 0.77\,\%$). All other fixed effects were less effective. BOUNDARY added $0.40\,\%$ explained data variance. Global SPEECH RATE explained $0.16\,\%$ of CPPS variance, while local SPEECH RATE had the smallest effect size ($Var = 0.05\,\%$).

Regarding the interaction models investigating the interaction between ID and the prosodic factors on CPPS we found the same significant interactions that were observed in the CPP model: the interaction between biphone SURPRISAL and word BOUNDARY had a significant negative effect on CPPS. The term added $0.09\,\%$ explained data variance to the model. Although the main effect of triphone SURPRISAL

Table 11.4: Cepstral peak prominence smoothed (CPPS): interaction of biphone surprisal of the preceding and triphone surprisal of the following context with prosodic factors. Only significant interactions are reported.

	Terms	Coeff.	SE	t-value	p-value
Biphone	Stress * Surprisal	0.34	0.11	3.23	= .001
	Boundary (phrase) * Surprisal	-6.85	4.20	-1.63	= .10
	Boundary (word) * Surprisal	-0.65	0.15	-4.23	< .001
	Global tempo * Surprisal	0.09	0.04	2.30	= .02
Triphone	Stress * Surprisal	0.46	0.005	9.21	< .001
	Local tempo * Surprisal	0.005	0.003	4.39	= .04

was not significant in the baseline model, interactions between the ID factor and other prosodic factors reached significance level. Triphone SURPRISAL interacted positively with STRESS on CPPS adding 0.39 % to the effect size of the model. In addition, the interaction model with an interaction term for triphone SURPRISAL and local SPEECH RATE performed better than the baseline model. The interaction had a positive effect on CPPS adding 0.02 % explained data variance. Biphone SURPRISAL also interacted positively with STRESS and showed a significant positive effect on CPPS ($Var = 0.13\,\%$). The interaction model including an interaction between biphone SURPRISAL and global SPEECH RATE had a better performance than the baseline model. However, the effect was positive: vowels in sentences with a high SPEECH RATE under high surprisal showed larger values in CPPS. The term added 0.01 % explained data variance (Table 11.4).

11.1.3 Discussion

The measures CPP and CPPS inform about the magnitude of peakedness in the harmonic structure of a sustained vowel or connected speech. The higher the values in these metrics, the more well defined is the harmonic structure. We therefore expected vowels in high surprisal contexts and low-frequency words to show higher values in CPP measurements than in low surprisal contexts and high-frequency words. This expectation was met for both CPP and CPPS with regards to biphone surprisal of the preceding context. Only for CPP, we also found a significant effect of triphone surprisal of the following context in the expected direction. Although non-significant as a main effect, triphone surprisal interacted with stress and speech rate leading to an improvement in model performance for the CPPS model. The working hypothesis was not confirmed regarding the impact of word frequency on these voice quality metrics. Although there was a tendency for a negative effect of word frequency on CPP and CPPS, it did not reach significance level in the models.

Prosodic boundary at word and phrase level had a significant effect on both CPP

and CPPS. Vowels immediately preceding these boundary positions showed significantly lower values in both metrics. The interaction between the factors biphone surprisal of the preceding context and prosodic boundary was significant for both CPP measurements. If a vowel stood in an open syllable as the last segment of a word and in a high surprisal context it showed significantly lower values in CPP and CPPS than at no boundary position under low surprisal. There was a tendency for the same interaction between phrase boundary and biphone surprisal, but it did not reach significance level, possibly because there was a great amount of overlap between phrase boundary and the random effect following phonological context which also included "pause" as a factor level. Despite its small effect on the perturbation metrics, we observed that the factor prosodic boundary impacted CPP and CPPS in a systematic way indicating that differences in voice quality can be interpreted as cues for prosodic boundaries at different levels of the prosodic hierarchy.

Primary lexical stress neither had a significant impact on CPP nor on CPPS. However, in interaction with triphone surprisal of the following context primary lexical stress had a positive effect on both CPP and CPPS, and for CPPS this positive interaction was also observed between biphone surprisal and stress. The lack of a significant effect for the main predictor can be explained by the weak collinearities that were found with other factors in the model, such as with surprisal or average vowel duration. In addition, we did not find large differences in either CPP or CPPS between unstressed and stressed vowels in the descriptive analysis. This finding showed that well-defined harmonic structure did not depend on the main effect of primary lexical stress, although some have suggested that harmonic richness and notions of prominence are correlated (Michaud, 2004).

There were conflicting results for global and local speech rate in the models for CPP and CPPS. Vowels in sentences at fast global speech rate showed lower CPP and CPPS values than at slow global speech rate. The opposite effect on both measures was observed for local speech rate. While the effect of global speech rate was expected because phonetic structures are known to expand in duration and are more distinct in their spectral characteristics at slow speech rate (Bell et al., 2009; Gahl et al., 2012; Turner et al., 1995), we did not expect the positive effect of local speech rate on both perturbation metrics. Also, the effect of global speech rate did not overrule local effects leading to these opposing results.

Average vowel duration had a significant effect on CPP leading to higher CPP values in longer vowels. It was the strongest fixed effect in the model. For CPPS, however, we only found a tendency for this effect which did not reach significance level. This result was due to the durational averaging that was part of the smoothing procedure applied for CPPS.

In both models, sentence position was predictive of variance in CPP and CPPS. Vowels in the last word of a sentence showed less well-defined harmonic structure than vowels in words with non-final position. In the CPP model, sentence position only had a small effect, while it was the second strongest predictor in the CPPS model.

Its effect was predicted based on studies on sentence intonation. Falling intonation patterns were reported for statements in German. Since the corpus material consists of newspaper read speech we assumed that the majority of the sentences contains statements. As pitch decreases towards the end of a sentence, glottal vibrations have a tendency to become irregular (Ferrer et al., 2002) which is then observed in a less well-defined harmonic structure in the vowels.

11.2 EGG analysis

In this section, we present voice quality estimated from the glottal source in relation to ID and prosody. The method, results and their discussion are given.

11.2.1 Method

The analysis of the electroglottographic (EGG) signal contained within the Siemens Synthesis corpus was performed using the peakdet algorithm implemented in Matlab by Michaud (2007). Peakdet takes a set of continuously voiced speech signals and calculates several open quotient (OQ) metrics and derivative-EGG open peak amplitude (DEOPA) based on the first derivative of the glottal waveform (DEGG) method.

We used the following settings: the EGG signal was reinterpolated at closing and opening peaks for accurate peak detection with the coefficient 100. The coefficient for recognition of double peaks in the signal was set at 0.5 (Henrich et al., 2004). If there was a multiple peak detected we chose to output the value between them (barycentre method). The amplitude threshold for peak detection was set automatically by the tool based on the sampling frequency of the EGG signals. The maximum F0 frequency was set at 300 Hz for male speakers. After acquiring the data, we discarded data points with F0 lower than 50 Hz because they were unlikely to be physiological productions by the speaker ($n = 413$). The following measures were extracted using peakdet:

- OQ: open quotient using detection by maxima on unsmoothed DEGG signal
- OQs: open quotient using detection by maxima on smoothed DEGG signal
- OQval: open quotient determined by peak detection on unsmoothed DEGG signal
- OQvals: open quotient determined by peak detection on smoothed DEGG signal
- DEOPA: amplitude of opening-peak of the DEGG signal

Table 11.5: EGG analysis: Pearson's correlation coefficients and tests ($\alpha = 0.05$) between EGG measurements and biphone of the preceding (BiSur) and following context (BiFolSur), as well as triphone of the preceding (TriSur) and following context (TriFolSur).

	BiFolSur	BiSur	TriFolSur	TriSur
OQ	-0.027***	0.075***	-0.042***	0.066***
OQs	-0.027***	0.080***	-0.042***	0.070***
OQval	-0.026***	0.075***	-0.046***	0.069***
OQvals	-0.027***	0.080***	-0.047***	0.072***
DEOPA	-0.014**	0.052***	-0.013*	0.074***

11.2.2 Results

Parallel to the analysis in Section 11.1 above, we excluded function words from the analysis, because the factor word class had a significant effect on all EGG measurements when tested in ANOVAS. We therefore reduced the data set to 42,414 vowel tokens in total. As mentioned above, we included the additional factor sentence position of the word (final vs. not final) in the statistical analysis for the EGG metrics.

The Matlab implemented peakdet algorithm calculates four metrics of OQ which are all strongly correlated. We therefore decided to only use one of the OQ metrics based on the correlations with the surprisal measurements and the correlations with DEOPA. Because of their strong correlation the OQ metrics showed highly similar Pearson's r correlation values with surprisal (Table 11.5). Overall, the smooth metrics had slightly higher negative correlations with surprisal than the unsmoothed OQ measurements. OQs and DEOPA ($r = 0.06, t(42412) = 12.21, p < 0.001$) were slightly less dependent on each other than OQvals and DEOPA ($r = 0.064, t(42412) = 13.35, p < 0.001$). For that reason, we decided to use OQs as the OQ metric in the following analysis (Table 11.5). In addition, we included an analysis on DEOPA and its relationship to ID and prosodic factors because it only showed weak relations to the other EGG measurements.

Descriptive statistics

Open quotient smoothed On average, OQs reached a value of 51.07 ($SD = 7.43$). We found differences in OQs depending on vowel height. Open vowels, such as /ɛː/ ($M = 49.24, SD = 6.20$), /ɛ/ ($M = 49.18, SD = 6.55$), and /aː/ ($M = 49.48, SD = 6.55$) showed the lowest values in OQs, while closed vowels had the highest values, e. g., /ʏ/ ($M = 56.75, SD = 8.61$) and /uː/ ($M = 55.84, SD = 7.11$) (Figure 11.3).

Stressed ($M = 51.13, SD = 7.49$) and unstressed vowels ($M = 51.03, SD = 7.39$) only differed slightly in their mean values of OQs. At phrase boundary, OQs was highest ($M = 52.62, SD = 6.60$), followed by no boundary position ($M = 51.16, SD =$

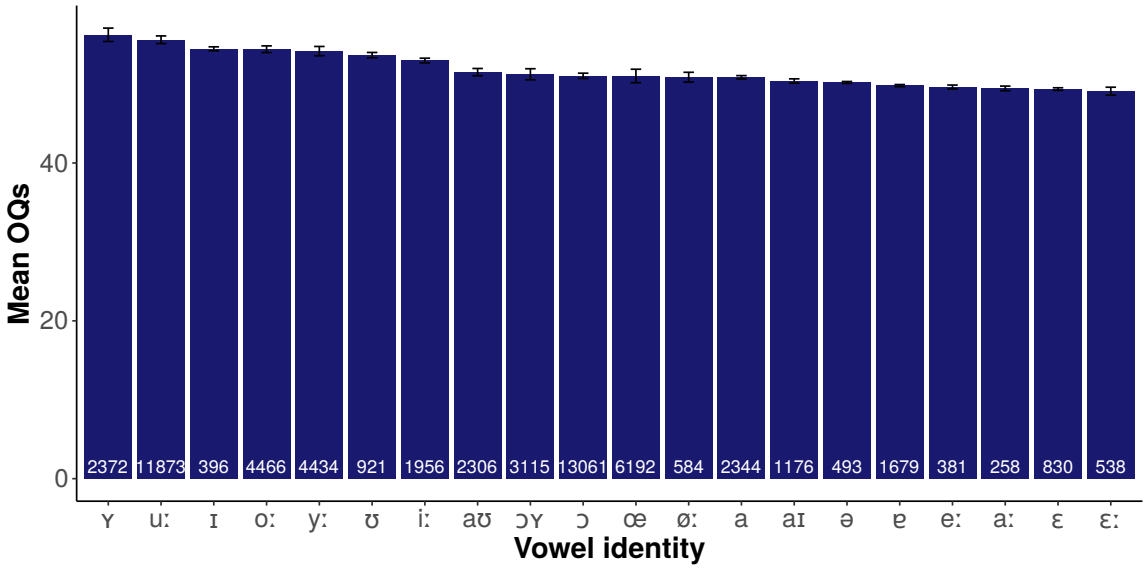


Figure 11.3: Mean OQs in German per vowel identity. Number of tokens per vowel are given at the bottom of individual bars.

7.50), and word boundary ($M = 50.04, SD = 6.90$). OQs decreased when the vowel stood in the last word of a sentence ($M = 49.14, SD = 6.74$) compared to vowels at no final sentence position ($M = 51.18, SD = 7.46$).

Derivative-EGG open peak amplitude As an overall mean value for DEOPA, we calculated -0.007 ($SD = 0.003$). Vowel identity had an influence on DEOPA values with similar tendencies as in OQs, and the CPP measurements. Open vowels showed a tendency for lower values in DEOPA indicating that they had a more well-defined harmonic structure compared to closed vowels. /ɛː/ ($M = -0.008, SD = 0.003$) and /ɛ/ ($M = -0.008, SD = 0.003$) showed the lowest DEOPA values of all vowel phonemes, while /oː/ ($M = -0.004, SD = 0.001$) and /ʊ/ ($M = -0.005, SD = 0.001$) had the highest values.

Vowels under primary lexical stress ($M = -0.0069, SD = 0.003$) showed similar values in DEOPA as unstressed vowels ($M = -0.0068, SD = 0.003$). If a phrase boundary ($M = -0.0065, SD = 0.003$) immediately followed the vowel DEOPA was higher than in vowels at no boundary ($M = -0.0068, SD = 0.003$) or word boundary position ($M = -0.0078, SD = 0.004$). There were also clear differences in mean DEOPA in German vowels depending on the sentence position of the word in which the vowel appeared: vowels in the last word of the sentence ($M = -0.0054, SD = 0.002$) showed higher DEOPA values than words in non-final positions ($M = -0.0070, SD = 0.003$).

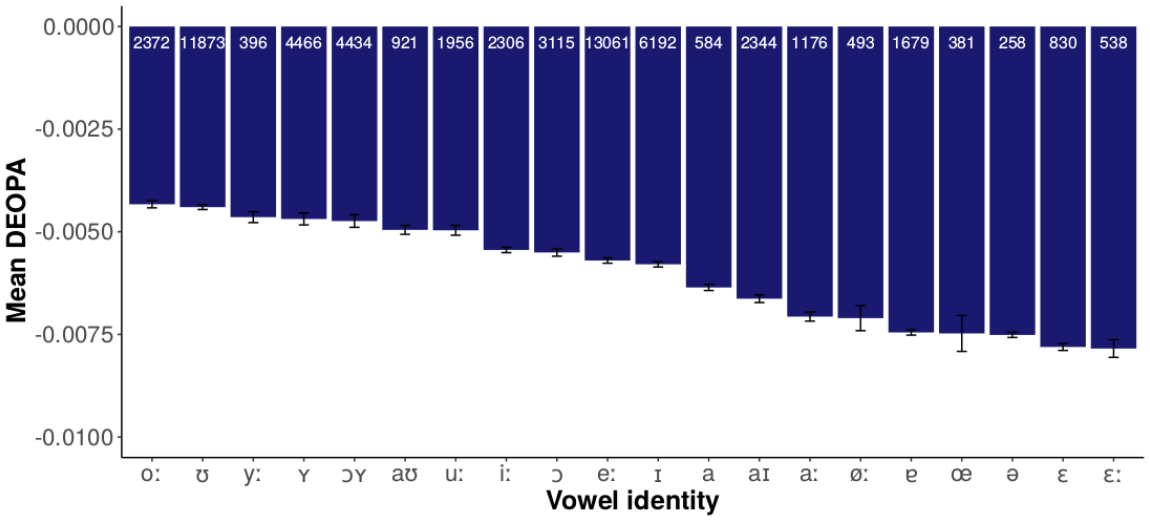


Figure 11.4: Mean DEOPA in German per vowel identity. Number of tokens per vowel are given at the top of individual bars.

Linear mixed-effects modeling

We tested Pearson's r correlations with biphone and triphone surprisal measures of both context directions (Table 11.5). For the following inferential statistics, we used triphone surprisal of the following context as the ID measure because it showed the highest negative correlations with the EGG metrics out of all the measures. Prior to model building we performed a collinearity analysis between the fixed effects. None of the correlations were above a weak level.

Open quotient smoothed For the OQs LMM, we used the same model structure that was used for the CPP analysis, except for the random slope for STRESS per WORD which had to be discarded because of convergence errors. Also, we did not include biphone SURPRISAL of the preceding context in the OQs model because it did not show a negative relationship with the dependent variable (Model structure 11.2).

$$\begin{aligned} OQs \sim\ & TriFolSur + Wordfreq + \\ & Stress + Boundary + GlobalTempo + LocalTempo + \\ & DurAverage + SentencePosition + \\ & (1|Speaker) + (1 + TriFolSur|Word) + \\ & (1 + Stress|Preceding) + (1|Following) + (1|Vowel) \end{aligned} \tag{11.2}$$

Results of the LMM analysis showed that high-frequency content words had significantly higher values in OQs than low-frequency words. Triphone SURPRISAL, however, only showed a tendency for a negative effect. Global and local SPEECH RATE both had a significant negative effect on OQs. In addition, we observed a sig-

Table 11.6: Open quotient smoothed (OQs): regression coefficients, standard error (SE) and statistical output of LMM analysis including triphone surprisal of the following context.

	Terms	Coeff.	SE	t-value	p-value
ID model	Surprisal	-0.11	0.08	-1.37	= .17
	Word frequency	0.09	0.02	4.32	< .001
Prosodic model	Global tempo	-0.24	0.06	-3.86	< .001
	Local tempo	-0.02	0.008	-2.74	< .001
	Boundary (phrase - none)	15.25	6.45	2.36	< .02
	Boundary (word - none)	-0.29	0.18	-1.65	= .10
	Stress (y - n)	0.35	0.13	2.64	< .01
Other control	Average vowel duration	-3.51	1.53	-2.30	< .03
	Sentence position (f - n)	-2.24	2.14	-10.45	< .001

nificant negative effect of SENTENCE POSITION, and average vowel DURATION, while STRESS and phrase BOUNDARY both shared a positive effect on OQs (Table 11.6).

Preceding and following obstruents and sonorants had a decreasing effect on OQs compared to the overall mean, while preceding and following pauses increased OQs values. For sibilants, we found that they had an increasing effect when immediately following, and a decreasing effect when they preceded the vowel.

The conditional pseudo-R^2 of the entire model for OQs added up to 33.08 % explained variance with only 1.85 % of the variance explained by the fixed effects. The strongest predictor of OQs was average vowel DURATION ($Var = 1.10\,\%$), followed by SENTENCE POSITION ($Var = 0.47\,\%$). STRESS ($Var = 0.14\,\%$), and WORD FREQUENCY ($Var = 0.11\,\%$). Both SPEECH RATEs ($Var = 0.03\,\%$) were much less effective in predicting OQs variance.

We also tested interactions between triphone surprisal and the prosodic factors as predictors for OQs in separate interaction models. None of the interaction models performed better than the baseline model.

Derivative-EGG open peak amplitude We used the same model structure for DEOPA as in the OQs model, except for the random slope of STRESS per PRECEDING CONTEXT because it showed a perfect correlation with the random intercept (Model

Table 11.7: Derivative-EGG open peak amplitude (DEOPA): regression coefficients, standard error (SE) and statistical output of LMM analysis including triphone surprisal of the following context.

	Terms	Coeff.	SE	t-value	p-value
ID model	Surprisal	0.00001	0.00002	0.12	= .90
	Word frequency	0.00001	0.00001	-0.81	= .42
Prosodic model	Global tempo	0.0001	0.00002	5.36	< .01
	Local tempo	-0.00003	0.00001	-12.06	< .01
	Boundary (phrase – none)	-0.003	0.002	-1.31	= .19
	Boundary (word – none)	0.0002	0.00006	3.89	< .01
	Stress (y – n)	-0.0002	0.00005	-4.79	< .01
Other control	Average vowel duration	-0.0008	0.0007	-1.12	= .28
	Sentence position (f – n)	0.001	0.00008	17.28	< .01

structure 11.3).

$$\begin{aligned} DEOPA \sim\ & TriFolSur + Wordfreq + \\ & Stress + Boundary + GlobalTempo + LocalTempo + \\ & DurAverage + SentencePosition + \\ & (1|Speaker) + (1 + TriFolSur|Word) + \\ & (1|Preceding) + (1|Following) + (1|Vowel) \end{aligned} \tag{11.3}$$

None of the ID factors, nor average vowel DURATION were significant predictors of DEOPA. The non-significant effect for triphone SURPRISAL on DEOPA was expected based on the very low correlation between both measures (Table 11.5). Global SPEECH RATE, word BOUNDARY, and SENTENCE POSITION all had a positive effect on DEOPA. These findings were expected: vowels at accelerated tempo, word BOUNDARY position or in the last word of the sentence showed less harmonic richness than vowels at slow tempo, no BOUNDARY position and non-final SENTENCE POSITION. Local SPEECH RATE, however, had the opposite effect on DEOPA. Vowels in words that were spoken at a fast tempo showed lower values in DEOPA than vowels in slowly spoken words. As expected, STRESS decreased DEOPA values leading to more harmonic richness in stressed vowels compared to unstressed vowels.

PRECEDING CONTEXT had the same impact on DEOPA as it had on OQs. All preceding contexts except for pause had a decreasing effect on the metric. With following obstruents, pauses, and sibilants DEOPA showed lower values than the average, while following sonorants had an increasing effect.

The model for DEOPA was far more effective than that for OQs. We reached a total of 61.31 % explained data variance for the entire model structure. Marginal

pseudo-R^2 indicated that the fixed effects contributed 1.72 % to the total effect size. SENTENCE POSITION was the strongest factor adding 1.04 % to the overall effect size. Local SPEECH RATE ($Var = 0.34\,\%$), BOUNDARY ($Var = 0.10\,\%$), STRESS ($Var = 0.07\,\%$), and global SPEECH RATE ($Var = 0.04\,\%$) added only little to the model, although they were significant predictors of variability in DEOPA.

We ran interaction models including interaction terms between triphone SURPRISAL and the prosodic factors in order to investigate whether SURPRISAL had a significant impact on DEOPA when interacting with one of the prosodic factors. STRESS interacted with triphone SURPRISAL complementing each other in their negative effect on DEOPA which decreased under STRESS and high SURPRISAL ($\beta = -0.0001, SE = 0.00005, t(736) = -2.35, p = 0.02$). The interaction added 0.02 % explained data variance to the model. All other interaction models did not perform significantly better than the baseline model.

11.2.3 Discussion

Triphone surprisal of the following context was neither predictive of OQs nor DEOPA. We found a tendency for an expected negative effect in the OQs model. This result was expected based on the low negative correlations between both measures. Surprisal had the expected negative effect on DEOPA in interaction with stress. This interaction model outperformed the baseline model. In previous analyses in this thesis, we also observed this phenomenon for other acoustic-phonetic measures: triphone surprisal of the following context was not significant in explaining segment duration, but in interaction with stress. Biphone surprisal of the preceding context showed the same null effect on F2 slope and VL, but was significant in interaction with stress.

On the word level, we found a significant positive effect for the ID measure word frequency on OQs. Vowels in high-frequency words showed higher values in OQs, i. e., these words were produced with decreased harmonic richness compared to low-frequency words. This effect was observed although we conducted the analysis only on content words which indicated that voice quality is used as a cue for word frequency, even after controlling for lexical class. We did not find an effect of word frequency on DEOPA, although strong peaks of its counterpart measure derivative-EGG closure peak amplitude (DECPA) have been reported for pragmatically important syllables of an utterance (Michaud, 2004).

Prosodic boundaries were marked in voice quality by different EGG metrics. At phrase boundaries, OQs significantly increased compared to no boundary position, while word boundaries were cued by a significant increase in DEOPA. Both effects shared the same expected direction: at prosodic boundary positions, the harmonic structure of vowels was less rich than at no boundary position. Voice quality was therefore used as a marker of prosodic boundaries at different levels. This was expected considering that creaky voice quality was reported as a marker for prosodic boundaries (Ferrer et al., 2002; Henton and Bladon, 1988).

Different stress environments were reported to have a strong effect on glottal excitation (Gobl, 1988). For primary lexical stress, we did not find a uniform effect for both EGG metrics. While stressed vowels showed significantly higher OQs, they also had lower DEOPA values. The main acoustic-phonetic correlates for stress are F0, duration and intensity. Stressed vowels are longer, and they are produced with higher F0 and amplitude than unstressed vowels (Lieberman, 1960). An increase in F0 for prominence marking leads to longer OQs which explained the positive effect of stress that we found in this model. Increased intensity of stressed vowels correlates with an increase in vocal effort, and thus more pronounced DEOPA. This effect accounted for the negative effect of stress on DEOPA (Michaud, 2004).

At increased local speech rate, OQs and DEOPA decreased. At the sentence level, we also found a negative effect of speech rate on OQs, but a positive effect on DEOPA. Fast speech rate exhibits lower F0 and less pitch movement than slow speech rate. Listeners use intonational cues to draw conclusions about speech rate deviations (Rietveld and Gussenhoven, 1987). The decreasing effect on OQs can be explained by this phenomenon. As pitch lowered as a result of an increase in speech rate, OQs also decreased. For DEOPA, we found more distinct peaks at faster local speech rate which can also stem from lower average F0 at increased speech rate. The opposite effect on DEOPA was observed for global speech rate. This result can also be confounded by other influences, such as a high amount of laryngealization at high speech rate. Laryngealization lowers the magnitude of open peaks in the DEGG signal (Henrich et al., 2005; Michaud, 2004). We have not controlled for these additional confounding variables in this analysis.

Average vowel duration only had a significant influence on OQs. Longer vowels were produced with decreased OQs values. Longer vowels also showed an increase in harmonic richness compared to short vowels in the previous CPP analysis.

For the factor sentence position, we found conflicting results on both EGG metrics. While there was a positive effect on DEOPA, vowels in the last word of the sentence displayed decreased OQs values. From a physiological standpoint, glottal vibration decreases in peakedness in their opening gesture at the end of a sentence because phonation ceases at this sentence position. For that reason, we found higher DEOPA values vowels in the last word of a sentence. Laryngealization towards the end of sentence can explain the negative effect of sentence position on OQs. This voice setting has been shown to lower OQ metrics compared to modal voice (Henrich et al., 2005; Michaud, 2004).

This chapter on voice quality as a function of ID and prosody concludes the production analyses in this thesis. The following chapter presents the perception experiment conducted using cross-splicing of target words in high and low ID contexts. While the production analyses focused on segments, the perception test was performed manipulating lexical items.

Chapter 12

Perceptual sensitivity of violated ID expectations

We have found that differences in ID have subtle effects on durational and spectral characteristics of speech. Based on these findings in the production data we have also tested listeners' awareness of differences in ID contexts in a cross-splicing experiment. Here, we did not manipulate on the phoneme level, but on the word level. As a post-hoc analysis of the experimental items, we estimated durational and spectral distances to their respective baselines. This chapter describes the experiment design and the conduction of the perception experiment, analyzes listeners' judgments based on d' values, compares stimuli to their baseline in a post-hoc analysis, and discusses the findings of these analyses.

12.1 Experiment design

The perception test was a cross-splicing experiment designed as a discrimination task. In a classic discrimination task, listeners are asked if two stimuli are the same or different. Since this was not the aim of the experiment we asked participants to judge whether the second stimulus sounded more natural than the first one. The task remained the same for the entire duration of the test. As controls, we included the order of presentation of the stimuli, i. e., we reversed the order of the stimuli for comparisons between the baseline and the crossed audio signals. The factor order of presentation consisted of two levels: crossed – base, and base – crossed.

An additional factor integrated in the experiment was the direction of cross-splicing. There were two different conditions: high surprisal context with low surprisal word integrated (hl), and low surprisal context with originally high surprisal word (lh). This added up to 80 trials of comparing crossed audio signals to the baseline (10 words $*$ 2 speakers $*$ 2 orders of presentation $*$ 2 surprisal crossings). Additionally,

there were 40 baseline comparisons included in the experimental setup (10 words $*$ 2 speakers $*$ 2 surprisal crossings). The total number of trials was 120. There was no option to repeat the trial. However, a goodness rating was included in the test (1 = sure, and 5 = unsure).

Before we conducted the experiment we ran a pre-test with two student assistants as participants. They gave feedback on the experiment design, and in particular on potentially problematic stimuli. These stimuli were then double-checked for artifacts of cutting. The perception experiment was conducted in a quiet seminar room in the Phonetics institute at Saarland University. The audio output was transmitted via high-quality headphones. The perception test was implemented in Praat (Boersma and Weenink, 2017) using the ExperimentMFC surface.

12.1.1 Experiment items

As materials, we used recordings from the Siemens Synthesis corpus (Schiel, 1997) (Section 3.1.1). This means, the speakers were not aware of differences in contextual predictability of lexical items in the corpus. Frequent content words in the corpus were identified that appeared at least 10 times. Surprisal values of word bigram, trigram, fourgram for the following and preceding context were extracted from the German word LM (Section 4.2.1). We used log-transformed surprisal values as in all previous analyses. Fourgram surprisal values of the preceding context had the highest mean standard deviation ($M = 0.89$) indicating the highest degree in variability between low and high values. Also, predictability estimates based on the following context did not have a significant influence on listeners' sensitivity towards differences in phonetic detail (Manker, 2017). We therefore decided to use fourgram surprisal of the preceding context as our predictability measure in this study.

Then, 10 words with fourgram surprisal values of the preceding context were identified as tokens for the perception test based on their high and low surprisal values with a maximum of difference between the two conditions (Table 12.1). Word tokens produced in a high surprisal context were cross-spliced into low surprisal context for the same word token, and vice versa. The stimuli consisted of the target word and a carrier with a maximum of three words preceding and following the target word (Appendix Table 3). This procedure reduced the duration of the stimuli from entire sentences to short passages. Stimuli were between 0.89 sec and 3.48 sec long. There were presented with an inter-stimulus interval of 1 sec. It follows that creating the stimuli included cutting the audio signal. The crossed stimuli were therefore checked for artifacts of manipulation before using them in the perception test.

12.1.2 Participants

29 native speakers of German ($m = 8, f = 21$) between the ages of 18 and 32 ($M = 23.93, SD = 4.28$) took part in the perception experiment. None of the par-

Table 12.1: Experiment items: mean standard deviation of surprisal per word token calculated based on the lexical items from the Siemens Synthesis corpus. NA (not applicable) denotes missing values.

Word token	Preceding context			Following context		
	Fourgram	Trigram	Bigram	Fourgram	Trigram	Bigram
machen	1.84	0.99	0.34	NA	NA	0.11
wieder	1.08	0.47	0.37	0.14	0.26	0.23
führen	1.07	0.44	0.12	0.01	0.28	0.13
zwischen	0.90	0.64	0.46	0.73	0.52	0.22
gegen	0.79	0.46	0.40	0.52	0.47	0.15
bereit	0.75	0.31	0.16	NA	0.39	0.16
deutschen	0.71	0.40	0.13	0.40	0.87	0.38
andere	0.70	0.59	0.13	NA	0.65	0.30
polnischen	0.47	0.18	0.08	NA	0	0.16
teil	0.37	0.16	0.22	NA	NA	0.12

ticipants reported past or present hearing issues. Almost all of them ($n = 24$) were students at Saarland University. Participants took about 20–25 min to complete the experiment. They received monetary compensation for taking part in the perception test. Participants were naive with regard to the purpose of the experiment.

12.2 Results

We used d' to measure the sensitivity in perceiving a difference between baseline and crossed audio stimuli (Green and Swets, 1966). d' is calculated as the z-scored distance between signal and noise. d' values of 0 indicate that participants made their choices at chance level, i.e., the signal cannot be distinguished from noise. Positive values of d' suggest that participants were able to distinguish differences between stimuli successfully. Negative values of d' were possible through response confusion, or misinterpretation of the task. They can also arise when participants are more prone to give false-alarms than hits (Stanislaw and Todorov, 1999). We used the R package "sensR" (Christensen and Brockhoff, 2017) for d' computation which also tests whether the difference between signal and noise was significant (Fisher's Exact test).

12.2.1 Descriptive statistics

Participants were not able to detect differences between baseline and crossed condition when these were presented in the order of base – crossed. d' calculation indicated that there was no significant difference between noise and signal for this particular order of

Table 12.2: Sensitivity (d') coefficients, standard error (SE), lower and upper values, as well as p-values from Fisher's Exact test for both orders of presentation of baseline and crossed stimuli per surprisal condition (hl or lh).

Condition	Coeff.	SE	Lower	Upper	p-value
base crossed hl	0.01	0.08	0.00	0.16	$=.50$
base crossed lh	-0.25	0.08	0.00	-0.09	$=.99$
crossed base hl	0.53	0.08	0.38	0.67	$< .001$
crossed base lh	0.55	0.07	0.40	0.70	$< .001$

Table 12.3: Sensitivity (d') coefficients, standard error (SE), lower and upper values, as well as p-values from Fisher's Exact test for both orders of presentation of baseline and crossed stimuli per surprisal condition (hl or lh) and per both male speakers (ai, wo).

Speaker	Condition	Coeff.	SE	Lower	Upper	p-value
ai	base crossed hl	0.09	0.11	0.00	0.31	$= .23$
	base crossed lh	-0.28	0.10	0.00	-0.07	$= .99$
	crossed base hl	0.52	0.11	0.31	0.73	$< .001$
	crossed base lh	0.46	0.11	0.25	0.67	$< .001$
wo	base crossed hl	-0.08	0.11	0.00	0.14	$= .80$
	base crossed lh	-0.22	0.11	0.00	0.004	$= .98$
	crossed base hl	0.53	0.11	0.32	0.75	$< .001$
	crossed base lh	0.66	0.11	0.45	0.88	$< .001$

presentation regardless of the surprisal condition of the stimuli. However, for the same stimuli in the order of presentation crossed – base we found that participants perceived that the crossed stimuli sounded less natural than the corresponding baseline. The difference between noise and signal for d' was significant for both surprisal conditions (Table 12.2).

Further analysis per speaker and condition revealed that speaker identity did not play a huge role in participants ability to successfully distinguish between crossed and baseline stimuli. We found slightly higher d' values for both crossed – base conditions for speaker *wo* compared to speaker *ai*. Apart from that, the analysis per speaker and condition for d' mirrored our previous analysis per condition (Table 12.3).

However, in a more detailed analysis, we found that item identity had a strong impact on d' values. Although crossed – base items had higher sensitivity values on average, for stimuli with the words "deutschen" and "gegen" there were higher values in d' for the order of presentation base – crossed. For "wieder" stimuli, we only found positive d' values in all four conditions. Crossed stimuli containing this word were successfully discriminated as sounding less natural compared to the baseline, regard-

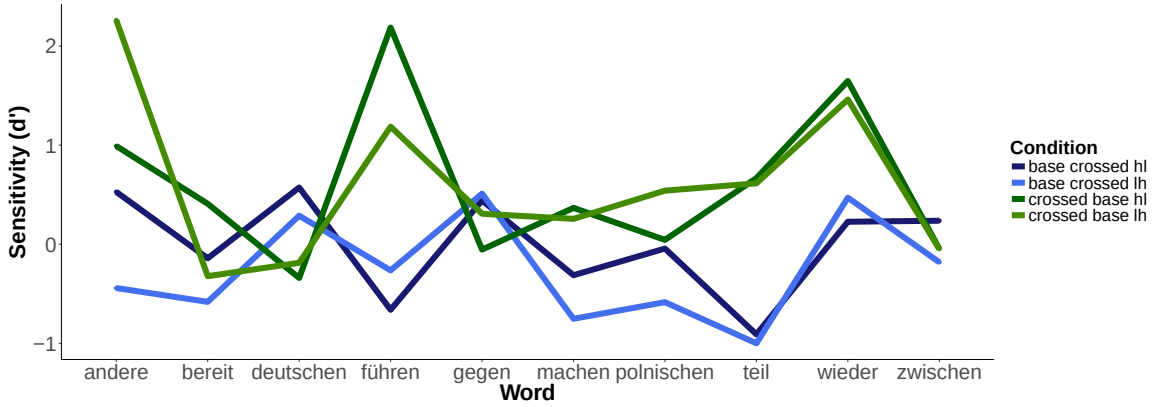


Figure 12.1: Sensitivity (d') of perceiving a difference between crossed and baseline stimuli per word.

less of the order of presentation or the surprisal condition. The biggest difference in participants' responses between same stimuli in different order of presentation were observed for the words "andere" (crossed – base lh = 2.26, base – crossed lh = -0.44) and "führen" (crossed – base hl = 2.19, base – crossed hl = -0.66). Participants were most sensitive to the difference between baseline and crossed condition for the word "andere" in crossed – base lh, while they were the least sensitive in discriminating baseline from crossed stimulus for the word "teil" in base – crossed lh (d' = -1.00) (Figure 12.1).

12.2.2 Linear mixed-effects model

In addition to the descriptive analysis of d' values, we ran a LMM with ORDER OF PRESENTATION and SURPRISAL CONDITION as fixed effects. Random effects included random intercepts for WORD and SPEAKER, as well as random slope for ORDER OF PRESENTATION and SURPRISAL CONDITION per WORD. Random slopes for the fixed effects per SPEAKER were excluded in the model selection process because they were perfectly correlated with the random intercept for SPEAKER (Model structure 12.1).

$$\begin{aligned} DPrime \sim\ & SurprisalCondition \\ & OrderOfPresentation + (1|Speaker) \\ & (1 + SurprisalCondition + OrderOfPresentation|Word) \end{aligned} \tag{12.1}$$

As results, we observed an expected significant effect of ORDER OF PRESENTATION. d' values for stimuli presented in the order of crossed – base were significantly higher than those for the order of base – crossed. SURPRISAL CONDITION was not significant in explaining variability in d' values. The LMM had a total effect size of 66.35 % of explained variability given by conditional pseudo-R^2 for both fixed and random effects. The fixed effects explained 20.99 % of variability in d' with ORDER OF PRESENTATION contributing 20.56 %.

Table 12.4: Sensitivity (d') model: regression coefficients, standard error (SE) and output of significance test.

Terms	Coeff.	SE	t-value	p-value
SurprisalCondition (lh-hl)	-0.12	0.12	-1.00	= .32
Order (crossed base - base crossed)	0.76	0.29	2.64	= .02

As expected from the descriptive analysis per speaker (Table 12.3), the random intercept for SPEAKER did not add explained variability to the model. The random intercept for WORD, however, was more informative in explaining differences in d' ($Var = 11.41\,\%$). This finding was expected based on the large amount of variability in sensitivity values that was observed in the by-item analysis (Figure 12.1).

High and low surprisal was defined per word token based on the occurrences and corresponding surprisal values in the Siemens Synthesis corpus. Some tokens showed larger differences between high and low surprisal condition than others (Table 12.1). Therefore, we tested the relationship between the distance between high and low surprisal condition and d' hypothesizing that higher surprisal differences coincided with higher sensitivity in perceiving a difference between crossed and baseline stimuli. We found a very weak tendency for a positive effect supporting our hypothesis. But the effect was not significant ($r = 0.03, t(78) = 0.28, p = 0.78$).

We have calculated participants' mean goodness rating per item and speaker. They used a rating between 1 (= sure) and 5 (= unsure). We assumed that participants showed higher sensitivity for differences between baseline and crossed condition when they were more confident in their decisions. Thus, a negative correlation between d' and goodness rating was expected (as 5 equals unsure). This assumption was confirmed. Sensitivity and goodness rating were significantly correlated ($r = -0.35, t(78) = -3.35, p = 0.001$).

12.2.3 Post-hoc analysis

Dynamic time warping

As a post-hoc analysis, we performed dynamic time warping (DTW) between the baseline stimuli as reference signals and the corresponding crossed stimuli used in the perception experiment. DTW is used to find an optimal alignment between two signals. This technique is typically used as an acoustic distance measures, e. g., in automatic speech recognition. Two time-dependent signals are warped in a non-linear fashion to match each other. DTW measures local cost of this time-alignment. Low values in DTW equal high similarity between acoustic signals, while high DTW values correspond to larger acoustic distances between two signals (Muda et al., 2010).

DTW calculation was performed using the inbuilt *dtw* function in SPTK 3.10 (Kobayashi et al., 2017). As preprocessing, .wav files were converted into .raw files

Table 12.5: Post-hoc analysis of perception test: DTW local cost values and MCD spectral distance between baseline and crossed signals per speaker, surprisal condition and word.

Word	Speaker ai				Speaker wo			
	high-low		low-high		high-low		low-high	
	DTW	MCD	DTW	MCD	DTW	MCD	DTW	MCD
machen	0.42	0.53	0.38	0.18	0.38	0.30	0.27	0.25
wieder	0.44	0.40	0.41	0.47	0.27	0.86	0.22	0.89
führen	0.36	0.21	0.35	0.49	0.37	0.25	0.36	0.50
zwischen	0.59	0.71	0.36	0.63	0.34	0.91	0.24	0.58
gegen	0.47	0.88	0.37	0.53	0.32	0.93	0.22	0.44
bereit	0.57	0.91	0.27	0.94	0.43	0.96	0.43	0.74
deutschen	0.32	0.45	0.31	0.76	0.33	0.22	0.25	0.59
andere	0.42	0.72	0.34	0.18	0.20	0.80	0.30	0.22
polnischen	0.62	0.89	0.32	0.44	0.44	0.59	0.32	0.48
teil	0.74	0.44	0.86	1.21	0.61	0.49	0.34	0.23

with a frame length of 320, and a frame period of 160. In a second step, mel-cepstral coefficients (.mcep) of the order 13, at frame length 512, and α of 0.42 were calculated from the .raw files. DTW calculation was performed on the .mcep files using a vector order of 13.

DTW values ranged between 0.20 and 0.86. Most values were, however, relatively low with a small standard deviation ($M = 0.40, SD = 0.14$). Crossed stimuli with a high surprisal context and a low surprisal word implemented (hl) had higher DTW values ($M = 0.43, SD = 0.13$) than crossed stimuli with low surprisal context and included high surprisal words ($M = 0.35, SD = 0.13$). We observed overall higher DTW acoustic distances between baseline and crossed signals for speaker *ai* ($M = 0.45, SD = 0.33$) than for speaker *wo* ($M = 0.33, SD = 0.10$) (Table 12.5).

We investigated whether low acoustic distances between baseline and crossed stimuli correlated with a lower sensitivity to perceive a difference in a discrimination task. Contrary to our hypothesis, we did not find a positive relationship between DTW and d', but a tendency for a negative correlation which was not significant ($r = -0.19, t(78) = -1.75, p = 0.08$). This non-significant relationship still held when we excluded negative d' values from the data set and ran the correlation test on a subset of items that were successfully discriminated by participants ($r = -0.17, t(40) = -1.12, p = 0.27$).

Mel-cepstral distortion

Since differences in time-alignment between baseline and crossed stimuli apparently were not related to participants' sensitivity in successfully discriminating both stim-

uli we ran an additional post-hoc analysis investigating potential spectral differences between baseline and respective crossed stimuli. We aimed to find that the magnitude of spectral differences correlated with participants' sensitivity to discriminate between baseline and crossed stimuli. We calculated mel-cepstral distortion (MCD) as a measure of acoustic spectral distance for the duration of the entire stimulus. Since comparison pairs differed in their duration (as shown in Section 12.2.3) they had to be time aligned prior to MCD calculation.

We used SPTK 3.10 for speech processing. Previously generated .raw files were transformed into MGCs with vector order of 13, a frame length of 512, an α value of 0.42, and a power parameter of 2. MCD distances between baseline and the respective crossed signals were calculated using the *cdist* function of SPTK with vector order of 13.

MCD values ranged between 0.18 and 1.21 with a mean of 0.58 ($SD = 0.27$). Crossed stimuli with a high surprisal context and a low surprisal word implemented (hl) did not only show higher DTW, but also MCD values ($M = 0.62, SD = 0.26$) than crossed stimuli with low surprisal context and included high surprisal words ($M = 0.54, SD = 0.27$). We also found the same tendency in higher acoustic distance for MCD for speaker *ai* ($M = 0.60, SD = 0.27$) than for speaker *wo* ($M = 0.56, SD = 0.26$) that was found in the previous DTW analysis (Table 12.5).

There was no significant correlation between MCD and d' ($r = -0.10, t(78) = -0.93, p = 0.36$). Running the correlation test on a subset of the data only including items with positive d' values replicated the non-significant relationship between MCD and d' ($r = -0.23, t(40) = -1.52, p = 0.14$). MCD and DTW showed, however, a weak positive correlation which was significant ($r = 0.28, t(78) = 2.56, p = 0.01$). Considering that the two measures were positively related it was not surprising to find the same non-significant relationship between both of them and d'.

12.3 Discussion

We tested whether listeners are sensitive towards violated ID expectations in the realization of specific target words. Stimuli were chosen from frequent words in the Siemens Synthesis corpus based on the distance in fourgram word surprisal of the preceding context. Words produced in a high surprisal context were cut and pasted into a low surprisal context, and vice versa. In a discrimination task we asked for the naturalness of the baseline compared to the crossed stimulus. In our experimental setup, we controlled for presentation of order, surprisal crossing condition, target word, and speaker. We found that listeners judged baseline stimuli to be more natural than crossed stimuli which violated ID expectations.

Although previous studies have shown that listeners are more sensitive to differences in phonetic detail when these occur in unpredictable compared to predictable contexts (Lieberman, 1963; Manker, 2017), so far there were no accounts of listeners'

sensitivity to violated ID expectations in phonetic detail. Therefore, this study added to the knowledge about speech perception and predictability.

However, we observed a time-order error (TOE). Participants only identified the crossed item successfully as the less natural one, when they first heard the crossed stimulus and then the baseline, and they failed to give accurate naturalness judgments in the opposite order of presentation. TOEs have been reported numerously for discrimination tasks in different domains (Schiefer and Batliner, 1991; Wherry, 1938; Wickelmaier and Choisel, 2006). As a positive (vs. negative) effect, it overestimates (vs. underestimates) the first relative to the second stimulus. The direction of the effect depends on numerous factors and cannot be predicted reliably (Hellström, 1985). Regarding our study, the TOE is interpreted as a positive effect on the first stimulus.

We tested for two cross-splicing conditions: high surprisal words produced in a low surprisal context, and low surprisal words in a high surprisal context. Based on the LMM analysis the surprisal condition had no significant influence on d' ratings. What was more informative in explaining differences in participants' sensitivity ratings was word identity. However, when we correlated surprisal difference between high and low condition with d' it was a non significant relationship with a very weak positive tendency in the expected direction.

Since the difference in surprisal was not informative in explaining item-related differences in d', we performed a post-hoc analysis on the crossed and baseline stimuli measuring their durational (DTW) and spectral distance (MCD). Neither one of the distance measures correlated positively with d' values indicating that participants did not rely on durational or spectral cues when giving their judgments. This finding mirrored to some extent the results reported in Le Maguer et al. (2016). There were no significant spectral distances between the baseline TTS synthesis system and the improved system including a descriptive predictability feature, but listeners showed a clear preference (72.5 %) for the updated speech synthesis system compared to the baseline in a forced AB preference task (Le Maguer et al., 2016). The authors of this study included additional distance measures, such as the root mean square error of F0 and duration, as well as voicing error rate, and were still not able to find significant differences between the two systems.

This chapter concludes the Part III of the thesis presenting all results from the production analyses and the perception experiment. All results are summarized in the following Table 12.6. In the following Part IV these results are discussed in more detail in relation to the ID variables and prosodic factors used in this thesis.

Table 12.6: Overview of the results of the thesis. Only significant results of the baseline models are reported. Arrows indicate the effect on the dependent variable: ⇑ equals increasing, ⇓ stands for decreasing. Results for biphone surprisal of the preceding (BiSur) and following context (BiFolSur), and triphone surprisal of the preceding (TriSur) and following context (TriFolSur) are given. DurAverage = average duration of segments, Wordfreq = word frequency, Wb = word boundary, Pb = phrase boundary, Syllfreq = syllable frequency, SentencePos = sentence position, PhStatus = phonemic status.

Corpus analysis	ID factors	Prosodic factors	Other controls
Segment duration	• TriSur ⇑	• Stress ⇑	• DurAverage ⇑
	• TriFolSur ⇑	• Pb ⇑	• Voicing ⇑
	• Wordfreq ⇓	• Wb ⇑	
		• Global tempo ⇓	
Segment deletion	• BiSur ⇓	• Stress ⇓	• Sound class ⇓
	• BiFolSur ⇓	• Pb ⇓	
	• Wordfreq ⇓	• Wb ⇓	
		• Global tempo ⇑	
VOT	• TriSur ⇑	• Stress ⇑	• Voicing ⇑
	• TriFolSur ⇑	• Global tempo ⇓	
		• Local tempo ⇓	
Vowel dispersion in German	• BiSur ⇑	• Stress ⇑	• DurAverage ⇑
	• Wordfreq ⇓	• Pb ⇓	
		• Wb ⇓	
		• Global tempo ⇓	
		• Local tempo ⇓	

Continued on next page

Table 12.6 Continued from previous page.

Corpus analysis	ID factors	Prosodic factors	Other controls
Vowel dispersion in 6 languages	• BiSur ⇑	• Normal – Fast ⇑	• /a/ – Mean ⇓
		• Slow – Fast ⇑	• /e/ – Mean ⇓
			• /i/ – Mean ⇑
Vowel dispersion of L2 speech			
L1 control group			• Tenseness ⇓
			• Word class ⇓
			• DurAverage ⇑
C2 speakers			• Tenseness ⇓
			• DurAverage ⇑
Dynamic formant trajectories			
VL	• Wordfreq ⇓	• Stress ⇑	• DurAverage ⇑
		• Wb ⇑	• PhStatus ⇓
		• Global tempo ⇓	
F1 slope	• BiSur ⇑	• Stress ⇑	• DurAverage ⇓
	• Wordfreq ⇓	• Wb ⇑	• PhStatus ⇓
		• Global tempo ⇓	
F2 slope	• BiSur ⇓	• Stress ⇑	• DurAverage ⇓
	• Wordfreq ⇓	• Wb ⇑	• PhStatus ⇓
VSL	• BiSur ⇑	• Stress ⇑	• DurAverage ⇑
	• Wordfreq ⇓	• Pb ⇓	• PhStatus ⇑

Continued on next page

Table 12.6 Continued from previous page.

Corpus analysis	ID factors	Prosodic factors	Other controls
		• Wb ⇓	
		• Local tempo ⇓	
F1 velocity	• BiSur ⇑	• Stress ⇑	• DurAverage ⇓
	• Wordfreq ⇓	• Wb ⇑	• PhStatus ⇓
		• Global tempo ⇓	
F2 DCT2	• BiSur ⇑	• Stress ⇑	• DurAverage ⇑
	• Wordfreq ⇓	• Pb ⇓	• PhStatus ⇓
		• Wb ⇓	
		• Global tempo ⇓	
		• Local tempo ⇑	
Spectral similarity	• BiSur ∗ ∗ ∗	• Stress ∗ ∗ ∗	
	• Syllfreq ∗ ∗ ∗	• Speech rate ∗ ∗ ∗	
Voice quality			
CPP	• BiSur ⇑	• Pb ⇓	• DurAverage ⇑
	• TriFolSur ⇑	• Wb ⇓	• SentencePos ⇓
		• Global tempo ⇓	
		• Local tempo ⇑	
CPPS	• BiSur ⇑	• Pb ⇓	• SentencePos ⇓
		• Wb ⇓	
		• Global tempo ⇓	
		• Local tempo ⇑	
OQs	• Wordfreq ⇑	• Stress ⇑	• DurAverage ⇓
		• Pb ⇑	• SentencePos ⇓
		• Global tempo ⇓	

Continued on next page

Table 12.6 Continued from previous page.

Corpus analysis	ID factors	Prosodic factors	Other controls
DEOPA		• Local tempo ⇓ • Stress ⇓ • Wb ⇑ • Global tempo ⇑ • Local tempo ⇓	• SentencePos ⇑
Perception experiment			• Order crossed – base ⇑

Part IV

Discussion

Structure Part IV: General discussion

All chapters in the preceding Part III contained interim discussions of the results presented in the respective chapter. In the following Part IV, we aim at discussing all results with regard to the ID and prosodic factors used in this thesis. We focus on the ID variables:

- n-phone surprisal
- Word frequency

In addition, we discuss the results regarding the prosodic factors:

- Stress
- Boundary
- Speech rate

In Section 13.3, we compare the effect sizes of the ID and prosodic factors in explaining segmental variability. We followed the hypothesis that prosodic factors are stronger predictors of segmental variability than ID factors. We conclude this Part of the thesis discussing the interactions between ID and prosody that significantly improved the performance of the baseline models. According to the SSR, effects of ID on phonetic structures are moderated through prosodic factors (Aylett and Turk, 2004, 2006).

Chapter 13

General discussion

The following part of this thesis presents a general discussion of the results of all analyses. It aims at giving a conclusive review of our findings in the light of the current state of the field. We discuss our results regarding the ID (surprisal, word frequency) and prosodic factors (stress, boundary, speech rate) used in this thesis in separate sections, in addition to a conclusive discussion of the effect sizes of the ID and prosodic model, as well as a discussion of the interaction between ID and prosody and its impact on the acoustic-phonetic measures that were the focus of this thesis.

13.1 Information density factors

We hypothesized that segments which are difficult to predict from the context are expanded in their spectral distinctiveness and increased in their durational features, while easily predictable segments are reduced spectrally and temporally. We tested this hypothesis conducting several production analyses of segment duration and deletion, VOT, vowel dispersion, dynamic formant trajectories, vocalic spectral distance, and voice quality. In addition, we tested listeners' sensitivity to violated ID expectations in a perception experiment.

As ID factors, we used surprisal values retrieved from n-phone LMs which were based on the preceding and following context of the segment, as well as word frequency and phoneme probability, both estimated on the LM corpus. Phoneme probability largely overlapped with surprisal estimates which was why we usually excluded this variable after our collinearity analyses. The following two sections discuss the results for surprisal and word frequency in more detail.

13.1.1 n-phone surprisal

Previous studies usually included predictability measured on the word level to explain segmental variability (e. g., Jurafsky, Bell, Gregory, et al., 2001; Tanner et al., 2017).

Only few studies have looked into predictability estimated on sub-lexical units (Aylett and Turk, 2004, 2006; Raymond et al., 2006). We argued that hierarchical structural information, such as syllable or word boundaries, which affect segmental properties are reflected in sequences of phones, especially if these sequences explicitly include word boundaries. This argument was supported by similar tendencies in Pearson's r correlation values between surprisal and vowel dispersion in German, irrespective of surprisal calculation based on a LM including or excluding word boundaries (Section 4.2.3). In addition, we followed Oh et al. (2015) in assuming that the relationship between ID and phonetic structure is best reflected by n-phone LMs: variability of local structures is supposed to be more accurately captured by local ID variables than by global ones.

***n*-phone size** We decided to use small n-phone orders as a result of an experimental analysis (Section 4.2.3). Small n-phones showed the highest Pearson's r correlation values with vowel dispersion in German, regardless whether the LM was built using word boundaries or not and whether the analysis was run on vowels from both content and function words or only on vowels from content words. Higher order of n-phone size led to a drop in the correlation values, while the "best" correlations were found for triphone surprisal and vowel dispersion.

One could argue that this result was due to the missing values for surprisal in higher order n-phones, or even due to the nature of the corpus that was used as training material for the LM. Both arguments do not hold based on our analyses. First, missing values for larger n-phones up to the size of 5 were relatively rare in the analysis of vowel dispersion in Bulgarian L2 speakers of German (Section 8.4). We still found the strongest correlations between triphone surprisal and vowel dispersion. Second, the German LM corpus for the analysis of vowel dispersion in six languages (Section 8.3) was updated from WebCelex (Max Planck Institute for Psycholinguistics, 2001), reported in Schulz et al. (2016), to Frankfurter Rundschau (Elsnet, 1992 – 1993) for this thesis. Correlation results differed only slightly ($r = 0.36$ vs. $r = 0.30$), and were therefore not a result of corpus-specific distributions of phoneme strings.

Context direction Traditionally, predictability is measured based on the preceding context of a linguistic unit which is based on the assumption that language processing functions incrementally (Aylett and Turk, 2006; Crocker et al., 2016; Roland et al., 2006). This approach has been challenged, to some extent, by studies which have shown that measures estimated on the following context of the linguistic unit also have an effect on linguistic variability (e. g., Hanique and Ernestus, 2011; Tanner et al., 2017). Especially in phonetics, both context directions impact the phonetic encoding of segments (Hillenbrand, Clark, et al., 2001; Moon and Lindblom, 1994). Therefore, we decided to use both surprisal of the preceding and following context in our production analyses. For the perception test, we limited ID estimations on the

preceding context because predictability based on the following context did not have a significant influence on listeners' sensitivity towards differences in phonetic detail (Manker, 2017).

Production analyses Regarding predictability based on n-phone surprisal we hypothesized that easily predictable segments show reduced duration (also in their VOT), higher deletion rates, reduced vowel dispersion and magnitude of formant change, as well as decreased periodicity and harmonic richness compared to segments that are difficult to predict.

We found the expected effect of triphone surprisal of the preceding and following context on segment duration and VOT. Duration values increased with increasing surprisal. In congruence with our hypothesis, German segments were less likely to delete when they stood in high surprisal contexts estimated from the preceding and following biphone. Also, vowel dispersion in German and in our cross-linguistic analysis of six languages increased significantly with increasing biphone surprisal of the preceding context. Surprisal was not significant, however, in our analysis of vowel dispersion in L2 speech. This was possibly due to the small number of different contexts of the vowels investigated which limited the range of n-phone surprisal values immensely compared to the analyses discussed above. The magnitude of formant change increased in vowels with high biphone surprisal of the preceding context in the measures F1 slope, F1 velocity, VSL, and F2 DCT2. In our analysis of spectral similarity, we found that biphone surprisal of the preceding context was a significant predictor, but it did not show the expected result of vowels in the same ID condition being more similar to each other than vowels in different ID contexts. Vowels showed an increase in harmonic richness in CPP and CPPS, when they were difficult to predict from the preceding biphone and following triphone (only for CPP).

We replicated previous results of the impact of predictability on duration, VOT, deletion rates, and vowel dispersion in studies predominantly performed with American English data for our German data. Easily predictable words or sub-lexical units were produced with shorter durations than units that were difficult to predict (Aylett and Turk, 2004; Bell et al., 2009; Jurafsky, Bell, Gregory, et al., 2001). We confirmed this finding for German segment duration. However, one should note that Polish segment durations were not affected by predictability expressed as surprisal (Malisz et al., 2018). The relation between duration and predictability can therefore be interpreted as language-dependent. Consonants were more prone to lenition when they were easily predictable from their context (Buz, Jaeger, and Tanenhaus, 2014; Cohen Priva, 2017; Manker, 2017) which was also supported by our findings for German read speech. The same surprisal values as in the segment duration analysis were predictive of VOT in German plosives. High predictability of segments led to higher likelihood of deletion (Buz, Jaeger, and Tanenhaus, 2014; Cohen Priva, 2017; Manker, 2017) which we confirmed for /t/ and /ə/ deletion in German. Easily predictable vowels in American English were less dispersed than vowels which were difficult to predict

(Aylett and Turk, 2006; Clopper and Pierrehumbert, 2008; Gahl et al., 2012; Jurafsky, Bell, Gregory, et al., 2001). We saw the same effect in our analysis of German read speech, and also for most of the six languages included in our cross-linguistic analysis.

In addition to replicating previous results, we also analyzed dynamic formant and voice quality measures in the context of ID and prosody in German. To the author's knowledge, these analyses were first in shedding light on the relationship between ID and formant trajectories and voice quality. F1 showed more magnitude in formant movement, F2 more curvature, and overall formant movement of F2 and F1 increased in unpredictable vowels compared to predictable ones. These effects were found in addition to known effects of prosodic factors on dynamic formant trajectories (Gay, 1978; Weismer and Berry, 2003).

Surprisal was predictive of voice quality as well. The cepstral voice quality metrics used in this thesis, CPP and CPPS, both increased under high surprisal indicating increased cepstral peakedness and harmonic richness in vowels that were difficult to predict from the context. This finding was in line with studies on cognitive load and voice quality, and cognitive load and surprisal: speakers produced less breathiness, visible in higher CPP and CPPS values, during a cognitively demanding task than under low cognitive load (Yap et al., 2011). Surprisal values were found to correlate positively with cognitive load (Demberg and Keller, 2008; Smith and Levy, 2013).

For four of the acoustic-phonetic measures, we found significant effects of surprisal of the following context in addition to surprisal of the preceding context in the same statistical model. Since both factors were not related, we chose to include them both in the same model to increase model performance. For segment duration, VOT, and CPP, triphone surprisal of the following context was a significant predictor. Biphone surprisal of the following context was significant in the segment and the /ə/ deletion models. To summarize, only durational measures or measures which depend strongly on duration, e. g., CPP, were affected by surprisal of the following context in addition to surprisal of the preceding context. The same pattern seems to be apparent in the literature: word durations (Bell et al., 2009; Gahl et al., 2012; Jurafsky, Bell, Gregory, et al., 2001; Tily et al., 2009) and coronal stop deletion rates (Raymond et al., 2006; Tanner et al., 2017) were affected by predictability calculated both based on preceding and following context. In studies focusing on spectral characteristics, only effects of ID estimated on the preceding context were reported, except for Jurafsky, Bell, Gregory, et al. (2001). Vowel reduction in American English function words were also predicted by the bigram of the following word. One should keep in mind that in this particular study vowel reduction was not estimated spectrally, but was based on narrow transcriptions. Also, this finding was only conclusive for function words, not for all words.

Perception experiment In the perception experiment (Chapter 12), we cross-spliced words from high ID contexts into low ID contexts, and vice versa. We pre-

sented these cross-spliced phrases alongside their respective baselines and asked listeners which one of the two recordings they perceived as more natural. Here, ID was measured estimated on a word LM on the preceding fourgram surprisal of the word. We found that listeners were, indeed, sensitive towards violated expectations of ID in phonetic encoding, but only when the items were presented in the order crossed – base.

Our post-hoc temporal and spectral distance analysis between cross-spliced items and their baselines revealed no significant differences. Similarly, Le Maguer et al. (2016) did not find significant durational nor spectral differences between their baseline TTS system and an updated version incorporating predictability values, while listeners preferred the updated system. Other studies have also observed that listeners were able to detect subtle differences between signals, and that they performed even better in distinguishing these differences when the targets were unpredictable from the context (Beaver et al., 2007; Manker, 2017).

Relation to phonotactic structure One of the main concerns regarding n-phone surprisal as a variable of predictability is that it mirrors phonotactic structure, especially if surprisal is calculated based on small n-phone sizes, as in this thesis. In order to control for this confound, we included preceding and following phonological context in the random structure of all the models, if at all possible due to convergence errors. Although these controls were included, we found significant effects of surprisal. This means that predictability of n-phone combinations still added information to explaining segmental variability, while contextual information were accounted for in the model.

The only exceptions to this are the deletion models for /ə/ and /t/ (Sections 6.2.2 and 6.2.3). Here, phonological context was included as a fixed effect in the models, either in the form of a normative deletion rule (i.e., Duden rule for /ə/ deletion) or as following context defined in relation to the phone under investigation (i.e., following context for /t/ deletion). Both predictors were the most effective fixed effect in their respective models, and led to non-significant results for biphone surprisal of the following context in the /t/ deletion model, and biphone surprisal of the preceding context in the /ə/ deletion model. The effect of surprisal was subsumed by the strong fixed effect for phonological context. This observation, however, was unique to segment deletion.

Relation to stress Surprisal and primary lexical stress were weakly related. Segments in syllables carrying lexical stress were more difficult to predict than segments in unstressed syllables. Since the relation was only weak, we included both factors in the statistical models, and also found significant effects for both. Research on focus marking (Turnbull et al., 2015) and intonation structure (Kakouros and Räsänen, 2016) has also found a connection between unpredictability and prosodic marking. In addition, the interaction between stress and predictability was effective in explaining

segmental variability in this thesis and previous studies (Aylett and Turk, 2006) (see Discussion 13.4).

In summary, surprisal had the expected effect on temporal and spectral characteristics of segments in the majority of our production analyses. While there is evidence from studies on (American) English for predictability effects on duration (Aylett and Turk, 2004; Bell et al., 2009; Jurafsky, Bell, Gregory, et al., 2001), VOT (Buz, Jaeger, and Tanenhaus, 2014; Cohen Priva, 2017; Manker, 2017), deletion rates (Cohen Priva, 2015; Raymond et al., 2006; Tanner et al., 2017), and vowel dispersion (Aylett and Turk, 2006; Clopper and Pierrehumbert, 2008; Gahl et al., 2012; Jurafsky, Bell, Gregory, et al., 2001), we replicated these results for German. Additionally, we broadened the scope of acoustic-phonetic measures to dynamic formant trajectories and voice quality metrics. For those, we could also find expected effects of surprisal for our German data.

13.1.2 Word frequency

We decided to include word frequency as an additional ID variable in our models because it is a well established predictor of segmental variability (e. g., Jurafsky, Bell, Gregory, et al., 2001; Pluymaekers et al., 2005b). In addition, we aimed at broadening the spectrum of ID variables from the sub-lexical level, i. e., phonemes, to the lexical level. Word frequency and n-phone surprisal values were largely independent and could therefore be entered in the same statistical model.

Word frequency was used in almost all statistical models of the production analyses, except for vowel dispersion in six languages and L2 speakers, as well as VOT. In these three studies, we included function and content words, and controlled for the factor word class in the two latter analyses. Since word class and word frequency are highly correlated, word frequency was excluded from these analyses.

We found significant effects of word frequency on almost all acoustic-phonetic measures, when this factor was included. One exception to this pattern was observed for MCD between vowels in same and different ID conditions (Section 10). Including random intercepts for word identity in the statistical model led to non-significant results for unigram word probability and corpus-specific word frequency. Word frequency apparently could not add to explaining differences in the spectral characteristics of vowels above and beyond the word identities from which the vowels were extracted. This finding, however, was very specific to a comparative spectral analysis since we included word identity as a random intercept in all other statistical models in this thesis, provided the model converged, and still found significant effects of word frequency.

German segments in high-frequency words were significantly longer and less likely to delete than segments in low-frequency words, as previously reported for American English words, syllables and segments (Aylett and Turk, 2006; Bell et al., 2009; Cohen

Priva, 2015; Gahl et al., 2012; Jurafsky, Bell, Gregory, et al., 2001; Zipf, 1949). As expected, vowels in high-frequency words were more dispersed (Gahl et al., 2012; Munson and Solomon, 2004; Munson, 2007; Pierrehumbert, 2000; Scarborough, 2006; Wright, 2004) and dynamic in the amount of formant change in a F1/F2 plane (VL, VSL), in absolute F1/F2 slope and F1 velocity, and showed increased curvature in F2 (F2 DCT2) than vowels in low-frequency words. In the voice quality analysis, only OQs was significantly effected by word frequency. On average, high-frequency words contained vowels with increased OQs, i. e., reduced harmonic richness, compared to vowels in low-frequency words.

To summarize, we confirmed our hypotheses regarding the effect of word frequency for segment duration and deletion, as well as for vowel dispersion and dynamic formant trajectories in German. Word frequency was neither included in the VOT, L2 speech nor the cross-linguistic analysis of vowel dispersion. In addition, we did not find expected significant effects of word frequency on most of the voice quality parameters, when function words were excluded from the data set. Word identity overruled word frequency effects in the analysis of vocalic spectral distances in different ID contexts.

13.2 Prosodic factors

Following Aylett and Turk (2006), we used stress and prosodic boundary as prosodic factors in the production analyses. In addition, we calculated articulation rate on the sentence level and on the target word. The following sections discuss the results of the production analyses regarding the prosodic factors stress, boundary and speech rate.

13.2.1 Stress

Stress was defined as a binary factor based on primary lexical stress. Monosyllabic function words were coded as unstressed, while monosyllabic content words were marked as stressed. Stress was included in all production studies of this thesis, except for the analysis of vowel dispersion in L2 speakers of German, and the analysis of /ə/ deletion. In both studies, vowels were either always in stressed position (L2 speech) or always unstressed (/ə/ deletion) which was why the factor was uninformative here.

We confirmed our initial hypothesis that stressed segments show increased duration (also in VOT), as reported in Aylett and Turk (2004), lower deletion rates (Cohen Priva, 2015), increased vowel dispersion (Aylett and Turk, 2006), and magnitude of formant change in German. Stress had a significant positive effect on all measures of formant change (Chapter 9). This was expected based on the lengthening effect of stress and increase in vowel dispersion for stressed vowels (Gahl et al., 2012; Malisz et al., 2018). Also, we saw expected results for stress in the analysis of vocalic spectral distance: vowels in both unstressed position were the most distant from each other

compared to all other stress conditions in the model, while vowels in both stressed condition were the most similar in MCD. This finding was related to larger variability in unstressed German vowels because unstressed syllables are produced with a higher degree of coarticulation (Mooshammer and Geng, 2008).

But there were also unexpected and non-significant effects of stress. Regarding the voice quality metrics, stress was non-significant in the cepstral measures, but significant in the EGG metrics. In American English, prominence was reported to have a strong effect on the voice quality of the entire word, and not only on the stressed syllable (Epstein, 2002). This phenomenon can explain a null effect of stress on the cepstral measures because we only investigated content words that carry lexical stress on one of their syllables. For the EGG metrics, we found an initially unexpected positive effect of stress on OQs, i. e., decreasing harmonic richness, and an expected negative effect of stress on DEOPA. We explained these findings by the effect of stress on the glottal excitation: stressed vowels are often produced with increased F0 (Lieberman, 1960) which presumably led to longer OQs in our data. Stress is also associated with increased muscular activity and vocal effort, and thus increased DEOPA values (Michaud, 2004).

In the cross-linguistic analysis of vowel dispersion, stress was non-significant (Section 8.3). This finding was in contrast to Schulz et al. (2016) who analyzed vowel dispersion in the BonnTempo corpus (Dellwo et al., 2004) using five languages (DEU, CES, POL, FIN, and FRA), while this study also included AE. The authors found a significant effect of stress across languages, but a non-significant effect for surprisal. This difference in findings might be due to the weak positive correlation between surprisal and stress ($r = 0.23$), and also due to the difference in the data set between the two analyses. It should be noted that the definition of stress across languages was not uniform. While we coded stress on the word level for AE, CES, DEU, FIN, POL, in FRA accent was marked on the last syllable of a phrase with a full vowel following Jun and Fougeron (2000).

In summary, prominence coded as a binary factor based on primary lexical stress proved to be a robust predictor of segmental variability in German in the durational and spectral domain. In addition, stress was an informative factor in the EGG voice quality analysis. When included in a cross-linguistic analysis with different definitions of stress and confounding variables, the factor did not reach significance level.

13.2.2 Boundary

Prosodic boundary was defined using three factor levels following Aylett and Turk (2006):

- No following prosodic boundary
- Following word boundary

- Following phrase boundary

It was not included in the VOT analysis because none of the consonants preceded a prosodic boundary, they all stood in word-initial position.

In accordance with our initial hypothesis, we observed that German segments immediately preceding a word or phrase boundary showed increased duration (also in VOT) (Turk, 2010; Wheeldon and Lahiri, 1997), and were less likely to delete due to the high number of /ə/ deletions in the main model which were not immediately preceding a boundary. Boundary was neither included in the /t/ nor in the /ə/ deletion model due to convergence errors.

We did not find an expected increasing effect of boundary on German vowel dispersion (Section 8.2) which was assumed because of the positive correlation between vowel dispersion and duration, and the known effect of prosodic lengthening (Wheeldon and Lahiri, 1997). In fact, vowels at word and phrase boundary were less dispersed than at no boundary position. This result was also in disagreement with Turk (2010) who claimed that there is a positive relation between prosodic boundary and acoustic salience which reduces coarticulatory effects at boundary position. However, for American English, E.-K. Lee and Cole (2006) found that spectral strengthening effects and durational lengthening effects of prosodic boundaries on vowels were independent from each other, and that the effect of spectral strengthening at boundary position depended on the vowel phoneme.

We found the same tendency of a negative effect of boundary in the analysis of vowel dispersion in six languages (Section 8.3). In addition to possible explanations given above this effect was also non-significant because of the cross-linguistic nature of the analysis. Apparently, not all languages showed the same tendency for reduced vowel dispersion at boundary position. In fact, Finnish vowel reduction is weakly expressed in the spectral domain due to the vowel harmony in that language (Bertram et al., 2004). Similarly, Polish vowel reduction is weakly pronounced spectrally, and highly depends on the presence of palatalized or palatal consonants (Nowak, 2006).

We expected a positive effect of boundary on formant movement, again motivated by the lengthening effect of prosodic boundaries on preceding segments (Byrd, 2000; Kohler, 1988), and the positive correlation between duration and formant change (Fox and Jacewicz, 2009). We confirmed this hypothesis with regard to some of the dynamic formant metrics: VL, F1/F2 slope, and F1 velocity significantly increased at word boundary compared to no boundary. VSL and F2 DCT2, on the other hand, decreased in vowels preceding a word or phrase boundary compared to no boundary position. These results were in line with Tabain (2003) who found that formant velocity and rate of change increased at weaker boundaries in French /aC/ sequences.

Regarding voice quality (Chapter 11) we expected to find decreased periodicity and harmonic richness at phrase boundary position compared to no boundary based on findings that creaky voice often occurred at phrase-final position (Ferrer et al., 2002; Henton and Bladon, 1988). We confirmed this hypothesis for both CPP and

CPPS, not only for phrase boundary, but also for word boundary. Differences in voice quality were systematically affected by boundary position. This was also apparent in the EGG metrics. Here, OQs increased at phrase boundary and DEOPA at word boundary compared to no boundary position indicating that the harmonic structure of the vowels was less rich at these positions.

We decided to include prosodic boundary factors of different hierarchical status (word vs. phrase) in our model in order to investigate potential differences in strength between them. There is some evidence that the effect of prosodic lengthening cumulates with successively higher order of prosodic boundary (Tabain, 2003; Turk, 2010). We therefore hypothesized that boundary effects were more pronounced at higher-level boundary (phrase level) than at a low-level boundary (word level). For those analyses with both significant effects of word and phrase boundary on the acoustic-phonetic measure, we confirmed this hypothesis, even though the direction of the effect was not always as initially expected (see above). Based on the estimates of the model outputs, segment duration was more lengthened at phrase boundary than at word boundary and segment deletion was less likely at phrase than at word boundary. The decreasing effect of boundary on CPPS, vowel dispersion, F2 DCT2, and VSL was more pronounced at phrase boundary than at word boundary.

To summarize, prosodic boundary had the expected lengthening effect on segment durations, but did not necessarily increase distinctiveness in the spectral domain. These findings can be explained by independent effects of boundary on lengthening and spectral strengthening (E.-K. Lee and Cole, 2006). When both phrase and word boundary had a significant effect on an acoustic-phonetic measure, we saw a stronger effect at the phrase boundary level compared to the word boundary which supported the idea of cumulative effects across prosodic constituents of different hierarchy levels.

13.2.3 Speech rate

We measured speech rate as articulation rate (produced phonemes per second) excluding pauses. Speech rate was estimated at the sentence (global) and word (local) level. The factor was entered as a continuous variable in our models, except for the cross-linguistic analysis of vowel dispersion (Section 8.3). Here, intended speech rate variations (normal, fast, slow) were used and compared. We used binned continuous speech rates (normal, fast, slow) in order to compare vocalic spectral distances in different speech rate conditions in the analysis presented in Chapter 10.

Speech rate acceleration was expected to have a decreasing effect on segment durations (Bell et al., 2009), vowel dispersion (Weiss, 2007), magnitude of formant change (Fox and Jacewicz, 2009), and harmonic richness (Rietveld and Gussenhoven, 1987), as well as lead to higher deletion rates (Cohen Priva, 2015). Accelerated global speech rate showed this expected decreasing effect on segment duration, VL, F1 slope and velocity, F2 DCT2, and also led to higher segment deletion rates. Both

local and global speech rate had the assumed decreasing impact on VOT, harmonic richness expressed as OQs, and vowel dispersion. Vowels at intended slow speech rate were more dispersed than at intended fast and normal speech rate. For VSL, we only found an expected decreasing effect of local tempo, but not for global speech rate. Inconclusive results were observed for F2 DCT2, CPP, CPPS, and DEOPA. Here, global speech rate had the expected effects on the measures, while local speech rate did not. For these cases, we assumed that local effects of speech rate deviation caused these unexpected effects, while we still observed expected tendencies in global measures in tempo. Global speech rate did not overrule those local effects which was in line with our initial hypothesis.

Based on the estimates of the LMM output we saw the same strength of effect of global and local speech rate both in the analysis of vowel dispersion and of VOT in German. In both models, the estimates for local speech rate showed less variation in standard error than for global speech rate. For OQs, the effect of global speech rate was stronger than that of local speech rate.

Regarding the results for speech rate in the analysis of spectral similarity of vowels (Chapter 10), we observed that vocalic spectral distance expressed as MCD between same vowel identities in contrasting conditions, such as slow – fast, was more pronounced than in non-contrasting conditions, such as slow – slow or fast – fast. This was contrary to our initial hypothesis assuming that fast – fast vowels were the most distant from each other. For speech rate, we found tendencies for a hierarchy of contrasting conditions showing more spectral distance than non-contrasting conditions. It was only a tendency because not all comparisons between these groups were significant and the two groups were not homogeneous.

We used intended speech rates (normal, fast, slow) in the cross-linguistic vowel dispersion analysis (Section 8.3) because speakers produced the same linguistic material at different intended speech rates. We hypothesized to observe a stable relation between vowel dispersion and ID across all intended speech rates. This assumption was based on the inverse relationship between speech rate and ID which was reported across several languages from different language families (Oh, 2015; Pellegrino et al., 2011). Our hypothesis was confirmed for all languages that showed a positive relationship between ID and vowel dispersion, i. e., all languages except Finnish: vowel dispersion increased with higher surprisal, irrespective of the intended speech rate.

In summary, we found that speech rate acceleration led to an expected decrease in temporal and spectral segmental features. The global estimate of speech rate showed expected effects on the acoustic-phonetic measures, while local speech rate effects showed unexpected results in the case of F2 DCT2, CPP, CPPS, and DEOPA. Intended speech rate deviations did not corrupt the relationship between segmental variability and ID in the case of vowel dispersion.

Table 13.1: Segmental variability: explained data variance in % of ID and prosodic model. Non-significant (n.s.) and non applicable effects (n.a.) are marked as such.

Measure	ID factors		Prosodic factors		
	Word frequency	Surprisal	Stress	Boundary	Speech rate
Duration	0.19	0.16	0.16	4.15	0.67
Deletion					
/t/ deletion	n.a.	3.00	n.s.	n.a.	0.50
/ə/ deletion	n.a.	2.00	n.a.	n.a.	n.s.
VOT	n.a.	1.32	0.14	n.a.	1.01
Vowel dispersion					
German	0.20	0.12	3.60	0.31	0.30
Six languages	n.a.	2.24	n.s.	n.s.	0.24
Dynamic trajectories					
VL	0.17	0.08	0.01	0.09	0.19
F1 slope	0.03	0.47	0.86	0.12	0.40
F2 slope	0.002	n.a.	0.34	0.005	n.s.
VSL	0.14	1.75	2.68	3.12	0.32
F1 velocity	0.01	0.001	0.42	0.07	0.07
F2 DCT2	0.20	1.99	2.25	0.002	0.14
Voice quality					
CPP	n.s.	1.00	0.04	0.21	0.43
CPPS	n.s.	3.00	n.s.	0.40	0.21
OQs	0.11	n.s.	0.14	n.s.	0.03
DEOPA	n.s.	n.s.	0.07	0.10	0.38

13.3 Effect sizes

In accordance with Aylett and Turk (2006), we expected to find small, but robust effects of ID on the acoustic-phonetic measures investigated in the production analyses. The prosodic factors, especially primary lexical stress, were assumed to have a stronger effect on these measures compared to the ID effect.

We confirmed this hypothesis for almost all measures, except for CPP and CPPS, /t/ and /ə/ deletion, as well as the vowel dispersion analysis in six languages (Table 13.1[1]). Surprisal was more effective in explaining variability in cepstral voice quality

[1]The analyses vowel dispersion in L2 speakers of German and spectral similarity of vowels are not included in the Table since their models were not constructed using comparable or identical factors.

metrics than all of the prosodic factors combined. In the /t/ deletion model, we did not include boundary as a factor, stress was non-significant and the effect of speech rate on /t/ deletion rates was very small. For these reasons, we found a surprisal to be a stronger predictor than the prosodic factors. Neither stress nor boundary were applicable in the /ə/ deletion analysis, while speech rate was non-significant. We therefore found surprisal to be more effective in this model than the prosodic factors. Surprisal was also stronger in predicting vowel dispersion in the cross-linguistic analysis than stress, boundary, and speech rate combined. Since surprisal was estimated on language-specific LMs it was an effective predictor of segmental variability across languages, while we used a uniform definition of the prosodic factors for all languages, when it was not necessarily always adequate, e. g., in the case of French stress definition.

Prosodic factors were more effective in explaining segmental variability in duration, vowel dispersion and dynamic formant trajectories in German, as well as in the EGG metrics of voice quality. In the VOT and the VL analyses, the ID and prosody model were similar in their strength. Thus, our results were, by and large, in line with previous accounts of effect sizes of prosodic and ID factors in models of segmental variability (Aylett and Turk, 2006).

13.4 Interaction between information density and prosody

According to the SSR hypothesis, the effect of language redundancy, or ID, is moderated by prosodic structure (Aylett and Turk, 2004, 2006). Therefore, we tested interactions between surprisal as our measure of ID and all prosodic factors in the model (stress, boundary, and speech rate) expecting to find that the interaction models outperformed the baseline additive models.

Contrary to our expectation, the interaction models did not always perform significantly better than the additive models. For F1 velocity, F1 slope, /ə/ deletion, /t/ deletion, vowel dispersion in six languages, and OQs, none of the tested interaction models yielded better model performance than the baseline models.

Surprisal * Stress Surprisal often interacted positively with stress on its effect on the measures investigated here. One should also keep in mind that stress and surprisal were weakly positive correlated which we found as a constant result in our collinearity analyses. Surprisal and stress complemented each other positively in their increasing effect on segment duration, vowel dispersion, VL, VSL, F2 DCT2, CPP and CPPS, as well as led to expected lower deletion rates and DEOPA values. The interaction between stress and surprisal neither improved model performance for F2 slope nor in the VOT model. These findings supported the SSR hypothesis which highlights

that prosody, in particular prominence expressed as primary lexical stress, moderates the effects of language redundancy on phonetic structures. The strong dependency between those two factors was also reflected in our results.

Surprisal * Boundary The dominating effect of an interaction between surprisal and boundary was in the same direction as the main effect of boundary. In some cases, there was a change in sign for the interaction compared to the main effect. For instance, triphone surprisal of the preceding context interacted negatively with word and phrase boundary in their effect on segment duration offering the conclusion that segments under high surprisal and at boundary position showed decreased duration values. This effect, of course, was contrary to the main effects, and after detailed visual inspection of interaction plots, we found that surprisal and boundary complemented each other positively, except for the significant interaction between triphone surprisal of the preceding context and word boundary. In this model, however, it was difficult to interpret the interaction term because of moderate to strong correlations with the respective main effects.

The main effects of surprisal and boundary showed contrary directions in the German vowel dispersion model (Section 8.2). The interaction between both terms showed that vowels under high surprisal at phrase boundary were more dispersed than vowels under low surprisal at no boundary position, i.e., prosodic boundaries only increased acoustic salience of the vowel when it appeared in a high surprisal context.

All other interactions of prosodic boundary and surprisal that significantly improved model performance had the same direction as the main effect of boundary. Segments at phrase boundary under high triphone surprisal of the following context showed lower durations. The interaction between phrase boundary and surprisal led to expected higher values in VL and F2 slope. Vowels at word boundary under high surprisal had decreased dispersion, CPP and CPPS values. In addition, the interaction between phrase boundary and surprisal also had a decreasing effect on CPPS.

Surprisal * Speech rate Since surprisal is a local predictability measure, it is not surprising that we mostly found model performance improvements when including an interaction for local speech rate and surprisal, and only in some rare cases also for global speech rate and surprisal. The interaction between local speech rate and surprisal had the same negative direction as the main effect of local speech rate on the measures vowel dispersion in German, VSL, and VOT. Triphone surprisal of the following context and local speech rate interacted positively on CPPS, meaning that these vowels showed increased harmonic richness. This result was in line with both main effects. For F2 DCT2, on the other hand, we found a negative effect of the interaction which was contrary to the main effect of local speech rate, but in line with initial expectations about the effect of speech rate on F2 DCT2: at high speech rate, we expected less curvature in F2. This expectation seemed to be confirmed only for vowels in high surprisal contexts. Regarding the interaction between global speech

rate and surprisal, we observed a negative effect on VOT in line with the main effect of global tempo in that model, as well as a positive effect on CPPS contrary to the main effect of global tempo.

In previous acoustic-phonetic analyses, we also observed non-significant main effects of surprisal but significant effects of surprisal in interaction with prosodic factors in the expected directions. Interactions between stress and surprisal, as well as between phrase boundary and surprisal led to an expected increase in VL, despite a non-significant surprisal effect. Also, the interaction term between surprisal and stress improved the DEOPA model significantly. The effect was negative, as expected, indicating higher harmonic richness for vowels under high surprisal and stress compared to unstressed vowels in low surprisal contexts. There was a positive interaction effect between stress and biphone surprisal of the preceding context on F2 slope, as well as between stress and triphone surprisal of the following context on segment duration.

To sum up, interaction models were often more effective than additive models in explaining segmental variability. Positive interactions between stress and surprisal were most frequent. Surprisal and boundary usually complemented each other in the same direction as the boundary effect. Local speech rate and surprisal interacted in the expected directions, even when the main effect of local speech rate did not show expected tendencies (e. g., for F2 DCT2).

We discussed the results of this thesis with regard to the ID and prosodic factors that were utilized. Main objectives of this thesis were to investigate the relative effect sizes and interaction effects of ID and prosody in explaining segmental variability. These research goals were discussed in separate sections. The following Part V concludes the thesis with a final summary.

Part V

Conclusion

13.5 Conclusion

This thesis investigated segmental variability as a function of ID and prosodic structure. We mainly focused on German read speech in our analyses. Previous studies found that phonetic structures lengthen in durational features and strengthen in their spectral features when they are difficult to predict from their context, whereas easily predictable phonetic structures are reduced spectrally and shortened. While most of these studies focused on American English (e. g., Aylett and Turk, 2006; Gahl et al., 2012; Jurafsky, Bell, Gregory, et al., 2001), with a few exceptions including Dutch, Russian or Finnish (van Son, Bolotova, et al., 2004), we broadened the scope of the field by investigating German data, as well as by our cross-linguistic analysis including Finnish, French, Czech and Polish in addition to German and American English. As ID factors, we used word frequency and n-phone surprisal calculated from LMs. As prosodic factors, primary lexical stress, prosodic boundary, and speech rate were included in the statistical models. For these factors, the main findings of the production analyses were:

- Surprisal
  - Easily predictable segments in German were shorter in duration and in their VOT, and were more likely to delete than segments that were difficult to predict.
  - Easily predictable German vowels were less dispersed, showed less formant change in their VSL, F1 slope and velocity, were less curved in their F2, and showed increased breathiness values in CPP and CPPS than German vowels that were difficult to predict from their context.
- Word frequency
  - German segments in high-frequency words were shorter in their overall duration, and were more likely to delete than segments in low-frequency words.
  - German vowels in high-frequency words were less dispersed, and showed less magnitude in formant change in the metrics VL, VSL, F1/F2 slope, and F1 velocity, less curvature in their F2 DCT2, as well as less harmonic richness in OQs than German vowels in low-frequency words.
- Stress
  - German segments in stressed syllables were longer in duration (also in their VOT), less likely to delete, more dispersed in their spectral characteristics, showed increased formant change (VL, VSL, F1/F2 slope, F1 velocity) and formant curvature in F2, longer OQs and more pronounced DEOPA values than segments in unstressed syllables.

- Boundary
  - German segments immediately preceding a prosodic boundary were longer in duration, less likely to delete, and showed more formant change in VL, F1/F2 slope, F1 velocity.
  - Boundary position led to reduced vowel dispersion, VSL, F2 DCT2, as well as decreased values in CPP and CPPS, and increased values in OQs and DEOPA indicating reduced periodicity compared to no boundary position.
- Speech rate
  - Speech rate acceleration led to reduced segment and VOT durations, higher segment deletion rates, reduced vowel dispersion and formant change in VL, VSL, F1 slope and velocity, and decreased OQs values.
  - There were opposing results of the effect of speech rate for global and local speech rate for F2 DCT2, CPP, CPPS, and DEOPA. The effect of global speech rate was expected for F2 DCT2, CPP, CPPS: vowels were less distinct in their spectral characteristics at fast global speech than at slow speech. For DEOPA, we found an expected effect for local speech rate: here, more distinct peaks at faster local speech rate can be explained by lower average F0 at increased speech rate.

In most cases, ID factors were less effective in explaining segmental variability than prosodic factors. This finding was in line with previous accounts of the impact of ID factors, such as frequency or predictability, on phonetic variability (e. g., Aylett and Turk, 2006).

If interactions of surprisal and prosodic factors improved model performance, the effect of the interaction was usually in the direction of the main effect of the prosodic factor. Stress and surprisal interacted positively and increased model performance in almost all production analyses, except for F2 slope and VOT. For instance, stressed vowels under high surprisal showed more vowel dispersion than unstressed vowels at low surprisal. Surprisal and prosodic boundary also interacted. For instance, segments preceding a prosodic boundary that stood in high surprisal contexts showed longer segment durations than segments at no boundary position in low surprisal contexts. Regarding interactions between speech rate and surprisal we mostly found model performance improvements when including an interaction for local speech rate and surprisal. This is not surprising considering that surprisal is a local predictability measure. For instance, plosives at high surprisal and accelerated local speech rate showed shorter VOTs than plosives in low surprisal context at slow speech rate.

We also calculated the spectral distance between same vowel identities in different ID and prosodic conditions. We found that surprisal and corpus-specific syllable frequency were significant in explaining spectral distance. We expected vowels to be

more similar when they stood in the same ID condition than vowels in different ID conditions. This hypothesis was confirmed for syllable frequency. Stress and speech rate were also significant predictors of spectral similarity. Increased variability in vocalic spectral characteristics induced by unstressed lexical position or accelerated speech rate led to expected results for these two factors in this analysis. Vowels in stressed syllables were less distant from each other than vowels in unstressed syllables, and when vowels in unstressed-stressed condition were compared. Non-contrasting speech rate conditions showed a clear hierarchy of spectral distance with lowest values for slow – slow, followed by normal – normal and than largest differences between vowels in sentences which were both produced at fast speech rate.

Effects of ID on vowel dispersion were robust across almost all of the six languages and the three intended speech rates (normal, fast, slow) that were included in the cross-linguistic analysis (Section 8.3). These findings were in line with previous cross-linguistic accounts which reported that there are universal tendencies across languages to reduce phonetic structures when they are easily predictable from the context (Pellegrino et al., 2011). Languages which do not have spectral, e. g., Finnish, or temporal reduction, e. g., Polish, did not show an ID effect on these phonetic variables (Malisz et al., 2018).

We did not only investigate the production of segmental variability as a function of ID and prosody, but also its perception. In a cross-splicing experiment, listeners were asked whether they perceived the manipulated or the baseline phrase as more natural. Manipulation entailed using high ID word productions in a low ID context, and vice versa. Participants identified the crossed items successfully as the less natural ones, but only when they first heard the crossed stimuli and then the baselines. This was an example of a time-order error (TOE) which has been reported numerously for discrimination tasks in different domains (Schiefer and Batliner, 1991; Wherry, 1938; Wickelmaier and Choisel, 2006).

13.6 Outlook

In this thesis, we exclusively analyzed read speech in German and other languages. Therefore, we cannot make inferences about the relationship between acoustic-phonetic measures across different registers of speech. From a practical standpoint, read speech is easier to segment and annotate automatically. These information are then easier to verify manually by human annotators. In addition, we were interested in information contained within the EGG signal with regard to our research question (Section 2.4). The author is not aware of a large German corpus of spontaneous speech containing EGG signals.

We argue that our results regarding the relationship between segmental variability and ID based on read speech can be transfered to spontaneous data because there is some evidence that these relations are apparent across different speech registers. For

other Germanic languages, such as American English and Dutch, previous studies have found that duration and vowel dispersion expand in low predictability contexts and reduce in high predictability contexts (Aylett and Turk, 2004, 2006; van Son, Bolotova, et al., 2004). Since German is closely related to both languages we would expect to find the same patterns for German spontaneous and read data. This assumption is worth investigating in future studies.

Can ID factors of the target language explain segmental variability in L2 speakers? We investigated this research question in a pilot study for vowel dispersion in Bulgarian speakers of German. We found a non-significant effect of surprisal in all speaker groups, but overall advanced L2 speakers were more aware of the effects of tenseness, lexical class and average duration on vowel dispersion in German than intermediate speakers. This study bears several potential points of improvement. We propose to include a prosodic model containing boundary, speech rate, different stress positions or other prosodic factors. In addition, the analysis would profit from including a larger number of vowel phonemes and phonological contexts to increase variability in the data, also with respect to the surprisal values obtained for these phonemes in different contexts.

13.7 Summary

In summary, this thesis has replicated previous findings of the impact of ID factors on segment duration, deletion and vowel dispersion which were mainly based on American English for another language, namely German. Our analyses also included dynamic formant measures, global vocalic characteristics, and voice quality metrics for German which have not been investigated in the context of ID yet. In addition, we found that listeners were sensitive towards violated ID expectations in phonetic detail. These findings added to previous accounts of listeners' increased sensitivity towards differences in phonetic detail when these occur in unpredictable compared to predictable contexts. In addition, we have gathered some first insight into the usefulness of ID variables in explaining variability in learners' speech and making inferences about their competence level. So far, our results were based on read speech only. We encourage future research to include other registers of speech.

List of Figures

List of Tables

List of Acronyms

AE	American English
B2	intermediate competence level
BLUP	Best Linear Unbiased Prediction
C2	advanced competence level
CES	Czech
COCA	Corpus of Contemporary American English
CPP	cepstral peak prominence
CPPS	cepstral peak prominence smoothed
CSD	coronal stop deletion
DCT	discrete cosine transformation
DECPA	derivative-EGG closure peak amplitude
DEGG	first derivative of the glottal waveform
DEOPA	derivative-EGG open peak amplitude
DEU	German
DeWaC	German Web-as-Corpus
DTW	dynamic time warping
EGG	electroglottographic
ERP	event-related potentials
FFT	Fast Fourier transform
FIN	Finnish
FRA	French

g2p	grapheme-to-phoneme
GLMM	generalized linear mixed-effects model
GToBI	German Tones and Break Indices
H1	first harmonic
H2	second harmonic
ID	information density
IPs	intonational phrases
L2	second language
LM	language model
LMM	linear mixed-effects model
MCD	mel-cepstral distortion
MFC	mel-frequency cepstral
MGC	mel-generalized cepstral
OOV	out of vocabulary
OQ	open quotient
OQs	open quotient smoothed
PND	phonological neighborhood density
POL	Polish
RMSE	root mean square error
roc	rate of change
SDeWaC	Stuttgart German Web-as-Corpus
SSR	Smooth Signal Redundancy
TL	trajectory length
TOE	time-order error
TTS	text-to-speech
UID	Uniform Information Density
VISC	vowel inherent spectral change

VL	vector length
VOT	voice onset time
VSL	vowel section length

Bibliography

Adank, P., R. Smits, and R. van Hout (2004). "A comparison of vowel normalization procedures for language variation research". In: *The Journal of the Acoustical Society of America* 116.5, pp. 3099–3107.

Adda-Decker, M., P. Boula de Mareuil, and L. Lamel (1999). "Pronunciation variants in French: schwa and liaison". In: *Proceedings of ICPhS*. San Francisco, pp. 2239–2242.

Ahmed, N., T. Natarajan, and K. Rao (1974). "Discrete cosine transform". In: *IEEE Transactions on Computers* C-23.1, pp. 90–93.

Allen, J. S., J. D. Miller, and D. DeSteno (2003). "Individual talker differences in voice-onset-time". In: *The Journal of the Acoustical Society of America* 113.1, pp. 544–552.

Andreeva, B., W. Barry, and J. Koreman (2013). "The Bulgarian stressed and unstressed vowel stystem. A corpus study". In: *Proceedings of Interspeech*. Lyon, France.

Aylett, M. and A. Turk (2004). "The Smooth Signal Redundancy Hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech". In: *Language* 47.1, pp. 31–56.

Aylett, M. and A. Turk (2006). "Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei". In: *The Journal of the Acoustical Society of America* 119, pp. 3048–3058.

Baayen, R. H., D. J. Davidson, and D. M. Bates (2008). "Mixed-effects modeling with cross random effects for subject and items". In: *Journal of Memory and Language* 59.4, pp. 390–412.

Baayen, R. H., R. Piepenbrock, and L. Gulikers (1995). *The CELEX Lexical Database (Release 2)*. Linguistic Data Consortium.

Babel, M., G. McGuire, and J. King (2014). "Towards a more nuanced view of vocal attractiveness". In: *PLoS One* 9.2.

Balasubramanium, R. K., J. S. Bhat, M. Srivastava, and A. Eldose (2012). "Cepstral analysis of sexually appealing voice". In: *Journal of Voice* 26.4, pp. 412–415.

Baroni, M. and A. Kilgarriff (2006). "Large linguistically-processed web corpora for multiple languages". In: *Proceedings of EACL*. Trento, Italy: Association for Computational Linguistics, pp. 87–90.

Barry, J. and M. Moyle (2011). "Covariation among vowel height effects on acoustic measure". In: *The Journal of the Acoustical Society of America* 130.5, pp. 365–371.

Bartoń, K. (2017). *MuMIn: Multi-Model Inference*. R package version 1.40.0. URL: `https://CRAN.R-project.org/package=MuMIn`.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). "Fitting linear mixed-effects models using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48.

Beaver, D. I., B. Clark, E. S. Flemming, T. F. Jaeger, and M. Wolters (2007). "When semantics meets phonetics: acoustical studies of second-occurrence focus". In: *Language* 83.2, pp. 245–276.

Bell, A., J. Brenier, M. Gregory, C. Girand, and D. Jurafsky (2009). "Predictability effects on durations of content and function words in conversational English". In: *Journal of Memory and Language* 60.1, pp. 92–111.

Bendixen, A., M. Scharinger, A. Strauß, and J. Obleser (2014). "Prediction in the service of comprehension: modulated early brain responses to omitted speech segments". In: *Cortex* 53.1, pp. 9–26.

Bertram, R., A. Pollatsek, and J. Hyönäc (2004). "Morphological parsing and the use of segmentation cues in reading Finnish compounds". In: *Journal of Memory and Language* 51.3, pp. 325–345.

Bigi, B. (2013). *SPPAS – Automatic Annotation of Speech*. Banque de données parole et langage (SLDR/ORTOLANG). URL: `http://www.lpl-aix.fr/~bigi/software.html`.

BNC Consortium (2007). *The British National Corpus, version 3 (BNC XML Edition)*. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

Boersma, P. and D. Weenink (2017). *Praat: doing phonetics by computer*. URL: `http://www.fon.hum.uva.nl/praat`.

Bondarko, L. V., A. Iivonen, L. C. W. Pols, and V. de Silva (2003). "Common and language dependent phonetic differences between read and spontaneous speech in Russian, Finnish and Dutch". In: *Proceedings of ICPhS*, pp. 2977–2980.

Bradlow, A. R., G. M. Torretta, and D. B. Pisoni (1996). "Intelligibility of normal speech. I. Global and fine-grained acoustic-phonetic talker characteristics". In: *Speech Communication* 20, pp. 255–272.

Brandt, E., B. Andreeva, and B. Möbius (2018). "Voice quality as a function of information density and prosodic factors". In: *Phonetik und Phonologie 14*. Vienna, Austria.

Brandt, E., F. Zimmerer, B. Andreeva, and B. Möbius (2017a). "Influence of information density on deletion rates in German". In: *Phonetik und Phonologie 13*. Berlin, Germany.

Brandt, E., F. Zimmerer, B. Andreeva, and B. Möbius (2017b). "Mel-cepstral distortion of German vowels in different information density contexts". In: *Proceedings of Interspeech*. Stockholm, Sweden, pp. 2993–2997.

Brandt, E., F. Zimmerer, B. Andreeva, and B. Möbius (2018). "Impact of prosodic structure and information density on dynamic formant trajectories in German". In: *Proceedings of Speech Prosody*. Poznan, Poland, pp. 119–123.

Brants, T. and A. Franz (2006). *Web 1T 5-gram Version 1*. Linguistic Data Consortium.

Broad, D. J. and H. Wakita (1977). "Piecewise–planar representation of vowel formant frequencies". In: *The Journal of the Acoustical Society of America* 62, pp. 1467–1473.

Burdin, R. S. and C. G. Clopper (2015). "Phonetic reduction, vowel duration, and prosodic structure". In: *Proceedings of the 18th ICPhS*.

Buz, E. and T. F. Jaeger (2016). "The (in)dependence of articulation and lexical planning during isolated word production". In: *Language, Cognition and Neuroscience* 31.3, pp. 404–424.

Buz, E., T. F. Jaeger, and M. K. Tanenhaus (2014). "Contextual confusability leads to targeted hyperarticulation". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 36. 36, pp. 1970–1975.

Bybee, J. (2002). "Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change". In: *Language Variation and Change* 14, pp. 261–290.

Byrd, D. (2000). "Articulatory vowel lengthening and coordination at phrasal junctures". In: *Phonetica* 57.1, pp. 3–16.

Chan, D. et al. (1995). "Eurom - a spoken language resource for the EU". In: *Proceedings of Eurospeech*. Madrid, pp. 867–870.

Chao, K. Y. and L. M. Chen (2008). "A cross-linguistic study of voice onset time in stop consonant productions". In: *International Journal of Computational Linguistics & Chinese Language Processing* 13.2, pp. 215–232.

Chen, S. F. and J. Goodman (1996). "An empirical study of smoothing techniques for language modeling". In: *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics* 13, pp. 310–318.

Childers, D. G. and C. K. Lee (1991). "Vocal quality factors: analysis, synthesis, and perception". In: *The Journal of the Acoustical Society of America* 90.5, pp. 2394–2410.

Christensen, R. H. B. and P. B. Brockhoff (2017). *sensR–An R-package for sensory discrimination*. R package version 1.5-0. URL: `http://www.cran.r-project.org/package=sensR/`.

Clopper, C. G., A. K. Carter, C. M. Dillon, L. R. Hernandez, D. B. Pisoni, C. M. Clarke, J. D. Harnsberger, and R. Herman (2002). *The Indiana Speech Project: an overview of the development of a multi-talker multi-dialect speech corpus*. Tech. rep. 25. Speech Research Laboratory, Indiana University.

Clopper, C. G. and J. B. Pierrehumbert (2008). "Effects of semantic predictability and regional dialect on vowel space reduction". In: *The Journal of the Acoustical Society of America* 124.3, pp. 1682–1688.

Clopper, C. G. and D. B. Pisoni (2006). "The nationwide speech project: A new corpus of American English dialects". In: *Speech Communication* 48, pp. 633–644.

Coetzee, A. W. and S. Kawahara (2013). "Frequency biases in phonological variation". In: *Natural Language & Linguistic Theory* 31.1, pp. 47–89.

Cohen Priva, U. (2017). "Informativity and the actuation of lenition". In: *Language* 93.3, pp. 569–597.

Cohen Priva, U. (2015). "Informativity affects consonant duration and deletion rates". In: *Laboratory Phonology* 6.2, pp. 243–278.

Crocker, M. W., V. Demberg, and E. Teich (2016). "Information Density and Linguistic Encoding (IDeaL)". In: *KI - Künstliche Intelligenz* 30.1, pp. 77–81.

Davies, M. (2008-). *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. URL: `https://corpus.byu.edu/coca/`.

Davis, S. B. and P. Mermelstein (1980). "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences". In: *IEEE Transactions on Acoustics, Speech and Signal Processing*. Vol. 28. 4, pp. 357–366.

Delattre, P. C., A. M. Liberman, and F. S. Cooper (1955). "Acoustic loci and transitional cues for consonants". In: *The Journal of the Acoustical Society of America* 27.4, pp. 769–773.

Dellwo, V., I. Steiner, B. Aschenberner, J. Dankovicova, and P. Wagner (2004). "BonnTempo-corpus and BonnTempo-tools: a database for the study of speech rhythm and rate". In: *Proceedings of Interspeech*. Jeju Island, Korea, pp. 777–780.

Delogu, F., H. Brouwer, and M. W. Crocker (2017). "Teasing apart coercion and surprisal: evidence from eye-movements and ERPs". In: *Cognition* 161, pp. 49–59.

Demberg, V. and F. Keller (2008). "Data from eye-tracking corpora as evidence for theories of syntactic processing complexity". In: *Cognition* 109, pp. 193–210.

Demberg, V., A. B. Sayeed, P. J. Gorinski, and N. Engonopoulos (2012). "Syntactic surprisal affects spoken word duration in conversational contexts". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 356–367.

Department of General Linguistics (1996–1998). *Finnish Parole Corpus*. University of Helsinki and Institute for the Languages of Finland. URL: `http://kaino.kotus.fi/sanat/taajuuslista/parole.php`.

Duddington, J. (2015). *eSpeak text to speech*. http://espeak.sourceforge.net/, retrieved on 1 February 2015. URL: `http://espeak.sourceforge.net/`.

Elsnet (1992 – 1993). *European Corpus Initiative Multilingual Corpus I (ECI/MCI): Frankfurter Rundschau*. URL: `http://www.elsnet.org/eci.html`.

Epstein, M. A. (2002). "Voice quality and prosody in English". PhD thesis. University of California, Los Angeles.

Faaß, G. and K. Eckart (2013). "SDeWaC - A corpus of parsable sentences from the web". In: *Language Processing and Knowledge in the Web*. Berlin, Heidelberg: Springer-Verlag, pp. 61–68.

Fant, G. (1960). *Acoustic theory of speech production*. Vol. 1. Mouton.

Ferrand, C. T. (2011). *Voice disorders: scope of theory and practice*. Vol. 1. Pearson Education.

Ferrer, L., E. Shriberg, and A. Stolcke (2002). "Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody". In: *Proceedings of Interspeech*. Denver, Colorado.

Fine, A. B., A. F. Frank, T. F. Jaeger, and B. van Durme (2014). "Biases in predicting the human language model". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 7–12.

Fischer-Jørgensen, E. (1954). "Acoustic analysis of stop consonants". In: *Miscellanea Phonetica*, pp. 42–59.

Fischer-Jørgensen, E. (1967). "Phonetic analysis of breathy (murmured) vowels in Gujarati". In: *Indian Linguistics* 28, pp. 71–139.

Fox, R. A. and E. Jacewicz (2009). "Cross-dialectal variation in formant dynamics of American English vowels." In: *The Journal of the Acoustical Society of America* 126.5, pp. 2603–18.

Gahl, S. (2008). "Time and thyme are not homophones: the effect of lemma frequency on word durations in spontaneous speech". In: *Language* 84.3, pp. 474–496.

Gahl, S. (2012). "Why so short? Competing explanations for variation". In: *Proceedings of the 29th West Coast Conference of Formal Linguistics*, pp. 1–10.

Gahl, S., Y. Yao, and K. Johnson (2012). "Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech". In: *Journal of Memory and Language* 66.4, pp. 789–806.

Gay, T. (1978). "Effect of speaking rate on vowel formant movements". In: *The Journal of the Acoustical Society of America* 63.1, pp. 223–230.

Gendrot, C. and M. Adda (2005). "Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German". In: *Proceedings of Interspeech*. Lisbon, Portugal, pp. 2453–2456.

Gobl, C. (1988). "Voice source dynamics in connected speech". In: *STL-QPSR*. Stockholm: Speech, Music and Hearing, Royal Institute of Technology, pp. 123–159.

Godfrey, J., E. Holliman, and J. McDaniel (1992). "SWITCHBOARD: telephone speech corpus for research and development". In: *Proceedings of ICASSP*, pp. 517–520.

Grabe, E., G. Kochanski, and J. Coleman (2007). "Connecting intonation labels to mathematical descriptions of fundamental frequency". In: *Language and Speech* 50, pp. 281–310.

Green, D. M. and J. A. Swets (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Grice, M. and S. Baumann (2002). "Deutsche Intonation und GToBI". In: *Linguistische Berichte 191*, pp. 267–298.

Guy, G. R. (1980). "Variation in the group and the individual: the case of final stop deletion". In: *Locating language in time and space*. Ed. by W. Labov. New York: Academic Press, pp. 1–36.

Haderlein, T. (2007). "Automatic evaluation of tracheoesophageal substitute voices". PhD thesis. Erlangen-Nürnberg: Friedrich-Alexander-Universität.

Haderlein, T., C. Moers, B. Möbius, F. Rosanowski, and E. Nöth (2011). "Intelligibility rating with automatic speech recognition, prosodic, and cepstral evaluation". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6836 LNAI.Tsd, pp. 195–202.

Hale, J. (2001). "A probabilistic early parser as a psycholinguistic model". In: *Proceedings of NAACL*. Stroudsburg, PA, pp. 1–8.

Hale, J. (2016). "Information-theoretical complexity metrics". In: *Language and Linguistics Compass* 10.9, pp. 397–412.

Hanique, I. and M. Ernestus (2011). "Final /t/ reduction in Dutch past-participles: the role of word predictability and morphological decomposability". In: *Proceedings of Interspeech*, pp. 2849–2852.

Harrington, J. and S. Cassidy (1994). "Dynamic and target theories of vowel classification: evidence from monophthongs and diphthongs in Australian English". In: *Language and Speech* 37.4, pp. 357–373.

Harrington, J. (2010). *The phonetic analysis of speech corpora*. Wiley-Blackwell.

Hazen, K. (2011). "Flying high above the social radar: coronal deletion in modern Appalachia". In: *Language Variation and Change* 23, pp. 105–137.

Hellström, A. (1985). "The time-order error and its relatives: mirrors of cognitive processes in comparing". In: *Psychological Bulletin* 97, pp. 35–61.

Heman-Ackah, Y. D., D. D. Michael, and G. S. Goding (2002). "The relationship between cepstral peak prominence and selected parameters of dysphonia". In: *Journal of Voice* 16.1, pp. 20–27.

Henrich, N., C. D'Alessandro, B. Doval, and M. Castellengo (2004). "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation". In: *The Journal of the Acoustical Society of America* 115.3, pp. 1321–1332.

Henrich, N., C. D'Alessandro, B. Doval, and M. Castellengo (2005). "Glottal open quotient in singing: measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency". In: *The Journal of the Acoustical Society of America* 117.3, pp. 1417–1430.

Henton, C. and A. Bladon (1988). "Creak as a sociophonetic marker". In: *Language, speech and mind: studies in honor of Victoria A. Fromkin*. Ed. by L. M. Hyman and C. N. Li. London: Routledge, pp. 3–29.

Hillenbrand, J. M., M. J. Clark, and T. M. Nearey (2001). "Effects of consonant environment on vowel formant patterns". In: *The Journal of the Acoustical Society of America* 109.2, pp. 748–763.

Hillenbrand, J. M. and R. T. Gayvert (1993). "Identification of steady-state vowels synthesized from the Peterson and Barney measurements." In: *The Journal of the Acoustical Society of America* 94.2 Pt 1, pp. 668–674.

Hillenbrand, J. M., R. A. Cleveland, and R. L. Erickson (1994). "Acoustic correlates of breathy vocal quality". In: *Journal of Speech, Language, and Hearing Research* 37.4, pp. 769–778.

Hillenbrand, J. M., L. A. Getty, M. J. Clark, and K. Wheeler (1995). "Acoustic characteristics of American English vowels". In: *The Journal of the Acoustical Society of America* 97.May, pp. 3099–3111.

Hillenbrand, J. M. and R. A. Houde (1996). "Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech". In: *Journal of Speech, Language, and Hearing Research* 39.2, pp. 311–321.

Hillenbrand, J. M. and R. A. Houde (2003). "A narrow band pattern-matching model of vowel perception." In: *The Journal of the Acoustical Society of America* 113.2, pp. 1044–1055.

Hume, E. (2004). "Deconstructing markedness: a predictability-based approach". In: *Annual Meeting of the Berkeley Linguistics Society*. Vol. 30. 1, pp. 182–198.

IPDS (1994). *The Kiel Corpus of Read Speech*. IPDS.

IPDS (1997). *The Kiel Corpus of Spontaneous Speech*. IPDS.

Jaeger, T. F. (2010). "Redundancy and reduction: speakers manage syntactic information density". In: *Cognitive Psychology* 61.1, pp. 23–62.

Jaeger, T. F. and E. Buz (2017). "Handbook of Psycholinguistic". In: ed. by E. M. Fernandez and H. M. I. Cairns. Oxford, United Kingdom: Wiley-Blackwell. Chap. Signal reduction and linguistic encoding, pp. 38–81.

Jenkins, J. J., W. Strange, and S. a. Trent (1999). "Context-independent dynamic information for the perception of coarticulated vowels." In: *The Journal of the Acoustical Society of America* 106.1, pp. 438–448.

Johnson, K. (2004). "Massive reduction in conversational American English". In: *Spontaneous speech: Data and analysis. Proceedings of the 1st session of the 10th international symposium*, pp. 29–54.

Joos, M. (1948). *Acoustic phonetics*. Language Monograph No 23. Baltimore: Linguistic Society of America.

Jun, S.-A. and C. Fougeron (2000). "A phonological model of French intonation". In: *Intonation*. Springer, pp. 209–242.

Jurafsky, D., A. Bell, and C. Girand (2002). "The role of the lemma in form variation". In: *Laboratory Phonology 7*. Ed. by C. Gussenhoven and N. Warner. Berlin: Mouton de Gruyter, pp. 1–34.

Jurafsky, D., A. Bell, M. Gregory, and W. D. Raymond (2001). "Probabilistic relations between words: evidence from reduction in lexical production". In: *Frequency and the Emergence of Linguistic Structure*. Ed. by J. Bybee and P. Hopper. Amsterdam: Benjamins, pp. 229–254.

Kakouros, S. and O. Räsänen (2016). "Perception of sentence stress in speech correlates with the temporal unpredictability of prosodic features". In: *Cognitive Science* 40.7, pp. 1739–1774.

Kalikow, D. N., K. N. Stevens, and L. L. Elliott (1977). "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability". In: *The Journal of the Acoustical Society of America* 61.5, pp. 1337–1351.

Kehoe, M. M., C. Lleó, and M. Rakow (2004). "Voice onset time in bilingual German-Spanish children". In: *Bilingualism: Language and Cognition* 7.1, pp. 71–88.

Keshet, J., M. Sonderegger, and T. Knowles (2014). *AutoVOT: a tool for automatic measurement of voice onset time using discriminative structured prediction [Computer program]*. Version 0.91. URL: `https://github.com/mlml/autovot/`.

Kessinger, R. H. and S. E. Blumstein (1997). "Effects of speaking rate on voice-onset time in Thai, French, and English". In: *Journal of Phonetics* 25.2, pp. 143–168.

Kisler, T., U. D. Reichel, and F. Schiel (2017). "Multilingual processing of speech via web services". In: *Computer Science and Language* 45, pp. 326–347.

Klatt, D. H. (1975). "Voice onset time, frication, and aspiration in word-initial consonant clusters". In: *Journal of Speech, Language, and Hearing Research* 18.4, pp. 686–706.

Klatt, D. H. (1980). "Speech perception: a model of acoustic-phonetic analysis and lexical access". In: *Perception and production of fluent speech*. Ed. by R. A. Cole. Hillsdale: Erlbaum, pp. 243–288.

Klatt, D. H. (1976). "Linguistic uses of segmental duration in English: acoustic and perceptual evidence". In: *The Journal of the Acoustical Society of America* 59.5, pp. 1208–1221.

Kleiner, S., R. Knöbl, and M. Mangold (2015). *Duden Aussprachewörterbuch*. Ed. by Dudenredaktion. Vol. 1. Duden.

Kłosowski, P. (2017). "Statistical analysis of orthographic and phonemic language corpus for word-based and phoneme-based Polish language modelling". In: *EURASIP Journal on Audio, Speech, and Music Processing* 5.1.

Kneissler, J. and D. Klakow (2001). "Speech recognition for huge vocabularies by using optimized subword units". In: *Proceedings of Eurospeech*.

Kobayashi, T., K. Tokuda, and K. Koishida (2017). *Speech Signal Processing Toolkit (SPTK)*. version 3.10.

Kohler, K. J. (1988). "Zeitstrukturierung in der Sprachsynthese". In: *ITG-Fachbericht* 105, pp. 165–170.

Kohler, K. J. and J. Rodgers (2001). *Schwa deletion in German read and spontaneous speech*. Tech. rep. Universität Kiel.

Kominek, J., T. Schultz, and A. W. Black (2008). "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion". In: *SLTU*.

Kubichek, R. (1993). "Mel-cepstral distance measure for objective speech quality assessment". In: *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*. Vol. 1, pp. 125–128.

Kuzla, C. and M. Ernestus (2011). "Prosodic conditioning of phonetic detail in German plosives". In: *Journal of Phonetics* 39.2, pp. 143–155.

Kuznetsova, A., P. B. Brockhoff, and R. H. B. Christensen (2017). "lmerTest package: tests in linear mixed effects models". In: *Journal of Statistical Software* 82.13, pp. 1–26.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Larson, M. and S. Eickeler (2003). "Using syllable based indexing features and language models to improve German spoken document retrieval". In: *Proceedings of Eurospeech*. Geneva, Switzerland.

Lasarcyk, E., H. Drenhaus, and B. Möbius (2015). "Experimente zur Wahrnehmung gezielt degradierter synthetischer Sprache". In: *Elektronische Sprachsignalverarbeitung 2015, Tagungsband der 26. Konferenz*. Eichstätt, pp. 112–119.

Lauwers, P. and D. Willems (2011). "Coercion: definition and challenges, current approaches, and new trends". In: *Linguistics* 49.6, pp. 1219–1235.

Laver, J. (1980). *The phonetic description of voice quality*. Vol. 1. Cambridge University Press.

Le Maguer, S., B. Möbius, and I. Steiner (2016). "Toward the use of information density based descriptive features in HMM based speech synthesis". In: *Proceedings of Speech Prosody*, pp. 1029–1033.

Lee, E.-K. and J. Cole (2006). "Acoustic effects of prosodic boundary on vowels in American English". In: *Proceedings from the Annual Meeting of the Chicago Linguistic Society*. Vol. 42. 1, pp. 181–195.

Lehiste, I. and G. E. Peterson (1961). "Transitions, glides, and diphthongs". In: *The Journal of the Acoustical Society of America* 33.3, pp. 268–277.

Levy, R. and T. F. Jaeger (2007). "Speakers optimize information density through syntactic reduction". In: *Advances in Neural Information Processing Systems* 19, pp. 849–856.

Levy, R. (2008). "A noisy-channel model of rational human sentence comprehension under uncertain input". In: *Proceedings of the 13th Conference on Empirical Methods in Natural Language Processing*. Waikiki, Honolulu, pp. 234–243.

Levy, R. (2011). "Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 1055–1065.

Levy, R. (2013). "Memory and surprisal in human sentence comprehension". In: *Sentence Processing* October, pp. 78–114.

Lieberman, P. (1960). "Some acoustic correlates of word stress in American English". In: *The Journal of the Acoustical Society of America* 32.4, pp. 451–454.

Lieberman, P. (1963). "Some effects of semantic and grammatical context on the production and perception of speech". In: *Language and Speech* 6.3, pp. 172–188.

Lindblom, B. (1963). "Spectrographic study of vowel reduction". In: *The Journal of the Acoustical Society of America* 35, pp. 1773–1781.

Lindblom, B., W. J. Hardcastle, and A. Marchal (1990). "Speech production and speech modelling". In: *Speech Production and Speech Modelling*. Ed. by W. J. Hardcastle and A. Marchal. Kluwer Academic Publishers. Chap. Explaining Phonetic Variation: A Sketch of the H&H Theory, pp. 403–439.

Lisker, L. (1978). "In qualified defense of VOT". In: *Language and Speech* 21.4, pp. 375–383.

Lisker, L. and A. S. Abramson (1964). "A cross-language study of voicing in initial stops: acoustical measurements". In: *WORD* 20.3, pp. 384–422.

Luce, P. A. and D. B. Pisoni (1998). "Recognizing spoken words: the neighborhood activation model". In: *Ear and Hearing* 19.1, pp. 1–36.

Malisz, Z., E. Brandt, B. Möbius, Y. Oh, and B. Andreeva (2018). "Dimensions of segmental variability: interaction of prosody and surprisal in six languages". In: *Frontiers in Communication*.

Manker, J. T. (2017). "Phonetic attention and predictability: how context shapes exemplars and guides sound change". PhD thesis. University of California, Berkeley.

Manning, C. D. and H. Schütze (1999). *Foundations of statistical natural language processing.*

Maryn, Y., N. Roy, M. de Bodt, and P. A. van Cauwenberge (2009). "Acoustic measurement of overall voice quality: a meta-analysis." In: *The Journal of the Acoustical Society of America* 126.5, pp. 2619–2634.

Max Planck Institute for Psycholinguistics (2001). *WebCelex German.* URL: `http://celex.mpi.nl/`.

McDougall, K. and F. Nolan (2007). "Discrimination of speakers using the formant dynamics of /u:/ in British English". In: *Proceedings of the 16th International Congress of Phonetic Sciences* August, pp. 1825–1828.

McMurray, B., V. M. Samelson, S. H. Lee, and J. B. Tomblin (2010). "Individual differences in online spoken word recognition: implications for SLI". In: *Cognitive Psychology* 60, pp. 1–39.

Michaud, A. (2007). *Software for electroglottographic analysis: <peakdet>, a script for calculations based on peak detection.* URL: `http://voiceresearch.free.fr/egg/softwares.htm`.

Michaud, A. (2004). "A measurement from electroglottography: DECPA, and its application in prosody". In: *Proceedings of Speech Prosody*, pp. 633–636.

Miller, J. L. and L. E. Volaitis (1989). "Effect of speaking rate on the perceptual structure of a phonetic category". In: *Perception & Psychophysics* 46.6, pp. 505–512.

Möbius, B. and J. von Santen (1996). "Modeling segmental duration in German text-to-speech synthesis". In: *Proceeding of ICSLP* 4, pp. 2395–2398.

Moers, C., B. Möbius, F. Rosanowski, E. Nöth, U. Eysholdt, and T. Haderlein (2012). "Vowel- and text-based cepstral analysis of chronic hoarseness". In: *Journal of Voice* 26.4, pp. 416–424.

Möhler, G., A. Schweitzer, M. Breitenbücher, and M. Barbisch (2000). *IMS German Festival (Version: 1.2-os).* University of Stuttgart: Institut für maschinelle Sprachverarbeitung (IMS).

Moon, S.-J. and B. Lindblom (1994). "Interaction between duration, context, and speaking style in English stressed vowels". In: *The Journal of the Acoustical Society of America* 96.1, pp. 40–55.

Mooshammer, C. and C. Geng (2008). "Acoustic and articulatory manifestations of vowel reduction in German". In: *Journal of the International Phonetic Association* 38.2, pp. 117–136.

Morgan, N. and E. Fosler-Lussier (1998). "Combining multiple estimators of speaking rate". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Vol. 2. IEEE, pp. 729–732.

Morrison, G. S. (2009). "Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs". In: *The Journal of the Acoustical Society of America* 125.4, pp. 2387–2397.

Muda, L., M. Begam, and I. Elamvazuthi (2010). "Voice recognition algorithms using Mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques". In: *Journal of Computing* 2.3, pp. 2151–9617.

Munson, B. and N. Solomon (2004). "The effect of phonological neighborhood density on vowel production". In: *Journal of Speech, Language, and Hearing Research* 47.5, pp. 1048–1058.

Munson, B. (2007). "Lexical access, lexical representation, and vowel production". In: *Laboratory phonology*, pp. 201–228.

Nearey, T. M. and P. F. Assmann (1986). "Modeling the role of inherent spectral change in vowel identification". In: *The Journal of the Acoustical Society of America* 80.5, pp. 1297–1308.

Neumeyer, V., J. Harrington, and C. Draxler (2010). "An acoustic analysis of the vowel space in young and old cochlear-implant speakers". In: *Clinical Linguistics & Phonetics* 24.9, pp. 734–741.

New, B., C. Pallier, L. Ferrand, and R. Matos (2001). "Une base de données lexicales du français contemporain sur internet: LEXIQUE 3.80". In: *L'Année Psychologique* 101, pp. 447–462. URL: `http://www.lexique.org`.

Ng, K. and V. W. Zue (1997). "Subword unit representation for spoken document retrieval". In: *Proceedings of Eurospeech*.

Nowak, P. (2006). "Vowel reduction in Polish". PhD thesis. University of California, Berkeley.

Oh, Y. M. (2015). "Linguistic complexity and information: quantitative approaches". PhD thesis. Lyon: Laboratoire Dynamique du Langage – CNRS - Université Lumière Lyon 2.

Oh, Y. M., Coupé Christophe, E. Marsico, and F. Pellegrino (2015). "Bridging phonological system and lexicon: insights from a corpus study of functional load". In: *Journal of Phonetics* 53, pp. 153–176.

Patterson, D., P. C. LoCasto, and C. M. Connine (2003). "Corpora analyses of frequency of schwa deletion in conversational American English". In: *Phonetica* 60, pp. 45–69.

Pätzold, M. and A. P. Simpson (1997). "Acoustic analysis of German vowels in the Kiel corpus of read speech". In: *The Kiel Corpus of Read/Spontaneous Speech Acoustic data base, processing tools and analysis results. Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK)* 32, pp. 215–247.

Pellegrino, F., C. Coupé, and E. Marisco (2011). "A cross-language perspective on speech information rate". In: *Language* 87.3, pp. 539–558.

Peterson, G. E. and I. Lehiste (1960). "Duration of syllable nuclei in English". In: *The Journal of the Acoustical Society of America* 32, pp. 693–703.

PHONDAT2 – PD2 (1995). version 2.8. University of Munich.

Piantadosi, S., H. Tily, and E. Gibson (2011). "Word lengths are optimized for efficient communication". In: *Proceedings of the National Academy of Sciences 108*, pp. 3526–3529. DOI: 10.1073/pnas.1012551108/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1012551108.

Pierrehumbert, J. B. (2000). "Exemplar dynamics: word frequency, lenition and contrast". In: *Frequency effects and the emergence of linguistic structure*. Ed. by J. Bybee and P. Hopper. Amsterdam: John Benjamins, pp. 1–19.

Pitt, M. (1998). "Phonological processes and the perception of phonotactically illegal consonant clusters". In: *Perception and Psychophysics* 60, pp. 941–951.

Pitt, M., K. Johnson, E. Hume, S. Kiesling, and W. D. Raymond (2005). "The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability". In: *Speech Communication* 45, pp. 90–95.

Pluymaekers, M., M. Ernestus, and R. H. Baayen (2005a). "Articulatory planning is continuous and sensitive to informational redundancy". In: *Phonetica* 62.2-4, pp. 146–159.

Pluymaekers, M., M. Ernestus, and R. H. Baayen (2005b). "Lexical frequency and acoustic reduction in spoken Dutch". In: *The Journal of the Acoustical Society of America* 118.4, pp. 2561–2569.

R Development Core Team (2008). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org.

Ramus, F. (2002). "Acoustic correlates of linguistic rhythm: perspectives". In: *Proceedings of Speech Prosody*.

Raymond, W., R. Dautricourt, and E. Hume (2006). "Word-internal /t,d/ deletion in spontaneous speech: modeling the effects of extra-linguistic, lexical, and phonological factors". In: *Language Variation and Change* 18.1, pp. 55–97.

Revelle, W. (2017). *psych: procedures for psychological, psychometric, and personality research*. R package version 1.7.8. Northwestern University. Evanston, Illinois. URL: `https://CRAN.R-project.org/package=psych`.

Rietveld, R. C. M. and C. Gussenhoven (1987). "Perceived speech rate and intonation". In: *Journal of Phonetics* 15, pp. 273–285.

Risdal, M. L. and M. E. Kohn (2014). "Ethnolectal and generational differences in vowel trajectories: evidence from African American English and the Southern vowel system". In: *University of Pennsylvania Working Papers in Linguistics* 20.2 Selected Papers from NWAV 42, Article 16.

Roland, D., J. L. Elman, and V. S. Ferreira (2006). "Why is that? Structural prediction and ambiguity resolution in a very large corpus of English sentences". In: *Cognition* 98.3, pp. 245–272.

Savin, H. B. (1963). "Word-frequency effect and errors in the perception of speech". In: *The Journal of the Acoustical Society of America* 35.2, p. 200.

Scarborough, R. (2006). "Lexical and contextual predictability: confluent effects on the production of vowels". In: *Laboratory Phonology Conference*. Paris, France.

Schiefer, L. and A. Batliner (1991). "Order effect and the order of accents". In: *Proceedings of the 12th ICPhS*. Vol. 3. Aix-en-Provence, pp. 86–89.

Schiel, F. (1997). *Siemens Synthesis Corpus - SI1000P*. URL: `https://www.phonetik.uni-muenchen.de/Bas/BasSI1000Peng.html`.

Schouten, M. E. H. and L. C. W. Pols (1979). "Vowel segments in consonantal contexts: a spectral study of coarticulation. Part I". In: *Journal of Phonetics* 7, pp. 1–23.

Schrumpf, C., M. Larson, and S. Eickeler (2005). "Syllable-based language models in speech recognition for English spoken document retrieval". In: *AVIVDiLib* November, pp. 196–205.

Schulz, E., Y. M. Oh, Z. Malisz, B. Andreeva, and B. Möbius (2016). "Impact of prosodic structure and information density on vowel space size". In: *Proceedings of Speech Prosody*. Boston, pp. 350–354.

Shannon, C. E. (1948). "A mathematical theory of communication". In: *The Bell System Technical Journal* 27, pp. 379–423, 623–656.

Simpson, A. P. and C. Ericsdotter (2007). "Sex-specific differences in f0 and vowel space". In: *Proceedings of the XVIth ICPhS*. August. Saarbrücken, pp. 6–10.

Slifka, J. (2003). "Tense/lax vowel classification using dynamic spectral cues". In: *15th ICPhS*, pp. 921–924.

Smith, N. J. and R. Levy (2013). "The effect of word predictability on reading time is logarithmic". In: *Cognition* 128.3, pp. 302–319.

Solé, M.-J. and E. Estebas (2000). "Phonetic and phonological phenomena: VOT. A cross-language comparison". In: *Proceedings of the XVII AEDEAN Conference*, pp. 437–444.

Sonderegger, M., M. Bane, and P. Graff (2017). "The medium-term dynamics of accents on reality television". In: *Language* 93.3, pp. 598–640.

Stanislaw, H. and N. Todorov (1999). "Calculation of signal detection theory measures". In: *Behavior Research Methods, Instruments, and Computer* 31.1, pp. 137–149.

Steinberg, J. and M. Scharinger (2018). "Processing of speech sounds depends on interactive context effects: evidence from electrophysiology with omissions". In: *in prep.*

Stevens, K. N. (1993). "Models for the production and acoustics of stop consonants". In: *Speech Communication* 13.3-4, pp. 367–375.

Stevens, K. N. (2002). "Toward a model for lexical access based on acoustic landmarks and distinctive features". In: *The Journal of the Acoustical Society of America* 111.4, pp. 1872–1891.

Stevens, K. N. and A. S. House (1963). "Perturbation of vowel articulations by consonantal context: an acoustical study". In: *Journal of Speech and Hearing Research* 6.2, pp. 111–128.

Stevens, K. N. and A. S. House (1961). "An acoustical theory of vowel production and some of its implications". In: *Journal of Speech, Language, and Hearing Research* 4, pp. 303–320.

Stolcke, A. (2002). "Srilm - an extensible language modeling toolkit". In: *Proceedings of Interspeech* 2.Denver, Colorado, pp. 901–904.

Strange, W. and O.-S. Bohn (1998). "Dynamic specification of coarticulated German vowels: perceptual and acoustical studies". In: *The Journal of the Acoustical Society of America* 104.1, pp. 488–504.

Tabain, M. (2003). "Effects of prosodic boundary on /aC/ sequences: acoustic results". In: *The Journal of the Acoustical Society of America* 113.1, pp. 516–531.

Tagliamonte, S. and R. A. M. Temple (2005). "New perspectives on an ol' variable: (t, d) in British English". In: *Language Variation and Change* 17, pp. 281–302.

Tanner, J., M. Sonderegger, and M. Wagner (2017). "Production planning and coronal stop deletion in spontaneous speech". In: *Laboratory Phonology* 8.1, pp. 1–39.

Tily, H., S. Gahl, I. Arnon, N. Snider, A. Kothari, and J. Bresnan (2009). "Syntactic probabilities affect pronunciation variation in spontaneous speech". In: *Language and Cognition* 1.2, pp. 147–165.

Toda, T., A. W. Black, and K. Tokuda (2004). "Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis". In: *ISCA Speech Synthesis Workshop*, pp. 31–36.

Tokuda, K., T. Kobayashi, T. Masuko, and S. Imai (1994). "Mel-generalized cepstral analysis - A unified approach to speech spectral estimation". In: *International Conference on Spoken Language Processing*. Vol. 3. Yokohama, Japan.

Torretta, G. M. (1995). *The 'easy-hard' word multi-talker speech database: an initial report*. Research on Spoken Language Processing Progress Report 20. Speech Research Laboratory, Indiana University.

Trouvain, J. (2004). "Words and beyond - The phonetic channel in communication." In: *Proceedings Conference on Foreign Language Teaching / Colóquio Ensino das Línguas Estrangeiras*. Porto, Portugal, pp. 69–82.

Turk, A. (2010). "Does prosodic constituency signal relative predictability? A Smooth Signal Redundancy hypothesis". In: *Laboratory Phonology* 1.2, pp. 227–262.

Turnbull, R. (2017). "The role of predictability in intonational variability". In: *Language and Speech* 60.1, pp. 123–153.

Turnbull, R., R. S. Burdin, C. G. Clopper, and J. Tonhauser (2015). "Contextual predictability and the prosodic realisation of focus: a cross-linguistic comparison". In: *Language, Cognition and Neuroscience* 3798, pp. 1–16.

Turner, G. S., K. Tjaden, and G. Weismer (1995). "The influence of speaking rate of vowel working space and intelligibility for individuals with amyotrophic lateral sclerosis". In: *Journal of Speech, Language, and Hearing Research* 38, pp. 1001–1013.

Tychtl, Z. and J. Psutka (1999). "Speech production based on the mel-frequency cepstral coefficients". In: *Sixth European Conference on Speech Communication and Technology*. Budapest, Hungary.

UCL (2013). *FormantPro*. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr. URL: `https://hdl.handle.net/11403/sldr000842/v1`.

van Bergem, D. R. (1993). "Acoustic vowel reduction as a function of sentence accent, word stress, and word class". In: *Speech Communication* 12, pp. 1–23.

van Son, R. J. J. H., D. Binnenpoorte, H. van den Heuvel, and L. C. W. Pols (2001). "The IFAcorpus: a phonemically segmented Dutch open source speech database". In: *Proceedings of Eurospeech*. Aalborg, pp. 2051–2054.

van Son, R. J. J. H., O. Bolotova, L. C. Pols, and M. Lennes (2004). "Frequency effects on vowel reduction in three typologically different languages (Dutch, Finnish, Russian)". In: *Proceedings of Interspeech*. Jeju Island, Korea, pp. 1277–1280.

van Son, R. J. J. H. and J. P. H. van Santen (2005). "Duration and spectral balance of intervocalic consonants: a case for efficient communication". In: *Speech Communication* 47.1-2, pp. 100–123.

van Son, R. J. J. H. and L. C. W. Pols (1990). "Formant frequencies of Dutch vowels in a text, read at normal and fast rate". In: *The Journal of the Acoustical Society of America* 88.4, pp. 1683–1693.

Watson, C. I. and J. Harrington (1999). "Acoustic evidence for dynamic formant trajectories in Australian English vowels". In: *The Journal of the Acoustical Society of America* 106.1, pp. 458–468.

Watson, D. G., J. E. Arnold, and M. K. Tanenhaus (2008). "Tic Tac TOE: effects of predictability and importance on acoustic prominence in language production". In: *Cognition* 106.3, pp. 1548–1557.

Wedel, A., N. Nelson, and R. Sharp (2018). "The phonetic specificity of contrastive hyperarticulation in natural speech". In: *Journal of Memory and Language* 100, pp. 61–88.

Weirich, M. and A. P. Simpson (2014). "Differences in acoustic vowel space and the perception of speech tempo". In: *Journal of Phonetics* 43.1, pp. 1–10.

Weismer, G. and J. Berry (2003). "Effects of speaking rate on second formant trajectories of selected vocalic nuclei". In: *The Journal of the Acoustical Society of America* 113.6, pp. 3362–3378.

Weiss, B. (2007). "Rate dependent vowel reduction in German". In: *Proceedings of the 12th SPECOM*. Moscow.

Wheeldon, L. and A. Lahiri (1997). "Prosodic units in speech production". In: *Journal of Memory and Language* 37.3, pp. 356–381.

Wherry, R. J. (1938). "Orders for the presentation of pairs in the method of paired comparison". In: *Journal of Experimental Psychology* 23, pp. 651–660.

Whiteside, S. P., L. Henry, and R. Dobbin (2004). "Sex differences in voice onset time: a developmental study of phonetic context effects in British English". In: *The Journal of the Acoustical Society of America* 116.2, pp. 1179–1183.

Whiteside, S. P. and J. Marshall (2001). "Developmental trends in voice onset time: some evidence for sex differences". In: *Phonetica* 58, pp. 196–210.

Wickelmaier, F. and S. Choisel (2006). "Modeling within-pair order effects in paired-comparison judgments". In: *Proceedings of Fechner Day* 22.1, pp. 89–94.

Winkler, R. and W. Sendlmeier (2006). "EGG open quotient in aging voices – Changes with increasing chronological age and its perception". In: *Logopedics Phoniatrics Vocology* 31.2, pp. 51–56.

Wright, R. (2004). "Factors of lexical competition in vowel articulation". In: *Papers in Laboratory Phonology VI*. Ed. by J. Local, R. Ogden, and R. Temple. Cambridge: Cambridge University Press, pp. 26–50.

Yannakoudakis, E. J. and P. J. Hutton (1992). "An assessment of n-phoneme statistics in phoneme guessing algorithms which aim to incorporate phonotactic constraints". In: *Speech Communication* 11.6, pp. 581–602.

Yao, Y. (2009). *Understanding VOT variation in spontaneous speech*. Tech. rep. University of California, Berkeley.

Yap, T. F., J. Epps, E. Ambikairajah, and E. H. Choi (2015). "Voice source under cognitive load: effects and classification". In: *Speech Communication* 72, pp. 74–95.

Yap, T. F., J. Epps, E. Ambikairajah, and E. H. Choi (2011). "Voice source features for cognitive load classification". In: *Proceedings of ICASSP*, pp. 5700–5703.

Zahorian, S. A. and A. J. Jagharghi (1993). "Spectral shape features versus formants as acoustic correlates for vowels". In: *The Journal of the Acoustical Society of America* 94.4, pp. 1966–1982.

Zeldes, A. (2008–2014). *Automatic phonetic transcription and syllable analysis*. Georgetown University. URL: `http://corpling.uis.georgetown.edu/amir/phon.php`.

Zhao, Y. and D. Jurafsky (2009). "The effect of lexical frequency and Lombard reflex on tone hyperarticulation". In: *Journal of Phonetics* 37, pp. 231–247.

Zimmerer, F. (2009). "Reduction in natural speech". PhD thesis. Johann Wolfgang Goethe-Universität.

Ziółko, B. and J. Gałka (2010). "Polish phones statistics". In: *Proceedings of the 2010 International Multiconference on Computer Science and Information Technology (IMCSIT)*. AGH Univesity of Science and Technology, Krakow.

Zipf, G. K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. London: George Routledge & Sons.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort: An introduction to human ecology*. New York: Addison-Wesley.

Zséder, A., G. Recski, D. Varga, and A. Kornai (2012). "Rapid creation of large-scale corpora and frequency dictionaries". In: *Proceedings of LREC*, pp. 1462–1465.

Zue, V. W. and M. Laferriere (1979). "Acoustic study of medial /t,d/ in American English". In: *The Journal of the Acoustical Society of America* 66.4, pp. 1039–1050.

Own publications listed in Section 1.3

Brandt, E., B. Andreeva, and B. Möbius (2018). "Voice quality as a function of information density and prosodic factors". In: *Phonetik und Phonologie 14*. Vienna, Austria.

Brandt, E., F. Zimmerer, B. Andreeva, and B. Möbius (2017a). "Influence of information density on deletion rates in German". In: *Phonetik und Phonologie 13*. Berlin, Germany.

Brandt, E., F. Zimmerer, B. Andreeva, and B. Möbius (2017b). "Mel-cepstral distortion of German vowels in different information density contexts". In: *Proceedings of Interspeech*. Stockholm, Sweden, pp. 2993–2997.

Brandt, E., F. Zimmerer, B. Andreeva, and B. Möbius (2018). "Impact of prosodic structure and information density on dynamic formant trajectories in German". In: *Proceedings of Speech Prosody*. Poznan, Poland, pp. 119–123.

Malisz, Z., E. Brandt, B. Möbius, Y. Oh, and B. Andreeva (2018). "Dimensions of segmental variability: interaction of prosody and surprisal in six languages". In: *Frontiers in Communication*.

Schulz, E., Y. M. Oh, Z. Malisz, B. Andreeva, and B. Möbius (2016). "Impact of prosodic structure and information density on vowel space size". In: *Proceedings of Speech Prosody*. Boston, pp. 350–354.

Appendices

A Appendix: German formant measurements

Table 2: Mean German formant values based on the Siemens Synthesis corpus measured at temporal midpoint and standard deviation (SD) per vowel identity.

Vowel	F1		F2		F3		F2_3	
/ə/	398	(55)	1577	(180)	2249	(128)	1912	(120)
/ɐ/	483	(87)	1380	(217)	2306	(163)	1843	(115)
/øː/	348	(33)	1419	(110)	2086	(90)	1752	(55)
/œ/	495	(48)	1352	(104)	2220	(96)	1787	(55)
/a/	622	(86)	1261	(99)	2365	(187)	1814	(80)
/aː/	634	(82)	1217	(93)	2393	(192)	1807	(76)
/eː/	349	(61)	1852	(139)	2500	(216)	2176	(160)
/ɛ/	466	(58)	1622	(180)	2296	(143)	1959	(135)
/ɛː/	442	(56)	1716	(127)	2311	(154)	2013	(128)
/iː/	303	(56)	1871	(138)	2638	(292)	2254	(200)
/ɪ/	341	(40)	1752	(133)	2400	(210)	2075	(151)
/oː/	392	(57)	925	(192)	2238	(181)	1584	(116)
/ɔ/	512	(61)	1056	(94)	2340	(194)	1700	(87)
/uː/	332	(52)	1028	(208)	2182	(145)	1604	(127)
/ʊ/	400	(64)	1032	(175)	2239	(169)	1636	(107)
/yː/	308	(46)	1560	(144)	2102	(137)	1830	(113)
/ʏ/	371	(39)	1356	(132)	2152	(122)	1752	(63)

B Appendix: EUROM-1 passages

Hallo, ist dort der telefonische Bestelldienst? Bei Ihnen scheint ein Fehler passiert zu sein. Ich hatte einen Römertopf aus dem Katalog bestellt, und es wurde mir ein elektrischer Rasenmäher berechnet. Dabei habe ich noch nicht mal einen Garten. Können Sie mich mit der zuständigen Stelle verbinden?

Kannst du mir sagen, was heute Abend im Fernsehen kommt? Ich hätte Lust auf etwas Leichtes und Amüsantes. Vielleicht kommt ja eine alte deutsche Komödie oder so etwas. Damals wurden die Filme natürlich anders gemacht als man das heute tun würde. Wer weiß, was die Leute von damals von unseren modernen Filmen halten würden, wenn sie sie sehen könnten.

Meine Frau braucht für den nächsten Monat gut aufeinander abgestimmte Bahnverbindungen. Könnten Sie mir bitte die jeweils günstigste Möglichkeit heraussuchen? Sie muss im Januar zu einer Reihe von Tagungen in Rom, Brüssel, Frankfurt, Köln und Mühlheim, jeweils von neun bis siebzehn Uhr. Würden Sie bitte Abendzüge und bequeme Hotels reservieren? Sie mag allerdings keine großen unpersönlichen Häuser.

Haben Sie die Möglichkeit, ein Essen für eine große Gruppe auszurichten? Wir müssen es so angeliefert bekommen, dass es sofort verzehrt werden kann. Es gibt nämlich keine Möglichkeit, das Essen in einem separaten Raum anzurichten. Wir denken da an belegte Brote, verschiedene Fleischsorten, Käse und Obst. Geben Sie eigentlich Rabatt für akademische Einrichtungen?

Geben Sie mir bitte Ihre Weihnachtsabteilung. Ich brauche vierundzwanzig Kisten Weihnachtsgebäck, fünf Schachteln gemischtes Gebäck und drei große Stollen. Die Anlieferungsadresse ist Apfelweg dreiundvierzig in Bielefeld. Können Sie mir garantieren, dass alles am dreiundzwanzigsten Dezember ankommt? Ich wäre ihnen dankbar, wenn Sie mir eine Bestätigung dieses Auftrags schicken würden.

C Appendix: Experiment items for perception test

Table 3: Experiment items: Extended list of experiment items of discrimination task between high and low surprisal with difference in surprisal (Diff.).

Word	**Fourgram preceding**		**Difference**
	high	low	
andere	wie für zahlreiche andere	Partei wie jede andere	0.99
bereit	jetzt nicht mehr bereit	<s> moskau sei bereit	1.33
deutschen	bischöfe an ihre deutschen	der bundesvereinigung der deutschen	2.50
führen	in vietnam zu führen	debatten im bundestag führen	2.06
gegen	sind vor allem gegen	verwahrte sich jedoch gegen	2.39
machen	zu stark zu machen	den garaus zu machen	4.91
polnischen	die botschaft der polnischen	dem brief der polnischen	0.66
teil	der sitzung nicht teil	an den verhandlungen teil	0.22
wieder	nach der wahl wieder	daß hin und wieder	2.88
zwischen	die vor allem zwischen	auf den zusammenhang zwischen	1.85