Observations on the dynamic control of an articulatory synthesizer using speech production data

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

vorgelegt von

Ingmar Michael Augustus Steiner

aus

San Francisco

Saarbrücken, 2010

Dekan:	Prof. Dr. Erich Steiner
Berichterstatter:	Prof. Dr. William J. Barry Prof. Dr. Dietrich Klakow
Datum der Einreichung:	14.12.2009
Datum der Disputation:	19.05.2010

Contents

Acknowledgments

My gratitude is due to many people who contributed to the success of this endeavor, notably to

- my advisor, William Barry, for his advice, support, and patience;
- Korin Richmond, for his close collaboration and extensive support;
- Peter Birkholz, for VocalTractLab, without which this thesis would not have taken its shape, as well as for support and granting me access to the source code;
- Sascha Fagel, Susanne Fuchs, Jörg Dreyer, and Phil Hoole, for generously providing me with their EMA data;
- `mngu0`, for providing his voice and vocal tract, and for unflinching endurance in the face of cruel and unusual recording techniques;
- Ian Marshall, for his support in making the MRI scanning a reality;
- Veena Singampalli and Philip Jackson, for help with the ACIDA code;
- Rob Clark, for supervising me during my first visit at CSTR;
- Hagen Fürstenau, for advice on mathematical notation and pythonic support;
- Eva Lasarcyk, who introduced me to VTL;
- Christoph Clodo, for amazing IT support;
- for helpful discussion, Bistra Andreeva, Alan Black, Stefan Breuer, Nick Campbell, Olov Engwall, Friedrich Faubel, Zhang Le, Simon King, Dietrich Klakow, Jacques Koreman, Bernd Kröger, Yves Laprie, John Ohala, Blaise Potard, Chao Qin, Jim Scobbie, Marc Schröder, Jürgen Trouvain, Alan Wrench, Junichi Yamagishi, and too many others who also deserve to be mentioned, but aren't.

For their various roles in providing me with funding over the years of my PhD studies, I am indebted to William Barry, Matt Crocker, Wolfgang Hess, Dietrich Klakow, Winfried Lenders, Manfred Pinkal, Steve Renals, Korin Richmond, Petra Wagner, Maria Wolters, and probably numerous others behind the scenes.

This work was funded by the DFG International Graduiertenkolleg 715 *Language Technology and Cognitive Systems* and Marie Curie Early Stage Training Site EdSST (MEST-CT-2005-020568).

And above all, thank you, Darja, for everything. This is for you!

List of Figures

List of Tables

List of Listings

Zusammenfassung und Überblick

Die Phonetik als Wissenschaft von der gesprochen Sprache steht vor einem ungewöhnlichen Problem: ihr Untersuchungsgegenstand läßt sich weder sehen noch greifen; ferner ist er weder zu sezieren, noch abzubauen oder ins Weltall zu schießen. Andererseits handelt es sich bei gesprochener Sprache auch nicht um ein abstraktes Modell, eine Annahme oder eine Theorie, die es mittels Kreidetafel oder Großrechner zu entwickeln oder lösen gilt. Das gesprochene Wort bildet die *Ausdrucksform* der menschlichen Sprache, und obzwar die Erforschung ersterer phonetische Phänomene beschreibt und nachbildet, besitzen diese Phänomene keinerlei Bedeutung, wenn sie nicht *beobachtet* werden können.

Wie allgemein bekannt, läßt sich gesprochene Sprache aufzeichnen und beliebig oft wiedergeben, und diese Möglichkeit besteht seit über 150 Jahren (Scott de Martinville, 1857). Allerdings stellt eine solche Aufnahme lediglich ein einziges Beispiel für gesprochene Sprache dar und besteht aus einer unfaßbaren Vielzahl winzigster Ereignisse und Vorgänge, die nicht voneinander getrennt untersucht werden können. Aus der Betrachtung einer einzigen Aufnahme kann keine wissenschaftliche Erkenntnis gewonnen werden.

Eine Lösung dieses Problems besteht in der Analyse großer Mengen von Aufnahmen in Form von *Korpora* gesprochener Sprache, sowie in der Zuhilfenahme empirischer Methoden und statistischer Ansätze, um daraus Belege für (oder wider) Theorien oder Hypothesen zu gewinnen. Auch wenn Zahlen vielleicht nicht lügen, ist auch hier wiederum jede bedeutsame Beobachtung untrennbar von zahllosen anderen Phänomenen durchwirkt, und jede Aufnahme ist einzigartig.

Ein anderer Ansatz wird gelegentlich gewählt, nicht selten mit fragwürdiger Rechtfertigung, indem Sprachaufnahmen nicht etwa in unabhängig erstellten (aber dennoch gegebenenfalls unausgewogenen) Korpora untersucht werden, sondern phonetische Phänomene in möglicherweise unnatürlicher Umgebung mit der klaren und häufig offensichtlichen Absicht erzeugt und aufgezeichnet werden, das Auftreten entsprechender Erscheinungen zu beobachten, bisweilen mit unbeabsichtigt zirkulären Ergebnissen.

Einen dritten und eigenen Forschungsweg bildet die *Analyse durch Synthese*, die künstliche Erzeugung gesprochener Stimuli, etwa mit dem Ziel der Beobachtung ihres wahrgenommenen Eindrucks auf Versuchspersonen unter kontrollierten Bedingungen. Schließlich handelt es sich bei den Versuchspersonen um Erzeuger und Benutzer von Sprache, und letztendlich sollte eines der Hauptziele phonetischer Forschung die Untersuchung gesprochener Sprache sowie ihres zwischenmenschlichen Gebrauchs sein.

Unter diesen Gesichtspunkten wurden und werden Methoden und Werkzeuge entwickelt, die es dem Phonetiker erlauben, synthetische Sprache zu erzeugen, entweder von Null an oder durch Manipulation von Aufnahmen, und sie gestatten es, mit chirurgischer Präzision jene Parameter zu kontrollieren, deren Funktion untersucht werden soll, während andere unverändert bleiben. Es muß jedoch ein Kompromiß bei der Auswahl solcher Hilfsmittel gefunden werden: Sollen die Stimuli von Grund auf erzeugt werden, so sollten sie dennoch wie natürliche Sprache klingen, ansonsten läuft man Gefahr, aufgrund ihrer Künstlichkeit den Eindruck der kontrollierten Variablen zu beeinflussen. Andererseits erhalten Stimuli, die aus natürlichen Sprachaufnahmen gewonnen wurden, teilweise oder ganz ihre ursprünglichen Eigenschaften (seien es Hintergrundgeräusche, Timbre, Geschlecht oder Herkunft des Sprechers, oder unzählige andere Faktoren), und ihre Manipulation kann leicht zu Störungen führen. (Diese Zweiteilung taucht noch einmal kurz in Kapitel 1 auf.)

Die vorliegende Arbeit ist aus dem Verlangen heraus entstanden, zu denjenigen Werkzeugen beizutragen, die die Simulation des Sprechens ermöglichen, indem sie die Sprechwerkzeuge in vollem Umfang nachbilden und dabei die grundsätzlichen Mittel bereitstellen, natürliche Sprache nachzuahmen, oder sogar noch grundlegendere Vorgänge der Spracherzeugung zu modellieren, die sich nicht durch oberflächliche Betrachtung untersuchen lassen. Diese *artikulatorischen Synthesesysteme* werden seit Jahrzehnten entwickelt – tatsächlich wurde bereits im ausgehenden achtzehnten Jahrhundert von von Kempelen (1791) eine „sprechende Maschine“ konstruiert und vorgeführt – aber ausgeklügelte Modelle, die in der Lage sind, realistische Äußerungen hervorzubringen, sind erst seit wenigen Jahren Dank der phantastischen Fortschritte in der Computertechnik möglich.

Eines der fortschrittlichsten verfügbaren artikulatorischen Sprachsynthesesysteme ist VocalTractLab (VTL)[1] (Birkholz, 2006), das aus einem dreidimensionalen geometrischen Vokaltraktmodell besteht, welches an die Anatomie eines echten Sprechers angepaßt werden kann, einer akustischen Komponente, die in der Lage ist, lebensechte Äußerungen zu erzeugen, sowie einer Steuerungsebene, die sich intuitiv verwenden läßt, um die Äußerungen einzuprogrammieren. Allerdings beschränkt sich die Synthese zur Zeit auf das Deutsche, und der Erfolg in der Bedienung von VTL hängt wesentlich von der Fachkunde des Benuzters ab. Obgleich diese Kenntnisse nicht überflüssig gemacht werden sollen, wäre es dennoch wünschenswert, eine geeignete Erleichterung der Steuerung des Systems zu bieten, sowie die Möglichkeiten des Benutzers so zu erweitern, daß ein Vergleich der Steuerungsanordnung mit den entsprechenden Parametern in natürlicher Sprache gezogen werden kann. Letzterer Gedanke bedarf womöglich einer Erläuterung: Bei den Parametern der Spracherzeugung handelt es sich um diejenigen, die in der Lage sind, auf angemessene Weise die artikulatorischen Bewegungen im menschlichen Vokaltrakt beim Sprechen zu beschreiben und vorherzusagen, was geeignete Beobachtungs- und Vermessungsmethoden voraussetzt.

Das erste Kapitel dieser Arbeit gibt einen kurzen Überblick über die verschiedenen möglichen Sprachsynthesetechniken, wobei die artikulatorische Synthese den übrigen Ansätzen gegenübergestellt wird, die den Schwerpunkt auf die Akustik des Sprachsignals legen, und nicht auf die zugrundeliegenden artikulatorischen Vorgänge. Auf diesen Über-

[1] `http://vocaltractlab.de/`

blick folgt eine vertiefende Beschreibung des VTL-Synthesizers, der in der vorliegenden Arbeit zum Einsatz kommt, so daß sich der Leser mit dessen Eigenschaften und Aufbau vertraut machen kann.

Kapitel 2 bietet eine Umschau der heutigen instrumentellen Techniken, die Beschaffenheit der Sprachproduktion auf unterschiedliche Weise unmittelbar zu beobachten; zu diesen gehören bildgebende Verfahren der Medizin, solche der Bewegungserfassung, und einige mehr. Dieses Kapitel schließt den Einführungsteil der Dissertation ab.

Damit ist die Bühne frei für Kapitel 3, das die Grundlagen dieser Dissertation einführt: den Vorgang der *artikulatorischen Resynthese*, der das artikulatorische Synthesesystem VTL mit Sprachproduktionsdaten kombiniert, die durch direkte Bewegungsmessungen an den Artikulatoren mittels der sogenannten elektromagnetischen Artikulographie (EMA) gewonnen wurden. Diese Technik erzeugt Daten, deren Beschaffenheit sich zum Vergleich mit der Animation von VTLs Vokaltraktmodell mithilfe einer intuitiven Zwischenschicht anbietet.

Der Ansatz der artikulatorischen Resynthese wird unter Verwendung von Algorithmen der dynamischen Programmierung entwickelt und in den Kapiteln 4 und 5 zum Zwecke der synthetischen Sprachnachbildung an zwei mit EMA aufgenommenen Korpora erprobt, die Äußerungen von ansteigender Komplexität enthalten, was im *artikulatorischen Bereich* geschieht. Das Ziel ist die Nachahmung der ursprünglichen Äußerungen mittels VTL, und zwar so, daß sich die Dynamik des Vokaltraktmodells der des menschlichen Sprechers annähert, unter der Annahme daß, sofern gewisse Bedingungen erfüllt sind, die synthetische Ausgabe auch im akustischen Bereich dem Vorbild nahekommt. Anschließend werden die Ergebnisse erörtert, und im Abschlußkapitel wird eine mögliche Anwendung für den Resyntheseansatz umrissen.

Die Dissertation schließt mit vorläufigen Ergebnissen der Erstellung eines Korpus von Sprachproduktionsdaten ab, das aus kernspintomographischen Aufnahmen eines Englischsprechers besteht, und welches den Weg für künftige Arbeit ebnet, in der VTL ans Englische angepaßt werden kann.

Summary and overview

Phonetics, the study of speech, faces an unusual problem: the object of its research cannot be seen, or held; it cannot be dissected or mined or shot into orbit. On the other hand, speech itself is not an abstract model, conjecture or theory, and cannot be developed or solved on blackboards or using supercomputers. Speech is the *manifestation* of language, and while the study of speech describes and models phonetic phenomena, these phenomena must be *observed* to have any relevance.

Everyone knows it is possible to record speech and play it back repeatedly, as it has been for over 150 years (Scott de Martinville, 1857). However, such a recording represents but a single instance of speech, representing an unfathomable multitude of tiny events and processes that cannot be inspected in isolation. No scientific insight can be gained from studying only one recording of speech.

One solution to this problem is the analysis of large amounts of recordings in the form of speech *corpora*, and the use of empirical methods and statistical approaches to distill from them evidence for (or against) theories or hypotheses. While there may be some truth in numbers, once again, each recorded observation of interest is inseparably intertwined with countless other phenomena, and each instance is unique.

Another approach sometimes taken, and not infrequently with questionable justification, is to analyze speech not from independently obtained (but nevertheless possibly biased) speech corpora, but to elicit and record phonetic phenomena in a potentially artificial setting with the distinct, and often overt, intention of observing specimens of interest, occasionally with inadvertently circular results.

A third, and somewhat different avenue of research lies in *analysis by synthesis*, the artificial creation of speech stimuli, for instance with the purpose of observing their perceptual effect on human subjects in controlled experiments. After all, since these subjects (or more generally, people) are producers and consumers of speech, it must ultimately be one of the main motivations for phonetic research to study speech and its effects in human interaction.

With this in mind, methods and tools have been, and continue to be, developed to allow phoneticians to create synthetic speech, either from scratch or by manipulating recordings, which allows them to control with surgical precision those parameters whose function they wish to investigate, while keeping others constant. However, a tradeoff must be made when selecting these tools: if stimuli are to be synthesized from scratch, they must still sound like natural speech, otherwise their artificial nature may interfere with the impression of the controlled parameter. On the other hand, stimuli harvested from recordings of natural speech will preserve some or all of the distinct characteristics of their origin (be it in the form of background noise, the speaker's timbre or gender or extraction, or countless other factors), and manipulating them may well introduce additional artifacts. (This dichotomy will briefly resurface in Chapter 1.)

The present thesis has evolved out of a desire to contribute to those tools which allow the simulation of speech, by modeling the organs of articulation in all their complexity, and which by doing so, provide the most elemental means of imitating natural speech, or even more fundamentally, model processes of speech production which cannot be investigated by superficial observation. Such *articulatory synthesizers* have been developed for several decades – in fact a "speaking machine" was constructed and demonstrated in the late 18th century by von Kempelen (1791) – but sophisticated models which are able to produce realistic speech output have only recently become a possibility in the wake of the fantastic developments in computing technology.

One of the most advanced articulatory synthesizers available, VocalTractLab[2] (Birkholz, 2006) consists of a three-dimensional (3D) geometric model of the vocal tract, which can be configured to match the anatomy of a real speaker, an acoustic component capable of producing lifelike output, and a control interface that can be used to program its utterances in an intuitive manner. However, synthesis is currently limited to German, and success in operating the VTL synthesizer depends critically on the expertise of the user. While this knowledge should not become superfluous, it would be desirable to provide the means of facilitating the synthesizer's control in an appropriate fashion, and extend the operator's possibilities by allowing him to compare the control structures to the corresponding parameters in the natural production of speech. This last aspect may warrant explanation: the parameters of speech production are those which are able to adequately describe and predict the articulatory movements in the human vocal tract during speech, and this requires the means to observe and measure those movements.

In the first chapter of this thesis, a brief overview is presented of various available techniques for the synthesis of speech, contrasting articulatory synthesis with other approaches which place emphasis on the acoustics of the speech signal, rather than on the underlying articulatory processes. Following this overview, the VTL synthesizer, which is used in this thesis, is described in depth, providing the reader with some familiarity with its features and composition.

Chapter 2 provides a survey of the instrumental techniques available today to observe the nature of speech production by various direct means; these include several modalities from the fields of medical imaging and motion capture, as well as a few others. This chapter completes the introductory part of the thesis.

The stage is set for Chapter 3, which introduces the main premise of the thesis: a process referred to as *articulatory resynthesis*, which combines the articulatory synthesizer VTL with speech production data obtained by directly measuring the movements of the articulators using a method called electromagnetic articulography (EMA). This technique provides data in a form that lends itself well to comparison with the animation

[2] `http://vocaltractlab.de/`

of VTL's vocal tract model, in an intuitive interface.

The articulatory resynthesis approach is developed using dynamic programming algorithms and put to the test in Chapter 4 and Chapter 5 by applying it to the synthetic reproduction of speech from two corpora recorded with EMA, containing utterances of increasing complexity, in the *articulatory domain*. The object is to imitate the original utterances using VTL in such a way that the dynamics of the vocal tract model closely resemble those of the human speaker, with the assumption that if certain conditions are satisfied, the synthetic output in the acoustic domain will match the original as well. The results are then discussed, and a potential application for the resynthesis approach is outlined in the final chapter.

The thesis concludes with preliminary results from the creation of a corpus of speech production data consisting of magnetic resonance imaging (MRI) scans of an English speaker, paving the road for future work adapting VTL to English.

Chapter 1

Speech synthesis methods

Some assembly required.

Proverb[1]

This chapter gives a short overview of several speech synthesis methods, focuses on articulatory synthesis, and finally discusses one articulatory synthesizer in particular.

1.1 Text-to-speech

Text-to-speech (TTS) can be loosely defined as a process by which written text is transformed into spoken text that can be understood by a person who may not have access to the original written input. The ubiquity of text in modern life presents a significant obstacle for anyone who is unable to utilize it, either temporarily (e.g. while driving or using a traditional telephone) or permanently (e.g. due to visual impairment or illiteracy), and so TTS systems are deployed in a wide range of settings for the benefit of their users. Detailed descriptions of TTS can be found in numerous publications and textbooks e.g. Klatt (1987); Dutoit (1997); Taylor (2009).

In discussions of TTS, it is sometimes neglected that the *input* (i.e. text) is not a perfect, unambiguous message, but rather a textual representation into which that

[1] The Greek word σύνϑεσις (*synthesis*), could be translated as "composition" or "assembly". The photo shows a LEGO® Storage Tray Unit (Item #851917), found at `http://www.1000steine.com/brickset/images/851917-1.jpg`.

message has been encoded, and that this encoding does not necessarily preserve the intended meaning, or certain other information available to the listener of a spoken utterance. Much of this meaning is normally conveyed over the speech channel by making use of prosody, which (whatever its exact nature) is not as straightforward to put into writing as the sequence of words that are spoken.

Consequently, there is much more to TTS than simply stringing together the correct pronunciation of the words in the input, which itself is not trivial; common problems include the disambiguation of homographs, the "normalization" of abbreviations, numbers, and other complex entities such as Uniform Resource Locators (URLs) or addresses, the resolution of out-of-vocabulary (OOV) tokens such as proper names not available in any lexicon, as well as the realization of multilingual material.

In addition to these issues, most of which can be solved at the segmental level by predicting the correct sequence of speech sounds, the greater challenge lies in correctly realizing more subtle phenomena, such as intonation, speech rhythm, expressivity, voice quality, or accent. These challenges transcend the somewhat ill-defined boundary between intelligibility and naturalness, two goals that are often used to describe the "quality" of TTS systems.

With the increasingly successful tackling of such problems, another bottleneck in TTS becomes apparent: the flexibility of acoustic output. Several techniques for waveform generation have been developed over the years, and exhibit different strengths and weaknesses in the production of audible output from the internal representation of a target utterance predicted by a TTS system. The main types of these waveform generation "back-ends" will be briefly described here.

1.1.1 Formant synthesis

Despite groundbreaking work by e.g. Chiba & Kajiyama (1941), it is usually Fant (1960) who is credited with presenting an *acoustic theory of speech production* in his eponymous monograph. By describing speech in terms of acoustic processes, a *source-filter* model of speech is proposed in which the glottis and vocal tract are represented by a source signal and a filter, respectively, which shapes its acoustic spectrum.

This model is implemented in a synthesis technique referred to as *formant* synthesis[2] to produce artificial signals that acoustically resemble human speech. In such a system, formant filters in a sequential or parallel arrangement are used to boost or dampen resonant frequencies of the source signal (which may be periodic, or noise for the synthesis of fricatives). Examples of such systems include Klatt (1980); Allen et al. (1987) and more recently Carlson et al. (2002), but many others too numerous to mention here have been developed.

Since the many parameters which control the output of a formant synthesizer, and the processes by which these parameters change over time to create the impression of speech, are quite complex and difficult to control directly, formant synthesizers typically make use of a set of acoustic-phonetic rules, painstakingly defined after analyzing natural

[2] also called *parametric* synthesis

speech.[3] Such rules govern both the values for individual acoustic parameters, as well as the dynamics, e.g. the duration of acoustic segments.

Formant-based synthesis offers very high flexibility in the synthesis of segmental and prosodic phenomena, and features a very small memory footprint, as the speech signals are generated "from scratch", unlike concatenative techniques (see below). Its drawbacks, however, include the limited naturalness of its acoustic output (e.g. Keller, 2002), and the requirement to explicitly define rules for almost every aspect of the synthesis process; the nature of these rules does not necessarily resemble the phonetic phenomena they are designed to emulate.

1.1.2 Concatenative synthesis

The most natural-sounding speech that can be artificially reproduced is of course speech itself, or rather, recordings of natural speech. By segmenting and recombining, or *concatenating* snippets, or *units* of speech recordings, new utterances can be created; this could be described as an extremely crude form of speech synthesis. Depending on the domain of application and the quality of the speech data, such a concatenative approach at the phrase or word level may even be sufficient for some purposes.

In domains that require inventories of more than a few dozen units, however, a more advanced approach is required, one that is able to concatenate smaller recorded units, increasing the number of their possible combinations and thereby, the utterances that can be synthesized, ideally to the generative level of language itself.

Diphone synthesis

Despite possible naïve intuitions about the nature of minimal phonetic units, one consequence of phenomena such as coarticulation is that the traditional acoustic segment does not lend itself well as a unit in concatenative synthesis. The boundaries of such segments crucially depend on their phonetic surroundings, and splicing them out of context produces output of objectionable quality.

However, most phones[4] exhibit a stable portion, and by segmenting with the stable centers as boundaries, it is possible to create units consisting of two such half-phones each, called *diphones*. These have turned out to be well-suited for concatenative synthesis, and the number of units in a typical diphone inventory is relatively small. While this concatenative approach is often referred to as *diphone synthesis*, it is also possible to concatenate units of different sizes (e.g. Portele et al., 1990; Stöber et al., 1999).

While the phonetic quality of the segments can be quite high in diphone synthesis, the output will be nevertheless barely intelligible, at best monotonous, unless suprasegmental phenomena are taken into account as well. This requires the application of signal processing techniques, such as the widely used pitch-synchronous overlap-add (PSOLA)

[3] for this reason, formant synthesis is also known as *synthesis-by-rule*

[4] with the exception of glides or transient consonants such as stops, but these can simply be modeled as two units

algorithm (Moulines & Charpentier, 1990), which allows the independent manipulation of segmental duration and fundamental frequency (F_0).

While popular diphone synthesis systems such as MBROLA (Dutoit & Leich, 1993; Dutoit et al., 1996) make good use of available signal manipulation techniques, the naturalness of the originally recorded units of natural speech does suffer from the inevitable artifacts. As a result, while this kind of synthesis is largely capable of acoustically rendering the various prosodic targets specified by a TTS system while maintaining the segmental quality, the overall result is clearly perceived as unnatural.

Unit-selection synthesis

Since in computing, memory is not nearly as expensive as it once was, an alternative concatenative approach consists of exploiting large amounts of recorded speech (up to several hours at once). Such data is of course highly redundant in terms of coverage at the segmental level (i.e. there will be hundreds, if not thousands of tokens for most unit types), but if designed, recorded, and processed appropriately, this database represents an inventory containing many of the more elusive phenomena desirable for natural-sounding speech synthesis. While the annotation and indexing of large speech corpora can be automated using automatic speech recognition (ASR), pitch tracking algorithms, etc., manual correction of the inevitable misalignments and errors, although an arduous and expensive process, is crucial for maintaining the high quality of subsequent synthesis using that inventory. An overview of these corpus-based methods is given by e.g. Möbius (2000).

In contrast to diphone synthesis, the required units must be *selected* at synthesis time from the multitude of available instances in the speech database. For each target unit, all possible candidates are identified in the inventory and evaluated by applying a *cost function* to calculate their suitability for the present requirements of the utterance to be synthesized. The cost function normally considers both segmental and suprasegmental features, as well as constraints of joining them together, and will select the optimal unit sequence from the search space of available candidates.

Since its development (Sagisaka, 1988; Black & Campbell, 1995), unit-selection has enjoyed increasing popularity and is currently widely used. Its main appeal lies in the high perceived naturalness at all levels of speech, but this "suspension of disbelief" on the part of the listener can be broken when no suitable unit is found in the inventory and a sub-optimal unit must be selected instead (Möbius, 2003; Mayo et al., 2005).

Unit-selection's solution to the challenges of speech synthesis lies in delegating responsibility to the database. As long as appropriate units are available, the output quality is very high. Consequently, the probability of searching for unavailable units can be reduced by taking great care in planning and recording the speech data, and by limiting the domain (e.g. a unit-selection voice designed for a telephone banking dialog system, which performs very well in that setting, may well be entirely unsuitable in a screen reader for the blind). Moreover, many unit-selection systems try to avoid a potential compromise of naturalness by forgoing any form of signal manipulation altogether.

As a consequence, unit-selection synthesis may be able to produce very natural speech, but it takes massive effort to prepare data required to utilize its full potential. Even then, explicitly controlling prosody and more subtle phonetic parameters tends to be either impossible or accompanied by spurious signal processing and selection artifacts. For phonetic research, such behavior may be even less desirable than quality that is lower, but predictably and consistently so, as with formant or diphone synthesis.

1.1.3 Statistical parametric synthesis

A synthesis approach combining the advantages of formant synthesis with corpus-based techniques has been maturing in recent years and is referred to as *statistical parametric* synthesis (Falaschi et al., 1989; Black, 2006; Zen et al., 2007a, 2009). Using statistical models such as hidden Markov models (HMMs), which are trained on annotated corpora of natural speech, these systems generate parameter trajectories which approximate the dynamics of real speech. The target values for individual speech units are essentially obtained by averaging the instances of these units in the training data, and such states are smoothly joined to produce the output parameter trajectories.

The major drawback of statistical parametric synthesis is the unnatural voice quality caused by vocoding techniques, but various improvements to the glottal source modeling have been applied (e.g. Kawahara et al., 1999; Cabral et al., 2008). As these problems become solved, this approach promises excellent flexibility combined with highly natural speech dynamics; it is however, not phonetically transparent, and therefore of limited use to speech production research.

1.1.4 Articulatory synthesis

Potentially the most complex, but theoretically ideal, approach to speech synthesis is referred to as *articulatory* synthesis, and attempts to directly simulate speech production. Although articulatory approaches offer a phonetically and physiologically intuitive structure and, when implemented accordingly, flexibility by reproducing the full range of variability of human speech, enabling fully expressive synthesis, singing, laughing, and other phenomena, the challenge of realizing these phonetic and physiological effects in the model is accompanied by the question of how to control it.

Shadle & Damper (2001) and Whalen (2003) present optimistic discussions of the potential of articulatory synthesis, and Kröger (2007) gives a more recent survey. While articulatory synthesis approaches have been explored for over fifty years with various levels of success, several factors have only recently rekindled mainstream research into articulatory techniques. Apart from the obvious advances in computing technology, these include the availability of volumetric imaging of the vocal tract (cf. Section 2.1.2) and prolific expansion in the field of audiovisual synthesis, with applications in e.g. systems with embodied agents (“talking heads”).

Following Kröger (2007), most articulatory synthesizers can roughly be subdivided into three main components: an acoustic model, a vocal tract model, and a control model. The many articulatory synthesis systems developed in the past have adopted various strategies in implementing these components, depending on the primary focus of the corresponding research, and not all have attempted to provide an integrated model of all three. The following sections sketch a brief outline of the components and some of the past solutions.

Acoustic models

Like formant synthesis, the acoustic model in articulatory synthesis is founded on Fant's (1960) acoustic theory of speech production and the source-filter model. However, instead of manipulating acoustic parameters directly, the source signal is generated by a glottis model, while the filter is obtained by calculating the transfer function of a vocal tract model.

The glottis model can be a self-oscillating model (Ishizaka & Flanagan, 1972) or a parametric model of glottal area (Titze, 1984) or glottal flow (Fant et al., 1985).

The vocal tract is modeled as a tube with non-uniform cross-sections[5], and its acoustic properties can be calculated using a modified version of Webster's (1919) horn equation (e.g. Weibel, 1955). Based on the approximate measurements of the adult male vocal tract and its diameter at the widest constriction, as well as the nature of sound waves, it is assumed that no transverse sound waves will form in the tube for frequencies up to ∼5 kHz[6], and so most acoustic models economize by simulating the sound wave with one-dimensional propagation only, along the centerline of the vocal tract model; this view permits the application of an efficient algorithm (Kelly & Lochbaum, 1962) in numerous synthesizers (e.g. Rubin et al., 1981; Meyer et al., 1989; Kröger, 1998). Alternative models are also used, based on an electrical transmission line circuit analog (e.g. Stevens et al., 1953; Flanagan et al., 1975; Maeda, 1982; Birkholz, 2006), hybrid time-frequency domain modeling (Sondhi & Schröter, 1987) or 3D simulation of sound propagation (e.g. El-Masri et al., 1996; Matsuzaki & Motoki, 2000).

Vocal tract models

The vocal tract model in an articulatory synthesizer serves as the basis for the acoustic model; specifically, its shape determines the vocal tract transfer function. The shape of the vocal tract model for target configurations in turn can be obtained from real vocal tract measurements and articulatory data recorded during speech production (cf. Section 2.1). With the availability of volumetric imaging, the 3D shape of the vocal tract can be directly observed, which has facilitated the development of three-dimensional vocal tract models. However, there are different paradigms in vocal tract modeling.

Geometric models (e.g. Mermelstein, 1973; Coker, 1976; Rubin et al., 1981, 1996; Fant & Båvegård, 1997; Boersma, 1998) consist of a manually defined vocal tract geometry with a number of degrees of freedom (DOF), or *control parameters*, which guide the articulators into specific target positions, determining the target shape of the vocal tract model. For two-dimensional (2D) vocal tract models, several methods have been proposed to predict the cross-sectional area along the centerline (e.g. Sundberg, 1969;

[5] in most cases, the vocal tract is treated as straight, ignoring its bent shape (Sondhi, 1986)

[6] the length of a 5 kHz sound wave in air at sea level is approximately 6.86 cm

Beautemps et al., 1995). In 3D geometric models (Engwall, 1999; Birkholz, 2006), such extrapolation is no longer required. The definition of such a geometric vocal tract model and its control parameters is an arduous task, and the parameter values must be precisely adapted to the shape and dynamics of appropriate data; once configured, such models are phonetically very intuitive to control.

Statistical models (e.g. Maeda, 1990; Badin et al., 1998, 2002; Beautemps et al., 2001; Engwall, 2002) rely on techniques such as guided principal component analysis (PCA) to derive control parameters automatically from speech production data. Special care must be taken, however, to ensure that the parameters correspond to actual articulatory DOF.

Finally, 3D models can be derived directly from the measured 3D vocal tract shape, using techniques such as the finite element method (FEM). Such vocal tract models are commonly known as *biomechanical* models (e.g. Perkell, 1974; Wilhelms-Tricarico, 1995, 2000; Dang & Honda, 2004; Fels et al., 2006), and while they are capable of realistic simulation of articulatory elasticity and activation, the computational complexity is formidable, and currently not efficient enough for use in speech synthesis applications.

Control models

The challenge of dynamically moving from one vocal tract configuration to the next is critical to articulatory synthesis, and several models have been proposed. In particular, the modeling of coarticulation, the movements of articulators or control parameters to their targets (taking into account aspects such as inertia and deformation), coordination, and feedback are some of the questions that must be addressed.

Notable models dealing with motor commands and articulatory gestures include Kelso et al. (1986); Saltzman & Munhall (1989); Saltzman (2003); Kröger (1998). The latter also provides a detailed review.

1.1.5 Physical articulatory synthesis

When discussing “models”, be they of acoustics or the vocal tract, it has been hitherto tacitly assumed that these models exist only in virtual form, or even more abstractly, as sets of mathematical formulae interpolating between points expressed in Cartesian coordinates and the like. Such models are implemented as software, using programming languages or scripting scientific computing frameworks. The most visual of such models are merely geometric functions rendered as lines or surfaces (cf. Figure 1.1a) and rendered into pixels on a screen or sheet of paper.

The laws of physics do not apply in such models, except where explicitly implemented, and even then only in somewhat simplified terms. Among the consequences of this (many of them desirable and intended) are the facts that articulators can move as quickly or as slowly as required, the glottal source requires no air to resonate, processing time depends on central processing unit (CPU) clock rate instead of real time, and results can be perfectly replicated.[7]

[7] with the possible exception of floating point precision and explicitly random components

Conversely, where physical phenomena obviously fundamental to acoustics or to the mechanics of articulation are modeled, it is certainly no trivial task to formulate and implement them in such a way that they perform as expected. This is one of the reasons that articulatory speech synthesis, especially when a biomechanical vocal tract model is employed, is still an active field of research.

Following these preliminary remarks, this section briefly outlines a fundamentally different approach to speech synthesis, which consists of the construction of a (literally) physical model of the vocal tract. This paradigm faces a separate set of challenges, such as mechanical control of the glottis and articulators, and the acoustic properties of the structures and materials used.

An obvious advantage of physical articulatory synthesis is that physical processes, in particular acoustics, are real and need not be elaborately simulated. However, this benefit is more than matched by the issues that arise to be solved from a mechatronics perspective, and only very few instances of such systems have been built.

Where physical vocal tract models (even those capable of producing only static vowels) have been realized, it is nevertheless pleasing that they can provide yet more experimental validation for acoustic theories of speech production. For instance, Kitamura et al. (2009) and Arai (2006, 2009) have built such models, the latter specifically for educational purposes.

Von Kempelen's speaking machine

Von Kempelen's (1791) speaking machine represents the earliest fully documented device capable of producing sound perceived to be similar to those of speech. Played like a musical instrument, it is driven by a bellows to produce audible vibration in a reed-based voice box; this source signal is filtered by wooden resonators. Additional apertures can provide secondary resonators and turbulence noise to augment the mainly vocalic sound repertoire with nasals and fricatives, and a skilled operator can elicit utterances such as "mama" and "papa" (cf. however Jakobson, 1960).

The speaking machine was discussed from a modern technical perspective by Dudley & Tarnóczy (1950). Over the years, a number of replicas have been built (most recently van den Broecke, 1983; Nikléczy & Olaszy, 2003; Brackhane & Trouvain, 2008), but have achieved only moderate success at producing particularly lifelike vocalizations.

Waseda Talker

Built on groundbreaking work by Umeda & Teranishi (1965), the most advanced physical articulatory synthesizer takes the form of an anthropomorphic talking robot developed at Waseda University in Tokyo. Dubbed the Waseda Talker, it has been improved over several generations, from the WT-1 (Nishikawa et al., 2000) and WT-1R (Nishikawa et al., 2003) to the more recent WT-6 (Fukui et al., 2007) and WT-7 (Fukui et al., 2008). However, despite significant research effort, its capabilities as a speech synthesizer are currently still restricted to vowels and simple consonant-vowel (CV) syllables.

1.2 VocalTractLab in depth

The articulatory synthesizer VTL was developed by Birkholz (2002, 2006) at the University of Rostock with a remarkable attention to detail at every level. It is characterized by its 3D model of the vocal tract, and by a high-level gesture-based control interface, as well as by the quality of its waveform generation. VTL is capable of synthesizing consonants as well as vowels, but due to the strictly geometric nature of the vocal tract model, which does not take fully into account the mass, elasticity, and texture of real articulators, some consonants (such as vibrants) cannot be produced.

The VTL synthesizer can be broken down into three subsystems, the vocal tract model, the gestural model, and the acoustic model, each of which are described below.

1.2.1 Vocal tract model

Extending the vocal tract model proposed by Mermelstein (1973) into the third dimension, VTL's model comprises a number of geometric surfaces whose position and deformation represent the overall shape of the supraglottal speech production system. These surfaces are formed by sets of vertices in a three-dimensional coordinate system, which are assembled into the vocal tract model, as illustrated in Figure 1.1a. The DOF of this model take the form of a set of *control parameters* carefully defined to permit the flexibility required to produce a large number of speech sounds, while prohibiting shapes that violate the structural constraints of the human anatomy.

These control parameters, whose values vary over time during synthesis, are described below in more detail. The basic dimensions of the vocal tract geometry, on the other hand, represent anatomical constants and therefore remain static, but they are defined by a second set of parameters which can be configured based on measurements obtained from a real speaker. This *speaker adaptation* represents a separate procedure and is described in Section 1.2.4.

The values of both the control parameters and the anatomical parameters are stored in a speaker definition file (in eXtensible Markup Language (XML)[8] format); this file is loaded once by VTL, before synthesis. Listing 1.1 illustrates the structure of such a speaker file. The anatomical parameters are defined under the `anatomy` node, while the `phoneList` node contains a list of `phone` definitions, each of which sets the target value for each control parameter for the respective phone.[9]

The glottis model is based on Titze (1984) and is schematically illustrated in Figure 1.1b. During waveform generation (cf. Section 1.2.3), the glottis and the vocal tract models act as a source and filter, respectively.

The remaining components of the vocal tract model include the nasal cavities, which

[8] `http://www.w3.org/XML/`

[9] It is important to differentiate between phones as phonetic units of speech and `phone`s in the sense of target configurations of the vocal tract model. The latter merely represent a shorthand notation for specifying a target value for every control parameter at once and are mnemonically named in reference to the phonetic entities approximated by this target configuration. From a phonological perspective, they most closely resemble phonemic allophones.

```
<speaker>
  <anatomy>
    <palate>
      <!-- several points defined here -->
    </palate>
    <jaw>
      <!-- several points defined here -->
    </jaw>
    <lips />
    <pharynx>
      <!-- several points defined here -->
    </pharynx>
    <larynx>
      <!-- several points defined here -->
    </larynx>
    <param index="0" name="HX" />
    <param index="1" name="HY" />
    <param index="2" name="JY" />
    <!-- and so on; value ranges and neutral value for each vocal tract
        parameter defined here -->
  </anatomy>
  <phoneList>
    <phone name="a:">
      <!-- target values for each vocal tract parameter for phone [a:]
          defined here -->
    </phone>
    <!-- and so on; remaining phones in phoneset defined here -->
  </phoneList>
</speaker>
```

Listing 1.1: XML speaker definition file used by VTL. Only the main elements are displayed, others have been replaced by comments. The parameters are detailed in Table 1.1.

are modeled following Dang et al. (1994); Dang & Honda (1996), and the subglottal airways, whose resonances (Fant et al., 1972) are modeled based on Weibel (1963). However, these do not influence the geometry of the 3D vocal tract model.

Control parameters

Once a speaker definition has been loaded, providing both constant anatomical parameters and the default values for the control parameters, the static shape of the vocal tract model is available. Dynamic control of the vocal tract model is enabled by overlaying the control parameters on the basic geometry; they can assume different values at different times, and at any given point in time, they determine the exact shape of the vocal tract model.

The set of control parameters contains 25 elements, a selection of which are illustrated in Figure 1.2; a full list is given in Table 1.1, extended by six glottis parameters critical for the generation of the glottal source signal.

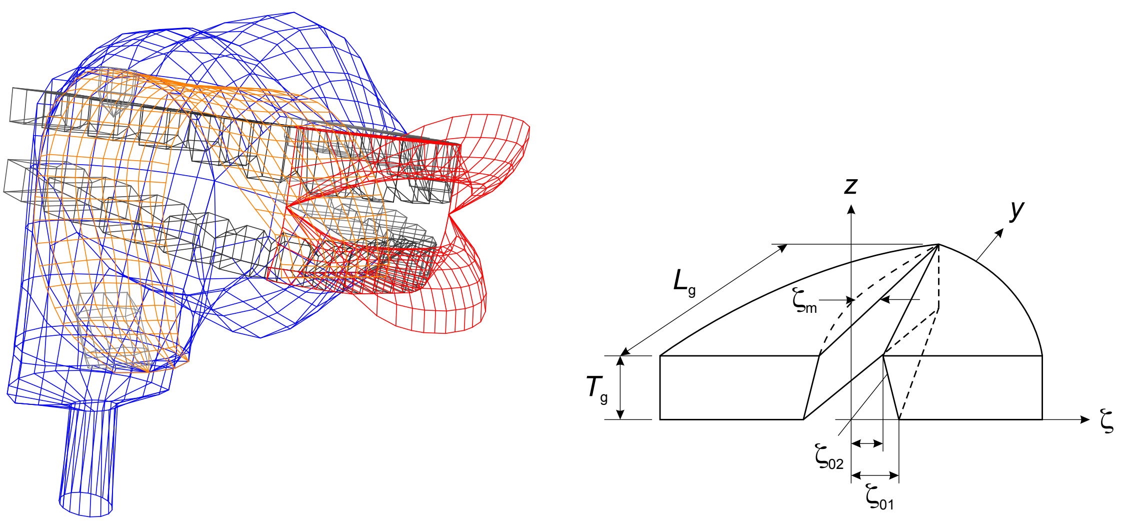
A crucial facet of the control concepts employed in VTL is that direct, low-level access of the control parameters governing the articulatory movements of the vocal tract is not directly exposed to the user in the interface. The rationale behind this design decision is that the separation of anatomy and gestural target configurations from dynamic control makes the high-level gestural control more manageable and prevents the user from deforming the vocal tract model in an unanticipated (and presumably undesired) way.

1.2.2 Gestural model

The dynamic control interface of VTL assumes the form of a *gestural score*, a concept taken from articulatory phonology (Browman & Goldstein, 1992a). Such a structure consists of a number of independent (i.e. autosegmental) *tiers* populated with discrete *gestures*, each of which represents movement toward a target configuration by the participating articulators, or, in the case of VTL, control parameters. These gestures are based on Saltzman & Munhall (1989); Kröger (1998).

At the onset of a gesture, each participating control parameter begins to move, accelerating smoothly, towards the specified target value (but never overshooting it). The exact position at any moment during this movement, and the time it takes to reach the target, are determined by Equation 1.1 (see below). This behavior is based on the target approximation (TA) model proposed by Xu & Wang (2001); Xu & Liu (2006) to model pitch movements. It remains to be investigated how successfully this model emulates the movement of physical articulators, since for instance it makes no provision for anticipatory movements or hyper-articulation (H&H theory, e.g. Lindblom, 1990), but the exact details of the equation could presumably be replaced by more realistic ones if required.

The function which models the parameter values takes several arguments, including the value and velocity at the offset of the previous gesture, and an *effort* variable, which influences the speed at which the new target value is approached. As a consequence of these variables and the duration of a given gesture, it is not uncommon for a parameter



(a) Wireframe view of 3D vocal tract model in VTL. The lips (red) are on the right, and two ridges of teeth can be seen towards the foreground. The tongue surface (orange) curves from the oral cavity (between the upper and lower cover drawn in blue) back down into the pharynx.

(b) Glottis model in VTL (from Birkholz (2006), based on Titze (1984))

Figure 1.1: Vocal tract model in VTL

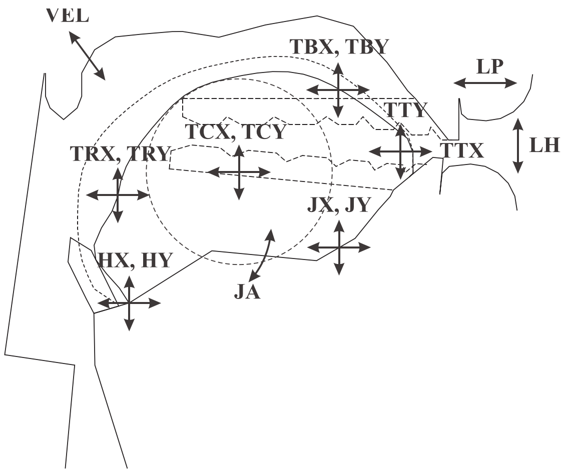


Figure 1.2: Selection of control parameters in the vocal tract model. Only those in the midsagittal plane are displayed. See Table 1.1 for explanation (from Birkholz et al., 2007b)

name	**description**
`HX`	hyoid bone, `x`,y position
`HY`	
`JX`	jaw, `x`,y position
`JY`	
`JA`	jaw aperture
`LP`	lip protrusion
`LH`	lip height
`VEL`	velic opening
`TCX`	tongue center, `x`,y position
`TCY`	
`TCRX`	tongue center, radius along `x`,y
`TCRY`	
`TTX`	tongue tip, `x`,y position
`TTY`	
`TBX`	tongue blade, `x`,y position
`TBY`	
`TRX`	tongue root, `x`,y position
`TRY`	
`TS1`	tongue side elevation at 4 points
`TS2`	
`TS3`	
`TS4`	
`MA1`	minimal area at 3 points near tongue tip
`MA2`	
`MA3`	
`F0`	fundamental frequency
`P_sub`	subglottal pressure
`x_top`	glottis width at top and bottom edge
`x_bot`	
`A_ary`	inter-arytenoid area
`lag`	phase difference between top and bottom glottis edges

Table 1.1: Names and descriptions of control and glottis parameters used in VTL (based on Birkholz, 2006). Note that the `TCR` parameters are treated as anatomy parameters. The `MA` parameters serve to guarantee a minimal area in consonant clusters (Birkholz, personal communication)

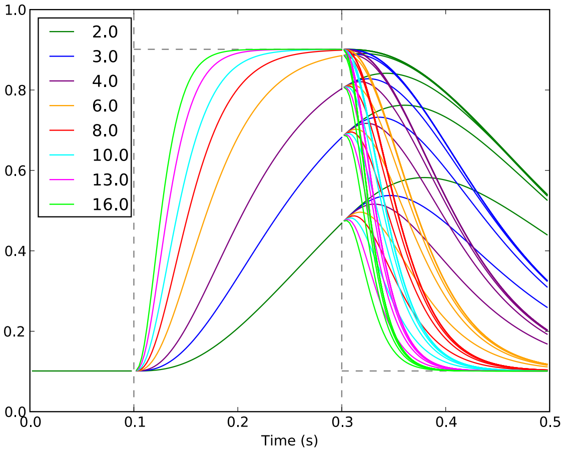


Figure 1.3: Example of one control parameter trajectory in VTL. Three gestures are shown, with the boundaries (vertical) and target values (horizontal) dashed in gray. Depending on which possible value from an arbitrary set (see legend) is chosen for each gesture's `effort` parameter, the resulting overall trajectory will assume different shapes.

to fail to reach a target value before the next gesture redirects it to a different target. This target undershoot corresponds to hypo-articulation in H&H theory.

With the exception of the `VEL`um (which has its own, dedicated tier), the vocal tract parameters are controlled not individually, but *en bloc*, through the elements of the `phoneList` as defined in the speaker file. These phone gestures serve as "macros" setting target values for all control parameters at once, and are placed on the `VOWEL` or `CONSONANT` tier, depending on their phonotactic role. A gesture on either of these two tiers can overlap in time with a gesture on the other, in which case the two gestures compete for control of the parameters. This competition is resolved by interpolation, and the respective weight of each competing parameter value is determined by a dominance value taken from the corresponding gesture's `phone` definition (Birkholz et al., 2006; Birkholz, 2007). This arrangement serves to implicitly model coarticulatory effects and is adapted from Öhman (1966, 1967); Gay (1977).

Variable articulatory `effort`

Gesture parameter trajectories in VTL are modeled by third-order cascaded systems, and the sample value y_i at time t in the i^{th} gesture can be calculated from Equation 1.1 (reproduced from Birkholz, 2007):

$$y_i(t) = (c_{1,i} + c_{2,i}t + c_{3,i}t^2)e^{-t/\tau_i} + b_i \tag{1.1}$$

Apart from the coefficients $c_{1,i}$, $c_{2,i}$, and $c_{3,i}$, which are obtained from the $(i-1)^{\text{th}}$ gesture to ensure continuity of articulator position and velocity, the freely definable parameters are the target position b_i and the speed at which the target position is

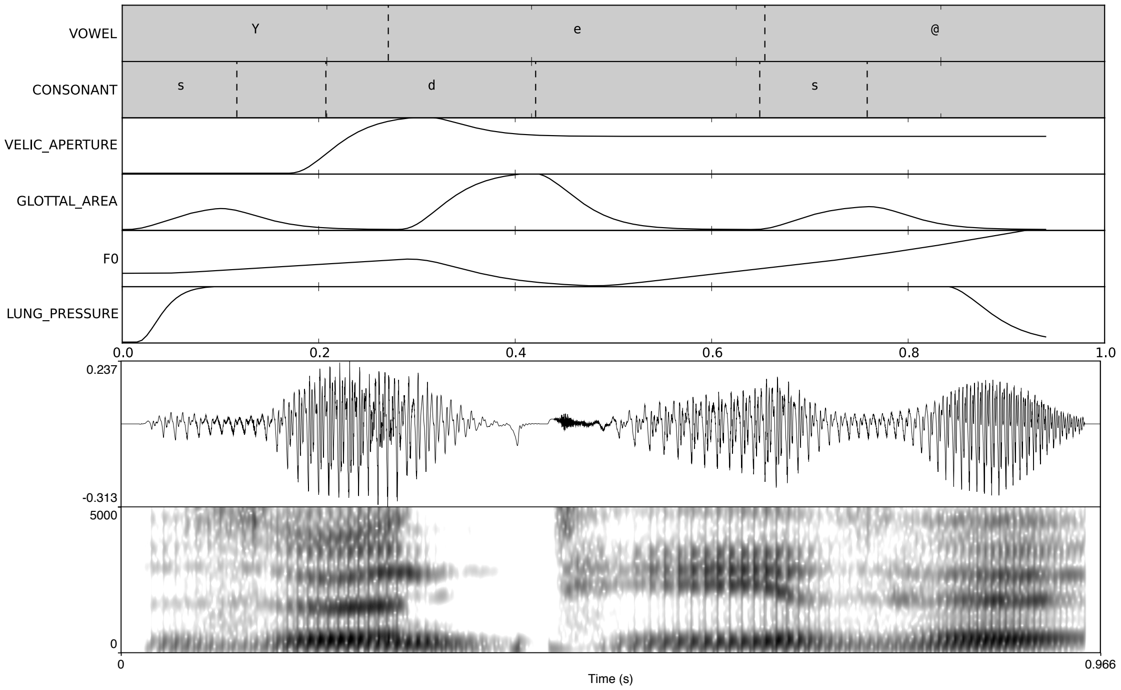


Figure 1.4: Example of a gestural score (top) and synthesis output (bottom). The utterance synthesized is the German word *Synthese* ("synthesis")

approached, the inverse of the time constant τ_i. This speed is used to model *articulatory* `effort`.

The result of setting the `effort` parameter to different values is illustrated in Figure 1.3.

Gestural scores and phonological discussion

To illustrate the dynamic control and the gestural score concepts used in VTL, one utterance will be discussed as an example, based on a score that was manually defined and synthesized. The gestural score and acoustic output are shown in Figure 1.4; the utterance consists of the German word *Synthese* [zʏnˈteːzə] ("synthesis"), with rising intonation.

In view of the canonical transcription of the target word, the names of the gestures on the `CONSONANT` tier (`s` and `d`) may demand some clarification. They could have been named differently, but this gestural score serves to highlight several of the characteristics of VTL's gestural model. Specifically, the velum is controlled by gestures on the `VELIC_APERTURE` tier, and is independent of the remaining control parameters. An apical occlusion (the `d` gesture) is therefore synthesized as a nasal in the presence of a velic opening gesture, and as a stop otherwise, or even (as in this example), as first one, then the other. Likewise, gestures on the `GLOTTAL_AREA` tier control the glottal parameters to produce phonation and other glottal settings; depending on the glottal gestures, an obstruent on the CONSONANT tier can be synthesized as voiced (the first s gesture) or voiceless (the second s gesture).

From a phonological perspective, gestures on the VELIC_APERTURE and GLOTTAL_AREA tiers determine the values of such distinctive features (e.g. Chomsky & Halle, 1968) as [±nasal] and [±voiced], respectively. In the context of a feature geometry such as that proposed by Clements (1985), the two tiers correspond to the SUPRALARYNGEAL feature [nasal] and the LARYNGEAL node, respectively; the CONSONANT gestures themselves are *underspecified* with respect to nasality and glottal setting. Of course, the feature geometry in turn contributed to the development of articulatory phonology, on which the gestural scores used in VTL are based; in this sense, VTL could be used to test certain phonological hypotheses, although this was probably not a primary motivation for its development.

Another aspect of the gestural model can be seen as well in this example; as a consequence of concurrent VOWEL and CONSONANT gestures, control parameters with a low dominance value in the definition of a CONSONANT phone are influenced more strongly by a VOWEL. In particular, the sibilant in the onset of the first syllable becomes rounded due to the simultaneous Y gesture's control of the lip parameters, while the second sibilant, during the schwa, remains unrounded. In this way, such coarticulatory effects are modeled elegantly and without the requirement for explicit commands.

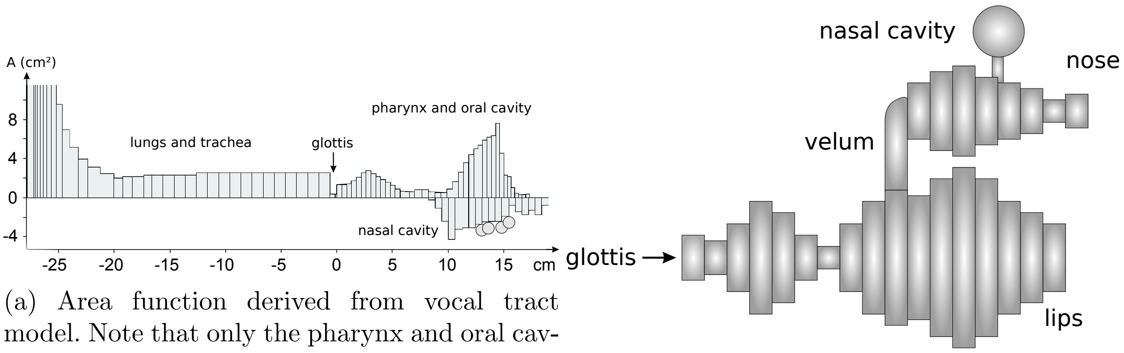
Gestural scores predicted by rules

An approach to the automatic generation of gestural scores was proposed by Birkholz et al. (2007b). Using the unit-selection TTS platform BOSS (Breuer, 2009; Breuer & Hess, 2010), which contains a durational component employing classification and regression trees (CARTs) (Breiman et al., 1984) trained on acoustically segmented speech, a working prototype of a TTS version of VTL was implemented. The acoustic-based onsets predicted by the CARTs were modified using a set of phasing rules (Browman & Goldstein, 1992a; Kröger, 1998), which moved the gestural onsets forward in time to the points where they would be expected to begin. The synthesis results of this rule-based prediction of gestural durations, however, were unsatisfactory, prompting further investigation such as that presented by the present thesis.

1.2.3 Acoustic model

The sophisticated aerodynamic-acoustical simulation used by VTL's acoustic model to generate waveforms will only briefly be outlined here. For technical details, the reader is refered to Birkholz & Jackèl (2004, 2006); Birkholz (2006).

For the purposes of acoustic synthesis, the vocal tract model can be interpreted as a complex tube whose non-uniform cross-sectional area is obtained by intersecting planes perpendicular to the curved centerline with the geometric surfaces of the 3D model. This yields the area function (Figure 1.5), which is sufficient to simulate sound traveling within the vocal tract modeled as a plane wave with one-dimensional propagation. Similarly to Ishizaka & Flanagan (1972); Flanagan et al. (1975); Maeda (1982), the length of the tube



(a) Area function derived from vocal tract model. Note that only the pharynx and oral cavity change dynamically

(b) Simplified schema of branched tube model used in acoustics model

Figure 1.5: Vocal tract area function and branched tube model in VTL (from Birkholz, 2006)

sections is variable, and the acoustic system is modeled using an electrical transmission line analog, and extended by consistent handling of the Bernoulli effect; damping and frication noise sources are added as well (Birkholz et al., 2007a).

1.2.4 Speaker adaptation

The creation of a speaker definition file may warrant further explanation. To derive anatomical measures, as well as the vocal tract target configurations for the `phoneList`, Birkholz (2006) used X-ray tracings of speech from a Russian speaker (Koneczna & Zawadowski, 1956), adapting the vocal tract model to the Russian phoneset. This was applicable to the synthesis of German only to a limited extent, and Birkholz & Kröger (2006) elaborate and extend the vocal tract adaptation process using MRI data from one male German speaker (Kröger et al., 2000, 2004). This section gives an overview of the procedure.

The first MRI corpus (Kröger et al., 2000; Kröger, 2000) contains volumetric scans of the speaker sustaining six long vowels of German, and in an initial step, the 3D vocal tract was configured to match the dimensions of the speaker. Special care was taken to model the dental ridges, whose configuration was obtained from a computed tomography (CT) scan of a plaster cast of the speaker's teeth.

The second MRI corpus (Kröger et al., 2004) consists of midsagittal scans of dynamic speech, acquired at 8 frames per second (fps); the speech data takes the form of repetitive production of the sequence [bVCV], with C ∈ {t, d, n, s, ʃ, l, x, ç, k, g, ŋ} in the vocalic context {aː, iː, uː}, as well as sustained realizations of these three vowels. Since the object of the dynamic scans was to acquire "snapshots" of intervocalic consonant production, only those frames were subsequently used that had been acquired during the stationary phase of consonant articulation.

The vocal tract contours were traced from the MRI scans using a combination of automatic edge detection and manual annotation. Since the two MRI corpora exhibited slightly different speaker postures, the tracings were warped, using a rotation point near the intersection of the tangent lines along the posterior pharyngeal wall and the roof of the palate parallel to the occlusal plane.

Using a dedicated adaptation component within the VTL graphical user interface (GUI), the MRI tracings were imported in scalable vector graphics (SVG) format[10], and overlaid with the vocal tract model contour. This allowed the vocal tract control parameters to be adjusted in a relatively quick manual process, fitting the model contour to each MRI tracing as closely as possible.

The first three formant frequencies of each vowel were subsequently improved in an iterative optimization routine, which attempts to match the formants measured in acoustic recordings of the target speaker, while changing the model parameter values as little as possible. This step further enhances the quality of vowel synthesis.

In a final step, the dominance values for consonant `phone`s were determined by adapting the vocal tract model to the tracings of the corresponding MRI scans for all three vocalic contexts, and comparing the values for the three versions of each `phone`. Those parameters that displayed resistance to cross-vocalic variation were judged critical to the production of the respective consonant and awarded a high dominance value, while the parameters exhibiting strong variation in the different contexts were considered less relevant and received low dominance values.

Overall, this adaptation procedure is obviously an arduous task and requires both appropriate data from a modality such as MRI and a concentrated effort to annotate it and adapt the vocal tract model to the tracings. On the other hand, the resulting speaker definition, combined with the vocal tract, gestural, and acoustic models of VTL certainly produce excellent results and present a valuable resource for speech production research.

[10] `http://www.w3.org/Graphics/SVG/`

Chapter 2

Speech production data

A variety of different techniques are in existence which are suitable for the observation and measurement of the vocal tract, as well as articulatory movements within the vocal tract during speech production. Several of these techniques are more commonly employed in the field of medical imaging, where they are referred to as *modalities*. This term will be adopted here to denote any method by which the shape or movement of the vocal tract or individual articulators can be measured, statically or dynamically, and by which corresponding data is produced.

This chapter presents a brief survey of available modalities for speech production analysis (an overview is given in Table 2.1), with a slight emphasis on those that are used in the remainder of this thesis. While a distinction is made between “full imaging” and “point-tracking” it would be correct to point out that the latter is technically a specialized version of the former, which reduces the dimensionality of raw data early in the acquisition process. Nevertheless, the distinction will be upheld to illustrate the practical repercussions of using data from these two modality groups for speech production analysis.

2.1 Full imaging

Depending on the modality employed, an object can be projected onto a plane in a 2D image, either by capturing visible light refracted off of the object (as in a photograph), or by accumulating rays passing *through* the object (e.g. X-rays). In each case, the resulting image is flat and can be sampled and processed digitally as a collection of *pixels*, each with a specific value (i.e. color, density). A number of such images can be acquired sequentially in time, becoming the frames of an animation.

In an alternative to projection, some modalities acquire a 2D *slice* through the object, and these slices can be sequenced as frames, or in parallel as a stack, in the latter case producing a volumetric representation of the object, where each slice’s pixels are processed as *voxels* with a third dimension.

In each of these cases, regardless of the form, the result is “raw” data in the sense that the entire field of view (FOV) is sampled equally, oblivious to the information it

modality	remarks
X-ray	static; 2D MIP; ionizing radiation
cineradiography	2D MIP; ionizing radiation
CT	static; ionizing radiation
MRI	static; no teeth
real-time MRI	2D; no teeth
UTI	2D; tongue body only
XRMB	2D; ionizing radiation
EMMA	2D
3D EMA	
optical tracking	lips only
EPG	tongue-palate contacts only
OPG	tongue only, prototype stage

Table 2.1: Overview of modalities for speech production analysis

contains. This also means that relatively large amounts of memory and processing power are required to store and manipulate the data.

Image data must therefore be *interpreted*; the information that is sought must be separated from the surrounding data by processes of *segmentation* or *annotation*. For quantitative analysis of speech production, the articulators must first be identified, and their shapes and positions extracted or measured. This process can be performed manually, but depending on the nature and amount of the image data, it may be prohibitively expensive in terms of time and effort. The alternative consists in automated approaches, which may not be as resilient to error as human experts normally are.

Such aspects must be borne in mind when employing imaging modalities for the analysis of articulatory movements.

2.1.1 X-ray and cineradiography

For most of the 20th century, the prevalent modality for investigating the internal shape of the vocal tract during speech production was projectional radiography, more commonly known as *X-ray*. While earlier studies used X-ray photography to produce still images, later work captures the dynamics of running speech by means of *cineradiography*. A detailed bibliography of this research, which includes such notable studies as Perkell (1969) and Fant (1970), is collected by Dart (1987). Furthermore, Munhall et al. (1994, 1995)[1] and Arnal et al. (2000) have collected and archived significant amounts of cineradiographic footage.

Since the articulatory contours are not always clearly visible in X-ray and cineradiographic data, numerous studies increase their visibility in the region of interest (ROI) by applying contrast agents such as barium sulphate paste to articulatory landmarks (Ericsdotter et al., 1999), or attaching reference markers (e.g. lead pellets) to the speaker's head

[1] `http://psyc.queensu.ca/~munhallk/05_database.htm`

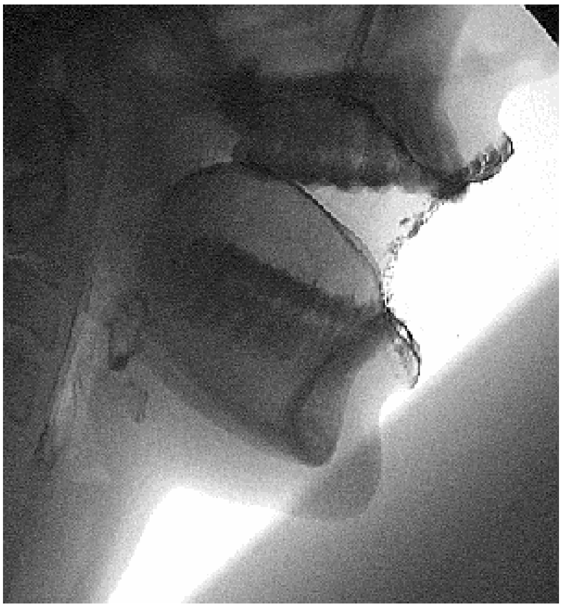


Figure 2.1: One frame of cineradiographic data from a female speaker. The teeth are clearly visible, as are the tongue and lips, which have been coated with a contrast paste (from `http://www.ling.su.se/STAFF/ericsdotter/projects/big_mouth.htm`)

or elsewhere in the frame. An example of such data is shown in Figure 2.1. These measures serve to facilitate subsequent processing and tracing of the articulator surfaces, originally by hand, but automatic techniques have been applied (Tiede & Vatikiotis-Bateson, 1994; Laprie & Berger, 1996; Thimm & Luettin, 1999; Fontecave & Berthommier, 2006). It must be borne in mind, however, that this transillumination presents a form of maximum intensity projection (MIP), flattening the three-dimensional structure of the vocal tract into 2D images. Despite the visibility of lateral tissue, the third dimension is essentially lost.

More recent projects recording new cineradiographic data include Branderud et al. (1998) and Connan et al. (2003), who make use of an advanced digital cineradiography facility in Strasbourg. Nevertheless, the exposure to ionizing radiation involved with X-ray and cineradiographic acquision prohibit these modalities for non-medical purposes, and consequently, they are no longer in use for speech production analysis.

Computed tomography

The limitations of the flattened projections can be overcome using computed tomography (CT). In this modality, a number of axial slices are acquired by rotating an X-ray camera around the ROI. The resulting scans are converted into a 3D representation of the subject's anatomy, which clearly shows all bones and soft tissue. A few studies (Perrier et al., 1992; Tom et al., 2001) have used CT for volumetric analysis of the vocal tract, and Story et al. (1998) compare CT and MRI scans of the vocal tract.

A major limitation of CT, however, is the fact that the subject must remain motionless for the duration of the scan. This makes CT unsuited for dynamic speech analysis, independent of the relatively high dose of ionizing radiation to which the subject is exposed, which again prohibits the use of this modality for non-medical purposes, especially for the collection of larger amounts of data.

2.1.2 Magnetic resonance imaging

Detailed technical introductions to MRI abound (e.g Hornak, 1996–2009; Wright, 1997), and so for reasons of brevity, this section will not attempt to provide another. Instead, only the bare essentials of the modality will be outlined, as far as they are relevant to MRI of the vocal tract.

In MRI, the subject of analysis (i.e. the speaker) is placed in a very strong homogeneous magnetic field[2], which causes the spin axes of atomic nuclei to align with the field's direction, or *polarize*. A second, transverse magnetic field is then temporarily created by a radio frequency (RF) excitation pulse, which causes the nuclear spin of atoms with an odd number of protons, such as the hydrogen in water molecules, to precess (or "wobble") at a frequency directly proportional to the strength of the static field. By applying a linear variation in the static field using *gradient* coils, the precession varies with location in the field. If the voltage induced in a receiver coil by the change in magnetic flux is measured, a complex nuclear magnetic resonance (NMR) signal can be obtained.

By applying a sophisticated combination of RF excitation pulse and gradient signal, jointly referred to as the *pulse sequence*, certain properties of the subject can be measured, for instance, the proton density at specific points in space, and this allows the three-dimensional, or *volumetric*, analysis of the human body, since different types of tissue have different chemical composition, and thereby hydrogen content.

For various technical reasons, a tradeoff must be made in MRI between the FOV, spatial resolution, and the time required to acquire a scan. Moreover, volumes are usually scanned as stacks of 2D slices, and a 3D MRI scan takes about as long as a sequence of 2D slices of the same area.

Since exposure to the magnetic fields used in MRI is generally considered harmless (Schenck, 2000), there are only a few practical limitations for this modality in the study of speech production. Potentially the most severe of these stems from the fact that MRI facilities are very expensive to install and maintain, and while they are available at tens of thousands of hospitals around the world, their use is normally restricted to medical applications. Nevertheless, MRI is rapidly gaining popularity for non-medical purposes, normally carried out as interdisciplinary research in cooperation with medical departments where such capacities are available.

Following early research applying MRI to the measurement of the vocal tract shape during vowel production (e.g. Rokkaku et al., 1986; Baer et al., 1987, 1991; Demolin et al., 1996; Lakshminarayanan et al., 1991; Story et al., 1996), as well as sustained consonant production (e.g. Narayanan et al., 1995; Masaki et al., 1996), the studies

[2] currently, field strengths of 1.5 T or (more recently) 3 T are common

using this modality for speech production analysis have become too many to enumerate, as MRI is becoming more advanced and widely available; Masaki et al. (2008) give a recent overview.

For the study of speech dynamics, it is of course necessary to acquire images of the vocal tract at a sufficient framerate. While the pulse sequence used for brain imaging (functional MRI) allows slices to be captured at over 30 fps, the FOV and heterogenous composition of the vocal tract (which contains substances as varied as bone, soft tissue, and air) is more difficult to scan quickly. Nevertheless, the benefits of “real-time”, or *cine*-MRI, are obvious and have motivated studies using rapid midsagittal MRI scans with framerates from as low as 4 fps (e.g. Demolin et al., 2000) to over 20 fps (Narayanan et al., 2004).

Moreover, a number of studies (Foldvik et al., 1993; Shadle et al., 1999; Masaki et al., 1999) have employed synchronized MRI scanning of repetitive speech utterances to animate dynamic 3D MRI sequences. And Stone et al. (2001) have used *tagged* cine-MRI to study the deformation of the lingual musculature during speech production.

In view of efforts to reduce the amount of data acquired without noticeably degrading the quality of the resulting scans (eg. Lustig et al., 2007; Lustig, 2008), the temporal and spatial resolution of MRI can be expected to improve further, and true 3D cine-MRI seems like a realistic possibility in the near future.

Acoustic recordings

The acoustic noise emitted by an MRI scanner during operation is very loud; sound pressure levels of over 110 dB are not uncommon. While various measures can be taken to reduce this noise somewhat (Hoiting, 2005), it nevertheless presents a significant problem for acoustic recording of speech *during* the MRI scanning. The fact that ferromagnetic materials and sensitive electronic equipment cannot safely be brought into proximity with the scanner’s powerful magnet compounds this problem even further.

Several studies (e.g. Bresch et al., 2006; NessAiver et al., 2006) have used noise-cancellation techniques with a fiber-optic (FO) microphone array to circumvent these problems and record synchronized audio during MRI scanning. In such a setup, one FO microphone is positioned close to the speaker’s mouth, while at least one other FO microphone is placed elsewhere in the scanning chamber, away from the speaker. The first microphone records a (very) noisy speech signal, while the second records only the scanner’s noise as a reference. The two recordings are then compared, and the noise from the reference recording “filtered out” of the speech recording, rendering it cleaner, but nowhere near studio quality.

Dental imaging and reconstruction

Another problem specific to MRI is the fact that air and bone (containing very little mobile hydrogen) are practically indistinguishable in the resulting scans. Bone-air transitions, specifically the teeth, are therefore invisible in most of the MRI data acquired during the scanning session. But of course visibility of the teeth is highly desirable for speech production studies.

One possible solution to this problem is to encase the teeth in an MRI-contrastive substance. This can be a pair of dental plates (either in themselves MRI-contrastive or containing a contrast agent; the plates are produced and fitted beforehand) such as those used by Wakumoto et al. (1997), or a non-toxic, malleable, adhesive contrast agent (e.g. peanut butter).

The problems with such methods are immediately apparent, however, since dental plates will interfere with articulation to a certain degree. Food-based adhesive contrast agents, on the other hand, will trigger increased salivation, forcing the speaker to swallow more often, which can cause some discomfort in a supine position, and will tend to interfere with the scanning procedure; additionally, the contrast agent itself will rub off, and dislodged amounts will interfere with articulation.

Another possibility is to fill thin, soft tubes with a contrast agent and attach them to the speaker's teeth (Ericsdotter, 2005). While this does not capture the dental surfaces, it does provide a landmark of sorts in the resulting MRI data and interferes no more with articulation than e.g. orthodontic braces would.

An alternative approach is to scan the teeth separately, obtaining a digital dental cast. Since the upper and lower teeth, respectively, are rigid, it is possible to reinsert this dental data into the MRI data of primary interest (i.e. speech production), a process known as *registration* (e.g. Gottesfeld Brown, 1992; Zitová & Flusser, 2003).

One such solution consists in the scanning of a dental cast. The subject has impressions of his teeth made using conventional orthodontic methods, resulting in a dental cast made of plaster or some equivalent material. This dental cast can then be scanned using any of the following modalities:

- MRI (most likely already available from scanning the vocal tract): the cast can be submerged in contrast fluid and scanned at a high resolution;

- CT: use of a contrast agent is not required, and since it is the cast (and not the subject) that is scanned, no harmful exposure to radiation is involved;

- laser scanning: a number of methods are available in orthodontics to optically scan dental casts; in fact, commercial 3D scanners are available specifically for this purpose from manufacturers such as 3Shape A/S[3], as well as more "homebrew" setups (e.g Adaškevičius & Vasiliauskas, 2008).

The advantage is that the unmoving, non-living, and portable cast can be scanned using methods that would be uncomfortable, hazardous, or simply impossible for the original subject. This way, a much higher-quality dental scan can be acquired.

Another solution is to scan the teeth *in vivo* using a contrast agent. Various contrast agents are commonly used in gastrointestinal MRI, where the subject swallows a quantity of the agent and the gastrointestinal tract, when filled by the agent, is then clearly visible

[3] `http://www.3shape.com/`

against the surrounding soft tissue (Low & Francis, 1997). In this context, contrast agents are compounds whose chemical composition gives either a strong (positive contrast) or very weak (negative contrast) signal in MRI.

Therefore, an alternative to manufacturing and scanning a dental cast is to scan the subject's teeth directly (in a separate scan) with MRI. The teeth are still "invisible" in the resulting images, but if they are tightly enclosed by a positive contrast agent, the dental surface can be reconstructed from the shape of that agent, i.e. the teeth-agent boundary (Olt & Jakob, 2004).

It is also possible to delineate the surface of the teeth by pressing the tongue and lips tightly against them. However, it may not always be possible to eliminate all pockets of air between the soft tissue and the teeth, which reduces the reliability of this method.

This problem can be minimized by using a liquid contrast agent. To this end, the subject lies prone (i.e. face-down) in the MRI scanner. Not only does this allow the subject to hold the fluid in the mouth (completely enveloping the teeth) for the required scanning duration, but also, any pockets of air will tend to rise to the surface of the contrast fluid, i.e. the back of the oral cavity, which enhances the quality of the dental surface in the resulting MRI data.

Although for this kind of scan, the oral contrast agent does not have to be ingested, it is common to use a food-based, and therefore completely harmless, agent to minimize discomfort for the speaker. A number of food-based contrast agents are known to provide improved contrast in MRI, among them blueberry juice, pineapple juice, and water (Hiraishi et al., 1995; Riordan et al., 2004).

2.1.3 Ultrasound tongue imaging

Ultrasound imaging has been used for decades in medical applications such as obstetric sonography or echocardiography, and is generally considered harmless to the subject. Non-line of sight (LOS) images of soft tissue are obtained in real time at frame rates from 29.97[4] to more than 120 fps, by holding an ultrasound probe to the surface of the imaged body.

The ultrasound probe contains one or more piezoelectric crystals, which emanate ultrasonic waves (in frequency ranges beyond 1 MHz), and which travel into the body, where they are reflected back to the probe by discontinuities in tissue density. A visual representation of the density along a planar FOV can be computed from the delay in the returning sound waves. Further details can be found in e.g. Hykes et al. (1985).

By adapting clinical ultrasound scanners to the requirements of tongue imaging, work by e.g. Sonies et al. (1981) has explored ultrasound tongue imaging (UTI) as a modality for speech production analysis. By holding the probe below a speaker's chin, the ultrasound waves are reflected from the tissue-air boundary[5] of the tongue surface, producing a 2D moving image of the tongue contour (although 3D UTI may soon become possible).

[4] the framerate of NTSC video

[5] or, in the case of palatal contact, the tissue-bone boundary (Epstein & Stone, 2005)

One limitation of UTI is the fact that air below the tongue tip prevents the tongue surface from being scanned at the tip, so the tongue contour is only visible from the dorsum to the blade. Another issue is the lack of visual landmarks to register the observed contour into a fixed frame of reference, a problem exacerbated by probe rotation in the sagittal plane during jaw or head movement. Possible solutions to this include restraining the speaker's head (e.g. Stone & Davis, 1995), rigid anchoring of the probe relative to the mandible using a helmet or headset (e.g. Scobbie et al., 2008), tracking the spatial positions of the probe and the speaker's head using optical (Whalen et al., 2005) or electromagnetic modalities (Aron et al., 2006), or a combination of such measures (Miller & Finch, in press).

Because of the relatively low information content (i.e. curves in video frames) in UTI, automated image processing is commonly used to extract the tongue contour for analysis (e.g. Parthasarathy et al., 2005; Li et al., 2005). A much more detailed overview is given by Stone (2005).

2.2 Point-tracking

An alternative data concept is the tracking of points by various means. These points are normally markers of some sort, which are attached to the fleshpoints of articulators whose movements are to be analyzed. The markers act as transducers of movement into signals (usually electrical voltages) which can be measured to obtain the position of the markers, and hence the fleshpoints, in space over time, producing *trajectories*. With an appropriate fleshpoint selection that captures the primary DOF of the articulators under investigation, these trajectories can be interpreted as a sparse representation of the motion and shape of the surfaces to which the markers are attached, facilitating the observation of the articulators themselves.

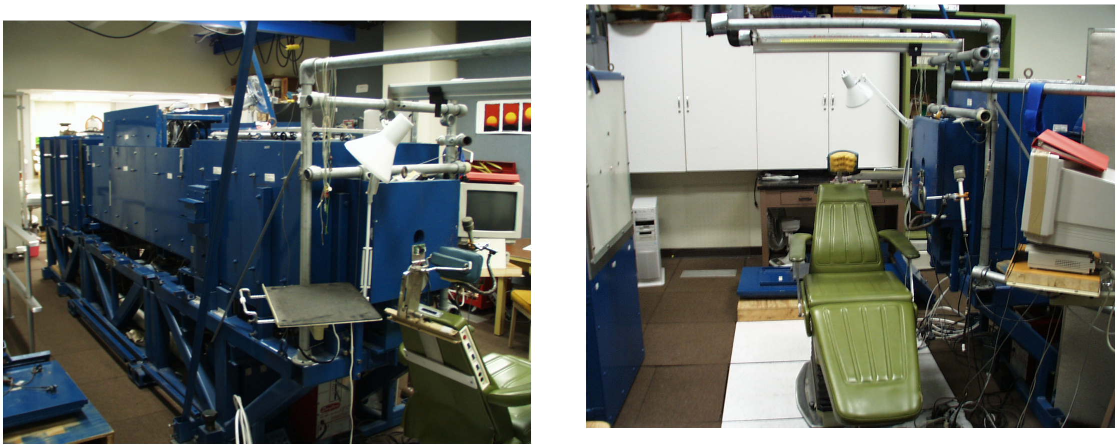
Point-tracking modalities have two main advantages over imaging techniques: they produce significantly smaller amounts of data, which speeds up the signal processing involved and permits much higher rates of acquisition, even with numerous trajectories in parallel; and secondly, the data obtained in this way can be interpreted as quasi-geometric, and therefore no manual segmentation or annotation is required to reduce its dimensionality for parametric analysis.

2.2.1 X-ray microbeam

The X-ray microbeam (XRMB) approach uses extremely thin X-ray beams (the eponymous microbeams[6], whose cross-section is no more than 36 mm^2) to track, through tissue and bone, the positions of pellets affixed to the inside of the speaker's mouth. The pellets are spherical and made of gold, giving them high contrast in the X-ray image.

Up to 12 pellets are used; in the standard layout, four are placed along the tongue (`T1-4`), two on the lips (`UL`, `LL`), two on the lower jaw (`MANi`, `MANm`), and three as reference points (two on the nose and one on the maxillary central incisors). The pellets are just

[6] occasionally abbreviated as *μbeams* or *ubeams*



(a) XRMB generator (blue). The subject chair can be seen in the foreground, on the right

(b) Position of the subject chair between the XRMB generator (right) and the detector (left)

Figure 2.2: The XRMB facility at the University of Wisconsin (Photos from the website of the Physical Sciences Laboratories at the University of Wisconsin-Madison, `http://www.psl.wisc.edu/projects/large/microbeam/more-microbeam`)

under 3 mm in diameter and are glued to the relevant articulators, but also attached to thin safety threads anchored outside of the speaker's mouth. While the pellets can be quite closely placed, a distance of 1 cm is advised to prevent overlapping trajectories, which could introduce errors in the prediction phase (see below) (Westbury, 1994).

The crucial difference between XRMB and cineradiography, which transilluminates the entire FOV, is that after an initialization scan to locate the pellets, only tightly focused X-ray microbeams are generated around the position of each observed pellet location. During pellet tracking, the previous observations are used to predict each pellet's next position, where the microbeam is then directed. A detector locates each pellet in the new local scans, yielding updated observed positions (Westbury, 1991).

The scan/prediction cycle is repeated at a rate of 720 Hz, but not every pellet is scanned at an equally high rate. In general, the sample rate for most pellets is 40 Hz, but the lower lip (`LL`) and posterior three tongue pellets (`T2-4`) are sampled at 80 Hz, while the anterior tongue pellet (`T1`), whose movements can be significantly faster than those of the other articulators, is sampled at 120 Hz (Westbury, 1994).

While the pellets need not be placed in the speaker's midsagittal plane, and a head held askew would be scanned at an oblique angle, the projection is always flattened to the image plane, and under these circumstances, a midsagittal placement will usually produce the most useful data.

XRMB background

An important fact of XRMB is that only a single facility exists worldwide, located at the Waisman Center at the University of Wisconsin, Madison. The equipment requires a voltage of 600 000 V and weighs about 20 000 kg; needless to say, it is immobile (see Figure 2.2). This unique installation was designed and built by the University's Physical Sciences Laboratories, from 1980 to 1985, and the first scan of a human subject was performed in 1987.

The facility at the University of Wisconsin is actually the second generation of XRMB; the "computer-controlled dynamic cineradiography" method was originally proposed and implemented by Fujimura et al. (1963, 1973) and Kiritani et al. (1975), who set up the first XRMB facility (built by the Japan Electron Optics Laboratory[7]) at the Research Institute of Logopedics and Phoniatrics at the University of Tokyo. This prototype was able to scan at an effective sample rate of 60 Hz and is apparently long retired (one of the last known studies is that of Kiritani et al. (1980)).

Because of this unusual situation, the term *XRMB* can, sometimes ambiguously, refer both to the modality itself and to the single site where it is employed. For a much more in-depth history of XRMB, in particular the details of the Wisconsin Facility, the reader is referred to Westbury (1994).

2.2.2 Electromagnetic articulography

This section presents a brief overview of electromagnetic articulography (EMA)[8] as a modality for transducing the movement of a set of fleshpoints. In general, such measurement devices require the subject to remain in a fairly small electromagnetic field, and the movements of fleshpoints on the speaker's articulators can be tracked within this measurement area, using small electromagnetic coils attached to thin, flexible wires.

Until the maturity of 3D EMA (see below) a few years ago, point-tracking was restricted to two dimensions, usually the speaker's midsagittal plane. Because of the necessity that the transducer coils remain in the measurement plane, the fixed transmitter coils are mounted on a helmet, and the speaker's head movements are restricted. A detailed overview of electromagnetic midsagittal articulography (EMMA) is given by Hoole & Nguyen (1997).

Permanent magnet systems

Jaw movement transducer systems, initially based on mechanical or optical signals (Lindblom & Bivner, 1966; Ohala et al., 1968; Sussman & Smith, 1970; Sonoda & Wanishi, 1982), led to the development of magnetic transducer systems, which use magnetometers to sense the distance to a small permanent magnet attached to the mandible (Radke, 1984). Since the magnetic field is not influenced by tissue or bone, such a system can also be used to track the movement *within* the mouth, e.g. that of the tongue (Sonoda, 1974; Sonoda & Ogata, 1992). In fact, commercially available jaw tracking systems (such as the K7/CMS Computerized Mandibular Scanner[9] produced by Myotronics-Noromed,

[7] `http://www.jeol.com/`

[8] also sometimes referred to as *electromagnetic articulometry*

[9] `http://www.myotronics.com/products/k7-evaluation-system/k7-cms-scanner.html`

Seattle, WA, or the JT-3D[10] from Bio Research Associates, Milwaukee, WI), which are used in orthodontics to diagnose temporomandibular conditions, can be used for this purpose (Dromey et al., 2006).

While the permanent magnets used with such devices can be wireless, it is not possible to measure the position of more than one or two magnets simultaneously (Sonoda & Ogata, 1993), so they must be combined with other modalities in order to capture the full range of articulatory dynamics (Ogata & Sonoda, 2001).

Early EMMA designs

Another approach more suited to the study of speech production employs electromagnetic fields alternating at different frequencies. In this scenario, several fixed transmitter coils are arranged to generate an inhomogeneous electromagnetic field. Small sensor coils placed within the field receive this electromagnetic radiation, which induces a voltage inversely proportional in strength to the cubed distance between transmitter and receiver coils. When this arrangement is properly calibrated (Kaburagi & Honda, 1997), the receiver coils can be fixed to articulators and act as transducers of articulatory movement.

This technique was pioneered at the University of Wisconsin, Madison, by Hixon (1971) with a single transducer[11] attached below the speaker's jaw, yielding trajectories of jaw aperture. Van der Giet (1977) designed and constructed a more elaborate device at the University of Bonn, introducing different frequencies for a fixed array of two opposing pairs of transmitter coils in a perpendicular configuration. This not only allows precise measurements (with an error margin of 0.1 mm) of the position of receiver coils in the midsagittal plane, but the tracking of several receiver coils simultaneously; van der Giet placed these coils on the upper and lower lip and jaw of two speakers, with an additional coil on the upper nose for reference, to correct for head movement.

Two-transmitter EMMA systems

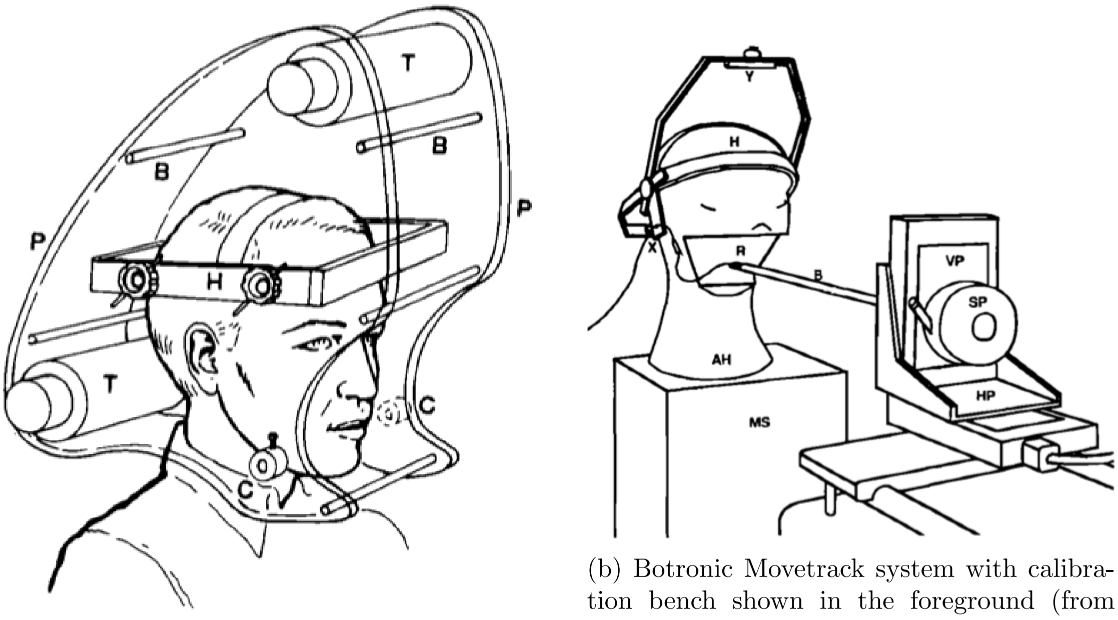
Perkell & Oka (1980) and their colleagues at MIT built a similar device using only two transmitter coils, and explicitly address the issue of rotational misalignment[12]. Using two transmitters, the simple triangulation of receiver coil positions in the measurement area requires the axes of transmitter and receiver coils to remain parallel. However, during speech, especially along the tongue, the axis of receiver coils can rotate to such a degree that the induced voltage is significantly reduced, resulting in measurement errors.

One solution to this problem is the design and use of biaxial receiver coils containing two ferrite cores arranged orthogonally (Perkell, 1982), which allow robust measurements of each receiver's position even in the presence of misalignment. The resulting system (depicted in Figure 2.3a) is capable of tracking up to 9 receiver coils in a 15 × 15 cm area, tolerating a deviation of up to 5 mm from the midsagittal plane (Perkell & Cohen,

[10] `http://www.biojva.net/products/jt_3d.php`

[11] Actually, Hixon's (1971) initial setup is inverse to that described above, as he uses fixed receivers and a single mobile generator coil, but the principle is the same.

[12] depending on the axis of rotation, this is referred to as either *tilt* or *twist*



(a) Two-transmitter MIT System (from Perkell et al., 1992)

(b) Botronic Movetrack system with calibration bench shown in the foreground (from Nguyen & Marchal, 1993)

Figure 2.3: Schematic drawings of the two-transmitter EMMA systems

1985), with measurement errors of less than 0.25 mm (Cohen & Perkell, 1985). It was extensively tested and can be used to collect substantial amounts of speech production data; an in-depth report is given by Perkell & Cohen (1986).

A similar device, but with standard receiver coils, was developed at the University of Stockholm by Branderud (1985) and subsequently marketed as Movetrack by Botronic in Hägersten, Sweden[13]. This system features a lightweight design (Figure 2.3b) and was evaluated by Nguyen & Marchal (1993), who found that despite its adequate precision (measurement error smaller than 0.4 mm) with up to 8 receiver coils, the lack of detection or correction of rotational misalignment can produce unreliable measurements.

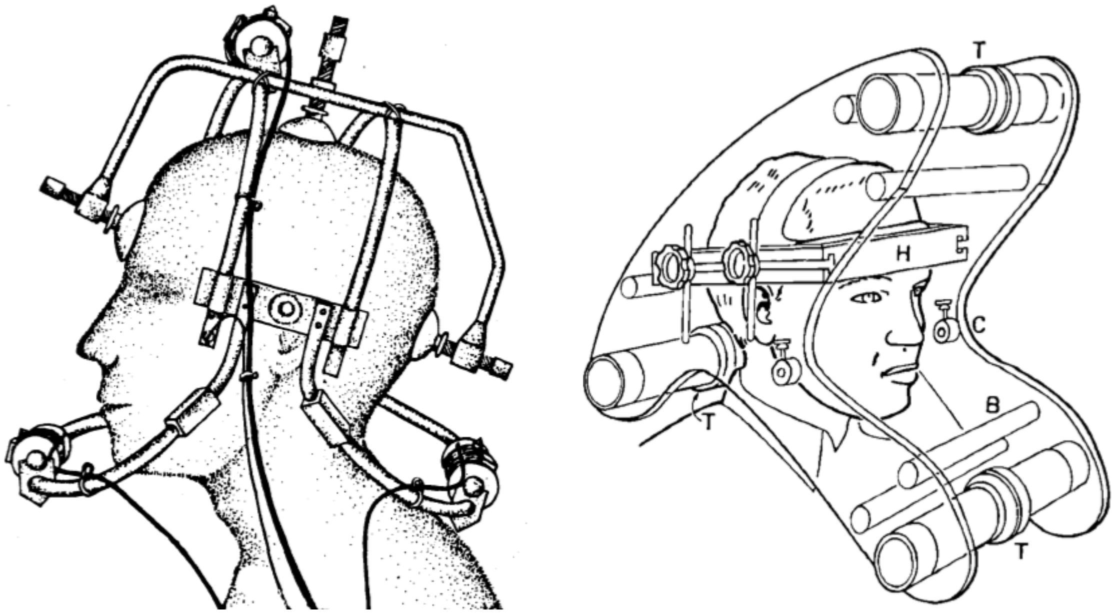
Three-transmitter EMMA systems

A different solution to the rotational misalignment problem was developed by Schönle et al. (1983, 1987) at the University of Göttingen, using single-core receiver coils, but adding a third fixed transmitter coil (see Figure 2.4a). The device was tested with German speakers by creating palate tracings and recording various CV syllables.

This three-transmitter system was further developed into the commercial Articulograph AG100, manufactured by Carstens Medizinelektronik[14] in Lenglern, Germany. It was evaluated by Tuller et al. (1990) who find a measurement error of less than 0.2 mm

[13] the company was apparently dissolved in 2008

[14] `http://articulograph.de/`



(a) Schönle et al.'s (1987) system, the precursor to the Carstens AG100 (from Höhne et al., 1987)

(b) Three-transmitter MIT System (from Perkell et al., 1992)

Figure 2.4: Schematic drawings of the three-transmitter EMMA systems

with 5 receiver coils. Hoole (1996) gives a detailed explanation of the calibration procedure. Along with its successor, the AG200, the AG100 has been installed at roughly 100 sites around the world[15], making it by far the most widely used EMMA system.

At MIT, Perkell et al. (1992) used this design to build a three-transmitter version of their previous system (Figure 2.4b), and systematically compare the performance of the two versions side by side. Their three-transmitter system has a slightly greater measurement error margin (0.4 mm), but the benefits of using lower electromagnetic field strengths, and cheaper, more robust receiver coils outweigh this drawback.

XRMB vs. EMMA

The tracking of points using XRMB and EMMA yields strikingly similar data, although the underlying technologies are fundamentally different. Both modalities measure the positions of a small number of marker points in the midsagittal plane during speech, at a sample rate high enough to capture virtually all articulatorily relevant movement. Byrd et al. (1995, 1999) compare XRMB and EMMA data (using the MIT system) from the same set of speakers and conclude that for all practical purposes, the results are essentially equivalent.

While both modalities require the transducer pellets to be attached to the speaker's

[15] according to `http://articulograph.de/useri.htm`

articulators, which may interfere with speech production to some degree (Weismer & Bunton, 1999), XRMB potentially causes less discomfort for the speaker than EMMA; the threads securing the pellets are thinner than EMA wires, and while subject head-movement is discouraged (and may require more initialization scans), the head is not fixed or restrained by a helmet of any sort (cf. Figure 2.2b).

Two major drawbacks of XRMB are, of course, the exposure to ionizing radiation, which is much lower than during cineradiography, but not negligible. In constrast, EMA is generally considered to be harmless (Perkell et al., 1992; Hasegawa-Johnson, 1998), even for patients with a cochlear implant (Katz et al., 2003) or implantable cardioverter-defibrillator (ICD) (Joglar et al., 2009).

The second, and perhaps more prohibitive, drawback is the fact that the Waisman Center at the University of Wisconsin is the only site in the world where this modality can be utilized, whereas EMMA systems are installed at many laboratories around the globe.

Under these circumstances, it is perhaps not surprising that the portion of speech production studies using XRMB has declined. Incidentally, it seems to have been not only the rise of EMMA that triggered the recession of XRMB, but administrative and funding issues caused by the sheer scale of the XRMB project. Nevertheless, it is debatable whether stronger support would have allowed XRMB to hold its ground against the alternative, but harm-free, and much less expensive[16], EMMA modality, which was already emerging even as the XRMB facility at the University of Wisconsin was under construction.

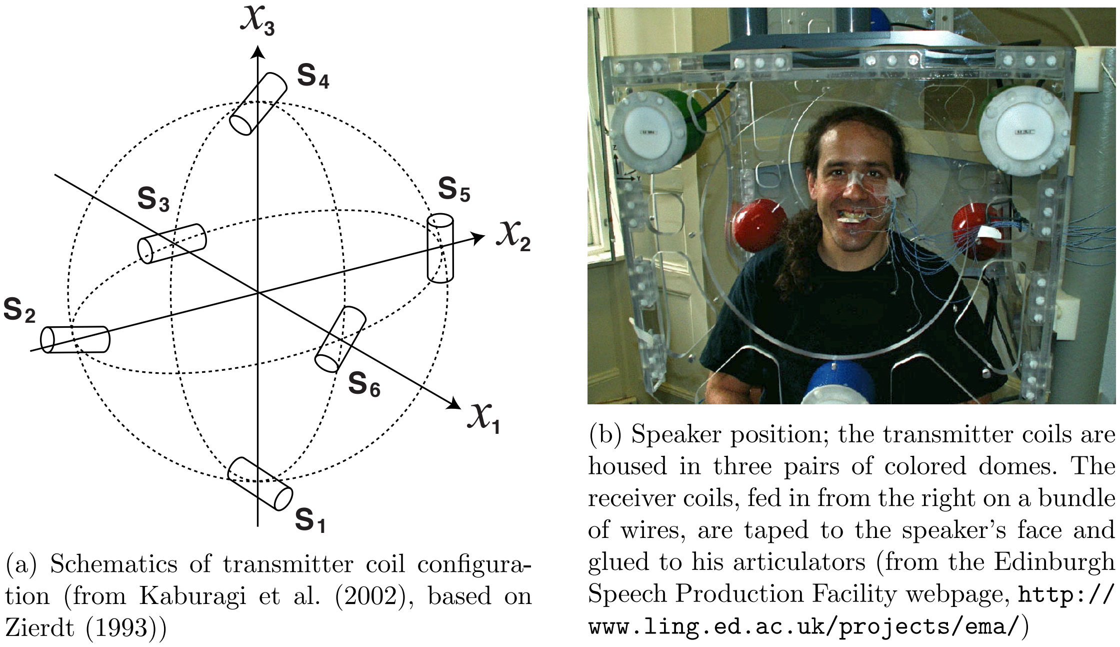
With the maturity of 3D EMA, both XRMB and EMMA can be considered superseded.

3D EMA

The main limitation of EMMA is the requirement for transducer coils to remain within the measurement plane, the violation of which can entail critical measurement errors. Consequently, the subject's head is firmly held in a rigid helmet, which in turn can cause some discomfort. In attempting to solve both of these issues, the possibility of three-dimensional EMA was explored as early as 1996.

Zierdt et al. (1999, 2000) developed a working prototype of such a 3D EMA system at LMU Munich, in close collaboration with Carstens Medizinelektronik. This device is constructed from similar components as the AG100/200, but the electromagnetic field forming the measurement volume is generated by six transmitter coils, arranged on the surface of a sphere at antipodal points, so that the lines from each transmitter to its antipode form the axes of a three-dimensional coordinate system (cf. Figure 2.5a). This arrangement ensures that each receiver coil within the sphere cannot be aligned perpendicularly to more than three transmitters at once, and thereby yields sufficient voltages to robustly recover its position; details of the measurement principle are given

[16] For comparison, the Carstens AG200 EMMA system last retailed for €52 800 (according to the manufacturer's website), while around $8 000 000 were spent on the XRMB grant over 13 years (Westbury, 1994).



(a) Schematics of transmitter coil configuration (from Kaburagi et al. (2002), based on Zierdt (1993))

(b) Speaker position; the transmitter coils are housed in three pairs of colored domes. The receiver coils, fed in from the right on a bundle of wires, are taped to the speaker's face and glued to his articulators (from the Edinburgh Speech Production Facility webpage, `http://www.ling.ed.ac.uk/projects/ema/`)

Figure 2.5: Transmitter coil layout and speaker position in Carstens AG500

by Kaburagi et al. (2005). The prototype was developed into the commercially available Articulograph AG500. As of 2009, over 80 units have been installed at sites around the world.[17]

The transmitter coils are attached to a cubic open frame made of acrylic glass, in which the speaker sits upright. While he is free to move his head (reference coils allow normalization for head movement), measurement accuracy is highest near the center of the measurement volume (Figure 2.5b). By virtue of the AG500's design, the receiver coils are sampled not only with respect to their position in three-dimensional space (three translational coordinates), but also to their orientation (two rotational coordinates), and these 5 DOF indicate that the AG500 is actually capable of tracking vectors, not points[18], which raises interesting possibilities for e.g. tongue surface tracking (Hoole et al., 2003).

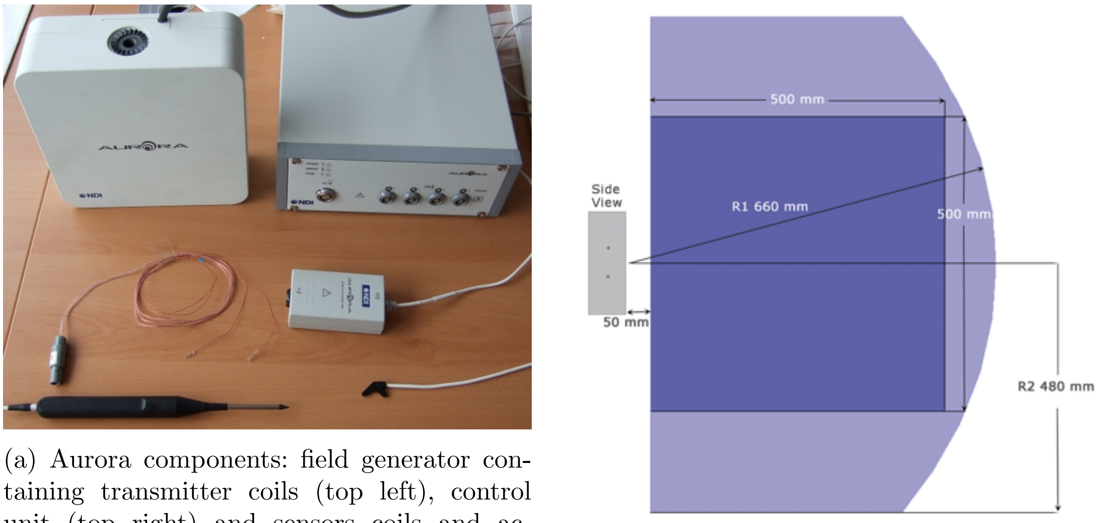
Calibration of the AG500 is not trivial (Zierdt, 2007), but the precision is adequate for speech production analysis (measurement error generally less than 0.5 mm) in a sphere of 15 cm radius around the center of the frame, and up to 12 receiver coils can be used (Kroos, 2008; Yunusova et al., 2009).

NDI systems

Electromagnetic tracking systems (EMTSs) used in image-guided surgery for non-LOS instrument tracking could also be used for point-tracking within the vocal tract, but

[17] again, according to `http://articulograph.de/useri.htm`

[18] this is sometimes referred to as *5D* EMA



(a) Aurora components: field generator containing transmitter coils (top left), control unit (top right) and sensors coils and accessories (such as the black palate trace instrument) in the foreground (from the LORIA MAGRIT website `http://magrit.loria.fr/Confs/Issp06/`)

(b) Aurora measurement volume, with the field generator on the left (from `http://www.ndigital.com/medical/aurora-techspecs-volume.php`)

Figure 2.6: NDI Aurora system

the requirements for temporal resolution and spatial accurracy, as well as for transducer design, are different, if not higher. Nevertheless, Northern Digital (NDI) of Waterloo, ON, Canada, have diversified by offering their Aurora EMTS as an alternative EMA system. According to the manufacturer's specifications[19], the Aurora system is capable of tracking up to 8 sensor coils with 5 DOF or 4 biaxial sensors with 6 DOF[20] within a $50 \times 50 \times 50$ cm cubic volume (cf. Figure 2.6b) at 40 Hz. The Aurora's performance for surgical applications was assessed with favorable results by Hummel et al. (2006). For purposes of speech production analysis, however, Kröger et al. (2008) find that despite its potential, the system's performance is clearly inferior to that of other EMA systems.

NDI has recently released an improved successor to the Aurora, branded as Wave and targeted specifically at speech production research. According to the manufacturer's specifications[21], the Wave system is capable of tracking up to 16 sensors with 6 DOF at 100 Hz. At the time of this writing, however, it has yet to be independently evaluated.

Despite actual and potential shortcomings, what sets the NDI systems apart from their direct competitor, the Carstens AG500, is the fact that the transmitter coils are housed within a compact ($20 \times 20 \times 7$ cm) encasing (see Figure 2.6a) positioned laterally near the speaker's vocal tract, but otherwise unobtrusive (Figure 2.6b). This design has significant benefits for multimodal acquisition arrangements where the acrylic cage of the AG500 would present a positional or visual obstacle (e.g. Aron et al. (2008); however,

[19] `http://www.ndigital.com/medical/aurora-techspecs.php`

[20] 6 DOF result from measuring 3D position as well as roll, pitch, and yaw for each sensor

[21] `http://www.ndigital.com/lifesciences/products-speechresearch-volume.php`

see Kroos (2007) for a different solution). This, combined with the advertised ease of integration with NDI's optical tracking products, can give these systems an advantage in such settings.

2.2.3 Optical tracking

Optical markers can be tracked using regular or infrared cameras (normally coupled in arrays of two or more), and, assuming proper calibration, the position in space can be calculated from the location of the markers in each video frame. This modality is widely used in motion capture and gait analysis, and on a finer scale to capture facial expressions (both most famously in the entertainment industry).

More recently, techniques such as stereophotogrammetry have become mature and allow the real-time tracking of 3D surfaces, such as the face, without the use of markers.

Face capturing is useful for the analysis of articulators (viz. the jaw, lips, and to a lesser extent, the tongue tip) with a direct LOS to the cameras, and it is certainly a central aspect of audiovisual speech analysis and synthesis. However, for the full analysis of speech production, the shape of the hidden vocal tract, in particular the tongue, plays a critical role, and therefore, optical tracking can at best complement other modalities.

2.2.4 Electropalatography

A technique to measure the dynamic palatal contacts of the tongue is electropalatography (EPG) (e.g. Hardcastle, 1972; Hardcastle et al., 1989), in which the speaker wears an artificial palate similar to the Hawley retainer commonly used as an orthodontic device. This EPG palate contains an array of electrodes, through which a weak electric current flows in the presence of lingual contact. Although the manufacture of an EPG palate is moderately expensive and the palate can only be used by the speaker for which it was produced, the arrangements of electrodes (numbering 62 or more, depending on the model) and sampling rates of 100 Hz or higher, allow precise characterizations of alveolar and palatal occlusions during speech.

While EPG is most widely used for speech therapy, it can also complement modalities such as UTI or EMMA, where measurements of tongue movements are restricted to the midsagittal plane, to allow a limited quasi-3D tracking of the tongue. A limitation, however, is that lateral movement can only be captured with EPG when contact with the artificial palate is made.

Optopalatography

An interesting approach to dynamic tongue surface tracking is optopalatography (OPG) (Wrench et al., 1996, 1997, 1998), a form of optical tracking in which the speaker wears an artificial palate similar to those used in EPG. The OPG palate however houses an array of infrared diodes and optical sensors to measure the vertical distance to the surface of the tongue at 16 points and a temporal resolution of 100 Hz. Unfortunately, development of this modality is still in a prototype stage and currently on hiatus (Wrench, personal communication).

Chapter 3

Articulatory resynthesis

This chapter gives a brief overview of the articulatory resynthesis approach used in this thesis. It introduces several key concepts and sets the stage for the resynthesis experiments that follow in the next chapters.

3.1 Concepts

A fully flexible yet phonetically intuitive speech synthesis system represents a valuable asset for research in speech science and applications in related fields. The potential of modern articulatory synthesizers, unfortunately, is offset by the significant effort of controlling them properly, that is, so that the output closely resembles natural speech. Clearly it would be desirable to integrate such a synthesizer with a TTS platform, so that laborious manipulation of sensitive control structures are no longer the sole responsibility of the user.

When exploring the possibilities of controlling the VTL synthesizer in such an automated way, without having to manually craft a gestural score for every utterance, it is important to ensure that relevant aspects of the vocal tract model are comparable to their counterparts in natural speech.

In particular, the articulatory movements of the vocal tract model must be examined in detail and compared to those of a human speaker, which they should approximate. This of course requires suitable speech production data. Such data should be comparable to that of the vocal tract model with minimal effort, and should satisfy conditions of temporal resolution, spatial precision (preferably in three dimensions), and phonetic coverage; a large amount of such data allows more powerful analysis. (Of course, minimizing discomfort of, and danger to, the speaker during the acquisition of such data is also a priority!)

Only a few of the modalities described in Chapter 2 satisfy these constraints. Those involving exposure to ionizing radiation for non-medical purposes must be ruled out, especially considering that large amounts of data would have to be collected using them, exacerbating the dosage. LOS modalities alone do not capture movements inside the oral cavity, and the acquisition rate of real-time MRI is too low, regardless of other limiting factors of MRI scanning.

One possible solution might be a combination of UTI (tongue body), EPG (for apical tongue-palate contact), and optical tracking (for lip and jaw movements), but the experimental overhead of setting up an appropriate, robust acquisition environment and the theoretical and practical problems of transforming the disparate data formats to the dimensionality of VTL are rather formidable.

The other solution would be EMA, which is not only a single proven system capable of collecting synchronized motion data from all of the relevant articulators simultaneously, but also available at many research sites. Moreover, EMA has the advantage of producing data which is directly comparable to that produced by tracking vertices on the vocal tract model surfaces. It therefore seems straightforward to use EMA for the comparison of articulatory movements in the model with that of a real human speaker.

This comparison should be controlled as much as possible, a condition which can be satisfied by selecting a corpus of appropriate EMA data and forcing VTL to mimic human speech production by synthesizing the same utterances as those in the corpus. This leads to an *analysis-by-synthesis* approach that uses EMA as the interface between synthesizer and human speech production. In what follows, this will be referred to as *articulatory resynthesis*.

Furthermore, the resynthesis should be performed in an automatic way, which makes it both objective and reproducible, and avoids the immense amount of labor otherwise required for manual control of the synthesizer.

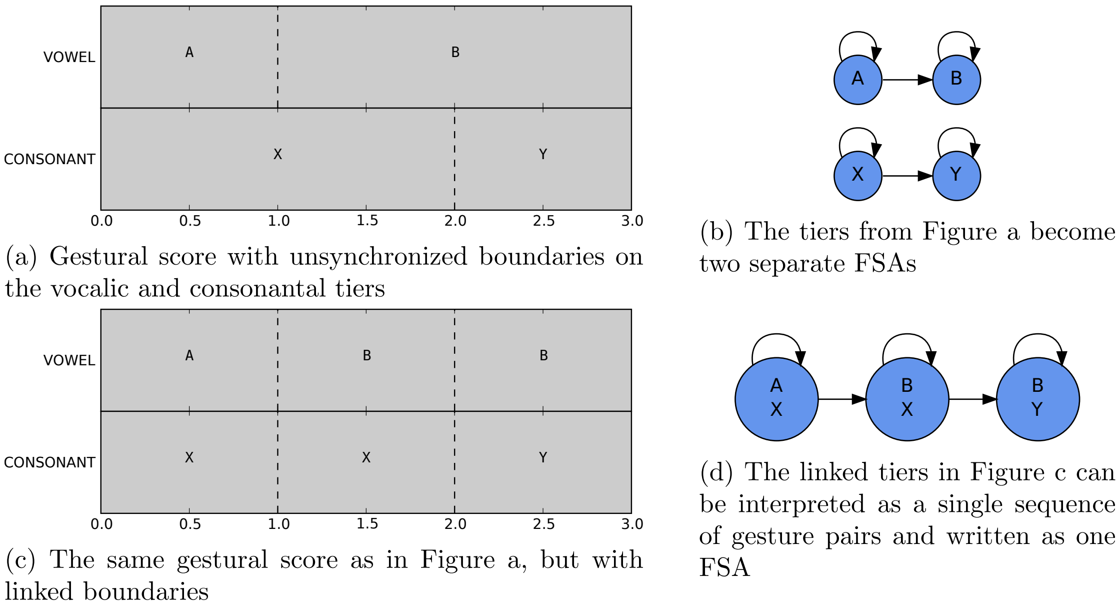
This part of the thesis will present and explain the automatic articulatory resynthesis process in depth by describing two resynthesis experiments of increasing complexity. The first of these should be viewed as a proof-of-concept under simplified conditions, while the second attempts to lay the foundations for one possible approach to articulatory TTS synthesis.

This last possibility represents further rationale for the resynthesis approach. It builds on the idea that once valid gestural scores can be produced automatically for a corpus of real articulatory data, the same gestural scores can generate training data for statistical models, which in turn allow the prediction of new gestural scores for unseen utterances.

3.2 Gestural score generation

To synthesize an utterance using VTL, a gestural score must be available. This means that the vocalic and consonantal tiers of the gestural score must be populated with appropriate gestures, and the same applies to the low-level control of the parameters associated with the velic, glottal, F_0, and pulmonary tiers (cf. Section 1.2.2). However, the resynthesis process outlined in this chapter focuses only on the former two tiers, assuming that the low-level gestures on the others can (by and large) be derived from the CV gestures.

The gestures on the vocalic and consonantal tiers do not necessarily have to be synchronized with respect to each other (i.e. a gesture boundary on one tier doesn't require a boundary on the other), but for present purposes, they are modeled as linked gesture



(a) Gestural score with unsynchronized boundaries on the vocalic and consonantal tiers

(b) The tiers from Figure a become two separate FSAs

(c) The same gestural score as in Figure a, but with linked boundaries

(d) The linked tiers in Figure c can be interpreted as a single sequence of gesture pairs and written as one FSA

Figure 3.1: A gestural score with vocalic gestures `A` and `B`, and consonantal gestures `X` and `Y`. Since two adjacent identical gestures on one tier can be combined into one (and vice versa), the gestural scores in Figure a and c are equivalent. However, while the former's tiers are encoded as two separate FSAs, discarding information on the timing of the unsynchronized boundaries with respect to each other (Figure b), the latter representation links the boundaries across both tiers and can be written as a single FSA, which additionally preserves the overlap between the `B` and `X` gestures (Figure d).

pairs, since this reduces the complexity of representing them as finite state automata (FSAs) (cf. Figure 3.1).

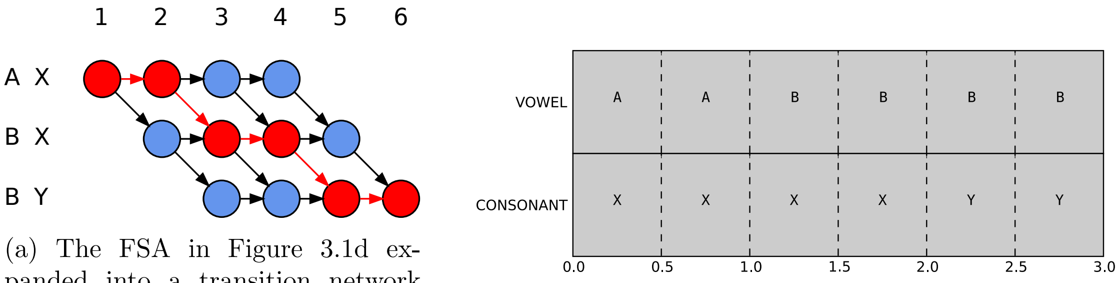
Existing gestural scores can be represented as FSAs (Section 3.2.1) and as transition networks (Section 3.2.2). However, to generate a new gestural score for a given utterance, these representations are used in a more complex process.

3.2.1 Gestural scores as finite state automata

At every given point in time over the gestural score, exactly one consonantal gesture and one vocalic gesture can be observed on the corresponding tier.[1] Such an observation can be regarded as a *state* of the gestural model, and each state is associated with exactly one consonantal and one vocalic gesture, i.e. one gesture *pair*.

Moving from one state of the gestural model to the next can be referred to as a state *transition*. The sequence of all states that are distinct from adjacent states can be visualized as nodes in a graph, with transitions drawn as arrows, or *edges*, between

[1] This is due to the nature of the gestural model, which defines these tiers as discrete sequences of gestures. Other gestural models are conceivable (and were implemented in an older version of VTL), but will not be discussed here.



(a) The FSA in Figure 3.1d expanded into a transition network with 6 frames; one of the 10 possible paths through the network is highlighted in red

(b) Gestural score generated by the path shown in Figure a

Figure 3.2: Assuming a frame duration of 0.5 s, the red path through the transition network in Figure a can be converted into the gestural score shown in Figure b; incidentally, this score is equivalent to the ones shown in Figure 3.1a and 3.1c.

nodes. If two adjacent states are not distinct, they are represented by a single node, but the transition from the former to the latter takes the form of an edge circling from the node back onto itself. Because each edge represents the transition from one state to the next *in time*, no edge may point from a node back to the previous node. This type of directed graph is also referred to as a state diagram, and is the graphical representation of a *finite state automaton (FSA)*.[2] An example of a gestural score and its corresponding FSA is shown in Figure 3.1.

A FSA can be thought of as a formal generator of a finite sequence of states, in this case, a sequence of CV gesture pairs.[3] This means that a given gestural score can be written as a FSA, but not vice versa, because the FSA itself does not provide information about the duration of each gesture. For a given gestural score, a sequence of states can be represented as a sequence of transitions between nodes representing these states, starting at the first state and ending at the last.

3.2.2 Gestural scores as transition networks

Given a gestural score and its corresponding FSA, a sequence of states in the gestural score can be selected in such a way that the time interval between adjacent states remains constant. This selection can be described as (uniform) *sampling* into contiguous *frames*, where each sample contains one frame. The time interval between two frames is the inverse of the *frame rate* and is equal to the duration, or *length*, of a frame.[4]

If the duration of a gesture is shorter than the frame length, it is possible that no sample falls within that gesture, so that it is not represented in the resulting FSA. Therefore, to ensure that every gesture is sampled, the frame length must be no greater

[2] also referred to as a finite state machine (FSM)

[3] This concept should not be confused with the sequential ordering of phonetic segments in a CV *syllable*; such a syllable would be represented as two gesture pairs, one for the onset, and one for the nucleus.

[4] Non-uniform sampling (i.e. frames with varying durations) will not be covered here.

than the duration of the shortest gesture. By extension, the number of frames must also be at least equal to the number of gestures.

Moreover, to accurately represent the duration of each gesture, the frame resolution (i.e. number of frames per time unit) must be at least twice the time resolution used to describe the gesture durations. Otherwise, a quantization error will be introduced, which will result in gesture boundaries being shifted to the nearest frame boundary. This condition is one form of the Nyquist-Shannon sampling theorem (Shannon, 1949).

Assuming that these sampling conditions are satisfied, the durations of the gestures in the gestural score can be reconstructed from the FSA if the sequence of transitions is known. This becomes evident from the fact that each edge in the graph represents the transition from one frame to the next, and thereby possesses a fixed temporal duration equal to the frame length. A representation that illustrates this point more clearly is the *transition network*.

If the number of state transitions for a given FSA is fixed, its state diagram can be redrawn as a network of transitions. In such a network, nodes are arranged in rows and columns. Assuming a left-to-right layout, with one row per gesture, each column represents one frame, and each node represents one possible state within that frame. Under the constraint of a fixed number of transitions through the FSA, all possible gestural scores generated by this FSA can be drawn as distinct paths through the transition network, from the start node to the end node.

Finally, given a specific path, the number of frames per gesture, and hence, the gestural durations become known, and the corresponding gestural score can be reconstructed from the path through the transition network. Figure 3.2a shows the transition network for the FSA in Figure 3.1d, highlighting one specific path.

3.2.3 Gestural scores from articulatory resynthesis

If no gestural score is available (i.e. the gesture durations are unknown), but the sequence and identities of gestures are given, a FSA can be used to generate all possible gestural scores for that gesture pair sequence. However, the total number of these gestural scores, the *search space*, is infinitely large, since the durations can be freely varied. Only once the number of transitions is fixed can the FSA be rewritten as a transition network with the corresponding number of frames, and this yields a finite number of paths. Each path corresponds to one gestural score, and so the search space becomes *discrete* (though still potentially very large).

Finding the best path through the transition network and generating its corresponding gestural score is impossible without some reference criterion, which evaluates the "appropriateness" of each path, and selects the one best satisfying any given constraints. These constraints can be formulated as an error metric, or *cost function*, which permits the calculation of a cost for each candidate path. The path with the lowest cost is selected as optimal.[5]

[5] Allusions to Optimality Theory (Prince & Smolensky, 1993) in this paragraph are deliberate; the concepts are fundamentally related.

The cost of a candidate path or gestural score is meaningless without an *interface* that allows comparison with a target, a spoken utterance. Since the nature of this target is quite different from that of the gestural score, the score must first be transformed into a domain that allows such direct comparison and hence, the application of the cost function.

By virtue of the fact that the gestural score is, somewhat orthogonally, a control interface for the movements of a vocal tract model, such transformation is actually rather straightforward. The VTL synthesizer can be used to generate, or *synthesize*, data comparable to that recorded from a speaker producing the target utterance when using a modality such as EMA. By analyzing both this reference target data and the synthesized data, a candidate score can be found that generates the best comparable data, rendering other candidate scores sub-optimal and discarding them.

This process can be described as an *analysis-by-synthesis* approach. However, to emphasize that in this paradigm, an existing target utterance is mimicked using "copy" synthesis, the term *resynthesis* will be used here.

Evaluation in the acoustic domain

It is of course possible to use VTL to synthesize an acoustic waveform from a gestural score, which could be compared to the acoustic recording of the target utterance. The cost function might then compute the distance of the mel-frequency cepstral coefficients (MFCCs) (Mermelstein, 1976; Davis & Mermelstein, 1980) or similar measures, traditionally used in ASR. From this perspective, the resynthesis task is essentially a *forced alignment* problem; the evaluation forces the selection of that candidate which best aligns with the target, according to the cost function.

The computational expense of synthesizing acoustics using VTL is, however, a somewhat prohibiting factor; waveform synthesis on current hardware takes roughly three or fourfold the time represented by the underlying gestural score. An exhaustive (single-threaded) search through the search space with reasonable parameter settings and frame rate would take weeks to complete for just a single utterance.

Evaluation in the articulatory domain

A promising alternative to acoustics is presented by evaluation in the articulatory domain. If appropriate articulatory data is available for the target utterance, this direct observation of the movements of the original speaker's articulators can be compared with the movement of the vocal tract model in VTL as a candidate gestural score is synthesized. Point-tracking data from modalities such as EMA can be compared to analogous data from the vocal tract model (after appropriate registration, cf. Section 4.1.2 and Section 5.1.3), which allows unrestricted observation of its movement. In fact, the synthesis of such motion data is computationally much less expensive than waveform synthesis, a fact that makes this approach feasible as the interface for articulatory resynthesis.

Perceptual evaluation

The final test for any speech synthesis is of course the perceptual evaluation by human users. Automatic acoustic evaluation may judge the synthetic speech signal to be equivalent to the original recording, and articulatory trajectories may match perfectly, and nevertheless, human listeners might perceive the synthesized speech to be significantly inferior to its natural ideal. Not all factors that influence such preference are easy to identify, isolate or improve, as they may range from obvious errors and audible artifacts of the synthesis process to literally subliminal effects of voice quality and similar phenomena.

The perceptual evaluation of synthetic speech on various levels has been the focus of enormous research efforts in the past and present, such as the ESPRIT-SAM projects (see Chan et al., 1995, for an overview) or more recently, the annual Blizzard Challenge[6] (Black & Tokuda, 2005), and will not be presented here. Performing such an evaluation of synthetic speech – provided the underlying system is able to produce such output – involves resources beyond the scope of this thesis.

3.2.4 Electromagnetic articulography as an interface

When exploring the possibility of using EMA data as the interface between human and synthetic articulation, the first issue which must be addressed is the exact nature of this interface.

More specifically, positional EMA data normally consists of multiple synchronized channels of data, referred to as *trajectories*, one for each dimension (x, y, and in the case of 3D EMA, z) of each transducer coil (attached to articulators, viz. the tongue tip, blade, dorsum, lower lip, jaw, etc). Each trajectory contains the measurement samples of the corresponding coil in the respective dimension over time. Taken together, the EMA trajectories represent the movements of all coils in a geometric plane or volume. The complexity of articulatory movements is thereby represented as the simultaneous movements of a set of points.

VTL's vocal tract model, on the other hand, consists of surfaces which in turn are defined by interpolating between points moving in a three-dimensional geometric space (cf. Section 1.2.1). Any or all of these points, or *vertices*, can be tracked over time, and the coordinates of each point along each of the three dimensions produces trajectories that are essentially commensurable with those in EMA data.

The striking similarity of trajectories derived from EMA and vertex tracking lends itself well to the comparison of the movements of articulators to which the coils and vertices are attached, or a part of, respectively. Given an EMA coil and a corresponding vertex, similar articulatory movements will produce trajectories or similar shape, but *only* if a number of preconditions are satisfied: geometric normalization, global and local similarity of the vocal tracts, and model adequacy.

[6] `http://www.synsig.org/index.php/Blizzard_Challenge`

Geometric normalization

The geometric spaces of the EMA data and the vocal tract model can be described as quasi-isometric. That is to say that the objects described by the EMA data and the model vertices are roughly the same shape and size, but the coordinate systems are not necessarily the same. It is therefore possible to map one set of data into the other's coordinate system by applying processes of reflection, rotation, and translation, *normalizing* the geometric space. If this is done correctly, the coordinate systems of both sets of data will converge and a direct comparison of corresponding trajectories will produce the most meaningful results.

In the context of biomedical imaging, this normalization is referred to as *registration*.

Global vocal tract similarity

While the synthesizer's vocal tract model is highly configurable, which is to say that many anatomical parameters can be freely defined, this definition will have a strong influence on the similarity between the vocal tract model's global shape and the anatomy of a given speaker's vocal tract. This means that any results drawn from the application of the EMA interface will have greater validity if the shape of the vocal tracts of the model and the human speaker are similar.

The vocal tract model's `anatomy` can be adapted to a speaker's vocal tract anatomy if sufficient suitable data is available to allow such adaptation to be performed. Such data is typically volumetric in nature, e.g. MRI or CT scans. Because of the high overall complexity of both the vocal tract model and such imaging data, the adaptation process itself must be carried out manually (although it is conceivable that semi-supervised or even unsupervised adaptation may be possible).

The anatomical parameters which are adapted in this process include such constant features as the shape and arrangement of the teeth, the palate shape, the dimensions of the oral cavity, basic vocal tract length, etc., none of which change radically during speech.[7]

Local vocal tract similarity

Because the vocal tract can assume many distinct shapes resulting from the DOF of the articulators, whose movements are in turn determined by the control parameters described in Section 1.2.1, it is not sufficient to adapt the overall anatomy of the vocal tract model to a static scan of a speaker. Rather, for each sound, or *phone*, in a phonetic inventory, the configuration of the vocal tract model must be individually adjusted to match the typical target position and shape of each articulator. Once again, the requirement is that suitable articulatory data is available, and just as for the global adaptation, this process must be performed manually.

[7] Of course, basic vocal tract length may vary with larynx height during real speech, but in the vocal tract model, it is a static dimension.

Model adequacy

The final condition for the applicability of the EMA interface is that regardless of the fulfillment of the preceding conditions, the vocal tract model's movements must still be as natural as possible. This means that even if the vocal tract model's target configurations for two `phone`s at two given points in time are equivalent to the vocal tract shape of a speaker producing the corresponding phones, the *transition* between these two target configurations should reflect the change in shape of the human vocal tract. This includes not only the shape and position of the articulators, but also their dynamic deformation and displacement during movements.

The exact manner in which these properties are modeled forms an intrinsic part of the VTL synthesizer, and is not exposed to, or controllable by, the user of the software (with the notable exception of the `effort` parameter, which corresponds to the speed of transition to a parameter target, cf. Section 1.2.2).

Chapter 4

Resynthesis of CV utterances

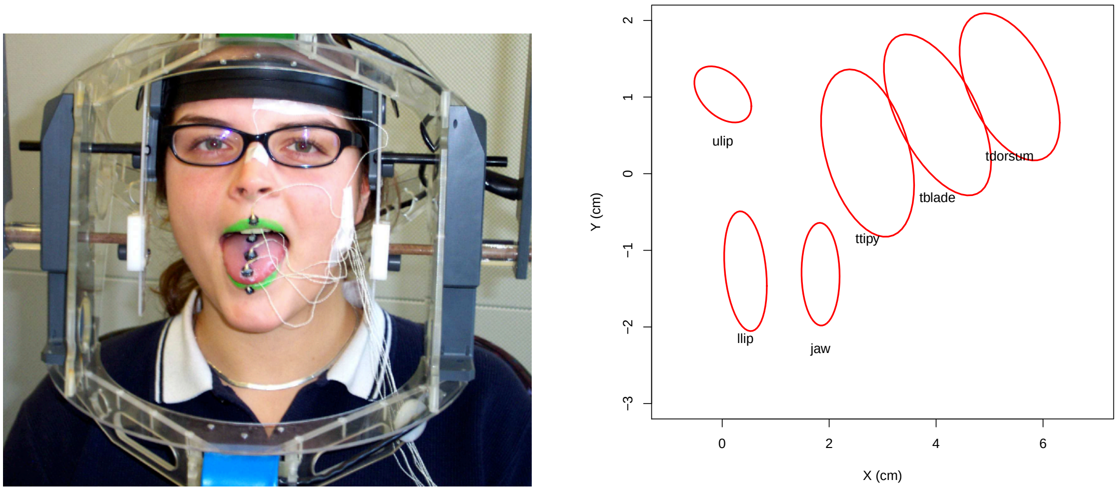
Takashi Murakami,
ZuZaZaZaZaZa, 1994[1]

In order to test the feasibility of the articulatory resynthesis approach introduced in Chapter 3, a simplified scenario was constructed in which the complexity of the utterances to be synthesized is restricted to a minimum. In particular, nonsense utterances containing repetitive sequences of CV syllables, recorded using EMA, are resynthesized. The simple structure of the utterances allows the use of a simple cost function which compares only the single most relevant trajectory pair in the EMA interface.

4.1 EMA data

To minimize the complexity of the automatic resynthesis, an EMA corpus was chosen which contains utterances with a phonotactically simple structure (no consonant clusters, open syllables only), with a single relevant articulatory trajectory. Fortunately, such data

[1] Acrylic, silkscreen on canvas mounted on board, 59 1/16 in. × 66 15/16 in. Takahashi Collection, Tomio Koyama Gallery, Tokyo. © 1994 Takashi Murakami/Kaikai Kiki Co., Ltd. All rights reserved.



(a) Arrangement of EMA coils and speaker in AG100 during the recording of the CV corpus (from Fagel, 2004)

(b) Distribution of EMA data in the CV corpus; the ellipses represent the 0.9 confidence level contours for all samples from each coil. Note that the reference coils are not shown.

Figure 4.1: Recording setup and EMA data distribution in CV corpus

was available and generously provided by Sascha Fagel at the Technical University of Berlin.

4.1.1 Corpus design and recording procedure

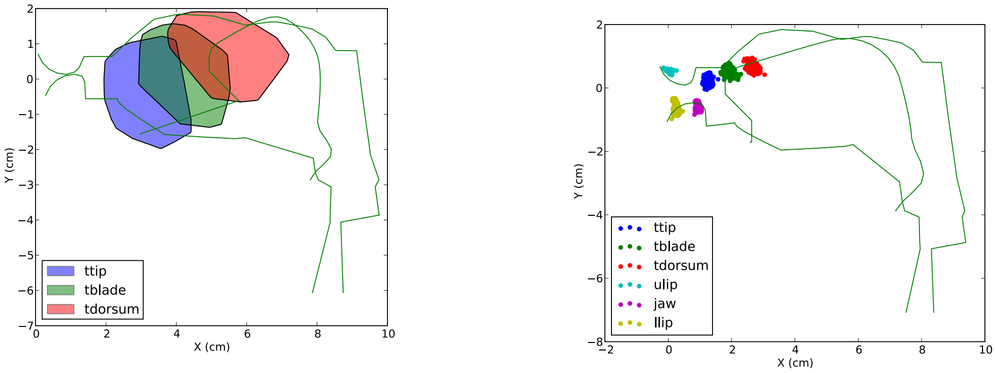
The EMA corpus consists of nonsense utterances of four repetitive CV syllables for a selection of phones from the German phoneset. The prompts were formed by combining each of the 9 consonants in the set [m, n, ŋ, v, z, ʒ, j, ʁ, l] with each of the 15 vowels in [a, e, i, o, u, ɛ, ø, y, ɪ, ɔ, ʊ, œ, ʏ, ɐ, ə]. In each utterance, the third CV syllable carries the primary stress, while the first is produced with secondary stress. The prompt list was recorded twice, first with neutral prominence, and again with stronger prominence on the third syllable. Two realizations of a final, non-nonsense utterance, *Ich habe "Rat" erwähnt.* ("I mentioned [the word] 'council'."[2]), complete the corpus.

A full listing of the utterances as recorded is given in Appendix 4.A.

Discounting the empty utterances and the final, well-formed utterance, this leaves 270 recorded utterances (135 of which are unique), each containing four identical CV syllables. While the range of vowels covers the complete inventory of Standard German monophthongs[3], the German consonant space was simplified for prompt design by grouping together all phones with the same place of articulation and selecting only one phone from each group. This reflects the intended application of the corpus in audiovisual

[2] or 'counsel', depending on which reading of the German polyseme *Rat* is preferred

[3] with the exception of [ɛː]



(a) Convex hulls of all EMA samples for the three tongue coils; the registered vocal tract model contour is superimposed in green, in the "neutral" configuration.

(b) Scatterplot of all sweep-initial EMA samples; the green vocal tract model is configured to this "rest" position

Figure 4.2: Speaker adaptation for CV corpus

speech synthesis (Fagel, 2004)[4].

A female native speaker of German was recorded at the Center for General Linguistics (ZAS)[5] in Berlin, in 2003, using a Carstens AG100 2D articulograph. All utterances were recorded in sweeps of 2 s duration; the EMA data was sampled at a rate of 200 Hz, while the audio was recorded at 16 kHz sampling rate. Additionally, the speaker was recorded using a video camera. Green pigment had been applied to the speaker's lips to facilitate lip tracking in the resulting video data (which was not used here). See Figure 4.1a for details.

EMA transducer coils were attached to the upper and lower lip (`ulip`, `llip`), lower incisors (`jaw`), and tongue tip (`ttip`), blade (`tblade`), and dorsum (`tdorsum`). Reference coils were placed on the bridge of the nose and the upper incisors. This coil layout is shown in Figure 4.1a.

Further details of the recording procedure and data are provided by Fagel & Clemens (2004).

4.1.2 Speaker adaption and data registration

While the vocal tract of VTL was adapted to a male speaker (cf. Section 1.2.4) and the CV EMA data was recorded from a female speaker, this obvious mismatch in vocal tract geometry was gracefully ignored, with the following rationale.

For the adaptation of VTL's vocal tract model to the anatomy and articulation of the male speaker, two MRI corpora had been used as reference. No such imaging data is available for the female speaker of the CV corpus. Furthermore, a full speaker adaptation,

[4] incidentally, consonants in VTL are handled in the same way

[5] `http://www.zas.gwz-berlin.de/`

even if the required articulatory data had been available, would be too extensive and laborious for a proof-of-concept experiment such as this.

Nevertheless, a few adjustments are necessary for proper comparison of the CV EMA and VTL vertex trajectories. In order to register the vertex data into the same coordinate system as the EMA data, the following process was used.

The vocal tract geometry in VTL is defined internally so that the model "faces" right, viz. the rear wall of the pharynx is on the left, the lips on the right. Since this conflicts with the EMA data's coordinate system, the synthesized vertex data is reflected horizontally.

In a second step, a reference point, or landmark, that remains stationary throughout the data should be identified and the vertex data translated to match it. An obvious choice for such a landmark would be the reference coil on the speaker's maxillary incisors, which corresponds to a front central vertex between the upper teeth in the vocal tract model. Unfortunately, data from the reference coils was not available, as it had been removed from the CV corpus. However, standard procedure for AG100 post-processing (at ZAS, in any case; cf. Section 5.1.1) implies that the location of this missing reference coil could be assumed to be near the origin of the coordinate system. Visual inspection of the EMA data distribution (Figure 4.1b) seems to place the origin some 1.5 cm or so forward of the expected position of the upper incisors. As a consequence of this observation, the vertex data was translated so that the front upper edge of the upper teeth lies approximately at the estimated position of the `upinc` reference coil.

In a final registration step, the vertex data could be rotated around the `upinc` reference coil (in this case, also the origin) so that the palate of the vocal tract model matches that of the speaker. This would require a palate contour, normally obtained with a palate trace in a separate recording sweep. However, this was also not available in the EMA data. As a substitute for the missing palate trace, the convex hulls of the three tongue coils (`ttip`, `tblade`, and `tdorsum`) were plotted, and compared to the palate of the vocal tract model; this is shown in Figure 4.2a. The fit was judged to be acceptable, which is not surprising, since the vocal tract's occlusal plane is parallel to the `x` axis, and the EMA data was presumably rotated to the same angle during post-processing. For this reason, no rotation of vertex data was deemed necessary. Nevertheless, the fit could not be expected to match so well, since the palate anatomy of the two original speakers could have differed rather more noticeably.

It turns out (cf. Section 4.3.3) that even without further adaptation or vocal tract normalization, the results of the resynthesis approach are encouraging enough for the follow-up experiment (Chapter 5), which avoids the issue altogether.

4.1.3 Rest position and articulatory setting

Given the opportunity, a speaker's articulators will tend to assume a "rest position" before and after speaking, which is occasionally referred to as inter-speech posture (ISP) (Gick et al., 2004). When breathing through the nose, the speaker's tongue will tend to be raised against the palate, while the lips may be either parted or closed. This potentially results in rarefaction of air in the post-lingual oral cavity, occasionally manifested in a smacking sound prior to speech, as the tongue loses contact with the palate. This could be described as a kind of click (Ladefoged & Traill, 1984) without linguistic function.

In preparation for speech, on the other hand, the articulators will assume another configuration, which may be not only speaker-specific, but also language-specific (Gick et al., 2004; Wilson, 2006; Schaeffler et al., 2008). This is sometimes called the "neutral position", although it may be influenced by the following phones and thereby subject to anticipatory assimilation. Conversely, it may also exert global influence on its surroundings, and another term for the phenomenon is *articulatory setting* (Honikman, 1964; Laver, 1994).

Whether the neutral position is equivalent to [ə] or represents a separate configuration has been a matter of some debate (see e.g. Barry's (1992) comments on Browman & Goldstein (1992b)). In VTL, the question has been answered in a pragmatic way; `<phone name="@"/>`[6] defines a target configuration of the vocal tract model which visually approximates the vocal tract shape during schwa production, and which has been acoustically optimized to produce a schwa-like vowel (Birkholz & Kröger, 2006).

In contrast, VTL also defines a "neutral" configuration which corresponds to a vocal tract configuration with the tongue bunched at the rear of the oral cavity. It can be assumed that this configuration is based on the vocal tract shape during a localization scan with no produced speech (Kröger et al., 2000), and that the retracted position of the tongue is due to the effects of muscle relaxation and gravity while the speaker lay supine in the MRI scanner. The resulting neutral configuration of the vocal tract model can be observed in Figure 4.2a.

The articulatory movements observed during the "silent" intervals before audible speech in the EMA recordings of the CV corpus exhibit two distinct positions; one is the initial rest position assumed between utterances, the ISP. The other appears to correspond to the preparatory neutral position described above. To achieve a better alignment during resynthesis, these two configurations were explicitly modeled by extending the phoneset of VTL with two "pseudo-phones", `sil`, and `_` (underscore).

The `sil` phone represents the rest position, but as no such configuration seems to be available in the MRI data used to configure VTL, the EMA data in the CV corpus was used to define the `sil` configuration of the vocal tract model. To this end, the initial sample of each EMA trajectory was extracted for all utterances in the CV corpus, and plotted alongside the vocal tract model (cf. Figure 4.2b). The vocal tract model was then adapted to this configuration by hand using the VTL GUI. The adaptation was not strict, since the two vocal tracts have different shapes, and the posture of the speakers in the underlying data is not the same (cf. Section 5.1.4). The lack of precision in this process is balanced by the fact that the `sil` phone is used only for alignment in the articulatory domain, and does not appear during audible speech.

The neutral phone `_` however, is merely an explicitly named alias of the vocal tract model's neutral configuration, and serves only to streamline the resynthesis process. Since VTL handles `neutral` gestures differently than those belonging to explicitly named

[6] `[@]` is the Speech Assessment Methods Phonetic Alphabet (SAMPA) notation for international phonetic alphabet (IPA) [ə] (Wells, 1997)

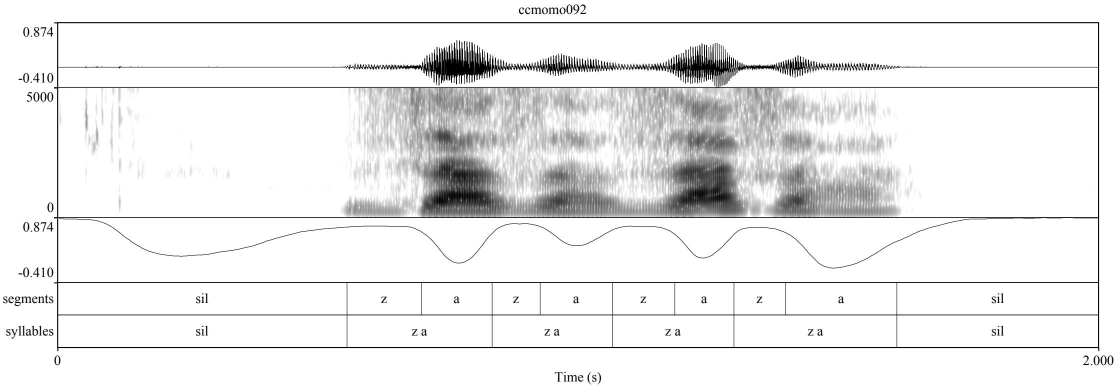


Figure 4.3: One utterance (`092`), [zazazaza], from the CV corpus; waveform, spectrogram, tongue tip height (`ttipY`) EMA trajectory, acoustic segmentation (*not* gestures)

`phone`s in the gestural model, this placeholder allows all gestures to be controlled in the same way, using entries from the `phoneList` (cf. Section 1.2.1).

As becomes evident from casual inspection of the EMA data (cf. Figure 4.3), the speaker's articulators assume these two distinct configurations during each of the initial and final silent intervals. To account for this behavior in the resynthesis approach, the sequence of canonical gestures is padded with leading and trailing neutral (`_`) and rest (`sil`) gestures.

4.1.4 Semi-automatic acoustic segmentation

Since no phonetic annotation was available for the acoustic data, the utterances in the corpus were semi-automatically segmented on the phone and syllable levels, in the following way.

1. In an initial step, the acoustic recordings were automatically segmented into speech and non-speech ("silent") intervals, with an intensity threshold of -25 dB and a minimal interval duration of 0.1 s.

2. Each speech interval was then divided into eight intervals of equal duration, corresponding to the segments of the four CV syllables...

3. ...and these were labeled using the corresponding entry in the prompt list (Appendix 4.A), which identifies the phones present in the respective utterance.

4. In a manual step, the phone boundaries were adjusted by hand, using both acoustic and visual cues (i.e. the spectrogram).

5. Finally, the phone-level segmentation was automatically augmented by adding a syllable-level tier and duplicating the relevant segment boundaries and labels.

Each of these steps was accomplished using Praat[7] (Boersma & Weenink, 1995–2010) with custom scripts. An example of the resulting annotation is shown in Figure 4.3.

4.1.5 F_0 parametrization

Since the VTL synthesizer allows low-level control of the `F0` parameter, it is possible to "transplant" the F_0 contour of an original utterance by encoding it as a sequence of appropriate gestures. This becomes useful for acoustic resynthesis, i.e. waveform generation for a gestural score produced by the articulatory resynthesis.

In preparation for this, the F_0 contours of all non-empty utterances in the corpus were extracted (with a pitch floor and ceiling of 75 Hz and 600 Hz, respectively) and saved as binary files using Praat. The resulting F_0 contours were visually inspected for octave errors, and in the handful of cases where such errors had occurred, the affected sections were manually corrected with a negligible amount of effort.

The F_0 contours were subsequently parameterized as `F0` gesture sequences using the following process in a Praat script:

1. From the `Pitch` object, the utterance was automatically segmented into voiced and unvoiced intervals to determine the start and end times of the F_0 contour, using a mean and maximum period of 10 ms and 20 ms, respectively (the default settings for Praat's `To TextGrid (vuv)...` command).

2. The F_0 contour was then smoothed with a bandwidth of 10 Hz to neutralize irrelevant micro-perturbations, and interpolated through any unvoiced intervals.

3. To encode the F_0 contour as a sequence of `F0` gestures, gesture boundaries were inserted at the inflection points of the F_0 curve, and the `F0` targets were defined as the asymptotes of the curve at these points, yielding the required `offset` and `slope` parameter values.

4. To accomodate the semitone (st) scale of VTL, which uses C_0 as the reference frequency, all values in Hz were converted into st using the formula

$$f_{\mathrm{st}} = \frac{12 \ln(f_{\mathrm{Hz}}/\mathrm{C}_0)}{\ln 2} \tag{4.1}$$

   where $\mathrm{C}_0 = 16.351\,597\,83\,\mathrm{Hz}$ (Young, 1939).

The `effort` parameter of the gestures (which controls the speed at which a vocal tract parameter approaches its target value, cf. Section 1.2.2) was kept constant at the default value of `8.0`. Optimizing this value would require an iterative approach in the absence of an analytic solution, but even before optimization, the fit of the parameterized `F0` trajectory to the original F_0 contour was generally considered acceptable.

One possible alternative to selectively placing gesture boundaries only at the inflection points of the F_0 contour is to split the utterance duration into a fixed number of

[7] `http://www.praat.org/`

frames, with one F0 gesture boundary at each frame boundary. Depending on the shape of the original contour, this method might yield a closer fit, but would typically require a greater number of gestures.

4.2 Resynthesis using dynamic time warping

An early approach to the EMA interface was explored using dynamic time warping (DTW). The idea was subsequently abandoned in favor of another dynamic programming solution, but will nevertheless be briefly outlined here.

DTW is a dynamic programming algorithm that yields information on how the timing in one sequence of events (or samples) can be manipulated, or *warped*, in the temporal domain in such a way that the sequence becomes more similar to another sequence used as a reference.

More specifically, given two time signals (in this case, trajectories), DTW attempts to find a *warp path* best suited to transforming one into the other. This warp path refers to the speeding up or slowing down of time, usually in a non-linear way. The technique is best used with signals expected to be similar (e.g. two different realizations of the same utterance), but not synchronized with respect to their temporal evolution. One traditional field of DTW application is ASR.

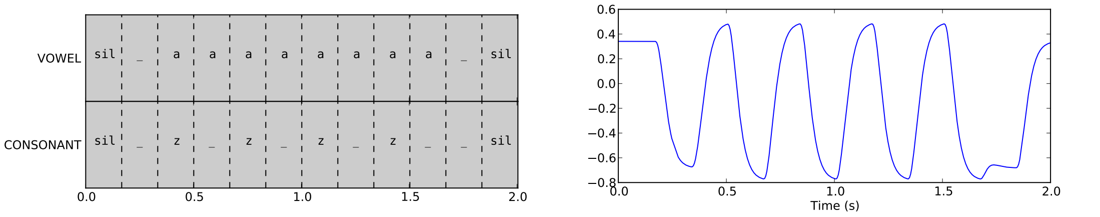
For the purposes of the present study, the DTW component of Praat was used, which implements the algorithm described by Sakoe & Chiba (1978).

As a case study, one realization of the utterance [zazazaza] was selected from the CV corpus, and the EMA trajectory for the tongue tip height (ttipY) was resynthesized in the following way.

The ttipY EMA trajectory exhibits the typical rest position at the start and end of the recording, so in addition to the canonical CV gesture pairs assumed to underlie the produced utterance, leading and trailing neutral and "rest" gestures were added to the respective gesture sequences (cf. Section 4.1.3).

Since the durations of the gestures were unknown, the gestural score was initialized, or *bootstrapped*, using equal durations for all gestures (Figure 4.4).

The resulting vertex trajectory and the original EMA trajectory were loaded in Praat



(a) Bootstrapped gestural score

(b) Bootstrapped ttipY vertex trajectory

Figure 4.4: Bootstrapped gestural score and synthesized ttipY vertex trajectory for [zazazaza]

and, after normalizing the sample values to the range $[-1;1]$, combined into a `DTW` object. The difference between the two sample values in each frame of each trajectory can be visualized as a distance map whose cells are shaded according to a color lookup table (CLUT); in this case, the CLUT maps distance values of 0 and 2 to white and black, respectively, and distance values between these extremes to a linear grayscale continuum.

A path is then found through the distance map, with the requirement that it lead from the first to the last frames, respectively, of both signals. Additionally, the path slope s is constrained to $\frac{1}{3} < s < 3$ to ensure that all gestures possess a positive duration after warping. The path is otherwise free to choose its way through the map, prefering to stay in the white areas as much as possible. This path is the *warp path*, and its deviation from a straight diagonal line (with slope $s = 1$) corresponds to the requirement for time to be compressed or expanded in order to transform the gestural timing which synthesized the vertex trajectory into that assumed to underlie the EMA trajectory. This is illustrated in Figure 4.5a.

The bootstrapped gestural score is converted into a `TextGrid`, with intervals corresponding to gestures, and loaded into Praat. The warp path from the `DTW` object is then applied to this `TextGrid`, which generates a new, warped `TextGrid`, whose interval (i.e. gesture) durations are those produced by the DTW (Figure 4.5b). The warped gestural score is then used to resynthesize the tongue tip height trajectory. The results of this resynthesis are shown in Figure 4.5c and exhibit a much closer temporal fit than the vertex trajectory synthesized by the initial bootstrapped score. This can be seen as evidence that the DTW approach holds some potential.

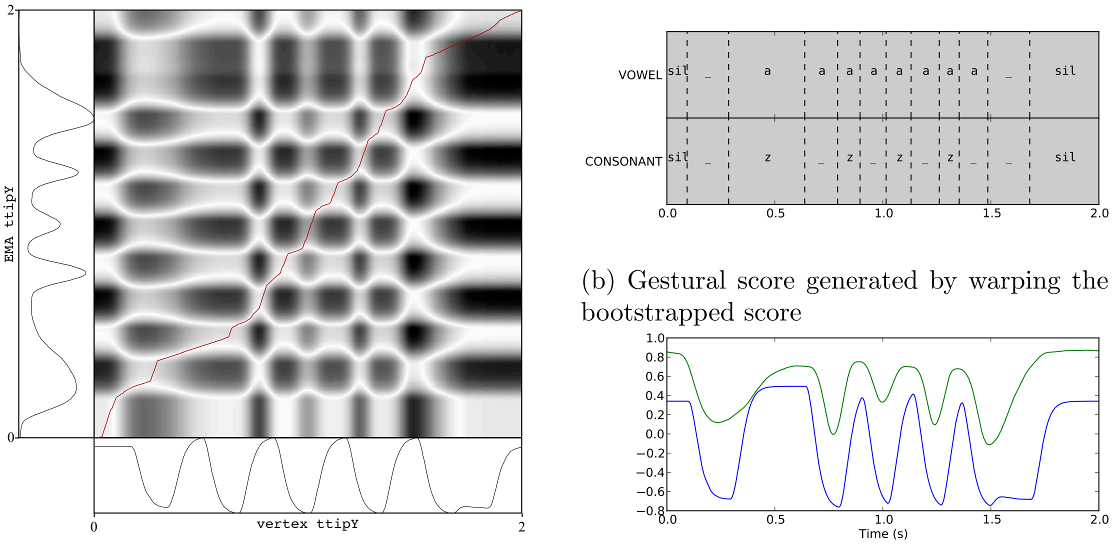
On the other hand, the warped gestural score does not quite seem to result in the expected trajectory; this may be due to the fact that the DTW does not take into account specific details of VTL's gestural model, such as the way in which the `effort` parameter influences the slopes of the resulting vertex trajectories (cf. Section 1.2.2).

At this point, it is conceivable to repeat the DTW process, but replacing the bootstrapped gestural score and its corresponding vertex trajectory by those generated by the DTW as described. Figure 4.6 shows this second step and its result, and it becomes evident that the temporal fit of the trajectories in Figure 4.6c has improved somewhat (mainly during the final 500 ms) in comparison to the input vertex trajectory.

This *iterative* DTW process, where the output of one application is used as the input to the next, can be repeated until a suitable condition is satisfied. For instance, an error metric such as a cost function (cf. Section 4.3.3) could be devised such that the iterations might terminate as soon as the cost becomes lower than some predetermined threshold.

For all the promise of the iterative DTW approach to the resynthesis task, there are several limitations which discouraged further exploration of this early avenue.

The opacity of VTL's gestural model to the DTW process has several serious repercussions. Compression in the temporal domain during DTW may lead to the near collapsing of two gesture boundaries, which will render the corresponding events in the trajectory undetectable. Conversely, an event “hidden” in this way will not become expanded, since the landmarks in the signal are missing, and so the DTW algorithm cannot align them. This phenomenon may be triggered whenever the input to a DTW iteration



(a) Distance map and warp path (red) for DTW of bootstrapped gestural score; synthesized `ttipY` vertex trajectory (bottom) plotted against original `ttipY` EMA trajectory (left).

(b) Gestural score generated by warping the bootstrapped score

(c) Original (green) and resynthesized (blue) tongue tip height trajectories from warped gestural score

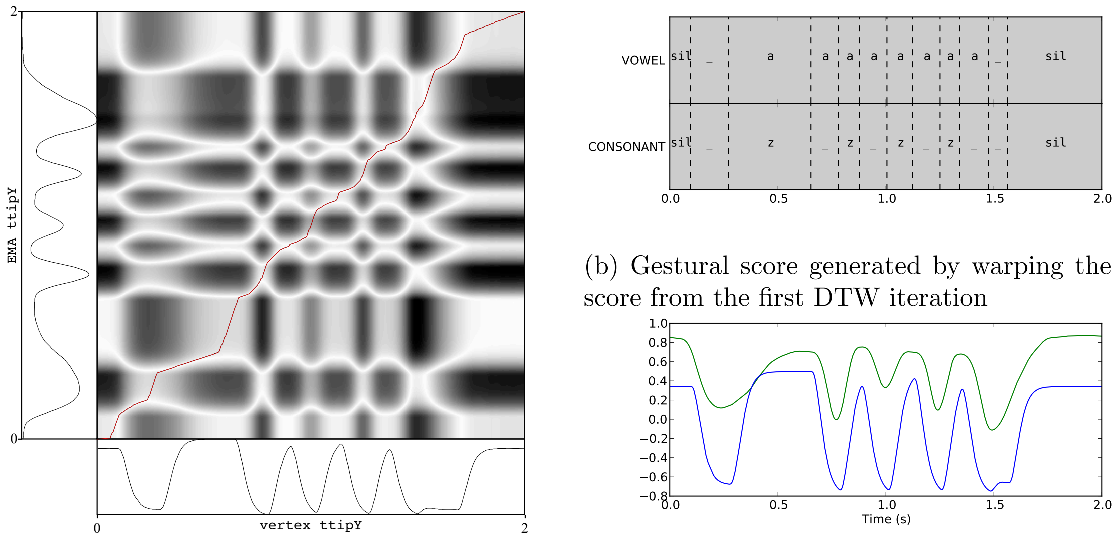
Figure 4.5: DTW of tongue tip height trajectory synthesized by bootstrapped gestural score (Figure a), as well as resulting warped gestural score (Figure b) and trajectory synthesized by warped score (Figure c)

contains relevant gestures that are too short to have a noticeable effect, and consequently, the input would have to be monitored and corrected by some suitable mechanism to avoid this problem.

Moreover, the increased complexity of resynthesizing *normal* utterances entails that multiple trajectories have to be processed in parallel (cf. Section 5.2). However, since resynthesis by iterative DTW is performed on a single trajectory pair, a separate warping would be required for each trajectory pair, and it can be expected that different trajectories would lead to different warp paths. This would lead to different, conflicting warped gestural scores, and it is not clear by which process these should be merged into a single one, as required by the resynthesis task. This aspect is potentially fatal to the feasibility of this form of resynthesis by DTW.

Incidentally, at the time the DTW approach was explored, the DTW code in Praat contained a programming bug that under certain conditions caused the last interval of a warped `TextGrid` to receive an invalid end time, requiring external mending of the resulting corrupt `TextGrid` and encumbering the resynthesis implementation further. The bug has been fixed since then.

The phase of experimenting with the iterative DTW approach ended once the source code of VTL became available to the author. This opened up many new possibilities, a



(a) Distance map and warp path (red) for DTW of warped gestural score; synthesized `ttipY` vertex trajectory (bottom) plotted against original `ttipY` EMA trajectory (left).

(b) Gestural score generated by warping the score from the first DTW iteration

(c) Original (green) and resynthesized (blue) tongue tip height trajectories from second-iteration warped gestural score

Figure 4.6: DTW of tongue tip height trajectory synthesized by *warped* gestural score (Figure a), as well as resulting second-iteration warped gestural score (Figure b) and trajectory synthesized by warped score (Figure c)

few of which were pursued with greater success than this one, and are described in the remainder of this thesis.

4.3 Brute-force search

Returning to the approach outlined in Section 3.2, as a proof of concept, and to provide a performance baseline, one of the unique utterances was selected from the prompt list of the CV corpus, and all possible gestural scores for the utterance's underlying gestures were explicitly generated at a reasonably high frame rate.[8] Every one of these scores was then synthesized, and the resulting vertex data was saved for subsequent analysis. This allowed the evaluation of different cost functions in preparation for a more efficient search algorithm.

This exhaustive approach is referred to as a *brute-force search*, because it calculates the cost for the entire discrete search space of possible gestural scores. Despite its minimal efficiency, it is guaranteed to yield the best gestural score in the search space for the given utterance and parameter constraints (i.e. frame rate, cost function).

[8] satisfying the tradeoff between high temporal resolution and manageable search space size, see below

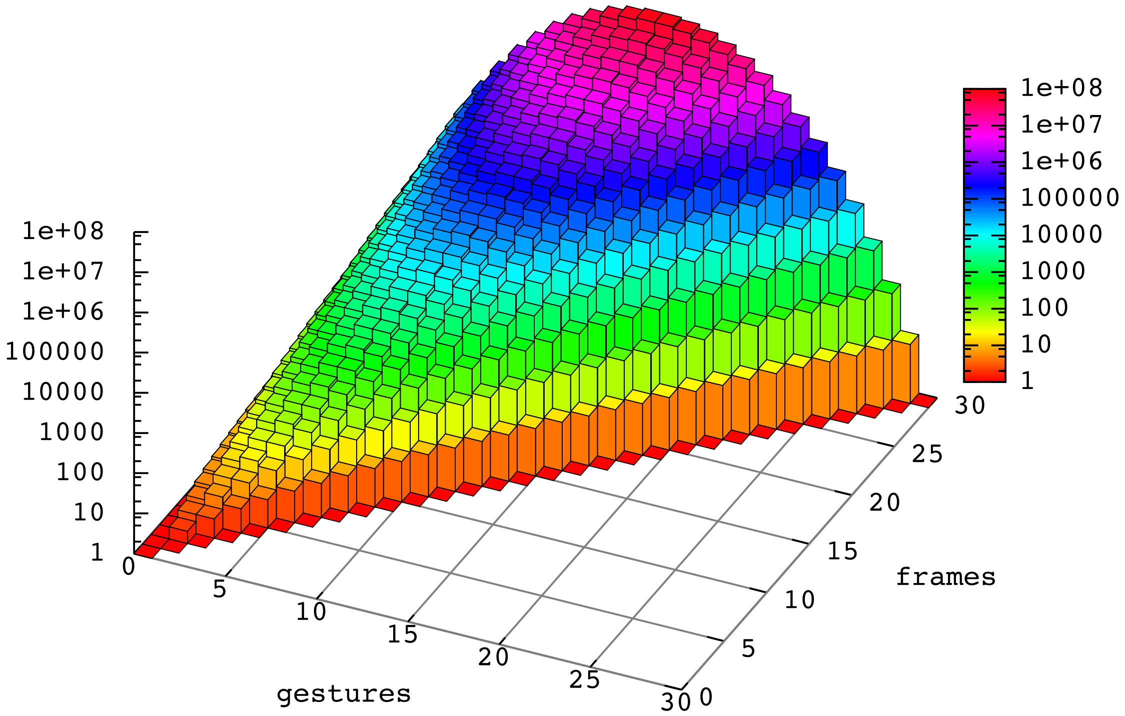


Figure 4.7: Number of possible gestural scores for f frames and g gestures; $1 \leq f, g \leq 30$. The height of each bar represents the number of scores, color-coded for additional clarity. Note the logarithmic scale, and that no scores are generated if $f < g$.

4.3.1 Search space size

The total number $n_{f,g}$ of possible paths through a transition network with g gestures and f frames is given by the recursive function

$$n_{f,g} = \begin{cases} \sum_{i=1}^{f-g+1} n_{f-i,g-1} & \text{for } g > 1 \\ 1 & \text{else} \end{cases} \tag{4.2}$$

As the number of gestures and/or frames increases, the number of possible paths, and hence gestural scores, quickly becomes astronomical, as illustrated in Figure 4.7.

By the nature of the prompts described in Section 4.1.1, each utterance in the CV corpus consists of 4 CV syllables, with leading and trailing silence. Since each syllable is open, requiring two CV gesture pairs, and the silent intervals contain one rest and one neutral position (cf. Section 4.1.3), the number of gestures required for one utterance is relatively low. An utterance of 4 CV syllables requires 12 gesture pairs to synthesize.

At a relatively low rate of 10 fps and an utterance duration of 2 s, this sets the number of frames $f = 20$ and the number of gestures $g = 12$. Applying Equation 4.2 returns the number of unique paths in the corresponding transition network, generating a search space of 75 582 gestural scores.

EMA trajectory	σ (092)	σ (232)
ttipX	0.112 399	0.095 499
tbladeX	0.077 196	0.090 174
tdorsumX	0.130 086	0.071 824
ulipX	0.018 068	0.023 889
jawX	0.067 099	0.045 103
llipX	0.068 545	0.065 039
ttipY	0.249 851	0.353 101
tbladeY	0.166 039	0.310 813
tdorsumY	0.142 339	0.228 095
ulipY	0.012 412	0.024 314
jawY	0.151 178	0.240 950
llipY	0.174 992	0.254 345

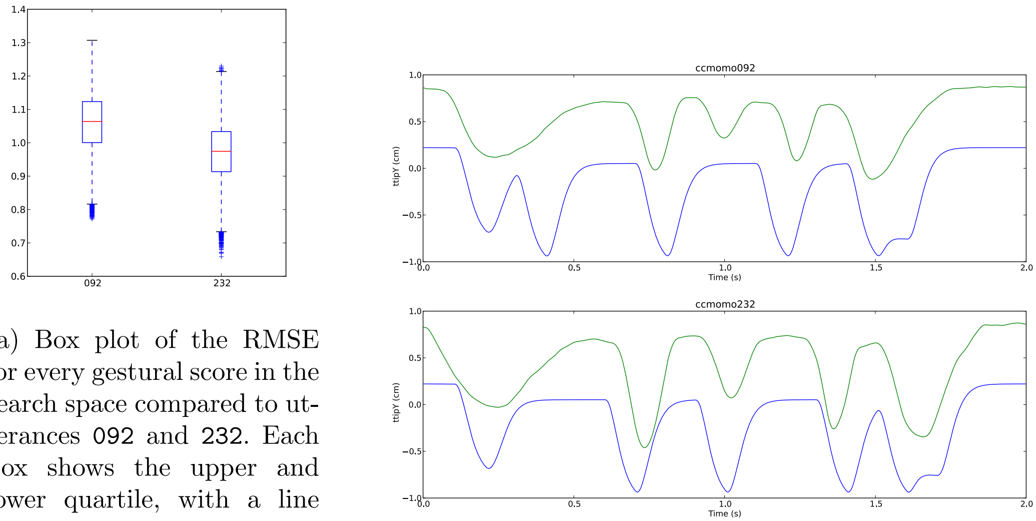
Table 4.1: Standard deviation (σ) of samples for each EMA trajectory in the two realizations (092 and 232) of the utterance [zazazaza] in the CV corpus. To (somewhat crudely) discard samples during silent intervals, the first and last 500 ms were excluded. The ttipY trajectory (shaded) exhibits the highest σ value in both realizations.

4.3.2 Brute-force resynthesis

Every utterance in the CV corpus involves only one type of CV syllable. If a trajectory is known to be highly relevant for the articulation of a specific consonant, then the resynthesis evaluation is significantly simplified, since only that one trajectory needs to be evaluated. With this in mind, the utterance [zazazaza] was selected for resynthesis, with the expectation that the syllable [za] would provide significant movement observable in the tongue tip, since the apical production of the alveolar obstruent [z] contrasts with the low configuration of the tongue blade during the open vowel [a]. This assumption is borne out by simple statistical analysis of all EMA trajectories, shown in Table 4.1. Consequently, the height of the tongue tip (ttipY) was selected as the relevant trajectory for resynthesis evaluation.

All gestural scores in the search space were explicitly generated as individual files in VTL's appropriate XML format. Each gestural score was then loaded, and vertex trajectories synthesized at a sampling rate of 200 Hz, matching that of the EMA data. The vertex tracking data for each gestural score was saved as a tabular text file to facilitate separate analysis in a subsequent step.

The synthesis process was implemented using parallel programming, so that it was possible to synthesize the vertex trajectories for all 75 582 gestural scores in approximately 160 min on a compute server with 8 CPUs clocked at 2.6 GHz. This processing time might be acceptable for such a one-shot evaluation of the resynthesis approach, but it becomes clear that a more efficient alternative is needed to make any resynthesis task feasible that requires more than a few utterances at a time.



(a) Box plot of the RMSE for every gestural score in the search space compared to utterances `092` and `232`. Each box shows the upper and lower quartile, with a line at the median. The whiskers extend to the most extreme values within 25 % to 75 % of the data range; outliers are shown as + marks.

(b) For utterances `092` (top) and `232` (bottom), EMA (green) and optimal vertex (blue) trajectories with respect to RMSE.

Figure 4.8: For all gestural scores in the search space, the synthesized vertex trajectory is compared to the EMA trajectory of utterances `092` and `232` in the CV corpus. RMSE analysis is shown in Figure a, and the lowest-scoring `ttipY` trajectory pair for each utterance is plotted in Figure b.

4.3.3 Cost function evaluation

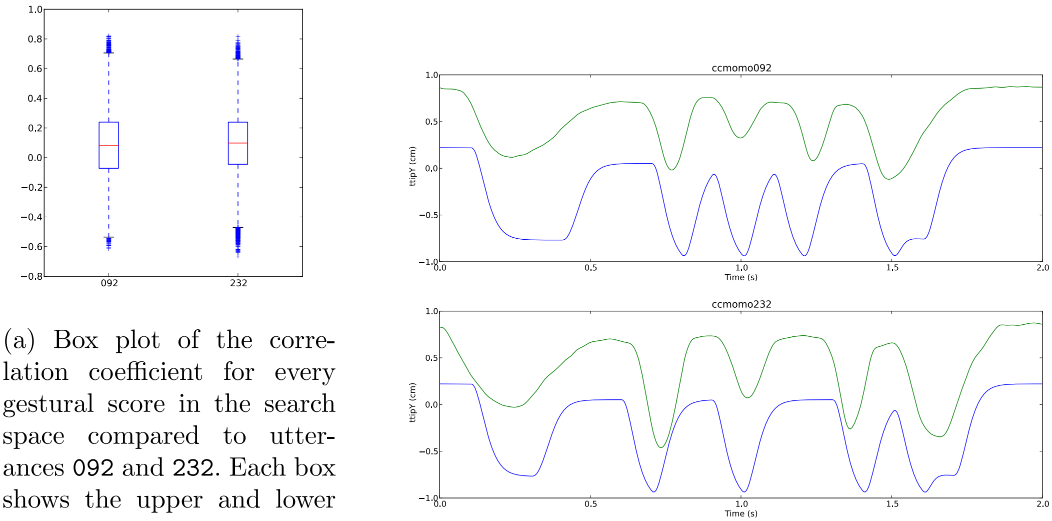
To evaluate the validity of the resynthesis approach, the synthesized vertex data was compared to the corresponding EMA data in *both* realizations of the target utterance [zazazaza] in the CV corpus (`092` and `232` in the prompt list, cf. Appendix 4.A). In each case, several cost functions were applied to identify the gestural score that had produced the best match.

Cost function based on RMSE

The root mean square error (RMSE)[9] can be used to measure the absolute difference between the values of two signals, in this case, the `ttipY` EMA trajectory x and vertex trajectory v; both are sampled at 200 Hz, yielding $N = 400$ samples. The RMSE of the two trajectories is given by

$$\mathrm{RMSE}(x, v) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n - v_n)^2} \tag{4.3}$$

[9] also referred to as root mean square deviation



(a) Box plot of the correlation coefficient for every gestural score in the search space compared to utterances `092` and `232`. Each box shows the upper and lower quartile, with a line at the median. The whiskers extend to the most extreme values within 25 % to 75 % of the data range; outliers are shown as + marks.

(b) For utterances `092` (top) and `232` (bottom), EMA (green) and optimal vertex (blue) trajectories with respect to correlation coefficient.

Figure 4.9: For all gestural scores in the search space, the synthesized vertex trajectory is compared to the EMA trajectory of utterances `092` and `232` in the CV corpus. Correlation coefficient analysis is shown in Figure a, and the highest-scoring `ttipY` trajectory for each utterance is plotted in Figure b.

This value was computed for each gestural score in the search space. Figure 4.8a shows the distribution of the RMSE values; most lie around the mean. This indicates that only a relatively small number of gestural scores produce a very low RMSE, and these are shown as the outliers.

The vertex trajectory synthesized by the score with the lowest RMSE with respect to each target utterance is plotted against that utterance's EMA trajectory in Figure 4.8b. Visual inspection prompts several observations. In utterance `092`, the gestures for the first syllable begin too early, during the neutral position. Likewise, the third syllable's gestures are timed to span both the second and third syllable in the target utterance. This result is clearly not felicitous. While the fit is better for utterance `232`, some of the slopes are significantly steeper and more abrupt than in the original trajectory.

It is obvious that a more robust cost function must be used, and that both the `effort` parameter (which controls the slopes) and the frame length (which determines the number of possible gesture boundaries) warrant greater flexibility.

Cost function based on correlation coefficient

An alternative error metric that focuses on the similarity of the overall *shape* of two signals is the *correlation coefficient* $r(x,v)$, which disregards the value ranges. A high $r(x,v)$ indicates that the two trajectories x and v vary in a synchronized manner; both rise and fall simultaneously. A negative $r(x,v)$ indicates that x rises while v falls, and vice versa. $r(x,v)$ close to 0 indicates that no clear relationship is visible between the two trajectories, which could be interpreted as a misalignment of the underlying gestures.

The correlation coefficient $r(x,v)$ of the trajectories is given by

$$r(x,v) = \frac{\sum\limits_{n=1}^{N}(x_n - \bar{x})(v_n - \bar{v})}{\sqrt{\sum\limits_{n=1}^{N}(x_n - \bar{x})^2 \sum\limits_{n=1}^{N}(v_n - \bar{v})^2}} \tag{4.4}$$

where $\bar{x}$ and $\bar{v}$ represent the mean of x and v, respectively.

$r_{x,v}$ was computed for each gestural score in the search space. Figure 4.9a shows the distribution of the correlation coefficient values; most lie around the mean. Once again, this indicates that only relatively small number of gestural scores produce a very high correlation, and these are shown as the (upper) outliers.

In analogy to the RMSE-based cost function in the previous section, the vertex trajectory synthesized by the score with the highest correlation coefficient with respect to each target utterance is plotted against that utterance's EMA trajectory in Figure 4.9b. The results seem more encouraging; however, the same observations can be formulated:

1. A significant amount of temporal misalignment is visible where the slopes of the gesture-induced vertex peaks are not synchronized with the corresponding slopes in the EMA trajectory. This can be attributed to the low frame rate enforced by constraints of search space size in the brute-force task. A slightly earlier or later gesture boundary would yield a better alignment, but is not possible because of the relatively wide spacing (0.1 s) of frame boundaries. In a more efficient search paradigm, a higher frame rate becomes feasible, which is very likely to remove this issue.

2. The slopes of the rest gestures (`sil`), as well as the onset of the first `z` gesture, are noticeably steeper in the vertex trajectories than in the target EMA trajectories; this can be seen in both utterances, but more so in `092`.

4.3.4 Conclusion

The brute-force search approach has demonstrated that it is possible to use an error function in such a way that it can be applied to generate (i.e. select from the search space represented by a FSA with a given frame rate) a gestural score that allows VTL to synthesize trajectories approximating the kinematics of a known target utterance for which EMA data is available. Several limitations inherent to this approach become evident, some of which can be solved by implementing a more efficient search algorithm.

To adequately generate gestural scores with a better fit, the frame rate must be increased to at least double that of the 10 Hz used here. This is clearly a requirement if the slopes in the original and synthetic trajectories are to be more tightly synchronized in a robust way.

Furthermore, the `effort` parameter should be more flexible to enable a closer fit of slopes, although this is secondary compared to the principal task of generating appropriate gestural durations. Nevertheless, the flexibility of matching different slopes in the original trajectories could help improve resynthesis results further. While it is possible to change such a slope by modifying the corresponding gesture's `effort` parameter, each additional value in a hypothetical set of `effort` parameter values would add another "layer" to the transition network, exponentially increasing the size of the search space. The performance constraints of the brute-force search prohibit such an additional level of complexity, but a more efficient resynthesis approach may allow reconsideration of this aspect.

These two issues are addressed in the following section, which implements a dynamic programming algorithm that makes the resynthesis vastly more efficient than the brute force search. It should be remembered, however, that the exhaustive search of the search space described in this section serves merely to show that optimal gestural scores can be generated in this way, and not to apply the brute force search repeatedly.

4.4 Viterbi search

The task of finding the optimal gestural score for a given utterance can be divided into a sequence of subtasks of finding optimal sub-scores for each frame-sized segment of the utterance. This allows a dynamic programming approach to be implemented which reduces the time required to find the optimal score by several orders of magnitude.

One such approach, known as a *Viterbi search* and proposed by Viterbi (1967), was implemented for the resynthesis task. In its simplest form, it relies on the assumption that the optimal path to a given node in a transition network will be an extension of the optimal path to the previous node in that path. By extension, if the optimal paths to all nodes in a given frame f_i of the transition network are known, the number of possibly optimal paths to the nodes in the next frame f_{i+1} will be among the possible extensions of these paths, the number of which is significantly smaller than the number of all possible paths to the nodes of f_{i+1} from the first frame f_1.

Put differently, while moving through the transition network frame by frame, the task in the Viterbi search consists of finding and storing the optimal path to every node in the current frame, and discarding all other paths to these nodes.

The benefit of applying the Viterbi algorithm to the search for the optimal gestural score in a resynthesis task is that the number of generated scores is equal to the number of transitions in the network. This reduces the exponentially large search space to one whose size increases linearly with the number of frames and gestures. In fact, only frame-sized sub-scores are generated, which further decimates the computational expense.

4.4.1 Implementation

As the Viterbi search moves from start to end through the transition network, generating the candidate gestural scores one frame at a time, the synthesized vertex trajectories differ only within the current frame. Each such segment of a score can itself be interpreted as a short gestural score, with the particularity that it does not begin at time 0, which means that additional variables must be known to synthesize trajectories in this frame-wise fashion.

Frame-wise gestural score generation

In the VTL gestural model (Birkholz, 2007), every gesture is initialized with information about the state of each control parameter in the vocal tract model at the gesture's start time; the parameter value must be supplied, as well as the first and second derivative (velocity and acceleration, respectively).[10] At the onset of a gestural score, these states are set to default initial values, but at a later point in time, they are the result of the preceding gestures and depend on their previous states.

Fortunately, this requirement presents no obstacle to the Viterbi search, since for every frame after the first, the parameter states at the end of the previous frame can be observed and provided to the new frame's gestures. For the implementation of the frame-wise gestural score generation, VTL's functions were extended accordingly, with the result that a gestural score can be synthesized in separate consecutive segments just as well as in one piece, and produce the same vertex trajectories under both conditions.

Implementation details

The transition network was implemented as a vector of frames, each of which contains a list of nodes. Each node contains a pointer to the previous node, which allows the reconstruction of the path leading up to it. Furthermore, each node possesses properties which return the number f of the frame it belongs to, the index g of its associated gesture in the globally known sequence of canonical gesture pairs, as well as the `effort` parameter value e of this gesture. These properties are used to generate a frame-sized gestural score with the corresponding gestures on the `CONSONANT` and `VOWEL` tiers, respectively. This score can be loaded into VTL, which allows the synthesis of vertex trajectories for that node.

The vertex data for a given node remains constant, so that it can be stored and retrieved whenever the vertex trajectories for a path containing that node are needed. Along with the corresponding segment of the original EMA trajectories, this data also allows for the calculation of the RMSE for each node, as well as the correlation coefficient

[10] These correspond to the coefficients $c_{1,i}$, $c_{2,i}$, and $c_{3,i}$ in Equation 1.1.

of the path leading up to it. Both of these values are stored as further properties of the node.

Finally, each node is able to generate possible next nodes in the transition network's next frame, depending on the frame number, the next element in the sequence of canonical gesture pairs, and the set of possible `effort` values. These next nodes are then assigned to the next frame.

Depending on its position in the transition network, a frame may have more than one node for each unique property tuple $\langle f, g, e \rangle$; the cost for each of these nodes is used to determine the single optimal node for each tuple, and all others are removed from the frame's node list. This reduction step collapses the paths leading into each frame and is largely responsible for the Viterbi search's efficiency.

To further improve the performance of the search implementation, parallel programming was used to synthesize the vertex data for several nodes simultaneously. This synthesis is carried out in subprocesses distributed evenly over all available CPUs of the compute machine performing the resynthesis.

4.4.2 Viterbi search results

The Viterbi search was performed using the same cost functions as in the brute force search. The main difference in applying them is the fact that only the vertex data in the current and previous frames is known; however, the results are essentially equivalent. What distinguishes the Viterbi search is that it generates a gestural score for an utterance in the CV corpus in a mere 6 s on the same compute server as was used for the brute force resynthesis (cf. Section 4.3.2). The resynthesized trajectories are shown in Figure 4.10a (RMSE) and Figure 4.10b (correlation). In the latter case, the onset of the first sibilant gesture in utterance `092` displays a slight mismatch with respect to the gestural score found by the same cost function in the brute force search space, but this can be accounted for by the low frame rate and consequences of the path reduction during the Viterbi search.

The very significant increase in performance over the baseline allows resynthesis parameters that were not realistic previously. The obvious next step lies in doubling the number of frames to 40. This increases the processing time to around 10 s per utterance, but the results indicate an improved fit of vertex trajectories to the respective EMA data. The results for resynthesis at 20 fps are shown in Figure 4.10c (RMSE) and Figure 4.10d (correlation). While the correlation based cost function appears to perform as expected, the RMSE based cost function appears to be critically sensitive to the lack of normalization between the trajectories.

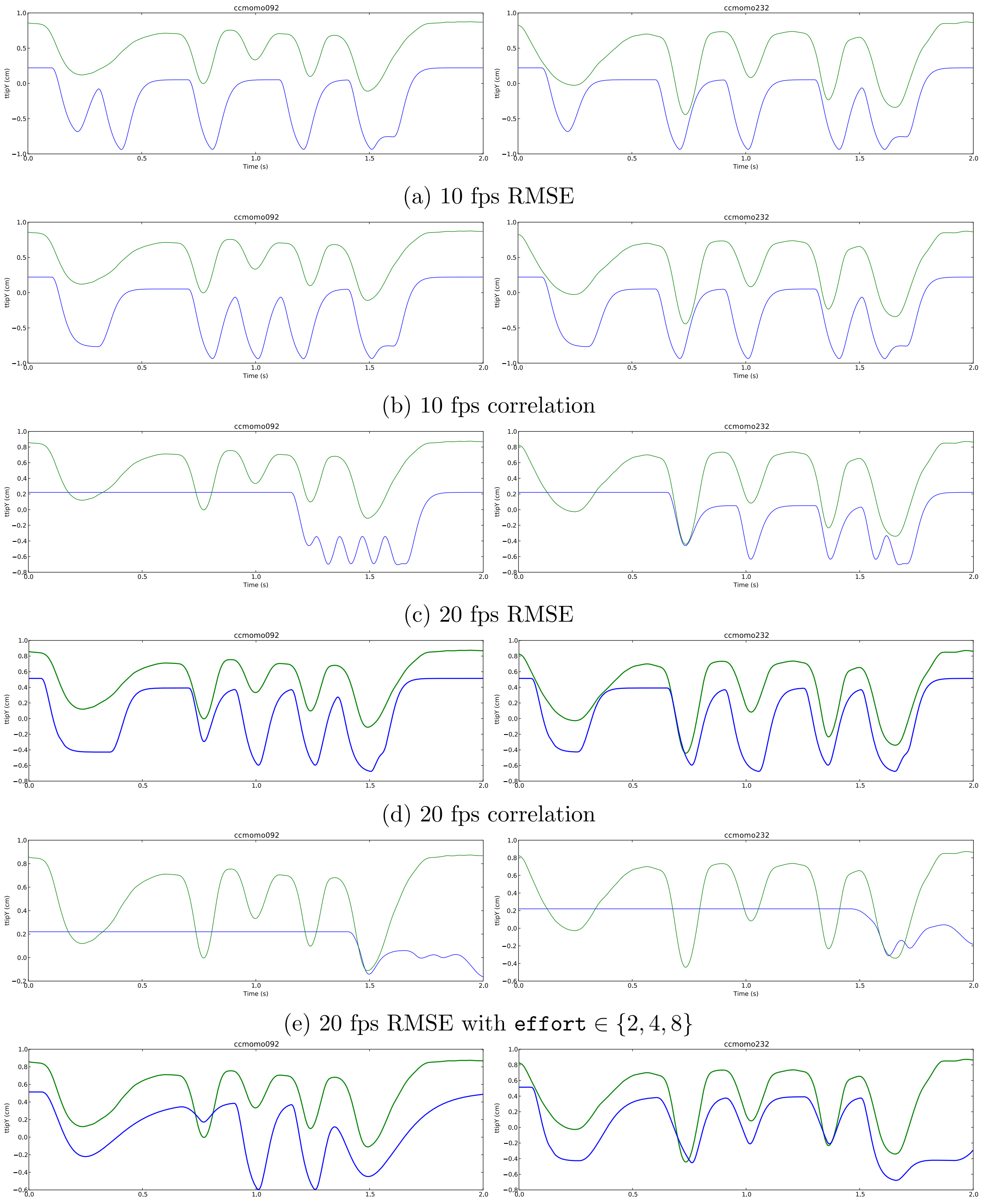
This phenomenon is akin to that observed by Zacks & Thomas (1994), who explore cost functions based on standard error and correlation to train neural networks to resynthesize XRMB trajectories for ASR purposes. They describe very similar problems with the purely error-based cost function, with the networks learning a straight line through the mean of the desired articulatory trajectory. It becomes apparent that a correlation (or similar) component is mandatory to avoid such issues.

Finally, the effect of variable `effort` parameter values for gestures was explored. Adding values of `2` and `4` to the default value `8` in the set of possible gesture effort values increases the complexity of the transition network significantly, raising the processing time per utterance to just under 40 s. The results, however, are not particularly encouraging, and are displayed in the final quadruplet of plots in Figure 4.10. The RMSE based cost function once again generates a straight line through most of the utterance, but the correlation based cost function seems to overzealously exploit its new ability to generate more gradual slopes, significantly undershooting the target values in several places.

4.4.3 Conclusion

The resynthesis of simple CV utterances produces useful results when a cost function based on correlation is used to compare the synthesized vertex trajectory to the desired EMA data. This carries over to an efficient dynamic programming approach that allows much higher frame rates and, consequently, better alignment through greater precision in gesture boundary placement. The mismatch in vocal tract geometry between the speaker of the CV corpus and the vocal tract model in VTL did not present a problem, but this can be attributed mainly to the nature of the correlation coefficient used in the successful cost functions, which does not take absolute difference into account when comparing trajectories.

The reliance on the comparison of but a single trajectory pair is however likely to pose a problem for the application of this resynthesis approach to more complex, natural speech material. A more elaborate cost function will be required which takes into account the fact that different places of articulation require focus on different articulatory trajectories during the corresponding times. This will be explored further in the next chapter.



(a) 10 fps RMSE

(b) 10 fps correlation

(c) 20 fps RMSE

(d) 20 fps correlation

(e) 20 fps RMSE with `effort` $\in \{2, 4, 8\}$

(f) 20 fps correlation with `effort` $\in \{2, 4, 8\}$

Figure 4.10: Results for Viterbi-based gestural score generation. Target realization `092` (left) and `232` (right) of prompt [zazazaza]; each plot shows original EMA (green) and synthesized vertex (blue) trajectories for `ttipY`. The cost function is either RMSE or correlation based. Conditions are 10 fps (Figures a, b), 20 fps (Figures c, d), and 20 fps with variable `effort` parameters (Figures e, f)

Appendix 4.A Prompt list for CV corpus

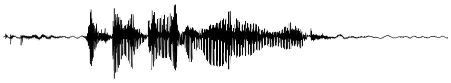
Note: Prompts are transcribed using the SAMPA (Wells, 1997). RUHE denotes an empty or invalid recording sweep.

001 RUHE
002 mamamama
003 memememe
004 mimimimi
005 momomomo
006 mumumumu
007 mEmEmEmE
008 m2m2m2m2
009 mymymymy
010 mImImImI
011 mOmOmOmO
012 mUmUmUmU
013 m9m9m9m9
014 mYmYmYmY
015 m@m@m@m@
016 m6m6m6m6
017 nananana
018 nenenene
019 nininini
020 nononono
021 nunununu
022 nEnEnEnE
023 n2n2n2n2
024 nynynyny
025 nInInInI
026 nOnOnOnO
027 nUnUnUnU
028 n9n9n9n9
029 nYnYnYnY
030 n@n@n@n@
031 n6n6n6n6
032 NaNaNaNa
033 NeNeNeNe
034 NiNiNiNi
035 NoNoNoNo
036 NuNuNuNu
037 NENENENE
038 N2N2N2N2
039 NyNyNyNy
040 NINININI
041 NONONONO
042 NUNUNUNU
043 N9N9N9N9
044 NYNYNYNY
045 N@N@N@N@
046 N6N6N6N6
047 jajajaja
048 jejejeje
049 jijijiji
050 jojojojo
051 jujujuju
052 jEjEjEjE
053 j2j2j2j2
054 jyjyjyjy
055 jIjIjIjI
056 jOjOjOjO
057 jUjUjUjU
058 j9j9j9j9
059 jYjYjYjY
060 j@j@j@j@
061 j6j6j6j6
062 RaRaRaRa
063 ReReReRe
064 RiRiRiRi
065 RoRoRoRo
066 RuRuRuRu
067 RERERERE
068 R2R2R2R2
069 RyRyRyRy
070 RIRIRIRI
071 RORORORO
072 RURURURU
073 R9R9R9R9
074 RYRYRYRY
075 R@R@R@R@
076 R6R6R6R6
077 vavavava
078 veveveve
079 vivivivi
080 vovovovo
081 vuvuvuvu
082 vEvEvEvE
083 v2v2v2v2
084 vyvyvyvy
085 vIvIvIvI
086 vOvOvOvO
087 vUvUvUvU
088 v9v9v9v9
089 vYvYvYvY
090 v@v@v@v@
091 v6v6v6v6
092 zazazaza
093 zezezeze
094 zizizizi
095 zozozozo
096 zuzuzuzu
097 zEzEzEzE
098 z2z2z2z2
099 zyzyzyzy
100 zIzIzIzI
101 zOzOzOzO
102 zUzUzUzU
103 z9z9z9z9
104 zYzYzYzY
105 z@z@z@z@
106 z6z6z6z6
107 ZaZaZaZa
108 ZeZeZeZe
109 ZiZiZiZi
110 ZoZoZoZo
111 ZuZuZuZu
112 ZEZEZEZE
113 Z2Z2Z2Z2
114 RUHE
115 ZyZyZyZy
116 ZIZIZIZI
117 ZOZOZOZO
118 ZUZUZUZU
119 Z9Z9Z9Z9
120 ZYZYZYZY
121 Z@Z@Z@Z@
122 Z6Z6Z6Z6
123 lalalala
124 lelelele
125 lililili
126 lolololo
127 lulululu
128 lElElElE
129 l2l2l2l2
130 lylylyly
131 lIlIlIlI
132 lOlOlOlO
133 lUlUlUlU
134 l9l9l9l9
135 lYlYlYlY
136 l@l@l@l@
137 l6l6l6l6
138 RUHE
139 mamamama
140 memememe
141 mimimimi

142 momomomo
143 mumumumu
144 mEmEmEmE
145 m2m2m2m2
146 mymymymy
147 mImImImI
148 mOmOmOmO
149 mUmUmUmU
150 m9m9m9m9
151 mYmYmYmY
152 m@m@m@m@
153 m6m6m6m6
154 nananana
155 nenenene
156 nininini
157 nononono
158 nunununu
159 nEnEnEnE
160 RUHE
161 n2n2n2n2
162 nynynyny
163 nInInInI
164 nOnOnOnO
165 nUnUnUnU
166 n9n9n9n9
167 nYnYnYnY
168 n@n@n@n@
169 n6n6n6n6
170 NaNaNaNa
171 NeNeNeNe
172 NiNiNiNi
173 NoNoNoNo
174 NuNuNuNu
175 NENENENE
176 N2N2N2N2
177 NyNyNyNy
178 RUHE
179 NINININI
180 NONONONO
181 NUNUNUNU
182 N9N9N9N9
183 NYNYNYNY
184 N@N@N@N@
185 N6N6N6N6
186 jajajaja
187 jejejeje
188 jijijiji
189 jojojojo
190 jujujuju
191 jEjEjEjE
192 j2j2j2j2
193 jyjyjyjy
194 jIjIjIjI
195 jOjOjOjO
196 jUjUjUjU
197 j9j9j9j9
198 jYjYjYjY
199 j@j@j@j@
200 j6j6j6j6
201 RaRaRaRa
202 ReReReRe
203 RiRiRiRi
204 RoRoRoRo
205 RuRuRuRu
206 RERERERE
207 R2R2R2R2
208 RyRyRyRy
209 RUHE
210 RIRIRIRI
211 RORORORO
212 RURURURU
213 R9R9R9R9
214 RYRYRYRY
215 R@R@R@R@
216 R6R6R6R6
217 vavavava
218 veveveve
219 vivivivi
220 vovovovo
221 vuvuvuvu
222 vEvEvEvE
223 v2v2v2v2
224 vyvyvyvy
225 vIvIvIvI
226 vOvOvOvO
227 vUvUvUvU
228 v9v9v9v9
229 vYvYvYvY
230 v@v@v@v@
231 v6v6v6v6
232 zazazaza
233 zezezeze
234 zizizizi
235 zozozozo
236 zuzuzuzu
237 zEzEzEzE
238 z2z2z2z2
239 zyzyzyzy
240 zIzIzIzI
241 zOzOzOzO
242 zUzUzUzU
243 z9z9z9z9
244 zYzYzYzY
245 z@z@z@z@
246 z6z6z6z6
247 RUHE
248 ZaZaZaZa
249 ZeZeZeZe
250 ZiZiZiZi
251 ZoZoZoZo
252 ZuZuZuZu
253 ZEZEZEZE
254 Z2Z2Z2Z2
255 ZyZyZyZy
256 ZIZIZIZI
257 ZOZOZOZO
258 ZUZUZUZU
259 Z9Z9Z9Z9
260 ZYZYZYZY
261 Z@Z@Z@Z@
262 Z6Z6Z6Z6
263 lalalala
264 lelelele
265 lilililI
266 lolololo
267 lulululu
268 lElElElE
269 l2l2l2l2
270 lylylyly
271 lIlIlIlI
272 lOlOlOlO
273 lUlUlUlU
274 l9l9l9l9
275 lYlYlYlY
276 l@l@l@l@
277 l6l6l6l6
278 IC ha:b@ Ra:t ?E6vE:nt
279 IC ha:b@ Ra:t ?E6vE:nt

Chapter 5

Resynthesis of natural German utterances



See Figure 5.1

Buiding upon the encouraging results from resynthesis experiments with CV utterances, this chapter explores the resynthesis of normal[1] utterances. To avoid introducing too much complexity into the resynthesis approach at once, VTL's default speaker (and phoneset) definition was retained for the resynthesis of utterances of German, and the EMA data used here was in fact recorded with the same speaker as was used by Birkholz & Kröger (2006) to create this speaker definition.

5.1 EMA data

The German EMA data used in this chapter was generously provided by Susanne Fuchs and Jörg Dreyer at the Center for General Linguistics (ZAS) in Berlin.

5.1.1 Corpus design and recording procedure

The data described here as the "German corpus" is actually a subset of a multispeaker EMA+EPG corpus recorded in 2000 at ZAS. The material was designed to investigate voicing constrast in the production of alveolar obstruents (Fuchs, 2005). Since one of these speakers was the same speaker who had been used for the MRI scans on which the vocal tract model in VTL is based (cf. Section 1.2.4), that speaker's EMA data comprises the corpus of regular German utterances used in this chapter.

In fact, the recordings were repeated for the relevant speaker in 2002 (due to problems with the EPG palate in the 2000 session), which produced twice the amount of EMA

[1] as opposed to nonsense CV

material. However, the EMA data in the 2002 session contains one less measurement coil (see below).

Each session contains two parts, one with nonsense words and a second with normal utterances. While only the latter were used in the resynthesis task, both parts are briefly described here for the sake of completeness.

Nonsense words

The prompts for the first part of each session are a set of nonsense words in a carrier phrase, *Ich habe "geCVCe" nicht "geCVC" erwähnt* ("I mentioned 'geCVCe', not 'geCVC'."). C and V represent consonants from the set [d, t, z, s] and vowels in [a, i, u], respectively. The consonants are voiced or unvoiced according to German grapheme-to-phoneme rules (depending on the position in the words in which they occur).

A set of 15 prompts was constructed from a simple template, permuting CV combinations, and additionally introducing two conditions for each vowel (except in the case of [d]), tense and long or lax and short. The set of prompts, listed with the corresponding ID codes, is shown here (the spelling of nonsense words follows Fuchs (2005)):

A+D Ich habe gedaade nicht gedaad erwähnt.
A+S Ich habe gesahse nicht gesahs erwähnt.
A+T Ich habe getaate nicht getaat erwähnt.
A-S Ich habe gesasse nicht gesass erwähnt.
A-T Ich habe getatte nicht getatt erwähnt.
I+D Ich habe gediede nicht gedied erwähnt.
I+S Ich habe gesiehse nicht gesiehs erwähnt.
I+T Ich habe getiehte nicht getieht erwähnt.
I-S Ich habe gesisse nicht gesiss erwähnt.
I-T Ich habe getitte nicht getitt erwähnt.
U+D Ich habe geduhde nicht geduhd erwähnt.
U+S Ich habe gesuhse nicht gesuhs erwähnt.
U+T Ich habe getuhte nicht getuht erwähnt.
U-S Ich habe gesusse nicht gesuss erwähnt.
U-T Ich habe getutte nicht getutt erwähnt.

These 15 prompts were repeated 10 times in randomized order, for a total of 150 utterances per session. Finally, each part was concluded with a palate trace created by the speaker sliding his tongue backward along the palate.

The utterances with nonsense words were however not used here. They could have served as a replacement for the CV data in the previous chapter, but the carrier phrase would have made the resynthesis, even if based solely on tongue tip movements, more complicated, potentially outweighing the benefit of using EMA data from the same speaker as VTL's vocal tract model. In any case, the CV corpus had been available to the author much earlier and was used for that reason, if no other.

Normal utterances

The second part of each recording session consists of 24 short, semantically meaningful utterances. Again, the prompts were constructed based on a carrier template, and are listed here, together with the corresponding ID codes:

FDXXA Ich habe mein Rad erkannt.
FDXXI Ich habe das Lied erkannt.
FSXXA Ich habe das Glas entleert.
FSXXI Ich habe das Verlies erspäht.
FSXXU Ich habe einen Fuß erspäht.
FTXXA Ich habe Rat erbeten.

`FTXXI` Ich habe den Kredit erhalten.
`FTXXU` Ich habe das Blut entfernt.
`IDMDU` Ich habe mehrere Duden gekauft.
10 `IDMTA` Ich habe Daten analysiert.
`IDMTI` Ich habe Dieter gesehen.
`ISMSU` Ich habe Suse getroffen.
`ISXXA` Ich habe Salz verstreut.
`ISXXI` Ich habe das volle Silo geleert.
15 `ITMTU` Ich habe eine Tute gekauft.
`ITXXA` Ich habe eine Tat vollbracht.
`ITXXI` Ich habe ihre Titel geändert.
`MDXXA` Ich habe Adam gesehen.
`MDXXI` Ich habe Lieder gesungen.
20 `MDXXU` Ich habe Sud entfernt.
`MSXXA` Ich habe eine Blase entfernt.
`MSXXI` Ich habe Lise gesehen.
`MTXXA` Ich habe den Atem angehalten.
`MTXXI` Ich habe den Liter getrunken.

The prompt list was recorded seven times in randomized order, which yielded 168 recorded utterances (7 realizations of each of the 24 prompts) per session, followed by a palate trace, just as is the nonsense prompt list.

A full listing of the utterances as recorded in this part of both sessions is given in Appendix 5.A.

Recording setup

The subject, a male native speaker of German, was recorded using a Carstens AG100 EMMA system. All utterances were recorded in sweeps of 3 s duration; the EMA signals were recorded at a sample rate of 500 Hz, subsequently downsampled to 200 Hz, while the audio was sampled at 16 kHz.

EMA measurement coils were attached to the lower lip (`llip`), jaw (`jaw`, i.e. lower incisors), and tongue tip (`ttip`), blade (`tblade`), dorsum (`tdorsum`), and back (`tback`). Reference coils were placed on the bridge of the nose (`nose`) and upper incisors (`upinc`). About one third into the recording of the nonsense part in the 2002 session, the `tback` coil became detached, and so the data from this coil is not available for the second session.

In addition to the EMA recording, linguopalatal contacts were simultaneously measured using EPG, so the speaker was wearing an artificial palate. However, the EPG data is not used here.[2]

Finally, during postprocessing, the EMA data was rotated so that the `x` axis lies in the occlusal (bite) plane of the speaker. The entire recording procedure is described in depth by Fuchs (2005).

Due to the focus of the research for which this corpus was designed (e.g. Fuchs & Perrier, 2003; Fuchs, 2005; Hamann & Fuchs, 2010), the coverage of the German phonetic inventory is of course far from complete. Nevertheless, the EMA data was deemed suitable for the resynthesis task at hand, as this shortcoming is greatly outweighed by the advantage that it largely eliminates the need for cross-speaker vocal tract normalization.

Figure 5.2a shows the distribution of EMA data for both recording sessions of the German corpus.

[2] Nevertheless, the presence of the palate seems to have influenced the speaker's articulation to a small degree (more so in the 2000 session than in the 2002 re-recording), at least based on subjective impression of listening to the audio.

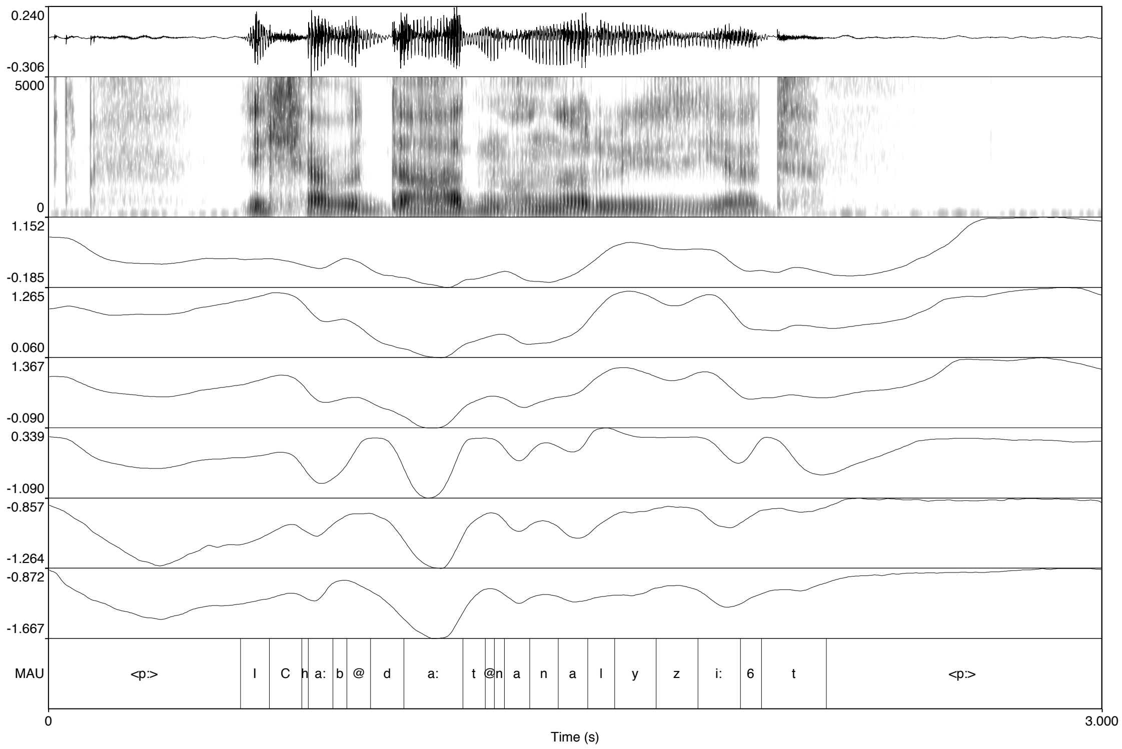


Figure 5.1: One utterance (032), *Ich habe Daten analysiert* ("I've analyzed data"), from session 1 of the German corpus; waveform, spectrogram, selected EMA trajectories (tbackY, tdorsumY, tbladeY, ttipY, jawY, llipY), acoustic segmentation (*not* gestures)

5.1.2 Automatic segmentation and manual correction

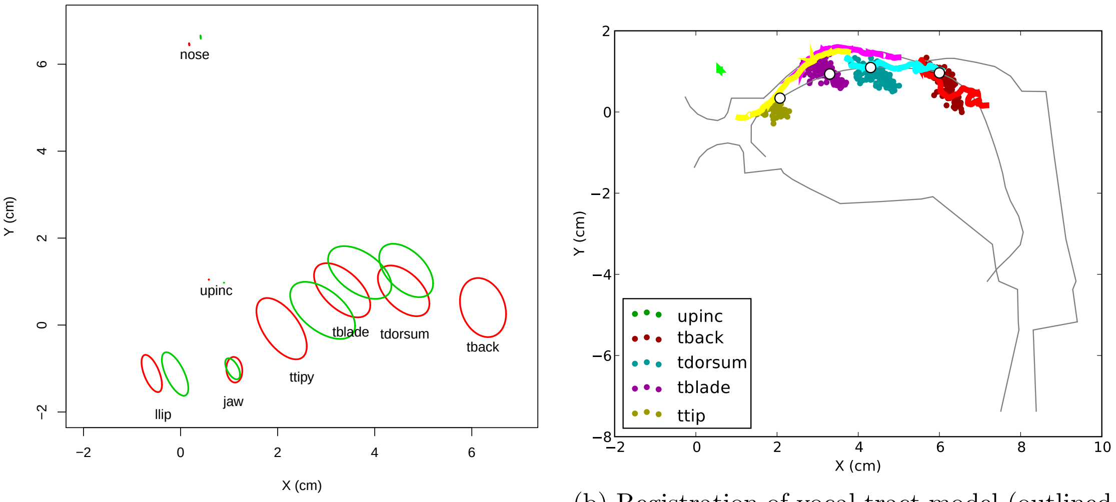
The utterances in the German corpus were automatically segmented using the Munich Automatic Segmentation System (MAUS)[3] v2.12, a C shell script (Joy, 1980) which makes use of pre-trained acoustic models for German, as well as several modules from the Hidden Markov Model Toolkit (HTK)[4] (Young et al., 2009) and signal processing through SoX[5]. Although MAUS is able to handle pronunciation variants (Kipp et al., 1996), non-prompted speech (Schiel, 1999), and iterative processing (Schiel, 2004), the present task was only one of forced alignment, since the material recorded had been read from the prompt list and monitored for mistakes.

Because of the relatively small number of prompts and their template-based design, it was straightforward to collect the word forms occurring in the corpus into a small custom dictionary and to transcribe their pronunciation by hand. By consulting this dictionary, a canonical phone string could be automatically generated from the prompt for each valid utterance, and this phone string was used as input for MAUS, along with

[3] http://phonetik.uni-muenchen.de/forschung/Verbmobil/VM14.7eng.html

[4] http://htk.eng.cam.ac.uk/

[5] http://sox.sourceforge.net/



(a) EMA data for the German corpus; the ellipses represent the 0.9 confidence level contours for all samples from each coil. Data from session 1 is shown in red, that from session 2, in green. Note the absence of the `tback` coil in the session 2 data. The apparent misalignment between data from the two sessions could be due to differences in fitting the EMA helmet, coil placement, rotation during postprocessing, or any combination of such factors.

(b) Registration of vocal tract model (outlined in gray) to EMA data, and adaptation to rest configuration (`sil`). The thick, brightly colored lines represent the EMA trajectories from the palate trace, while the darker scatter marks show the initial samples from all normal utterances in session 1. Vertices on the vocal tract model's tongue surface were selected to approximate these positions (see text) and are shown as white circles.

Figure 5.2: Distribution and registration of EMA data in the German corpus

the corresponding audio file.

While the process of automatic segmentation was very fast and worked “out of the box”, cursory visual inspection of the resulting label files alongside the acoustic signals indicated that some manual correction of boundary placement was required, which was then performed by a phonetically trained transcriptionist. Due to the size of the German corpus, however, the amount of labor was relatively small. Additionally, in most cases, the automatic segmentation proved a very good starting point for the manual adjustment.

5.1.3 Data registration and rest position

In contrast to the EMA data used for the initial resynthesis experiments in the previous chapter, the speaker in the German corpus is the one whose MRI data was used to adapt the vocal tract model and configure the phoneset in VTL. This is of course the primary reason for using the EMA data from the German corpus.

The main advantage of this situation is that the same vocal tract forms the basis of both the vocal tract model and the EMA data. Nevertheless, a few adjustments were necessary before the EMA trajectories from the German corpus could be compared to the vertex trajectories synthesized by VTL. The main differences to the registration of the CV data (cf. Section 4.1.2) are the availability of reference coils and palate traces in the EMA data, as well as the fact that the anatomy of the vocal tract is the exactly the same.

For these reasons, it was straightforward to reflect and translate the vocal tract model's vertices to match the EMA data, using the position of the `upinc` reference coil and the palate trace as a guide; the result is illustrated in Figure 5.2b.

To obtain a VTL `phone` for the speaker's rest position (reducing problems from the posture effect, see below and Section 4.1.3), the target configuration of the vocal tract model's tongue contour for the `sil` entry in the `phoneList` was manually adjusted to roughly correspond to the position of the tongue coils in the initial samples of each recording sweep in the EMA data. In a final step, the selection of tongue vertices tracked for the EMA interface was updated to coincide approximately with the centers of distribution of these initial samples. These vertices are also shown in Figure 5.2b, along with the initial EMA samples. This was necessary to improve the expected trajectory fit in the resynthesis task; the match in vocal tract geometry must be combined with the appropriate correspondence of EMA coil arrangement and vertex selection.

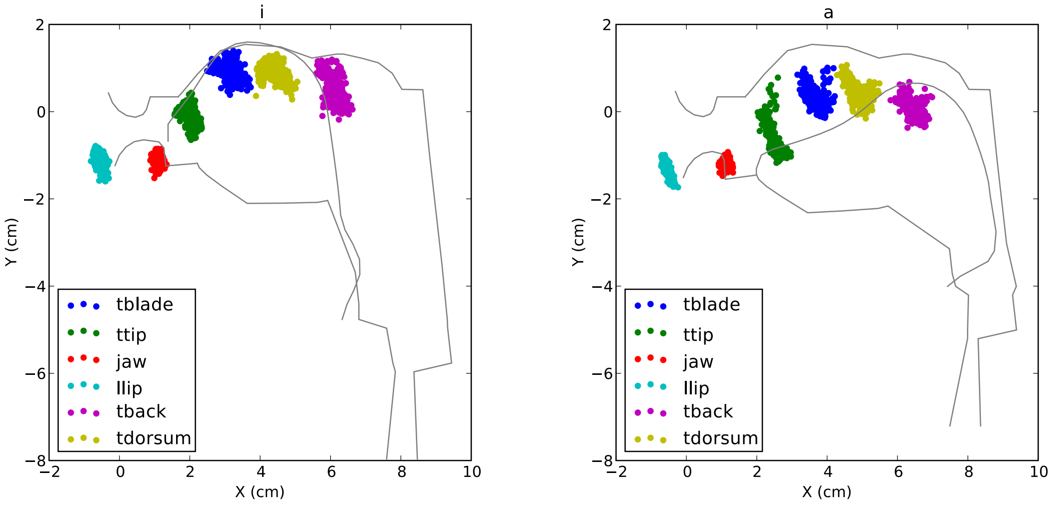
The vertex selection is limited by the number and spacing of available tongue vertices, as well as by the fact that a `ttip` vertex too close to the tip of the model's tongue will behave quite unnaturally when the tongue surface is stretched by raising its tip. To avoid such geometric artifacts, a posterior vertex must be selected.

5.1.4 Posture effect

When combining, or *registering*, articulatory data sets recorded under different conditions, a number of factors may influence the kinematics and target positions observed during speech production. Because of the architecture of the recording or scanning apparatus, the conditions may enforce a specific posture to be assumed by the subject. Specifically, modalities such as MRI usually require a supine posture, while EMA usually requires the subject to be seated in an upright position. The effects of gravity and head posture on the vocal tract are more than likely to influence the shape of the vocal tract and the position of the articulators, so these effects must be taken into account when data from modalities with different recording postures are registered.

A number of previous studies have attempted to explore this *posture effect*, using several different modalities. This research includes that of Whalen (1990), who compared velum height during five English vowels using upright cineradiography and supine MRI; Tiede et al. (1997), who compared vowels and running speech using upright and supine EMA; Tiede et al. (2000), who compared Japanese vowels, CV sequences, and running speech using upright and supine XRMB; Stone et al. (2002, 2007), who compared English word sequences using upright and supine UTI; Kitamura et al. (2005), who compared Japanese vowels using upright and supine MRI in an *open* scanner; and Engwall (2006), who compared Swedish vowels and sustained consonants using supine and prone MRI.

The conclusions drawn by these studies suggest that the magnitude of the posture effect is principally speaker-specific, and can become quite prominent. Furthermore, it seems that speakers usually attempt to adjust for the effects of gravity and head posture



(a) Vowel [i] and `phone i`

(b) Vowel [a] and `phone a`

Figure 5.3: Vocal tract model contours for two vowels. The scatter marks represent the EMA samples at the midpoints of all *acoustic* segments labeled with the respective vowel.

on their articulators in such a way that the acoustics of their speech are preserved, while this adjustment may influence the actual configuration of their articulators to varying degrees.

It can be assumed that when speakers are *not* speaking while in lying in a supine position, they will tend to allow gravity and head position to affect their articulators, which retracts the tongue body into the pharynx and (to a lesser extent) shifts the mandible backwards. This will affect the rest position much more than the neutral position and articulation itself, during which most speakers attempt to compensate for the posture effect to some degree.

The consequences of the posture effect become an issue for the resynthesis approach, since the EMA data was recorded with the speaker sitting upright in the AG100, while the target configurations of the vocal tract model were adapted to MRI scans acquired with the speaker in a supine position. Figure 5.3 shows vocal tract model configuration for the `phones` corresponding to two vowels, superimposed on a scatter plot showing the distribution of EMA data samples in the session 1 data, at the midpoints of all acoustic segments labeled with the respective vowel. The posture effect is manifested as a noticeable mismatch in vocal tract configuration despite the fact that the same configuration would be expected, which unduly increases the error value returned by distance-based cost functions when comparing corresponding trajectory pairs in the EMA interface.

5.2 Multi-trajectory cost function

For successful resynthesis, the utterances in the German corpus require a more elaborate cost function than the CV corpus. Whereas the CV prompts were designed to contain only a single consonant type in each utterance, this is not the case for the normal utterances of the German corpus. Despite the fact that the prompts are based on a template and designed to allow analysis of apical obstruent production, they contain a variety of different phones and are much less constrained phonotactically. These phones are of course not necessarily all produced with the same articulators, and so data from more than one EMA trajectory must be taken into account.

The cost function at the core of the Viterbi search based resynthesis, which was introduced in Section 4.3.3, is suited only to the comparison of a single pair of trajectories, which was sufficient in the CV paradigm. However, in the context of the increased complexity of the German data, and the necessity of comparing all available trajectory pairs to handle it, the cost function must be reformulated and extended to accommodate this multi-trajectory scenario. In this case, the multi-trajectory cost function is a combination of single-trajectory cost functions.

In the single-trajectory paradigm, the mean distance $d(x, v)$ between the EMA trajectory x and the corresponding vertex trajectory v, both vectors containing N samples, can be written as

$$d(x, v) = \frac{1}{N} \sum_{n=1}^{N} |x_n - v_n| \tag{5.1}$$

When comparing T trajectory pairs, $d(x, v)$ is given by

$$d(x, v) = \frac{1}{N \cdot T} \sum_{n=1}^{N} \sum_{t=1}^{T} |x_{n,t} - v_{n,t}| \tag{5.2}$$

which is equivalent to Equation 5.1 when $T = 1$.

Similarly, in the context of multiple trajectories, $\bar{r}(x, v)$ (cf. Equation 4.4) can be formulated as the mean of a set of correlation coefficients for the trajectories:

$$\bar{r}(x, v) = \frac{1}{T} \sum_{t=1}^{T} r(x_t, v_t) \tag{5.3}$$

However, since not every articulatory trajectory is equally relevant to the production of a given phone (this is true in a particularly obvious way for obstruents), the single-trajectory cost functions are combined in a way that reflects this uneven distribution of relevance; each trajectory is *weighted* according to its relevance to the production of the phone in question. This requires the consultation of some sort of look-up table or *weight matrix*, from which the appropriate weight for each combination of trajectory and phone can be retrieved as required.

To integrate the weighting, Equation 5.2 can be rewritten as

$$d(x, v) = \frac{1}{N \cdot T} \sum_{n=1}^{N} \sum_{t=1}^{T} \omega(n, t) |x_{n,t} - v_{n,t}| \tag{5.4}$$

where $\omega(n, t)$ is a weighting function (cf. Equation 5.9).

For correlation, a $N \times T$ matrix $R(x, v)$ can be defined, grouping the N samples into G segments of N_g samples, corresponding to the gestures, where $(N_1, N_2, \ldots, N_G)$ sum to N, such that

$$R_{n,t}(x, v) = r(x_{g,t}, v_{g,t}) \tag{5.5}$$

where $x_{g,t}$ and $v_{g,t}$ are the g^{th} segment of the t^{th} EMA and vertex trajectory, respectively, and g is determined from n. The correlation component of a cost function is then the matrix mean $\bar{R}(x, v)$.

Finally, the two components are combined into a cost function

$$C(x, v) = a \cdot d(x, v) + b \cdot (1 - \bar{R}(x, v)) \tag{5.6}$$

where a and b are scaling factors, to obtain the overall cost of multiple trajectories in the EMA interface.

5.2.1 Weight matrix

For the German resynthesis task, a weight matrix was defined manually, based on the system of the IPA (International Phonetic Association, 1999); more specifically, the place of articulation. Accordingly, the production of (for instance) labial consonants, which involves the lips, assigns a large weight value to trajectories from transducer coils attached to the lips for such phones. The same applies to dental and alveolar consonants with respect to the tongue tip and blade coils, etc. Vowels are not as easy to identify with specific coil configurations, but in general, the posterior tongue coils, as well as the lip (for rounded vowels) and jaw coils (for open vowels) are relevant.

In the case of the resynthesis approach detailed here, syllables and phones are processed as a sequence of gestures[6], i.e. CV tuples; a syllable σ is expanded into a gesture sequence as follows:

$$\sigma \longmapsto [C_1 V_1 C_2] \longmapsto ((\mathtt{C}_1, \mathtt{V}_1), (\mathtt{\_}, \mathtt{V}_1), (\mathtt{C}_2, \mathtt{V}_1)) \tag{5.7}$$

The weight matrix has the size $\Phi \times T$, where Φ is the number of phones in the phoneset (including `sil` and `_`) and T is the number of EMA trajectories. Each cell in the matrix contains a value in the range $[0, 1]$. One possible such weight matrix is shown in Listing 5.3.

Given a weight matrix and a gesture, the phones are unpacked from the gesture's CV tuple and used to retrieve a pair of weight vectors (w_c, w_v) from the matrix, each with one weight value per trajectory. A global scaling factor s is applied to combine these into the single vector

$$w_g = \frac{s \cdot w_c + (1 - s) \cdot w_v}{2} \tag{5.8}$$

The weighting function

$$\omega(n, t) \longmapsto w_{g,t} \tag{5.9}$$

[6] actually, linked *pairs* of gestures, cf. Section 3.2.1

```
label  llipX  llipY  jawX   jawY   ttipX  ttipY  tbladeX tbladeY tdorsumX tdorsumY tbackX tbackY
@      0      0      0      0      0      0      0       0       0.25     0.25     0.25   0.25
6      0      0      0      0      0      0      0       0       0        0        0      0
a      0      0      0      0      0      0      0       0       0.25     0.25     0.25   0.25
a:     0      0      0      0      0      0      0       0       0.25     0.25     0.25   0.25
b      0      1.00   0      0      0      0      0       0       0        0        0      0
C      0      0      0      0      0      0      0       0       0        1.00     0      0
d      0      0      0      0      0      1.00   0       0       0        0        0      0
e      0      0      0      0      0      0      0       0       0        0        0      0
e:     0      0      0      0      0      0      0       0       0        0        0      0
E      0      0      0      0      0      0      0       0       0        0        0      0
E:     0      0      0      0      0      0      0       0       0        0        0      0
f      0      0      0      0      0      0      0       0       0        0        0      0
g      0      0      0      0      0      0      0       0       0        0        0      0
h      0      0      0      0      0      1.00   0       0       0        0        0      0
i:     0      0      0      0      0      0      0       0       0        0        0      0
I      0      0      0      0      0      1.00   0       0       0        0        0      0
k      0      0      0      0      0      0      0       0       0        0        0      0
l      0      0      0      0      0      1.00   0       0       0        0        0      0
m      0      0      0      0      0      0      0       0       0        0        0      0
n      0      0      0      0      0      1.00   0       0       0        0        0      0
N      0      0      0      0      0      0      0       0       0        0        0      0
o      0      0      0      0      0      0      0       0       0        0        0      0
O      0      0      0      0      0      0      0       0       0        0        0      0
p      0      0      0      0      0      0      0       0       0        0        0      0
r      0      0      0      0      0      0      0       0       0        0        0      0
s      0      0      0      0      0      0      0       0       0        0        0      0
S      0      0      0      0      0      0      0       0       0        0        0      0
t      0      0      0      0      0      1.00   0       0       0        0        0      0
u:     0      0      0      0      0      0      0       0       0        0        0      0
U      0      0      0      0      0      0      0       0       0        0        0      0
x      0      0      0      0      0      0      0       0       0        0        0      0
y:     1.00   0      0      0      0      0      0       0       0.25     0.25     0.25   0.25
z      0      0      0      0      0      1.00   0       0       0        0        0      0
sil    0      0      0      0      0      0.25   0       0.25    0        0.25     0      0.25
_      0      0      0      0      0      0.25   0       0.25    0        0.25     0      0.25
```

Listing 5.3: Baseline weight matrix for the German corpus phoneset and EMA coils for session 1

maps sample n to the sample's corresponding gesture g and returns the weight for trajectory t according to Equation 5.8.

While the weights are applied to the trajectory distances as in Equation 5.4, the correlation component can be determined by defining a $N \times T$ matrix $W(h)$ where $h = (g_1, g_2, \ldots, g_G)$ such that

$$W_{n,t}(h) = \omega_{g,t} \tag{5.10}$$

in analogy to Equation 5.5. $W(h)$ is then multiplied with $R(x, v)$ before the matrix mean is calculated to obtain the correlation cost component.

Automatic determination of articulatory relevance

Of course, it would be preferable if the weight matrix was populated with appropriate values determined by a robust automatic process. This however requires a process able to recover measures for the relevance of individual articulatory trajectories from this kind of data.

One approach that may be able to accomplish this task is described by Jackson & Singampalli (2009). Their Acida tools[7] use EMA data from two speakers (`msak0` and `fsew0`) in the MOCHA-TIMIT corpus[8] (Wrench & Hardcastle, 2000; Wrench, 2000), as well as the accompanying acoustic segmentation[9], to compute the Kullback-Leibler divergence (KLD) for each EMA trajectory at the midpoints of all tokens of each segment type. The results can be interpreted to yield a measure of the articulatory relevance of each trajectory, as well as some information about the dependence relations between the individual trajectories.

Applying the same process to the German session 2 data, however, did not yield satisfactory results (see Table 5.1). In numerous cases, the trajectories determined to be relevant for a phone defy phonetic knowledge; in others, no trajectories are identified at all. Such obvious mismatches between the expected phone/trajectory correspondences and the results of this automatic approach can have a number of reasons, but among the most likely are the automatic approach's sensitivity to segmentation errors (the uncorrected labels were used), the small size of the German corpus, and the lack of phonetic balance; several phones appear exclusively in a single phonetic context.

While this approach may hold promise for the automatic determination of phonetic relevance in a larger EMA corpus with high-quality segmentation, the present resynthesis task resorts to phonetic expert knowledge to manually tune the values in the weight matrix.

[7] `http://personal.ee.surrey.ac.uk/Personal/P.Jackson/Dansa/Acida/`

[8] `http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html`

[9] Numerous misalignments in the segmentation of the MOCHA-TIMIT corpus had to be manually corrected (Singampalli, 2009).

phone	identified articulatory trajectories					
ə	ttipY					
ɐ	tdorsumY					
a	ttipY	ttipX	tdorsumY	jawX	llipY	
eː	jawY	tbladeY	llipY			
g	tbladeX	jawY	ttipY			
h	ttipX	tdorsumY	ttipY	jawY		
iː	tbladeX	ttipY	llipY	ttipX	tdorsumY	
l	tbladeX					
m	jawY	tdorsumY	ttipY	llipY	tdorsumX	
o	tdorsumY	tbladeX	ttipY	llipX	jawY	
ʁ	tdorsumX	ttipY	jawY	llipX	ttipX	tdorsumY
ʊ	ttipY	tdorsumY	tdorsumX	jawX	llipY	
x	jawY					
yː	jawY					
z	tbladeY					

Table 5.1: Sample results of applying the algorithm presented in Jackson & Singampalli (2009) to the session 2 EMA data. For each phone (in IPA notation) the articulatory trajectories identified as relevant are listed. Phones for which no relevant trajectories were found have been omitted here. See text for discussion.

5.3 Resynthesis using weighted multi-trajectory cost function

For the resynthesis of German utterances, the session 1 EMA data from the German corpus was used, since the trajectories from the tback EMA coil (missing from session 2) were expected to contribute to, and improve, performance. All EMA data was smoothed using a 11-sample Hann window. Initial results, however, are ambiguous and seem to indicate that the scaling factors in the multi-trajectory cost function, and above all, the values in the weight matrix, are critical to the resynthesis outcome.

5.3.1 Example

As an example, one utterance (032) was selected from session 1 of the German corpus and resynthesized; the utterance is one realization of the prompt *Ich habe Daten analysiert* ("I've analyzed data"), whose canonical phonetic transcription, [ʔɪç haːbə daːtən ʔanalyːziːɐ̯t], is modeled as the canonical gesture sequence[10]

```
sil _ . I C . h a: . b @ . d a: . t @ n . a . n a . l y: . z i: 6 t . _ sil
```

Resynthesis was carried out using the Viterbi search described in Section 4.4 with the cost function detailed in Section 5.2 as well as a weight matrix. Gestural score

[10] the CV gestures have been collapsed into a single flat sequence for reasons of legibility; syllable boundaries are represented by a period .

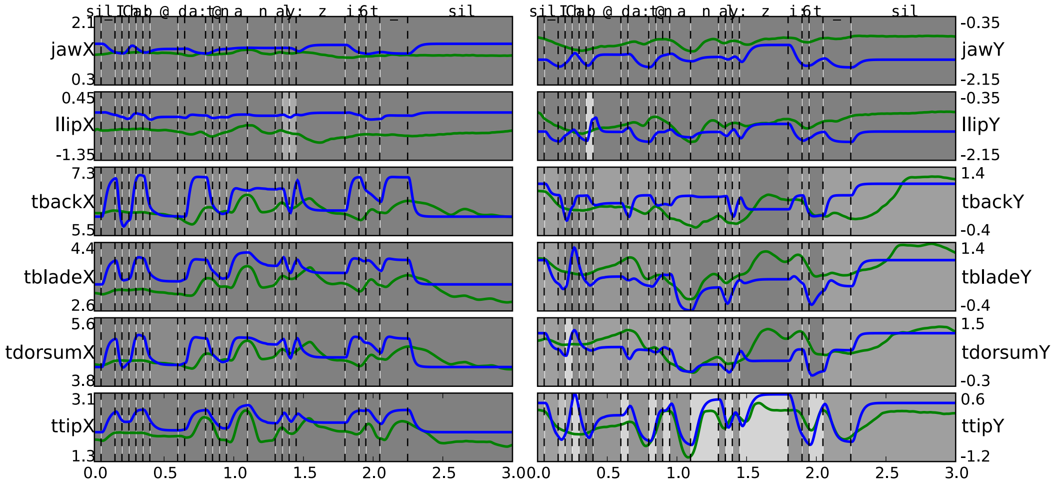


Figure 5.4: Resynthesis results at 20 fps with baseline weight matrix for utterance `032` (*Ich habe Daten analysiert*, "I've analyzed data") in session 1 data. The trajectory labels are displayed on the outer margins beside the trajectories. In each plot, the green curve shows the target EMA trajectory, while the blue curve represents the corresponding vertex trajectory synthesized by the gestural score selected as optimal by the multi-trajectory cost function in the Viterbi search. The vertical dashed lines represent the gesture boundaries, and the corresponding gesture names are shown as labels along the top plot row, with CV gestures flattened into a single canonical gesture sequence. The plot background is shaded to represent the weight for each gesture and trajectory; a dark gray corresponds to a weight of 0, while white represents an overall weight of 1; shades of gray between these extremes correspond to fractional weight values.

generation for one utterance can take anywhere from 70 s to over 20 min on the same compute server that was used in the previous resynthesis approaches, depending on the number of canonical gestures, the frame rate, the number of `effort` parameter values to consider, etc.

With the minimal weight matrix shown in Listing 5.3 as a baseline, the result of the resynthesis is displayed in Figure 5.4. The initial gestures occur much earlier than intended, and the resulting mismatches are evident across all trajectory plots, indicating that constraints imposed by the original timing are largely ignored.

Since the utterance under consideration features several consonants with an apical place of articulation,[11] concentration on the tongue tip holds some potential. Accordingly, in order to probe the impact this has on the weighting, the weight matrix was modified to give full weight exclusively to the `ttipY` trajectory for all consonants (even those not produced with the tongue tip, `b`, and `C`). The resynthesis was then repeated, with the

[11] which is not surprising, considering the original purpose of the corpus (cf. Section 5.1.1)

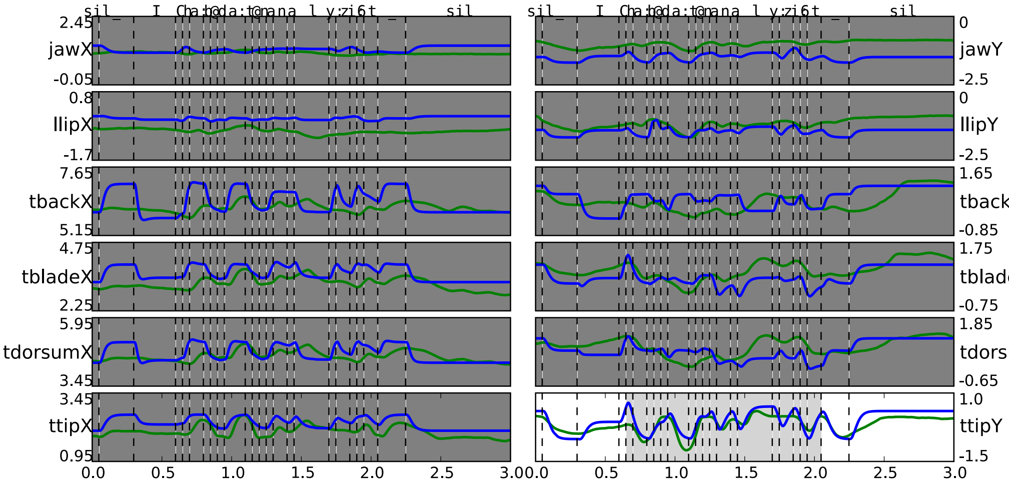


Figure 5.5: Resynthesis results at 20 fps with modified weight matrix for utterance 032 in session 1 data. For details, refer to Figure 5.4

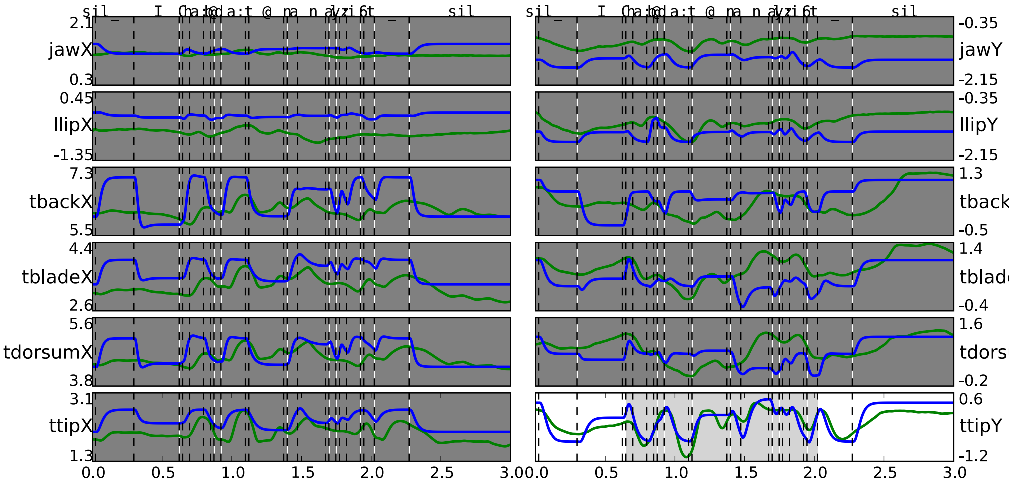


Figure 5.6: Resynthesis results at 40 fps with modified weight matrix for utterance 032 in session 1 data. For details, refer to Figure 5.4

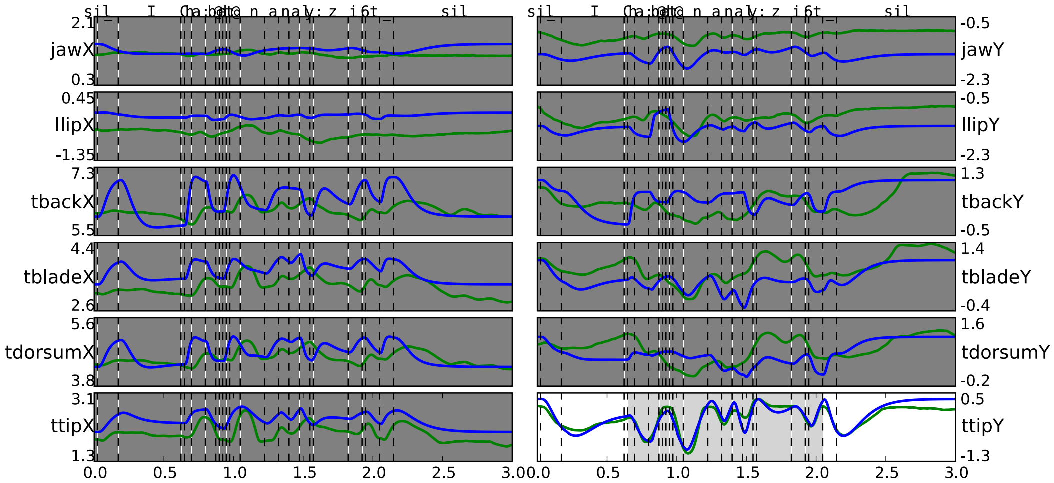


Figure 5.7: Resynthesis results at 40 fps and `effort` $\in \{2, 4, 8\}$ with modified weight matrix for utterance `032` in session 1 data. For details, refer to Figure 5.4

results shown in Figure 5.5.

While the outcome of using this modified weight matrix is certainly an improvement over the baseline weight matrix, the implications are somewhat sobering. Using only a single trajectory for the resynthesis seems like a step back to the CV resynthesis of the last chapter, and furthermore can only be expected to produce usable gestural scores when that one trajectory captures all relevant articulatory movements, as in the case of the utterance under consideration here.[12]

A further unexpected result is that the alignment with the original EMA seems to deteriorate slightly when the framerate is doubled to 40 fps (Figure 5.6). However, also extending the set of available `effort` value parameters to include $\{2, 4, 8\}$ produces the closest trajectory fit yet. The tradeoff is a processing time of over 25 min, with up to 271 nodes per frame in the 120 frame transition network. The result of this last resynthesis condition is presented in Figure 5.7. While the trajectory fit for `ttipY` is very good with these resynthesis parameters, the overall timing of the gestures, as well as the mismatch for the other trajectories, seem to indicate that other factors must be taken into account as well.

[12] There is one caveat: for an utterance such as this, the initial articulatory movements are insufficient to determine the onset of the utterance as perceived acoustically, which necessitates the consultation of the glottal source signal (through electroglottography (EGG) or pitch detection in the acoustic recording).

5.3.2 Results

Acoustic evaluation

To generate the waveform, using VTL, for a gestural score obtained by articulatory resynthesis, the gestures on the `CONSONANT` and `VOWEL` tiers of the score must be accompanied by appropriate gestures on the `VELIC_APERTURE`, `GLOTTAL_AREA`, and `F0` tiers. Since these tiers are irrelevant for the resynthesis described here, no gestures are generated, but it is fairly straightforward to add them post-hoc, using a set of rules applied to the CV gestures.

Specifically, velic opening gestures are inserted on the `VELIC_APERTURE` tier to co-occur with nasals (`m`, `n`, `N`) on the `CONSONANT` tier. Despite evidence that velic and lingual/labial gestures are not necessarily synchronized in this way (e.g. Kiritani et al., 1980; Engelke et al., 1996), this solution seems practical in the absence of EMA data from a velic coil. For stops (`p`, `t`, `k`, `b`, `d`, `g`) on the `CONSONANT` tier, velic gestures are inserted which seal the velum more tightly, preventing a pressure leak into the nasal cavity during the occlusion phase.

On the `GLOTTAL_AREA` tier, gestures controlling the degree of glottal abduction are inserted to co-occur with voiceless obstruents (`f`, `s`, `S`, `C`, `x`, `h`) on the `CONSONANT` tier; likewise, gestures controlling phonation are inserted during voiced consonants and vowels on the `CONSONANT` and `VOWEL` tiers, respectively.

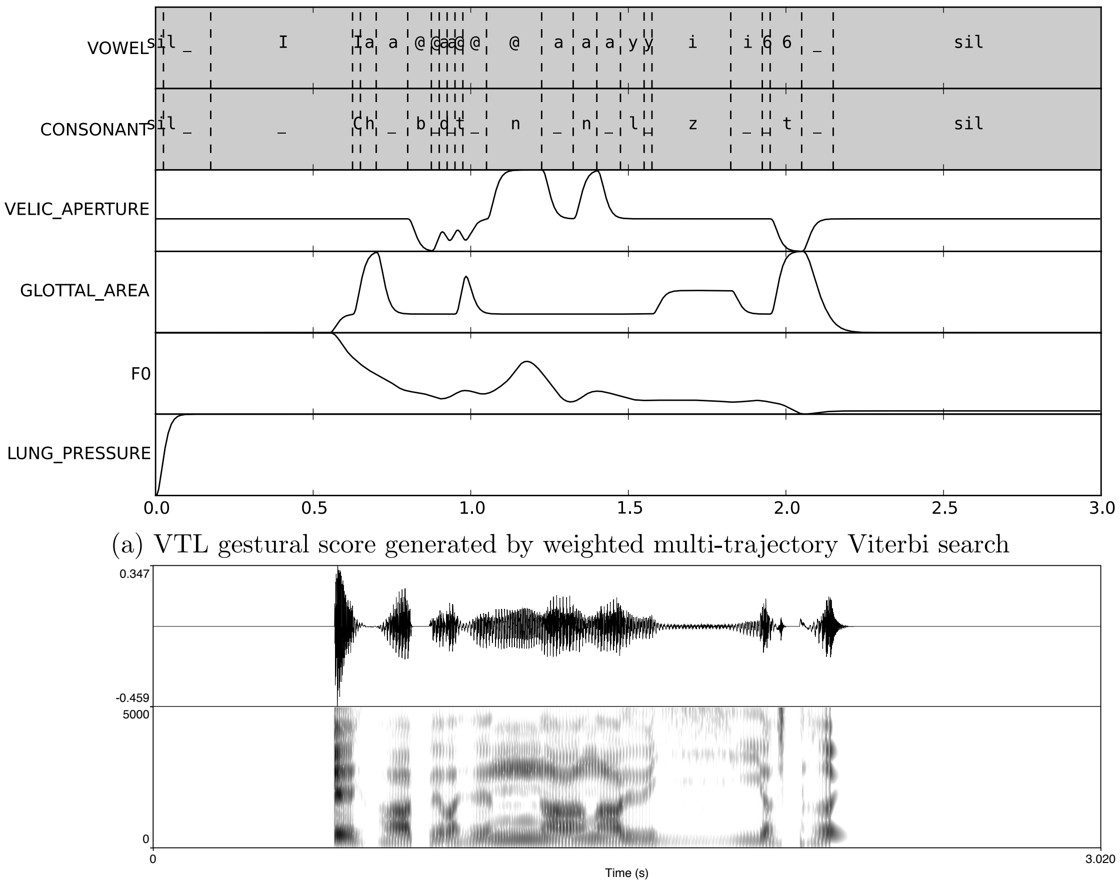
The F_0 contour from the original utterance is recreated as a sequence of appropriate gestures on the `F0` tier as described in Section 4.1.5. Since the voicing defined by gestures on the `GLOTTAL_AREA` tier is initially derived from articulatory movements preparatory to, and usually earlier than, the onset of voicing in the original recording, the corresponding glottal gestures are subsequently corrected according to the timing derived from the F_0 contour.

A final modification must be made to the gestures on the `CONSONANT` and `VOWEL` tiers; the close fit of the vertex trajectories to the EMA data does not automatically guarantee that the articulatory precision of the original speaker will be matched by VTL. In fact, while the timing of the CV gestures was generated using various `effort` parameter values to produce such a close fit, informal auditory evaluation suggests that lower `effort` during short gestures results in a significant target undershoot that reduces the intelligibility of the synthesis output considerably.

Figure 5.8a shows the gestural score generated by the resynthesis and enriched in the manner just described. The resulting waveform generated by acoustic synthesis with VTL is displayed in Figure 5.8b, which can be compared to the original recording shown in Figure 5.1.

5.3.3 Discussion

Many variables affect the resynthesis process described in this chapter in a complex interplay whose nature is not made entirely clear. Consequently, the results of the resynthesis are mixed; some utterances are resynthesized quite successfully, for others, the approach



(a) VTL gestural score generated by weighted multi-trajectory Viterbi search

(b) Waveform and spectrogram for the synthesis output of the gestural score in Figure 5.8a

Figure 5.8: Gestural score and acoustic output of resynthesis

does not perform well at all. This section recapitulates a selection of factors that exert influence over the resynthesis process at all levels.

Vocal tract model

VTL's vocal tract model is quite refined but nevertheless restricted to a certain degree by its parametric nature. This means that certain movements produced naturally but unanticipated by the model's design[13] cannot be replicated in quite the same way, or at all, if they are not allowed for by control parameters or the deformability of surfaces.

The adaptation to a single speaker has the advantage of modeling that speaker's vocal tract very accurately, but not exactly.[14] The nature and limitations of the MRI

[13] For example, the model's tongue blade contour, which is defined by a Bézier curve, seems to have trouble matching some observed retroflex tongue shapes.

[14] On the other hand, adaptation to another speaker (using appropriate data) may reveal whether the

data used to adapt the vocal tract model to the target speaker entail that not all speech sounds are available to VTL, and some, such as vibrants, cannot be produced at all.

Since the surfaces of the vocal tract model are stretched and deformed in a purely linear manner, and no conservation of mass is modeled, the trajectories obtained from tracking vertices on these surfaces do not necessarily match those of fleshpoints tracked in a real vocal tract using EMA.

Another potential problem is posed by the posture effect, which is rather noticeable in the target configuration of several `phone`s. This is not surprising, since they are derived from MRI scans of the speaker in a supine position, while the EMA data was recorded with the speaker sitting upright. Consequently, the adequacy of the EMA interface as configured is weakened somewhat, especially where gravity had affected the position of the supine speaker's articulators, as was noted during non-speech intervals.

The gestural model and its adequacy in describing the trajectories of real articulators is another variable that might be replaced with one more complex in nature, with a direct impact on the resynthesis process as detailed here. In addition, the `effort` parameter, when varied, does not seem to contribute to the resynthesis quite in the manner that might be naïvely expected.

EMA data

The recording of articulatory data using EMA is a fairly mature and robust procedure when carried out by experienced technicians, and post-processing, as well as smoothing, contribute to the reliability and well-formedness of the resulting data. Nevertheless, it is not impossible for things to go wrong when transducer coils are faulty, become detached, or simply interfere with a speaker's normal articulation.

In fact, even if the speaker's articulation were completely unaffected by the EMA recording, it is only natural that he will economize in articulation, and phenomena of reduction and coarticulation are certainly more complex in nature than their treatment in this context can accurately reflect.

The layout of transducer coils is another issue, and while certain arrangements are more likely to capture relevant articulatory movements, there are limitations regarding the proximity of adjacent coils, as well as the placement of the coils themselves (e.g. the `ttip` coil cannot be placed directly on the tip of the tongue). In addition, the total number of EMA channels restricts the number of coils, and hence fleshpoints, whose motion is tracked, and not all articulatory movement can be losslessly encoded as a set of a dozen or so trajectories.

Furthermore, the axes of the Cartesian coordinate system used to represent the EMA data is not necessarily aligned with the primary dimensions of articulatory movements, which are also different for each articulator. It might be possible to achieve greater descriptive and predictive power by transforming both EMA and vertex trajectories using PCA or a similar process.

vocal tract parameters are speaker-independent or rather better suited to some speakers than others.

All of this means that the selection of vertices in the vocal tract model must be optimized to match the EMA coil layout, but doing so does not guarantee the same trajectories, and neither will fully capture all aspects of real articulation.

Viterbi search and cost function

Potential deficiencies and oversimplification in the EMA interface are compounded in the automatic resynthesis approach by issues in the implementation of the gestural score generation. In particular, the Viterbi search is certainly not free of programming errors, or maximally efficient, and many alternative cost functions are conceivable, potentially better suited to generating the gestural scores than the cost function used in this chapter.

Even with the Viterbi search and cost function assembled to perform as described, several variables and scaling factors contribute to the preference for some paths through the search space over others. The systematic variation of such coefficients would be necessary to effectively "debug", and improve the performance of, the resynthesis approach.

Finally, the weight matrix integrated into the resynthesis exerts critical control over the outcome. Determining optimal weight values reveals itself to be an arduous task, and the structure of the entire process is extremely sensitive to the weight values in the matrix.

Conclusion

In spite of the many open questions that remain with respect to the articulatory resynthesis approach presented, the results are not entirely discouraging. The approach has shown that it is indeed possible to generate a gestural score from articulatory data alone, suitable for use with the VTL synthesizer, and while this approach may not be the best one, or the most robust, it appears to reassert the potential benefits of incorporating articulatory data into the dynamic control of an articulatory synthesizer.

Appendix 5.A Prompt lists for German corpus

Note: RUHE denotes an empty recording sweep; prompt ID codes ending in R indicate a recording that was repeated.

5.A.1 Session 1, normal utterances

```
001 RUHE1 RUHE
002 ISXXI1 Ich habe das volle Silo geleert.
003 FTXXU1 Ich habe das Blut entfernt.
004 MSXXI1 Ich habe Lise gesehen.
005 FSXXI1 Ich habe das Verlies erspäht.
006 ITXXI1 Ich habe ihre Titel geändert.
007 MDXXI1 Ich habe Lieder gesungen.
008 ITXXA1 Ich habe eine Tat vollbracht.
009 IDMTA1 Ich habe Daten analysiert.
010 ISXXA1 Ich habe Salz verstreut.
011 ISMSU1 Ich habe Suse getroffen.
012 MTXXI1 Ich habe den Liter getrunken.
013 MTXXA1 Ich habe den Atem angehalten.
014 FTXXA1 Ich habe Rat erbeten.
015 FTXXI1 Ich habe den Kredit erhalten.
016 FDXXI1 Ich habe das Lied erkannt.
017 IDMTI1 Ich habe Dieter gesehen.
018 FDXXA1 Ich habe mein Rad erkannt.
019 MDXXU1 Ich habe Sud entfernt.
020 FSXXA1 Ich habe das Glas entleert.
021 IDMDU1 Ich habe mehrere Duden gekauft.
022 MDXXA1 Ich habe Adam gesehen.
023 FSXXU1 Ich habe einen Fuß erspäht.
024 ITMTU1 Ich habe eine Tute gekauft.
025 MSXXA1 Ich habe eine Blase entfernt.
026 RUHE2 RUHE
027 IDMTI2 Ich habe Dieter gesehen.
028 ISXXA2 Ich habe Salz verstreut.
029 MSXXI2 Ich habe Lise gesehen.
030 FSXXI2 Ich habe das Verlies erspäht.
031 MDXXI2 Ich habe Lieder gesungen.
032 IDMTA2 Ich habe Daten analysiert.
033 ITXXI2 Ich habe ihre Titel geändert.
034 FTXXI2 Ich habe den Kredit erhalten.
035 FDXXI2 Ich habe das Lied erkannt.
036 FDXXA2 Ich habe mein Rad erkannt.
037 ITMTU2 Ich habe eine Tute gekauft.
038 FTXXA2 Ich habe Rat erbeten.
039 MDXXA2 Ich habe Adam gesehen.
040 MTXXA2 Ich habe den Atem angehalten.
041 IDMDU2 Ich habe mehrere Duden gekauft.
042 ITXXA2 Ich habe eine Tat vollbracht.
043 ISXXI2 Ich habe das volle Silo geleert.
044 MTXXI2 Ich habe den Liter getrunken.
045 MSXXA2 Ich habe eine Blase entfernt.
046 ISMSU2 Ich habe Suse getroffen.
047 FSXXA2 Ich habe das Glas entleert.
048 FSXXU2 Ich habe einen Fuß erspäht.
049 FTXXU2 Ich habe das Blut entfernt.
050 MDXXU2 Ich habe Sud entfernt.
051 RUHE3 RUHE
052 FDXXI3 Ich habe das Lied erkannt.
053 ITXXI3 Ich habe ihre Titel geändert.
054 MSXXI3 Ich habe Lise gesehen.
055 MDXXI3 Ich habe Lieder gesungen.
056 ITMTU3 Ich habe eine Tute gekauft.
057 FTXXA3 Ich habe Rat erbeten.
058 FTXXI3 Ich habe den Kredit erhalten.
059 FSXXA3 Ich habe das Glas entleert.
060 IDMTI3 Ich habe Dieter gesehen.
061 IDMDU3 Ich habe mehrere Duden gekauft.
062 ISXXA3 Ich habe Salz verstreut.
063 ISXXI3 Ich habe das volle Silo geleert.
064 FSXXI3 Ich habe das Verlies erspäht.
065 MDXXA3 Ich habe Adam gesehen.
066 ISMSU3 Ich habe Suse getroffen.
067 ITXXA3 Ich habe eine Tat vollbracht.
068 IDMTA3 Ich habe Daten analysiert.
069 FDXXA3 Ich habe mein Rad erkannt.
070 MTXXI3 Ich habe den Liter getrunken.
071 FTXXU3 Ich habe das Blut entfernt.
072 MTXXA3 Ich habe den Atem angehalten.
073 FSXXU3 Ich habe einen Fuß erspäht.
074 MDXXU3 Ich habe Sud entfernt.
075 MSXXA3 Ich habe eine Blase entfernt.
076 RUHE4 RUHE
077 FDXXI4 Ich habe das Lied erkannt.
078 MTXXI4 Ich habe den Liter getrunken.
079 FTXXA4 Ich habe Rat erbeten.
080 ISMSU4 Ich habe Suse getroffen.
081 ISXXIR Ich habe das volle Silo geleert.
082 ISXXI4 Ich habe das volle Silo geleert.
083 MTXXA4 Ich habe den Atem angehalten.
084 IDMDU4 Ich habe mehrere Duden gekauft.
085 ITMTU4 Ich habe eine Tute gekauft.
086 FSXXA4 Ich habe das Glas entleert.
087 IDMTA4 Ich habe Daten analysiert.
088 FSXXI4 Ich habe das Verlies erspäht.
089 MSXXIR Ich habe Lise gesehen.
090 MSXXI4 Ich habe Lise gesehen.
```

091 FTXXI4 Ich habe den Kredit erhalten.
092 FSXXU4 Ich habe einen Fuß erspäht.
093 IDMTI4 Ich habe Dieter gesehen.
094 MDXXA4 Ich habe Adam gesehen.
095 FDXXA4 Ich habe mein Rad erkannt.
096 MDXXU4 Ich habe Sud entfernt.
097 ITXXI4 Ich habe ihre Titel geändert.
098 ISXXA4 Ich habe Salz verstreut.
099 FTXXU4 Ich habe das Blut entfernt.
100 MSXXA4 Ich habe eine Blase entfernt.
101 ITXXA4 Ich habe eine Tat vollbracht.
102 MDXXI4 Ich habe Lieder gesungen.
103 RUHE5 RUHE
104 FTXXU5 Ich habe das Blut entfernt.
105 FSXXI5 Ich habe das Verlies erspäht.
106 ISXXA5 Ich habe Salz verstreut.
107 FSXXU5 Ich habe einen Fuß erspäht.
108 ITMTU5 Ich habe eine Tute gekauft.
109 IDMTA5 Ich habe Daten analysiert.
110 FSXXA5 Ich habe das Glas entleert.
111 FTXXI5 Ich habe den Kredit erhalten.
112 FDXXI5 Ich habe das Lied erkannt.
113 ISMSU5 Ich habe Suse getroffen.
114 FTXXA5 Ich habe Rat erbeten.
115 MTXXA5 Ich habe den Atem angehalten.
116 ITXXA5 Ich habe eine Tat vollbracht.
117 MDXXAR Ich habe Adam gesehen.
118 MDXXA5 Ich habe Adam gesehen.
119 ISXXI5 Ich habe das volle Silo geleert.
120 MDXXI5 Ich habe Lieder gesungen.
121 MTXXI5 Ich habe den Liter getrunken.
122 MDXXU5 Ich habe Sud entfernt.
123 IDMDU5 Ich habe mehrere Duden gekauft.
124 FDXXA5 Ich habe mein Rad erkannt.
125 MSXXI5 Ich habe Lise gesehen.
126 MSXXA5 Ich habe eine Blase entfernt.
127 IDMTI5 Ich habe Dieter gesehen.
128 ITXXI5 Ich habe ihre Titel geändert.
129 RUHE6 RUHE
130 FDXXA6 Ich habe mein Rad erkannt.
131 FTXXU6 Ich habe das Blut entfernt.
132 MDXXI6 Ich habe Lieder gesungen.
133 FSXXU6 Ich habe einen Fuß erspäht.
134 FTXXA6 Ich habe Rat erbeten.
135 IDMTA6 Ich habe Daten analysiert.
136 ITXXI6 Ich habe ihre Titel geändert.
137 MDXXA6 Ich habe Adam gesehen.
138 FSXXI6 Ich habe das Verlies erspäht.
139 MSXXI6 Ich habe Lise gesehen.
140 ISXXI6 Ich habe das volle Silo geleert.
141 IDMDU6 Ich habe mehrere Duden gekauft.
142 IDMTI6 Ich habe Dieter gesehen.
143 FDXXI6 Ich habe das Lied erkannt.
144 ISXXA6 Ich habe Salz verstreut.
145 MTXXI6 Ich habe den Liter getrunken.
146 MSXXA6 Ich habe eine Blase entfernt.
147 FSXXA6 Ich habe das Glas entleert.
148 ITMTU6 Ich habe eine Tute gekauft.
149 ISMSU6 Ich habe Suse getroffen.
150 FTXXI6 Ich habe den Kredit erhalten.
151 MTXXA6 Ich habe den Atem angehalten.
152 ITXXA6 Ich habe eine Tat vollbracht.
153 MDXXU6 Ich habe Sud entfernt.
154 RUHE7 RUHE
155 FDXXA7 Ich habe mein Rad erkannt.
156 MTXXIR Ich habe den Liter getrunken.
157 MTXXI7 Ich habe den Liter getrunken.
158 IDMTI7 Ich habe Dieter gesehen.
159 IDMDU7 Ich habe mehrere Duden gekauft.
160 MDXXI7 Ich habe Lieder gesungen.
161 FTXXU7 Ich habe das Blut entfernt.
162 MSXXI7 Ich habe Lise gesehen.
163 MDXXA7 Ich habe Adam gesehen.
164 FTXXA7 Ich habe Rat erbeten.
165 ITMTU7 Ich habe eine Tute gekauft.
166 FSXXU7 Ich habe einen Fuß erspäht.
167 MDXXU7 Ich habe Sud entfernt.
168 FSXXA7 Ich habe das Glas entleert.
169 FDXXI7 Ich habe das Lied erkannt.
170 ITXXI7 Ich habe ihre Titel geändert.
171 MSXXA7 Ich habe eine Blase entfernt.
172 FSXXI7 Ich habe das Verlies erspäht.
173 MTXXA7 Ich habe den Atem angehalten.
174 ISXXI7 Ich habe das volle Silo geleert.
175 IDMTA7 Ich habe Daten analysiert.
176 ISXXA7 Ich habe Salz verstreut.
177 ITXXA7 Ich habe eine Tat vollbracht.
178 ISMSU7 Ich habe Suse getroffen.
179 FTXXI7 Ich habe den Kredit erhalten.
180 FLAA2 [palate trace]
181 RUHE8 RUHE

5.A.2 Session 2, normal utterances

001 RUHE1 RUHE
002 RUHE1 RUHE
003 ISXXI1 Ich habe das volle Silo geleert.
004 FTXXU1 Ich habe das Blut entfernt.
005 MSXXI1 Ich habe Lise gesehen.
006 FSXXI1 Ich habe das Verlies erspäht.

```
007 ITXXI1 Ich habe ihre Titel geändert.
008 MDXXI1 Ich habe Lieder gesungen.
009 ITXXA1 Ich habe eine Tat vollbracht.
010 IDMTA1 Ich habe Daten analysiert.
011 ISXXA1 Ich habe Salz verstreut.
012 ISMSU1 Ich habe Suse getroffen.
013 MTXXI1 Ich habe den Liter getrunken.
014 MTXXA1 Ich habe den Atem angehalten.
015 FTXXA1 Ich habe Rat erbeten.
016 FTXXI1 Ich habe den Kredit erhalten.
017 FDXXI1 Ich habe das Lied erkannt.
018 IDMTI1 Ich habe Dieter gesehen.
019 FDXXA1 Ich habe mein Rad erkannt.
020 MDXXU1 Ich habe Sud entfernt.
021 FSXXA1 Ich habe das Glas entleert.
022 IDMDU1 Ich habe mehrere Duden gekauft.
023 MDXXA1 Ich habe Adam gesehen.
024 FSXXU1 Ich habe einen Fuß erspäht.
025 ITMTU1 Ich habe eine Tute gekauft.
026 MSXXA1 Ich habe eine Blase entfernt.
027 RUHE2 RUHE
028 IDMTI2 Ich habe Dieter gesehen.
029 ISXXA2 Ich habe Salz verstreut.
030 MSXXI2 Ich habe Lise gesehen.
031 FSXXI2 Ich habe das Verlies erspäht.
032 MDXXI2 Ich habe Lieder gesungen.
033 IDMTA2 Ich habe Daten analysiert.
034 ITXXI2 Ich habe ihre Titel geändert.
035 FTXXI2 Ich habe den Kredit erhalten.
036 FDXXI2 Ich habe das Lied erkannt.
037 FDXXA2 Ich habe mein Rad erkannt.
038 ITMTU2 Ich habe eine Tute gekauft.
039 FTXXA2 Ich habe Rat erbeten.
040 MDXXA2 Ich habe Adam gesehen.
041 MTXXA2 Ich habe den Atem angehalten.
042 IDMDU2 Ich habe mehrere Duden gekauft.
043 ITXXA2 Ich habe eine Tat vollbracht.
044 ISXXI2 Ich habe das volle Silo geleert.
045 MTXXI2 Ich habe den Liter getrunken.
046 MSXXA2 Ich habe eine Blase entfernt.
047 ISMSU2 Ich habe Suse getroffen.
048 FSXXA2 Ich habe das Glas entleert.
049 FSXXU2 Ich habe einen Fuß erspäht.
050 FTXXU2 Ich habe das Blut entfernt.
051 MDXXU2 Ich habe Sud entfernt.
052 RUHE3 RUHE
053 FDXXI3 Ich habe das Lied erkannt.
054 ITXXI3 Ich habe ihre Titel geändert.
055 MSXXI3 Ich habe Lise gesehen.
056 MDXXI3 Ich habe Lieder gesungen.
057 ITMTU3 Ich habe eine Tute gekauft.
058 FTXXA3 Ich habe Rat erbeten.
059 FTXXI3 Ich habe den Kredit erhalten.
060 FSXXA3 Ich habe das Glas entleert.
061 IDMTI3 Ich habe Dieter gesehen.
062 IDMDU3 Ich habe mehrere Duden gekauft.
063 ISXXA3 Ich habe Salz verstreut.
064 ISXXI3 Ich habe das volle Silo geleert.
065 FSXXI3 Ich habe das Verlies erspäht.
066 MDXXA3 Ich habe Adam gesehen.
067 ISMSU3 Ich habe Suse getroffen.
068 ITXXA3 Ich habe eine Tat vollbracht.
069 IDMTA3 Ich habe Daten analysiert.
070 FDXXA3 Ich habe mein Rad erkannt.
071 MTXXI3 Ich habe den Liter getrunken.
072 FTXXU3 Ich habe das Blut entfernt.
073 MTXXA3 Ich habe den Atem angehalten.
074 FSXXU3 Ich habe einen Fuß erspäht.
075 MDXXU3 Ich habe Sud entfernt.
076 MSXXA3 Ich habe eine Blase entfernt.
077 RUHE4 RUHE
078 FDXXI4 Ich habe das Lied erkannt.
079 MTXXI4 Ich habe den Liter getrunken.
080 FTXXA4 Ich habe Rat erbeten.
081 ISMSU4 Ich habe Suse getroffen.
082 ISXXI4 Ich habe das volle Silo geleert.
083 MTXXA4 Ich habe den Atem angehalten.
084 IDMDU4 Ich habe mehrere Duden gekauft.
085 ITMTU4 Ich habe eine Tute gekauft.
086 FSXXA4 Ich habe das Glas entleert.
087 IDMTA4 Ich habe Daten analysiert.
088 FSXXI4 Ich habe das Verlies erspäht.
089 MSXXI4 Ich habe Lise gesehen.
090 FTXXI4 Ich habe den Kredit erhalten.
091 FSXXU4 Ich habe einen Fuß erspäht.
092 IDMTI4 Ich habe Dieter gesehen.
093 MDXXA4 Ich habe Adam gesehen.
094 FDXXA4 Ich habe mein Rad erkannt.
095 MDXXU4 Ich habe Sud entfernt.
096 ITXXI4 Ich habe ihre Titel geändert.
097 ISXXA4 Ich habe Salz verstreut.
098 FTXXU4 Ich habe das Blut entfernt.
099 MSXXA4 Ich habe eine Blase entfernt.
100 ITXXA4 Ich habe eine Tat vollbracht.
101 MDXXI4 Ich habe Lieder gesungen.
102 RUHE5 RUHE
103 FTXXU5 Ich habe das Blut entfernt.
104 FSXXI5 Ich habe das Verlies erspäht.
105 ISXXA5 Ich habe Salz verstreut.
106 FSXXU5 Ich habe einen Fuß erspäht.
107 ITMTU5 Ich habe eine Tute gekauft.
108 IDMTA5 Ich habe Daten analysiert.
109 FSXXA5 Ich habe das Glas entleert.
110 FTXXI5 Ich habe den Kredit erhalten.
111 FDXXI5 Ich habe das Lied erkannt.
112 ISMSU5 Ich habe Suse getroffen.
```

113 FTXXA5 Ich habe Rat erbeten.
114 MTXXA5 Ich habe den Atem angehalten.
115 ITXXA5 Ich habe eine Tat vollbracht.
116 MDXXA5 Ich habe Adam gesehen.
117 ISXXI5 Ich habe das volle Silo geleert.
118 MDXXI5 Ich habe Lieder gesungen.
119 MTXXI5 Ich habe den Liter getrunken.
120 MDXXU5 Ich habe Sud entfernt.
121 IDMDU5 Ich habe mehrere Duden gekauft.
122 FDXXA5 Ich habe mein Rad erkannt.
123 MSXXI5 Ich habe Lise gesehen.
124 MSXXA5 Ich habe eine Blase entfernt.
125 IDMTI5 Ich habe Dieter gesehen.
126 ITXXI5 Ich habe ihre Titel geändert.
127 RUHE6 RUHE
128 FDXXA6 Ich habe mein Rad erkannt.
129 FTXXU6 Ich habe das Blut entfernt.
130 MDXXI6 Ich habe Lieder gesungen.
131 FSXXU6 Ich habe einen Fuß erspäht.
132 FTXXA6 Ich habe Rat erbeten.
133 IDMTA6 Ich habe Daten analysiert.
134 ITXXI6 Ich habe ihre Titel geändert.
135 MDXXA6 Ich habe Adam gesehen.
136 FSXXI6 Ich habe das Verlies erspäht.
137 MSXXI6 Ich habe Lise gesehen.
138 ISXXI6 Ich habe das volle Silo geleert.
139 IDMDU6 Ich habe mehrere Duden gekauft.
140 IDMTI6 Ich habe Dieter gesehen.
141 FDXXI6 Ich habe das Lied erkannt.
142 ISXXA6 Ich habe Salz verstreut.
143 MTXXI6 Ich habe den Liter getrunken.
144 MSXXA6 Ich habe eine Blase entfernt.
145 FSXXA6 Ich habe das Glas entleert.
146 ITMTU6 Ich habe eine Tute gekauft.
147 ISMSU6 Ich habe Suse getroffen.
148 FTXXI6 Ich habe den Kredit erhalten.
149 MTXXA6 Ich habe den Atem angehalten.
150 ITXXA6 Ich habe eine Tat vollbracht.
151 MDXXU6 Ich habe Sud entfernt.
152 RUHE7 RUHE
153 FDXXA7 Ich habe mein Rad erkannt.
154 MTXXI7 Ich habe den Liter getrunken.
155 IDMTI7 Ich habe Dieter gesehen.
156 IDMDU7 Ich habe mehrere Duden gekauft.
157 MDXXI7 Ich habe Lieder gesungen.
158 FTXXU7 Ich habe das Blut entfernt.
159 MSXXI7 Ich habe Lise gesehen.
160 MDXXA7 Ich habe Adam gesehen.
161 FTXXA7 Ich habe Rat erbeten.
162 ITMTU7 Ich habe eine Tute gekauft.
163 FSXXU7 Ich habe einen Fuß erspäht.
164 MDXXU7 Ich habe Sud entfernt.
165 FSXXA7 Ich habe das Glas entleert.
166 FDXXI7 Ich habe das Lied erkannt.
167 ITXXI7 Ich habe ihre Titel geändert.
168 MSXXA7 Ich habe eine Blase entfernt.
169 FSXXI7 Ich habe das Verlies erspäht.
170 MTXXA7 Ich habe den Atem angehalten.
171 ISXXI7 Ich habe das volle Silo geleert.
172 IDMTA7 Ich habe Daten analysiert.
173 ISXXA7 Ich habe Salz verstreut.
174 ITXXA7 Ich habe eine Tat vollbracht.
175 ISMSU7 Ich habe Suse getroffen.
176 FTXXI7 Ich habe den Kredit erhalten.
177 FLAA2 [palate trace]
178 RUHE8 RUHE

Chapter 6

Discussion and Outlook

This chapter presents a brief discussion of the articulatory resynthesis approach, as well as a proposed application. In addition, preliminary results toward the adaptation of VTL to English are outlined.

6.1 Articulatory resynthesis

The previous chapters have demonstrated the possibility of automatically synthesizing utterances using dynamic speech production data to control the articulatory synthesizer VTL. The synthesizer is not controlled directly; instead, gestural control structures are generated in an analysis-by-synthesis approach by means of an interface which allows the comparison in the articulatory domain of EMA data and the trajectories of corresponding vertices in the wireframe mesh of the synthesizer's vocal tract model.

The EMA interface was implemented using a Viterbi search with an error metric based on a multi-trajectory cost function weighted by articulatory relevance. In this arrangement, the definition of the weighting was shown to have a very strong impact on resynthesis performance. Further work investigating the nature of the interplay between human speaker and EMA data, vocal tract model and vertex selection, cost function and weighting, could systematically improve the reliability and performance of the resynthesis technique.

Of course, the challenge of the premise of articulatory resynthesis could be solved using altogether different means. An early approach using DTW was outlined, and more sophisticated techniques are conceivable using HMMs or Kalman filters.

Whichever strategy is pursued, the immediate result of the articulatory resynthesis is the gestural score that best satisfies all imposed constraints. However, this score is merely a schematic representation of the underlying parameters controlling the synthesizer's vocal tract model. From this representation, the control *parameter trajectories* can be obtained, and these define the dynamic shape of the vocal tract model as it changes over the course of the utterance.

6.1.1 Parameter-based articulatory TTS

If every utterance in a corpus of appropriate articulatory data is resynthesized in this manner, then each utterance is transformed into a control parametric representation. This resulting corpus of parameter trajectory data could then be used to train statistical models, based on the underlying utterances that produced the original articulatory data. These models might then be used to generate parameter trajectories directly for unseen utterances, for which no articulatory data is required. A similar approach was taken by Zhang & Renals (2008); Zhang (2009), who use trajectory HMMs (Zen et al., 2007b) to predict EMA trajectories from a textual representation.

This concept is essentially a slightly modified scenario for statistical parametric synthesis (cf. Section 1.1.3). In this case, instead of training models like HMMs on the parameters of acoustic recordings (such as MFCCs and their dynamic features), the parameter trajectories obtained via articulatory resynthesis are used as training data. And instead of predicting acoustic parameter trajectories for new utterances during synthesis and using them in a vocoder (or similar) component for waveform generation, the predicted trajectories are control parameter trajectories, which are then "injected" into the vocal tract model of VTL, bypassing the control model, and (assuming that the parameter synthesis has performed as expected) allowing the acoustic model of VTL to generate a waveform.

This synthesis pipeline could be implemented, with no significant difficulty expected, using an existing modular TTS platform, such as Festival[1] (Black et al., 1999; Clark et al., 2007) or MARY[2] (Schröder & Trouvain, 2003). Both of these platforms provide the necessary framework and means to enrich textual input with the required symbolic segmental and suprasegmental information, and both platforms can be used with the statistical parametric synthesis engine HTS[3] (Tokuda et al., 2002; Zen et al., 2007a), which is based on HTK[4] (Young et al., 2009). This design represents a prototype TTS system with an articulatory synthesis back-end.

6.2 Adapting VTL to English

As an articulatory synthesizer, VTL is theoretically fully capable of multilingual synthesis. However, while the phonetic inventory of German is available from adaptation to articulatory data recorded from a German speaker (cf. Section 1.2.4), to synthesize speech in any other language, say English, this phoneset must be extended or replaced.

A "quick and dirty" solution could consist of taking the existing German `phoneList` and modifying its elements to include the missing English phones (i.e. the phoneset difference, containing dental fricatives, etc.). This might be accomplished using acoustic adaptation (in the VTL GUI) with reference to subjective auditory perception or (more

[1] `http://www.cstr.ed.ac.uk/projects/festival/` or `http://festvox.org/`

[2] `http://mary.dfki.de/`

[3] `http://hts.sp.nitech.ac.jp/`

[4] `http://htk.eng.cam.ac.uk/`

elaborately) spectral analysis of appropriate recordings. However, such an approach is laborious and prone to error, and may fail to take into account more subtle phonetic details, resulting in a German-biased English `phoneList`. In the absence of speech production data, it also fails to take into account the dynamics of running speech, such as those simulated by articulatory resynthesis.

For a more robust and objective solution, the full adaptation process described in Section 1.2.4 could be applied using articulatory data from a speaker of English. This approach is preferred in potential future work, using MRI data and a 3D dental scan of an English speaker, as well as applying the articulatory resynthesis approach to a large corpus of 3D EMA data from the same speaker. Furthermore, this corpus is phonetically balanced, which may allow fine-tuning of the weight matrix using an automatic approach (cf. Section 5.2.1). The posture effect (cf. Section 5.1.4) may also be reduced by integrating the EMA data into the adaptation procedure, and if possible, might be investigated more thoroughly for the speaker by recording a small set of appropriate EMA data in upright and supine position.

6.3 EMA data

A corpus of English utterances was recorded with 3D EMA in 2007 at LMU Munich by Phil Hoole and Korin Richmond (CSTR, University of Edinburgh). The utterances were designed to provide a large amount of phonetically balanced speech.

The subject, a male native speaker of British English and professional actor, was recorded using a Carstens AG500 EMA system. The utterances were recorded in sweeps of variable length, at a sample rate of 200 Hz, while the audio was sampled at 16 kHz. Video recordings were made as well.

The utterances were recorded in two sessions with the same prompt list (in randomized order), featuring English utterances randomly selected from newspaper texts, controlling for phonetic balance. Session 1 consists of 1 263 utterances, while session 2 contains 800 utterances.

In session 1, EMA measurement coils were attached to the upper and lower lip (`upperlip`, `lowerlip`), lower incisors (`jaw`), tongue tip (`T1`), blade (`T2`), and dorsum (`T3`). Reference coils were placed on the nose and upper incisors, and laterally on both sides of the head. The coil layout in session 2 was identical, except for the tongue dorsum coil (`T3`), which was attached to the velum instead.

6.4 MRI data

In 2008, the same speaker was scanned using volumetric and real-time MRI. The scanning was carried out together with Korin Richmond and Ian Marshall and his staff at the SFC[5] Brain Imaging Research Centre (SBIRC)[6] located in the Division of Clinical

[5] Scottish Funding Council

[6] `http://www.sbirc.ed.ac.uk/`

hit	[ɪ]	*fin*	[f]
pet	[ɛ]	*thin*	[θ]
hat	[æ]	*sin*	[s]
hot	[ɒ]	*shin*	[ʃ]
hut	[ʌ]	*mock*	[m]
put	[ʊ]	*knock*	[n]
heat	[i]	*thing*	[ŋ]
hoot	[u]	*ring*	[r]
hurt	[ɜ]	*long*	[l]
hart	[a]	*loch*	[x]
ought	[ɔ]	*ball*	[ɫ]
about	[ə]	⟨p⟩	[p]
there	[ɛː]	⟨t⟩	[t]
		⟨k⟩	[k]

(a) Sustained production; vowels are shown in the left column, consonants in the right. The last three prompts represent the speaker holding the occlusion of the corresponding stop.

apa	*ipi*	*upu*	[p]
ata	*iti*	*utu*	[t]
aka	*iki*	*uku*	[k]
afa	*ifi*	*ufu*	[f]
atha	*ithi*	*uthu*	[θ]
asa	*isi*	*usu*	[s]
asha	*ishi*	*ushu*	[ʃ]
ara	*iri*	*uru*	[ɾ]
ala	*ili*	*ulu*	[l]
ama	*imi*	*umu*	[m]
ana	*ini*	*unu*	[n]
anga	*ingi*	*ungu*	[ŋ]
acha	*ichi*	*uchu*	[x]
ara	*iri*	*uru*	[ɹ][a]
awa	*iwi*	*uwu*	[w]
aya	*iyi*	*uyu*	[j]

(b) Dynamic production. Each prompt represents an individual scan, yielding the vocal tract configuration for the target phone in the vocalic context of [ɑ], [i], and [u].

[a] To elicit the approximant [ɹ], the speaker was instructed to produce these prompts using an "American pronunciation".

Table 6.1: Prompt lists for the MRI scanning session. The prompts are emphasized, with the underlined letter(s) corresponding to the target phone, along with the target phone itself in IPA notation.

Neurosciences (DCN) of Western General Hospital, University of Edinburgh.

To replicate the adaptation procedure described in Birkholz & Kröger (2006) and Section 1.2.4, a prompt list was designed consisting of sustained vowels and consonants (Table 6.1a) and dynamic VCV transitions, elicited by repetitive production of CV syllables (Table 6.1b).

6.4.1 Acoustic reference recordings

Since no equipment required for simultaneous acoustic recording (such as a FO microphone array, cf. Section 2.1.2) was available at the MRI facility, a separate acoustic recording session was run on the evening before the MRI scanning. This not only produced high-quality audio recordings of the speaker uttering the prompts, but gave the speaker opportunity to familiarize himself with the prompt list and the general procedure in an informal, non-clinical environment with no time pressure.

The acoustic recordings took place in the studio suite of the Informatics Forum at the University of Edinburgh. The speaker was located in a sound-proofed room, and the prompts were recorded using a DPA[7] Type 4035 microphone mounted on a headset. Since construction of the building had only recently been completed at the time and recording equipment had not yet been installed in the studio, a digital recording setup was improvised by recording directly onto hard disk using an EDIROL[8] UA-25 audio interface connected to a laptop computer. The recordings were made with a 96 kHz sampling rate at 24 bit quantization. Unfortunately, the particular audio interface hardware used was subsequently discovered to be faulty in that it introduced artifacts in the form of spectral spikes at 13 kHz and its harmonics. However, these artifacts lie outside the frequency range most relevant to phonetic analysis and can therefore be disregarded.

The speaker read out the prompt list twice, once standing upright, and again in supine position (cf. Section 5.1.4). This was done to allow comparison in the acoustic domain between these two postures. The supine recordings were made to ensure that the audio matched the articulatory configuration shown in the MRI scans (where the speaker would be lying in the same position within the scanner).

Trained as a professional voice actor, the speaker was able to produce sustained vowels for 20 s on average, and the repetitive utterances for around 10 s to 15 s. This was valuable for the MRI scanning, since each scan would take approximately that long, and the speaker would be capable of producing each prompt over the entire duration of the scan.

6.4.2 MRI scanning

The Scanner used was a GE Medical Systems Signa HDx 1.5T. The speaker was placed in the scanner in supine position and fitted with a head and neck RF coil. The scans were acquired as listed in Table 6.1. All of the scans were completed in the allotted 120 min session as planned, with a short break between the sustained and dynamic scans.

The static scans of sustained vowels and consonants were acquired in 26 sagittal slices each with a resolution of 256 × 256 pixels and a thickness of 4 mm.

The dynamic scans were acquired as midsagittal slices with a resolution of 256 × 256 pixels, at a rate of approximately 4 fps. The speaker synchronized production of the target consonants to the scans by timing their stable phase with the noise emitted by the MRI scanner.[9] An example of the dynamic data (for nasals) is displayed in Figure 6.1.

The ROI is clearly visible, extending from the lips to the rear wall of the pharynx in anterior-posterior direction, from the larynx to the nasal cavity in inferior-superior direction, and (in the volumetric scans) laterally between the mandibular joints.

Unfortunately, an aliasing artifact[10] is visible in the resulting image data; a segment of the back of the subject's head and neck appears at the anterior edge of the images, overlapping with the subject's nose. Such artifacts can occur when the imaged anatomy

[7] `http://www.dpamicrophones.com/`

[8] `http://www.edirol.net/`

[9] Although no instructions had been given to the speaker with regard to timing syllable production, this unconscious behavior turned out to significantly increase the number of usable frames.

[10] also referred to as foldover, wraparound, phase wrapping, etc.

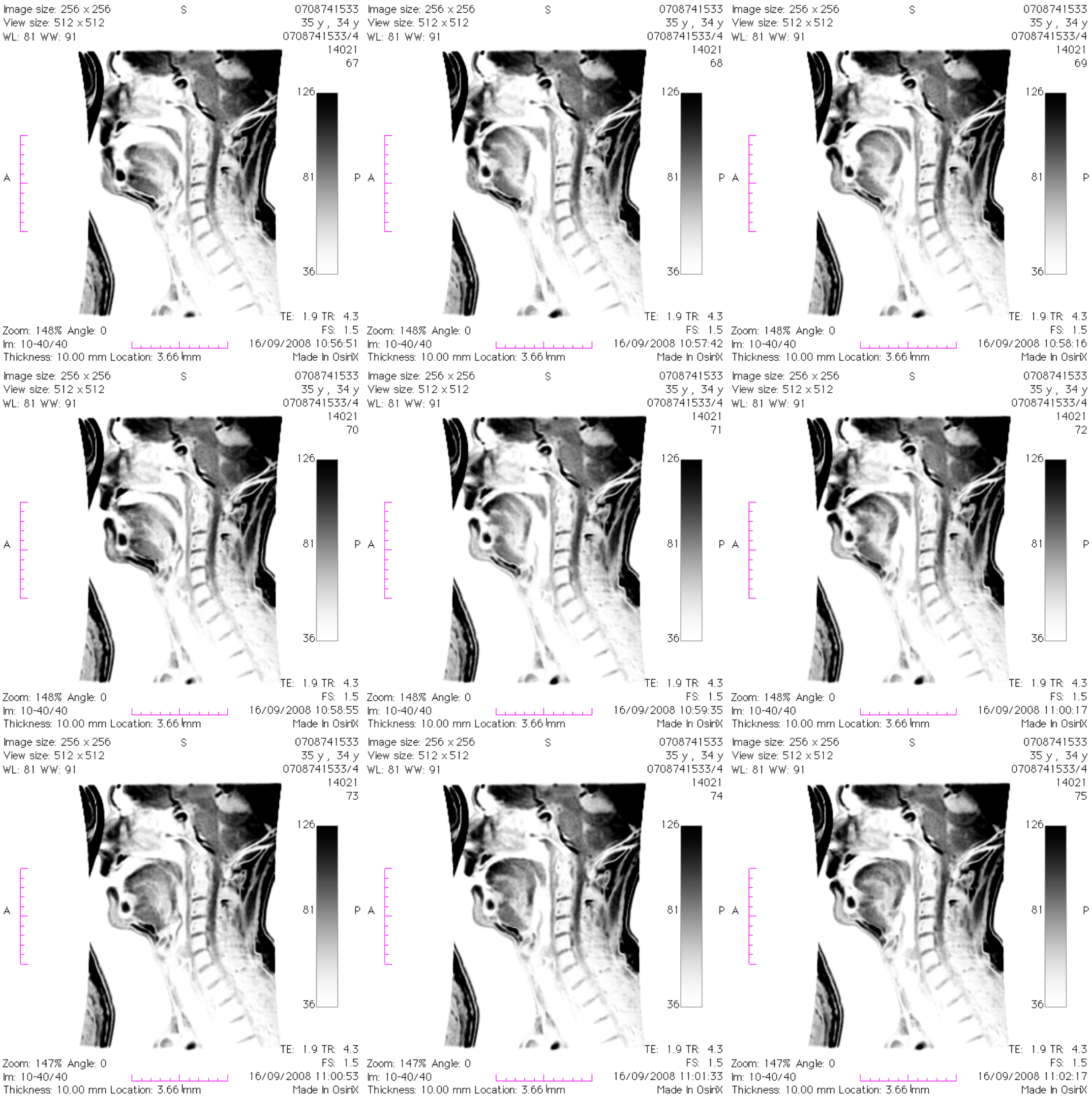


Figure 6.1: Maximum intensity projection of 30 midsagittal frames of [m] (top), [n] (middle), and [ŋ] (bottom) dynamically produced in vocalic context [ɑ] (left), [i] (center) and [u] (right). The visible aliasing artifact overlaps with the speaker's nose, but the image is not degraded in the region of interest.

extends outside the FOV. However, the aliasing does not impact subsequent analysis and can therefore be safely ignored, since it does not degrade the image quality in the ROI.

6.4.3 Dental reconstruction

For subsequent dental reconstruction, the speaker's teeth were acquired in a separate scan, using a contrast agent (cf. Section 2.1.2). Due to its high manganese content (which has desirable NMR characteristics), blueberry juice was chosen. Since no undiluted blueberry juice could be procured, fresh blueberries were pressed instead at a local juice bar, obtaining approximately one liter of pure blueberry juice.

For a dental scan at the end of the MRI session, the speaker lay prone in the scanner, and filled his mouth as fully as possible with the juice (which had fairly high viscosity) by sucking it from a bottle through a section of oxygen tubing. While this did produce images clearly showing the teeth, there was no time left in the session for a long acquisition, and the spatial resolution was therefore too low to enable reliable reconstruction of the dental surfaces.

For this reason a dental plaster cast was taken of the speaker's teeth in 2009. Unfortunately, the dental cast was improperly packaged and damaged during transport, resulting in shattered maxillary incisors.[11] The plaster cast will be scanned in the near future, using a laser triangulation scanner at the Institute of Perception, Action and Behaviour (IPAB) in Edinburgh. This is expected to produce a high-resolution 3D mesh of the speaker's dental surface, which can be registered with the volumetric MRI data to improve the vocal tract model adaptation in VTL.

6.5 Concluding remarks

The use of articulatory synthesis greatly benefits speech production research and related fields, using realistic integrated 3D models of the vocal tract and acoustic simulation. While imaging techniques such as MRI have advanced to allow direct observation of the vocal tract during speech production, and the safety of this modality places few constraints on the collection of such data, the full dynamics of articulatory movements in natural speech production are more successfully captured by other means, such as EMA.

This thesis has explored some of the potential uses of such speech production data in the dynamic control of a high-quality, 3D articulatory synthesizer, highlighting similarities between artificial and real vocal tracts.

Several promising approaches have been outlined, though not fully explored, yet it can be hoped that they somehow contribute to the development of dynamic control in articulatory synthesis.

[11] However, it might be possible to recover the missing crown surfaces using post-processing techniques such as that described by Buchaillard et al. (2007).

Bibliography

Adaškevičius, Rimas & Arunas Vasiliauskas (2008) "3D multicamera dental cast scanning system". *Electronics and Electrical Engineering*, **82**(2):49–52. URL `http://www.ktu.lt/lt/mokslas/zurnalai/elektros_z/z82/10_ISSN_1392-1215_3D%20Multicamera%20Dental%20Cast%20Scanning%20System.pdf`.

Allen, Jonathan, M. Sharon Hunnicutt & Dennis H. Klatt (1987) *Text-to-speech: The MITalk system*. Cambridge University Press.

Arai, Takayuki (2006) "Sliding three-tube model as a simple educational tool for vowel production". *Acoustical Science and Technology*, **27**(6):384–388. doi:10.1250/ast.27.384.

Arai, Takayuki (2009) "Sliding vocal-tract model and its application for vowel production". In: *Interspeech*, pp. 72–75. Brighton, England: ISCA.

Arnal, Alain, Pierre Badin, Gilbert Brock, Pierre-Yves Connan, Evelyne Florig, Noël Perez, Pascal Perrier, Pela Simon, Rudolph Sock, Laurent Varin, Béatrice Vaxelaire & Jean-Pierre Zerling (2000) "Une base de données cinéradiographiques du français". In: *XXIIIèmes Journées d'Etude sur la Parole*, pp. 425–428. Aussois, France.

Aron, Michael, Erwan Kerrien, Marie-Odile Berger & Yves Laprie (2006) "Coupling electromagnetic sensors and ultrasound images for tongue tracking: acquisition set up and preliminary results". In: *7th International Seminar on Speech Production*, pp. 435–450. Ubatuba, Brazil.

Aron, Michael, Anastasios Roussos, Marie-Odile Berger, Erwan Kerrien & Petros Maragos (2008) "Multimodality acquisition of articulatory data and processing". In: *16th European Signal Processing Conference*. Lausanne, Switzerland: EURASIP. URL `http://www.eurasip.org/Proceedings/Eusipco/Eusipco2008/papers/1569104643.pdf`.

Badin, Pierre, Gérard Bailly, M. Raybaudi & Christoph Segebarth (1998) "A three-dimensional linear articulatory model based on MRI data". In: *Third ESCA/COCOSDA Workshop on Speech Synthesis (SSW3)*, pp. 249–254. Jenolan Caves House, Blue Mountains, NSW, Australia: ESCA/COCOSDA. URL `http://www.isca-speech.org/archive/ssw3/ssw3_249.html`.

Badin, Pierre, Gérard Bailly, Lionel Revéret, Monica Baciu, Christoph Segebarth & Christophe Savariaux (2002) "Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images". *Journal of Phonetics*, **30**(3):533–553. doi:10.1006/jpho.2002.0166.

Baer, Thomas, John C. Gore, S. Boyce & Patrick W. Nye (1987) "Application of MRI to the analysis of speech production". *Magnetic Resonance Imaging*, **5**(1):1–7. doi:10.1016/0730-725X(87)90477-2.

Baer, Thomas, John C. Gore, L. C. Gracco & Patrick W. Nye (1991) "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels". *Journal of the Acoustical Society of America*, **90**(2):799–828. doi:10.1121/1.401949.

Barry, William J. (1992) "Comments on Chapter 2". In: Gerard J. Docherty & D. Robert Ladd (eds.), *Gesture, Segment, Prosody*, number 2 in Papers in Laboratory Phonology, pp. 65–67. Cambridge University Press.

Beautemps, Denis, Pierre Badin & Gérard Bailly (2001) "Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling". *Journal of the Acoustical Society of America*, **109**(5):2165–2180. doi:10.1121/1.1361090.

Beautemps, Denis, Pierre Badin & Rafael Laboissière (1995) "Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: A new model for vowels and fricative consonants based on experimental data". *Speech Communication*, **16**(1):27–47. doi:10.1016/0167-6393(94)00045-C.

Birkholz, Peter (2002) *Entwicklung eines dreidimensionalen Artikulatormodells für die Sprachsynthese*. Diploma thesis, University of Rostock, Rostock, Germany. URL `http://vcg.informatik.uni-rostock.de/assets/publications/theses_mas/DA_Birkholz2002.pdf`.

Birkholz, Peter (2006) *3D-Artikulatorische Sprachsynthese*. Berlin, Germany: Logos. ISBN 978-3-8325-1092-3. PhD thesis.

Birkholz, Peter (2007) "Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets". In: *Interspeech*, pp. 2865–2868. Antwerp, Belgium: ISCA. URL `http://www.isca-speech.org/archive/interspeech_2007/i07_2865.html`.

Birkholz, Peter & Dietmar Jackèl (2004) "Simulation of flow and acoustics in the vocal tract". In: *CFA/Deutschen Jahrestagung für Akustik*, pp. 895–896. Strasbourg, France.

Birkholz, Peter & Dietmar Jackèl (2006) "Noise sources and area functions for the synthesis of fricative consonants". *Rostocker Informatik Berichte*, **30**:17–23.

Birkholz, Peter, Dietmar Jackèl & Bernd J. Kröger (2006) "Construction and control of a three-dimensional vocal tract model". In: *International Conference on Acoustics, Speech, and Signal Processing*, p. 873–876. Toulouse, France: IEEE. doi:10.1109/ICASSP.2006.1660160.

Birkholz, Peter, Dietmar Jackèl & Bernd J. Kröger (2007a) "Simulation of losses due to turbulence in the time-varying vocal system". *IEEE Transactions on Audio, Speech, and Language Processing*, **15**(4):1218–1226. doi:10.1109/TASL.2006.889731.

Birkholz, Peter & Bernd J. Kröger (2006) "Vocal tract model adaptation using magnetic resonance imaging". In: *7th International Seminar on Speech Production*, pp. 493–500. Ubatuba, Brazil.

Birkholz, Peter, Ingmar Steiner & Stefan Breuer (2007b) "Control concepts for articulatory speech synthesis". In: Petra Wagner, Julia Abresch, Stefan Breuer & Wolfgang Hess (eds.), *Sixth ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)*, pp. 5–10. Bonn, Germany: ISCA. URL `http://www.isca-speech.org/archive/ssw6/ssw6_005.html`.

Black, Alan W. (2006) "CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling". In: *Interspeech*, pp. 1762–1765. Pittsburgh, PA, USA: ISCA. URL `http://www.isca-speech.org/archive/interspeech_2006/i06_1394.html`.

Black, Alan W. & W. Nick Campbell (1995) "Optimising selection of units from speech databases for concatenative synthesis". In: *Eurospeech*, volume 1, pp. 581–584. Madrid, Spain: ESCA. URL `http://www.isca-speech.org/archive/eurospeech_1995/e95_0581.html`.

Black, Alan W., Paul Taylor & Richard Caley (1999) *The Festival Speech Synthesis System*. University of Edinburgh, 1.4 edition. URL `http://www.cstr.ed.ac.uk/projects/festival/manual/`.

Black, Alan W. & Keiichi Tokuda (2005) "The Blizzard Challenge - 2005: Evaluating corpus-based speech synthesis on common datasets". In: *Interspeech*, pp. 77–80. Lisbon, Portugal: ISCA. URL `http://www.isca-speech.org/archive/interspeech_2005/i05_0077.html`.

Boersma, Paul (1998) *Functional Phonology. Formalizing the interactions between articulatory and perceptual drives.* Ph.D. thesis, University of Amsterdam. URL `http://www.fon.hum.uva.nl/paul/diss/diss.html`.

Boersma, Paul & David Weenink (1995–2010) "Praat: doing phonetics by computer". Computer program. URL `http://www.praat.org/`.

Brackhane, Fabian & Jürgen Trouvain (2008) "What makes "mama" and "papa" acceptable? - Experiments with a replica of von Kempelen's speaking machine". In: Rudolph Sock, Susanne Fuchs & Yves Laprie (eds.), *8th International Seminar on Speech Production*, pp. 329–332. Strasbourg, France: LORIA. URL `http://issp2008.loria.fr/Proceedings/PDF/issp2008-76.pdf`.

Branderud, Peter (1985) "Movetrack – a movement tracking system". In: *French–Swedish Symposium on Speech*, pp. 113–122. Grenoble, France.

Branderud, Peter, Hans-Jerker Lundberg, Jaroslava Lander, Hassan Djamshidpey, Ivar Wäneland, Diana Krull & Björn E. F. Lindblom (1998) "X-ray analyses of speech: Methodological aspects". In: Peter Branderud & Hartmut Traunmüller (eds.), *Proc. FONETIK 98*, pp. 168–171. Stockholm, Sweden. URL `http://www.ling.su.se/staff/peter/Pb_Bli.html`.

Breiman, L., J. H. Friedman, R. A. Olshen & C. J. Stone (1984) *Classification and Regression Trees.* Belmont, CA: Wadsworth.

Bresch, Erik, Jon Nielsen, Krishna Nayak & Shrikanth S. Narayanan (2006) "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans". *Journal of the Acoustical Society of America*, **120**(4):1791–1794. doi:10.1121/1.2335423.

Breuer, Stefan (2009) *Multifunktionale und multilinguale Unit-Selection-Sprachsynthese.* Ph.D. thesis, Universität Bonn. URL `http://nbn-resolving.de/urn:nbn:de:hbz:5-16507`.

Breuer, Stefan & Wolfgang Hess (2010) "The Bonn Open Synthesis System 3". *International Journal of Speech Technology*, **13**(2):75–84. doi:10.1007/s10772-010-9072-2.

van den Broecke, Marcel (1983) "Wolfgang von Kempelen's speaking machine as a performer". In: Marcel van den Broecke, Vincent van Heuven & Wim Zonneveld (eds.), *Sound Structures. Studies for Antonie Cohen.*, pp. 9–19. Dordrecht: Foris. ISBN 978-90-7017693-8.

Browman, Catherine P. & Louis M. Goldstein (1992a) "Articulatory phonology: An overview". *Phonetica*, **49**:155–180.

Browman, Catherine P. & Louis M. Goldstein (1992b) ""Targetless" schwa: an articulatory analysis". In: Gerard J. Docherty & D. Robert Ladd (eds.), *Gesture, Segment, Prosody*, number 2 in Papers in Laboratory Phonology, pp. 26–65. Cambridge University Press.

Buchaillard, Stéphanie I., Sim Heng Ong, Yohan Payan & Kelvin Foong (2007) "3D statistical models for tooth surface reconstruction". *Computers in Biology and Medicine*, **37**(10):1461–1471. doi:10.1016/j.compbiomed.2007.01.003.

Byrd, Dani, Catherine P. Browman, Louis M. Goldstein & Douglas N. Honorof (1995) "EMMA and X-ray microbeam comparison". *Journal of the Acoustical Society of America*, **97**(5):3365. doi:10.1121/1.412700.

Byrd, Dani, Catherine P. Browman, Louis M. Goldstein & Douglas N. Honorof (1999) “Magnetometer and X-ray microbeam comparison”. In: John J. Ohala, Yoko Hasegawa, Manjari Ohala, Daniel Granville & Ashlee C. Bailey (eds.), *14th International Congress of Phonetic Sciences*, volume 1, pp. 627–630. San Francisco, CA, USA.

Cabral, João, Steve Renals, Korin Richmond & Junichi Yamagishi (2008) “Glottal spectral separation for parametric speech synthesis”. In: *Interspeech*, pp. 1829–1832. Brisbane, Australia: ISCA. URL `http://www.isca-speech.org/archive/interspeech_2008/i08_1829.html`.

Carlson, Rolf, Tor Sigvardson & Arvid Sjölander (2002) “Data-driven formant synthesis”. *KTH Department for Speech, Music and Hearing Quarterly Progress and Status Report*, **44**(1):121–124. URL `http://www.speech.kth.se/prod/publications/files/qpsr/2002/2002_44_1_121-124.pdf`.

Chan, Dominic, Adrian Fourcin, Dafydd Gibbon, Björn Granström, Mark Huckvale, George Kokkinakis, Knut Kvale, Lori Lamel, Børge Lindberg, Asunción Moreno, Jiannis Mouropoulos, Franco Senia, Isabel Trancoso, Corin ’t Veld & Jerome Zeiliger (1995) “EUROM - a spoken language resource for the EU - the SAM projects”. In: *Eurospeech*, pp. 867–870. Madrid, Spain: ESCA. URL `http://www.isca-speech.org/archive/eurospeech_1995/e95_0867.html`.

Chiba, Tsutomu & Masato Kajiyama (1941) *The Vowel: Its Nature and Structure*. Tokyo, Japan: Tokyo-Kaiseikan.

Chomsky, Noam & Morris Halle (1968) *The Sound Pattern of English*. New York, NY: Harper & Row.

Clark, Robert A. J., Korin Richmond & Simon King (2007) “Multisyn: Open-domain unit selection for the Festival speech synthesis system”. *Speech Communication*, **49**(4):317–330. doi:10.1016/j.specom.2007.01.014.

Clements, G. N. (1985) “The geometry of phonological features”. *Phonology Yearbook*, **2**:225–252. URL `http://www.jstor.org/stable/4419958`.

Cohen, Marc H. & Joseph S. Perkell (1985) “Evaluation of an alternating magnetic field system for transducing articulatory movements in the midsagittal plane”. *Journal of the Acoustical Society of America*, **77**(S1):S99. doi:10.1121/1.2022629.

Coker, Cecil H. (1976) “A model of articulatory dynamics and control”. *Proceedings of the IEEE*, **64**(4):452–460.

Connan, Pierre-Yves, Gilbert Brock, Johanna-Pascale Roy & Fabrice Hirsch (2003) “Using digital cineradiography to study anticipatory labial activity in French”. In: *15th International Congress of Phonetic Sciences*, pp. 3153–3156. Barcelona, Spain.

Dang, Jianwu & Kiyoshi Honda (1996) “Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation”. *Journal of the Acoustical Society of America*, **100**(5):3374–3383. doi:10.1121/1.416978.

Dang, Jianwu & Kiyoshi Honda (2004) “Construction and control of a physiological articulatory model”. *Journal of the Acoustical Society of America*, **115**(2):853–870. doi:10.1121/1.1639325.

Dang, Jianwu, Kiyoshi Honda & Hisayoshi Suzuki (1994) “Morphological and acoustical analysis of the nasal and the paranasal cavities”. *Journal of the Acoustical Society of America*, **96**(4):2088–2100. doi:10.1121/1.410150.

Dart, Sarah N. (1987) “A bibliography of x-ray studies of speech”. *UCLA Working Papers in Phonetics*, **66**:1–97. URL `http://escholarship.org/uc/item/8bd688fw`.

Davis, Steven B. & Paul Mermelstein (1980) “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **28**(4):357–366.

Demolin, Didier, Thierry Metens & Alain Soquet (1996) “Three-dimensional measurement of the vocal tract by MRI”. In: *4th International Conference on Spoken Language Processing*, volume 1, p. 272–275. Philadelphia, PA, USA: ISCA. URL `http://www.isca-speech.org/archive/icslp_1996/i96_0272.html`.

Demolin, Didier, Thierry Metens & Alain Soquet (2000) “Real time MRI and articulatory coordinations in vowels”. In: Phil Hoole (ed.), *5th Seminar on Speech Production*, pp. 93–96. Kloster Seeon, Germany. URL `http://www.phonetik.uni-muenchen.de/institut/veranstaltungen/SPS5/abstracts/21_abs.html`.

Dromey, Christopher, Shawn Nissen, Petrea Nohr & Samuel G. Fletcher (2006) “Measuring tongue movements during speech: Adaptation of a magnetic jaw-tracking system”. *Speech Communication*, **48**(5):463–473. doi:10.1016/j.specom.2005.05.003.

Dudley, Homer & Thomas H. Tarnóczy (1950) “The speaking machine of Wolfgang von Kempelen”. *Journal of the Acoustical Society of America*, **22**(2):151–166. doi:10.1121/1.1906583.

Dutoit, Thierry (1997) *An Introduction to Text-To-Speech Synthesis.* Number 3 in Text, speech, and language technology. Springer. ISBN 978-0792344988.

Dutoit, Thierry & Henri Leich (1993) “MBR-PSOLA: Text-To-Speech synthesis based on an MBE resynthesis of the segments database”. *Speech Communication*, **13**(3-4):435–440. doi:10.1016/0167-6393(93)90042-J.

Dutoit, Thierry, Vincent Pagel, N. Pierret, F. Bataille & O. van der Vrecken (1996) “The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes”. In: *4th International Conference on Spoken Language Processing*, pp. 1393–1396. Philadelphia, PA, USA: ISCA. URL `http://www.isca-speech.org/archive/icslp_1996/i96_1393.html`.

El-Masri, Samir, Xavier Pelorson, Pierre Saguet & Pierre Badin (1996) “Vocal tract acoustics using the transmission line matrix (TLM) method”. In: *4th International Conference on Spoken Language Processing*, pp. 953–956. Philadelphia, PA, USA: ISCA. URL `http://www.isca-speech.org/archive/icslp_1996/i96_0953.html`.

Engelke, Wilfried, Thomas Bruns, Mark Striebeck & Gerhard Hoch (1996) “Midsagittal velar kinematics during production of VCV sequences”. *Cleft Palate-Craniofacial Journal*, **33**(3):236–244. doi:10.1597/1545-1569(1996)033<0236:MVKDPO>2.3.CO;2.

Engwall, Olov (1999) “Vocal tract modeling in 3D”. *KTH Department for Speech, Music and Hearing Quarterly Progress and Status Report*, **40**(1-2):31–38. URL `http://www.speech.kth.se/prod/publications/files/qpsr/1999/1999_40_1-2_031-038.pdf`.

Engwall, Olov (2002) *Tongue Talking: Studies in Intraoral Speech Synthesis.* Ph.D. thesis, KTH, Stockholm, Sweden. URL `http://www.speech.kth.se/prod/publications/files/828.pdf`.

Engwall, Olov (2006) “Assessing Magnetic Resonance Imaging measurements: Effects of sustenation, gravitation, and coarticulation”. In: Jonathan Harrington & Marija Tabain (eds.), *Speech Production: Models, Phonetic Processes, and Techniques*, Macquarie Monographs in Cognitive Science, chapter 17, pp. 301–313. New York, NY: Psychology Press.

Epstein, Melissa A. & Maureen Stone (2005) “The tongue stops here: Ultrasound imaging of the palate”. *Journal of the Acoustical Society of America*, **118**(4):2128–2131. doi:10.1121/1.2031977.

Ericsdotter, Christine (2005) *Articulatory-Acoustic Relationships in Swedish Vowel Sounds.* Ph.D. thesis, Stockholm University. URL `http://www2.ling.su.se/staff/ericsdotter/thesis/`.

Ericsdotter, Christine, Björn E. F. Lindblom & Johan Stark (1999) "Articulatory coordination in coronal stops: Implications for theories of coarticulation". In: John J. Ohala, Yoko Hasegawa, Manjari Ohala, Daniel Granville & Ashlee C. Bailey (eds.), *14th International Congress of Phonetic Sciences*, volume 3, pp. 1885–1888. San Francisco, CA, USA.

Fagel, Sascha (2004) *Audiovisuelle Sprachsynthese. Systementwicklung und -bewertung.* Number 2 in Mündliche Kommunikation. Berlin: Logos. ISBN 978-3-8325-0742-8. PhD thesis.

Fagel, Sascha & Caroline Clemens (2004) "An articulation model for audiovisual speech synthesis: Determination, adjustment, evaluation". *Speech Communication*, **44**(1-4):141–154. doi:10.1016/j.specom.2004.10.006.

Falaschi, Alessandro, Massimo Giustiniani & Massimo Verola (1989) "A hidden Markov model approach to speech synthesis". In: *Eurospeech*, volume 2, pp. 187–190. Paris, France: ESCA. URL `http://www.isca-speech.org/archive/eurospeech_1989/e89_2187.html`.

Fant, Gunnar (1960) *Acoustic theory of speech production.* Den Haag, Netherlands: Walter de Gruyter. ISBN 978-90-2791600-6.

Fant, Gunnar (1970) *Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations*, volume 2 of *Description and analysis of contemporary standard Russian.* Den Haag, Netherlands: Walter de Gruyter, second edition.

Fant, Gunnar & Mats Båvegård (1997) "Parametric model of VT area functions: vowels and consonants". *KTH Department for Speech, Music and Hearing Quarterly Progress and Status Report*, **38**(1):1–20. URL `http://www.speech.kth.se/prod/publications/files/qpsr/1997/1997_38_1_001-020.pdf`.

Fant, Gunnar, K. Ishizaka, Lindqvist-Gauffin J & Johan Sundberg (1972) "Subglottal formants". *KTH Department for Speech, Music and Hearing Quarterly Progress and Status Report*, **13**(1):1–12. URL `http://www.speech.kth.se/prod/publications/files/qpsr/1972/1972_13_1_001-012.pdf`.

Fant, Gunnar, Johan Liljencrants & Qi guaq Lin (1985) "A four-parameter model of glottal flow". *KTH Department for Speech, Music and Hearing Quarterly Progress and Status Report*, **26**(4):1–13. URL `http://www.speech.kth.se/prod/publications/files/qpsr/1985/1985_26_4_001-013.pdf`.

Fels, Sidney, Florian Vogt, Kees van den Doel, John E. Lloyd, Ian Stavness & Eric Vatikiotis-Bateson (2006) *ArtiSynth: A Biomechanical Simulation Platform for the Vocal Tract and Upper Airway.* Technical Report 10, Computer Science Dept., University of British Columbia, Vancouver, BC, Canada. URL `http://hct.ece.ubc.ca/publications/pdf/TR-2006-10.pdf`.

Flanagan, James L., Kenzo Ishizaka & K.L. Shipley (1975) "Synthesis of speech from a dynamic model of the vocal cords and vocal tract". *Bell System Technical Journal*, **45**:485–506.

Foldvik, Arne Kjell, Ulf Kristiansen & Jørn Kværness (1993) "A time-evolving three-dimensional vocal tract model by means of magnetic resonance imaging (MRI)". In: *Eurospeech*, pp. 557–558. Berlin, Germany: ESCA. URL `http://www.isca-speech.org/archive/eurospeech_1993/e93_0557.html`.

Fontecave, Julie & Frédéric Berthommier (2006) "Semi-automatic extraction of vocal tract movements from cineradiographic data". In: *Interspeech*, pp. 569–572. Pittsburgh, PA, USA: ISCA. URL `http://www.isca-speech.org/archive/interspeech_2006/i06_1439.html`.

Fuchs, Susanne (2005) "Articulatory correlates of the voicing contrast in alveolar obstruent production in German". *ZAS Papers in Linguistics*, **41**(41):1–252. URL `http://www.zas.gwz-berlin.de/fileadmin/material/ZASPiL_Volltexte/zp41/zaspil41.pdf`. PhD thesis.

Fuchs, Susanne & Pascal Perrier (2003) "An EMMA/EPG study of voicing contrast correlates in German". In: *15th International Congress of Phonetic Sciences*, pp. 1057–1060. Barcelona, Spain.

Fujimura, Osamu, Shigeru Kiritani & Haruhisa Ishida (1963) "Computer controlled dynamic cineradiography". *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics*, **2**:6–10. URL `http://www.umin.ac.jp/memorial/rilp-tokyo/R02/R02_006.pdf`. Faculty of Medicine, University of Tokyo.

Fujimura, Osamu, Shigeru Kiritani & Haruhisa Ishida (1973) "Computer controlled radiography for observation of movements of articulatory and other human organs". *Computers in Biology and Medicine*, **3**(4):371–378. doi:10.1016/0010-4825(73)90003-6.

Fukui, Kotaro, Yuma Ishikawa, Takashi Sawa, Eiji Shintaku, Masaaki Honda & Atsuo Takanishi (2007) "New anthropomorphic talking robot having a three-dimensional articulation mechanism and improved pitch range". In: *IEEE International Conference on Robotics and Automation*, pp. 2922–2927. Rome, Italy: IEEE. doi:10.1109/ROBOT.2007.363915.

Fukui, Kotaro, Yuma Ishikawa, Eiji Shintaku, Keisuke Ohno, Nana Sakakibara, Atsuo Takanishi & Masaaki Honda (2008) "Vocal cord model to control various voices for anthropomorphic talking robot". In: Rudolph Sock, Susanne Fuchs & Yves Laprie (eds.), *8th International Seminar on Speech Production*, pp. 341–344. Strasbourg, France: LORIA. URL `http://issp2008.loria.fr/Proceedings/PDF/issp2008-79.pdf`.

Gay, Thomas (1977) "Articulatory movements in VCV sequences". *Journal of the Acoustical Society of America*, **62**(1):1977. doi:10.1121/1.381480.

Gick, B., Ian L. Wilson, Karsten Koch & Clare Cook (2004) "Language-specific articulatory settings: Evidence from inter-utterance rest position". *Phonetica*, **61**(4):220–233. doi:10.1159/000084159.

van der Giet, Gerhard (1977) "Computer-controlled method for measuring articulatory activities". *Journal of the Acoustical Society of America*, **61**(4):1072–1076. doi:10.1121/1.381376.

Gottesfeld Brown, Lisa M. (1992) "A survey of image registration techniques". *ACM Computing Surveys*, **24**(4):325–376. doi:10.1145/146370.146374.

Hamann, Silke & Susanne Fuchs (2010) "How can voiced retroflex stops evolve: Evidence from typology and an articulatory study". *Language and Speech*, **53**(2):181–216. doi:10.1177/0023830909357159.

Hardcastle, William J. (1972) "The use of electropalatography in phonetic research". *Phonetica*, **25**(4):197–215. doi:10.1159/000259382.

Hardcastle, William J., W. Jones , C. Knight , A. Trudgeon & G. Calder (1989) "New developments in electropalatography: A state-of-the-art report". *Clinical Linguistics & Phonetics*, **3**(1):1–38. doi:10.3109/02699208908985268.

Hasegawa-Johnson, Mark (1998) "Electromagnetic exposure safety of the Carstens Articulograph AG100". *Journal of the Acoustical Society of America*, **104**(4):2529–2532. doi:10.1121/1.423775. URL `http://www.ifp.illinois.edu/~hasegawa/emma/emma.html`.

Hiraishi, Kumiko, Isamu Narabayashi, Osamu Fujita, Kazuhiro Yamamoto, Akihiko Sagami, Yoichi Hisada, Yoshinori Saika, Itaru Adachi & Hideo Hasegawa (1995) "Blueberry juice: Preliminary evaluation as an oral contrast agent in gastrointestinal MR imaging". *Radiology*, **194**(1):119–123. URL `http://radiology.rsna.org/content/194/1/119.full.pdf`.

Hixon, Thomas J. (1971) "An electromagnetic method for transducing jaw movements during speech". *Journal of the Acoustical Society of America*, **49**(2B):603–606. doi:10.1121/1.1912395.

Höhne, Jörg, Paul W. Schönle, Bastian Conrad, H. Veldscholten, Peter Wenig, H. Fakhouri, N. Sandner & G. Hong (1987) “Direct measurement of vocal tract shape – articulography”. In: *European Conference on Speech Technology*, volume 2, pp. 230–232. Edinburgh, Scotland: ISCA. URL `http://www.isca-speech.org/archive/ecst_1987/e87_2230.html`.

Hoiting, Gerke Jan (2005) *Measuring MRI noise.* Ph.D. thesis, Rijksuniversiteit Groningen, Groningen, The Netherlands. URL `http://dissertations.ub.rug.nl/FILES/faculties/science/2005/g.j.hoiting/thesis.pdf`.

Honikman, Beatrice (1964) “Articulatory settings”. In: David Abercrombie, Dennis B. Fry, Peter A. D. MacCarthy, Norman C. Scott & John L. M. Trim (eds.), *In Honour of Daniel Jones. Papers contributed on the occasion of his eightieth birthday, 12 September 1961*, pp. 73–84. Longman.

Hoole, Phil (1996) “Issues in the acquisition, processing, reduction and parameterization of articulographic data”. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München*, **34**:158–173. URL `http://www.phonetik.uni-muenchen.de/~hoole/pdf/pd_ema.pdf`.

Hoole, Phil & Noël Nguyen (1997) “Electromagnetic articulography in coarticulation research”. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München*, **35**:177–184. URL `http://www.phonetik.uni-muenchen.de/forschung/FIPKM/vol35/f35_ph_5.ps`.

Hoole, Phil, Andreas Zierdt & Christian Geng (2003) “Beyond 2D in articulatory data acquisition and analysis”. In: *15th International Congress of Phonetic Sciences*, pp. 265–268. Barcelona, Spain.

Hornak, Joseph P. (1996–2009) “The basics of MRI”. Online. URL `http://www.cis.rit.edu/htbooks/mri/`.

Hummel, Johann B., Michael L. Figl, Wolfgang W. Birkfellner, Michael R. Bax, Ramin Shahidi, Jr. Calvin R. Maurer & Helmar Bergmann (2006) “Evaluation of a new electromagnetic tracking system using a standardized assessment protocol”. *Physics in Medicine and Biology*, **51**(10):N205–N210. doi:10.1088/0031-9155/51/10/N01.

Hykes, David L., Wayne R. Hedrick & Dale E. Starchman (1985) *Ultrasound physics and instrumentation.* Churchill Livingstone. ISBN 9780443084072.

International Phonetic Association (1999) *Handbook of the International Phonetic Association.* Cambridge University Press. doi:10.2277/0521637511.

Ishizaka, Kenzo & James L. Flanagan (1972) “Synthesis of voiced sounds from a two-mass model of the vocal cords”. *Bell System Technical Journal*, **51**(6):1233–1268.

Jackson, Philip J.B. & Veena D. Singampalli (2009) “Statistical identification of articulation constraints in the production of speech”. *Speech Communication*, **51**(8):695–710. doi:10.1016/j.specom.2009.03.007.

Jakobson, Roman (1960) “Why “mama” and “papa”?” In: Bernard Kaplan & Seymour Wapner (eds.), *Perspectives in Psychological Theory: Essays in Honor of Heinz Werner*, pp. 124–134. New York: International Universities Press.

Joglar, Jose A., Carol Nguyen, Diane M. Garst & William F. Katz (2009) “Safety of electromagnetic articulography in patients with pacemakers and implantable cardioverter-defibrillators”. *Journal of Speech, Language, and Hearing Research*, **52**(4):1082–1087. doi:10.1044/1092-4388(2009/08-0028).

Joy, William (1980) *An Introduction to the C shell.* University of California, Berkeley, CA, USA. URL `http://docs.freebsd.org/44doc/usd/04.csh/paper.html`.

Kaburagi, Tokihiko & Masaaki Honda (1997) "Calibration methods of voltage-to-distance function for an electromagnetic articulometer (EMA) system". *Journal of the Acoustical Society of America*, **101**(4):2391–2394. doi:10.1121/1.418255.

Kaburagi, Tokihiko, Kohei Wakamiya & Masaaki Honda (2002) "Three-dimensional electromagnetic articulograph based on a nonparametric representation of the magnetic field". In: John H. L. Hansen & Bryan Pellom (eds.), *7th International Conference on Spoken Language Processing*, pp. 2297–2300. Denver, CO, USA: ISCA. URL `http://www.isca-speech.org/archive/icslp_2002/i02_2297.html`.

Kaburagi, Tokihiko, Kohei Wakamiya & Masaaki Honda (2005) "Three-dimensional electromagnetic articulography: A measurement principle". *Journal of the Acoustical Society of America*, **118**(1):428–443. doi:10.1121/1.1928707.

Katz, William F., Sneha V. Bharadwaj, Gretchen J. Gabbert, Philipos C. Loizou, Emily A. Tobey & Oguz Poroy (2003) "EMA compatibility of the Clarion 1.2 cochlear implant system". *Acoustics Research Letters Online*, **4**(3):10–105. doi:10.1121/1.1591712.

Kawahara, Hideki, Ikuyo Masuda-Katsuse & Alain de Cheveigné (1999) "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds". *Speech Communication*, **28**(3-4):187–207. doi:10.1016/S0167-6393(98)00085-5.

Keller, Eric (ed.) (2002) *Improvements in speech synthesis: COST 258: the naturalness of synthetic speech*. Wiley-Blackwell. ISBN 9780471499855.

Kelly, John L. & Carol C. Lochbaum (1962) "Speech synthesis". In: *4th International Congress on Acoustics*, pp. G42:1–4. Copenhagen, Denmark: International Acoustics Commission.

Kelso, J. A. Scott, Elliot L. Saltzman & Betty Tuller (1986) "The dynamical perspective on speech production: Data and theory". *Journal of Phonetics*, **14**(1):29–59.

von Kempelen, Wolfgang (1791) *Über den Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine*. Vienna, Austria: J. B. Degen.

Kipp, Andreas, Maria-Barbara Wesenick & Florian Schiel (1996) "Automatic detection and segmentation of pronunciation variants in German speech corpora". In: *4th International Conference on Spoken Language Processing*, pp. 106–109. Philadelphia, PA, USA: ISCA. URL `http://www.isca-speech.org/archive/icslp_1996/i96_0106.html`.

Kiritani, Shigeru, Hajime Hirose & Masayuki Sawashima (1980) "Simultaneous X-ray microbeam and EMG study of velum movement for Japanese nasal sounds". *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics*, **14**:91–100. URL `http://www.umin.ac.jp/memorial/rilp-tokyo/R14/R14_091.pdf`. Faculty of Medicine, University of Tokyo.

Kiritani, Shigeru, Kenji Itoh, Hiroshi Imagawa, Hiroya Fujisaki & Masayuki Sawashima (1975) "Tongue pellet tracking and other radiographic observations by a computer controlled X-ray microbeam systems". *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics*, **9**:1–14. URL `http://www.umin.ac.jp/memorial/rilp-tokyo/R09/R09_001.pdf`. Faculty of Medicine, University of Tokyo.

Kitamura, Tatsuya, Hironori Takemoto, Seiji Adachi & Kiyoshi Honda (2009) "Transfer functions of solid vocal-tract models constructed from ATR MRI database of Japanese vowel production". *Acoustical Science and Technology*, **30**(4):288–296. doi:10.1250/ast.30.288.

Kitamura, Tatsuya, Hironori Takemoto, Kiyoshi Honda, Yasuhiro Shimada, Ichiro Fujimoto, Yuko Syakudo, Shinobu Masaki, Kagayaki Kuroda, Noboru Oku-Uchi & Michio Senda (2005) "Difference in vocal tract shape between upright and supine postures: Observations by an open-type MRI scanner". *Acoustical Science and Technology*, **26**(5):465–468. doi:10.1250/ast.26.465.

Klatt, Dennis H. (1980) “Software for a cascade/parallel formant synthesizer”. *Journal of the Acoustical Society of America*, **67**(3):971–995. doi:10.1121/1.383940.

Klatt, Dennis H. (1987) “Review of text-to-speech conversion for English”. *Journal of the Acoustical Society of America*, **82**(3):737–793. doi:10.1121/1.395275.

Koneczna, Halina & Witold Zawadowski (1956) *Obrazy rentgenograficzne głosek rosyjskich*. Warsaw, Poland: Państwowe Wydawnictwo Naukowe.

Kröger, Bernd J. (1998) *Ein phonetisches Modell der Sprachproduktion*. Tübingen, Germany: Niemeyer. ISBN 3-484-30387-5.

Kröger, Bernd J. (2000) “Analyse von MRT-Daten zur Entwicklung eines vokalischen Artikulationsmodells auf der Ebene der Areafunktion”. In: Klaus Fellbaum (ed.), *Elektronische Sprachsignalverarbeitung*, number 20 in Studientexte zur Sprachkommunikation, pp. 201–208. Cottbus, Germany.

Kröger, Bernd J. (2007) “Perspectives for articulatory speech synthesis”. In: Petra Wagner, Julia Abresch, Stefan Breuer & Wolfgang Hess (eds.), *Sixth ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)*, p. 391. Bonn, Germany: ISCA. URL `http://www.isca-speech.org/archive/ssw6/ssw6_391.html`.

Kröger, Bernd J., Phil Hoole, Robert Sader, Christian Geng, Bernd Pompino-Marschall & Christiane Neuschaefer-Rube (2004) “MRT-Sequenzen als Datenbasis eines visuellen Artikulationsmodells”. *HNO*, **52**(9):837–843. doi:10.1007/s00106-004-1097-x.

Kröger, Bernd J., Marianne Pouplier & Mark K. Tiede (2008) “An evaluation of the Aurora system as a flesh-point tracking tool for speech production research”. *Journal of Speech, Language, and Hearing Research*, **51**(4):914–921. doi:10.1044/1092-4388(2008/067).

Kröger, Bernd J., Ralf Winkler, Christine Mooshammer & Bernd Pompino-Marschall (2000) “Estimation of vocal tract area function from magnetic resonance imaging: Preliminary results”. In: Phil Hoole (ed.), *5th Seminar on Speech Production*, pp. 333–336. Kloster Seeon, Germany. URL `http://www.phonetik.uni-muenchen.de/institut/veranstaltungen/SPS5/abstracts/10_abs.html`.

Kroos, Christian (2007) “The AHAA! auditory-visual speech production lab at MARCS”. URL `http://www.ag500.de/wiki/emaws/EmaWs/ahaa_lab.pdf`. Presented at XVIth ICPhS EMA workshop.

Kroos, Christian (2008) “Measurement accuracy in 3D electromagnetic articulography (Carstens AG500)”. In: Rudolph Sock, Susanne Fuchs & Yves Laprie (eds.), *8th International Seminar on Speech Production*, pp. 61–64. Strasbourg, France: LORIA. URL `http://issp2008.loria.fr/Proceedings/PDF/issp2008-9.pdf`.

Ladefoged, Peter & Anthony Traill (1984) “Linguistic phonetic descriptions of clicks”. *Language*, **60**(1):1–20. URL `http://www.jstor.org/stable/414188`.

Lakshminarayanan, A. V., Sungbok Lee & Martin J. McCutcheon (1991) “MR imaging of the vocal tract during vowel production”. *Journal of Magnetic Resonance Imaging*, **1**(1):71–76. doi:10.1002/jmri.1880010109.

Laprie, Yves & Marie-Odile Berger (1996) “Extraction of tongue contours in X-ray images with minimal user interaction”. In: *4th International Conference on Spoken Language Processing*, pp. 268–271. Philadelphia, PA, USA: ISCA. URL `http://www.isca-speech.org/archive/icslp_1996/i96_0268.html`.

Laver, John (1994) *Principles of Phonetics*. Cambridge Textbooks in Linguistics. Cambridge University Press. ISBN 978-0-52145655-5.

Li, Min, Chandra Kambhamettu & Maureen Stone (2005) “Automatic contour tracking in ultrasound images”. *Clinical Linguistics & Phonetics*, **19**(6/7):545–554. doi:10.1080/02699200500113616.

Lindblom, Björn E. F. (1990) "Explaining phonetic variation: A sketch of the H&H theory". In: William J. Hardcastle & Alain Marchal (eds.), *Speech Production and Speech Modeling*, pp. 403–439. Dordrecht: Kluver.

Lindblom, Björn E. F. & P. Bivner (1966) "A method for continuous recording of articulatory movement". *KTH Department for Speech, Music and Hearing Quarterly Progress and Status Report*, **7**(1):14–16. URL `http://www.speech.kth.se/prod/publications/files/qpsr/1966/1966_7_1_014-016.pdf`.

Low, Russell N. & Isaac R. Francis (1997) "MR Imaging of the gastrointestinal tract with IV gadolinium and diluted barium oral contrast media compared with unenhanced MR Imaging and CT". *American Journal of Roentgenology*, **169**(4):1051–1059. URL `http://www.ajronline.org/cgi/reprint/169/4/1051`.

Lustig, Michael (2008) *Sparse MRI*. Ph.D. thesis, Stanford University. URL `http://www-mrsrl.stanford.edu/~mlustig/thesis.pdf`.

Lustig, Michael, David L. Donoho & John M. Pauly (2007) "Sparse MRI: The application of compressed sensing for rapid MR imaging". *Magnetic Resonance in Medicine*, **58**(6):1182–1195. doi:10.1002/mrm.21391.

Maeda, Shinji (1982) "A digital simulation method of the vocal-tract system". *Speech Communication*, **1**(3-4):199–229. doi:10.1016/0167-6393(82)90017-6.

Maeda, Shinji (1990) "Compensatory articulation during speech: Evidence from the analysis of vocal-tract shapes using an articulatory model". In: William J. Hardcastle & Alain Marchal (eds.), *Speech Production and Speech Modeling*, pp. 131–149. Dordrecht: Kluver.

Masaki, Shinobu, Reiko Akahane-Yamada, Mark K. Tiede, Yasuhiro Shimada & Ichiro Fujimoto (1996) "An MRI-based analysis of the English /r/ and /l/ articulations". In: *4th International Conference on Spoken Language Processing*, volume 3, pp. 1581–1584. Philadelphia, PA, USA: ISCA. URL `http://www.isca-speech.org/archive/icslp_1996/i96_1581.html`.

Masaki, Shinobu, Yukiko Nota, S. Takano, Hironori Takemoto, Tatsuya Kitamura & Kiyoshi Honda (2008) "Integrated magnetic resonance imaging methods for speech science and technology". URL `http://lpp.univ-paris3.fr/productions/conferences/2008/SpeechProductionWorkshop_2008/abstract/OptionalPaperASAParis%28Masaki%29.pdf`.

Masaki, Shinobu, Mark K. Tiede, Kiyoshi Honda, Yasuhiro Shimada, Ichiro Fujimoto, Yuji Nakamura & Noboru Ninomiya (1999) "MRI-based speech production study using a synchronized sampling method". *Journal of the Acoustical Society of Japan*, **20**(5):375–379. URL `http://ci.nii.ac.jp/naid/110003106213/en/`.

Matsuzaki, Hiroki & Kunitoshi Motoki (2000) "FEM analysis of 3-D vocal tract model with asymmetrical shape". In: Phil Hoole (ed.), *5th Seminar on Speech Production*, pp. 329–332. Kloster Seeon, Germany. URL `http://www.phonetik.uni-muenchen.de/institut/veranstaltungen/SPS5/abstracts/32_abs.html`.

Mayo, Catherine, Robert A. J. Clark & Simon King (2005) "Multidimensional scaling of listener responses to synthetic speech". In: *Interspeech*, pp. 1725–1728. Lisbon, Portugal: ISCA. URL `http://www.isca-speech.org/archive/interspeech_2005/i05_1725.html`.

Mermelstein, Paul (1973) "Articulatory model for the study of speech production". *Journal of the Acoustical Society of America*, **53**(4):1070–1082. doi:10.1121/1.1913427.

Mermelstein, Paul (1976) "Distance measures for speech recognition, psychological and instrumental". In: Chi-Hau Chen (ed.), *Pattern Recognition and Artificial Intelligence*, pp. 374–388. New York, NY: Academic Press.

Meyer, Peter, Reiner Wilhelms & Hans Werner Strube (1989) "A quasiarticulatory speech synthesizer for German language running in real time". *Journal of the Acoustical Society of America*, **86**(2):523–539. doi:10.1121/1.398232.

Miller, Amanda L. & Kenneth B. Finch (in press) "Corrected high-speed anchored ultrasound with software alignment". *Journal of Speech, Language, and Hearing Research*. URL `http://faculty.arts.ubc.ca/amiller/CHAUSA_AMiller_5.24.09.pdf`.

Möbius, Bernd (2000) "Corpus-based speech synthesis: methods and challenges". *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung*, **6**(4):87–116. URL `http://www.ims.uni-stuttgart.de/phonetik/aims.html`.

Möbius, Bernd (2003) "Rare events and closed domains: Two delicate concepts in speech synthesis". *International Journal of Speech Technology*, **6**(1):57–71. doi:10.1023/A:1021052023237.

Moulines, Eric & Francis Charpentier (1990) "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones". *Speech Communication*, **9**(5-6):453–467. doi:10.1016/0167-6393(90)90021-Z.

Munhall, Kevin G., Eric Vatikiotis-Bateson & Y. Tohkura (1994) *Manual for the X-ray film database*. Technical Report TR-H-116, ATR.

Munhall, Kevin G., Eric Vatikiotis-Bateson & Y. Tohkura (1995) "X-ray film database for speech research". *Journal of the Acoustical Society of America*, **98**(2):1222–1224. doi:10.1121/1.413621.

Narayanan, Shrikanth S., Abeer A. Alwan & Katherine Haker (1995) "An articulatory study of fricative consonants using magnetic resonance imaging". *Journal of the Acoustical Society of America*, **98**(3):1325–1347. doi:10.1121/1.413469.

Narayanan, Shrikanth S., Krishna Nayak, Sungbok Lee, Abhinav Sethy & Dani Byrd (2004) "An approach to real-time magnetic resonance imaging for speech production". *Journal of the Acoustical Society of America*, **115**(4):1771–1776. doi:10.1121/1.1652588.

NessAiver, Moriel S., Maureen Stone, Vijay Parthasarathy, Yuvi Kahana & Alex Paritsky (2006) "Recording high quality speech during tagged cine-MRI studies using a fiber optic microphone". *Journal of Magnetic Resonance Imaging*, **23**(1):92–97. doi:10.1002/jmri.20463.

Nguyen, Noël & Alain Marchal (1993) "Assessment of an electromagnetometric system for the investigation of articulatory movements in speech production". *Journal of the Acoustical Society of America*, **94**(2):1152–1155. doi:10.1121/1.406964.

Nikléczy, Péter & Gábor Olaszy (2003) "A reconstruction of Farkas Kempelen's speaking machine". In: *Eurospeech*, pp. 2453–2456. Geneva, Switzerland: ISCA. URL `http://www.isca-speech.org/archive/eurospeech_2003/e03_2453.html`.

Nishikawa, Kazufumi, Kouichirou Asama, Kouki Hayashi, Hideaki Takanobu & Atsuo Takanishi (2000) "Development of a talking robot". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 3, pp. 1760–1765. Takamatsu, Japan: IEEE/RSJ. doi:10.1109/IROS.2000.895226.

Nishikawa, Kazufumi, Satoshi Imai, Takayuki Ogawara, Hideaki Takanobu, Takemi Mochida & Atsuo Takanishi (2003) "Development of a talking robot for vowels and consonant sounds". *Acoustical Science and Technology*, **24**(1):32–34. doi:10.1250/ast.24.32.

Ogata, Kohichi & Yorinobu Sonoda (2001) "Articulatory measuring system by using magnetometer and optical sensors". *Acoustical Science and Technology*, **22**(2):141–147. doi:10.1250/ast.22.141.

Ohala, John J., Shizuo Hiki, Stanley Hubler & Richard Harshman (1968) "Photoelectric methods of transducing lip and jaw movements in speech". *UCLA Working Papers in Phonetics*, **10**:135–144. URL `http://escholarship.org/uc/item/05m564xz`.

Öhman, Sven E. G. (1966) "Coarticulation in VCV utterances: Spectrographic measurements". *Journal of the Acoustical Society of America*, **39**(1):151–168. doi:10.1121/1.1909864.

Öhman, Sven E. G. (1967) "Numerical model of coarticulation". *Journal of the Acoustical Society of America*, **41**(2):310–320. doi:10.1121/1.1910340.

Olt, Silvia & Peter M. Jakob (2004) "Contrast-enhanced dental MRI for visualization of the teeth and jaw". *Magnetic Resonance in Medicine*, **52**(1):174–176. doi:10.1002/mrm.20125.

Parthasarathy, Vijay, Maureen Stone & Jerry L. Prince (2005) "Spatiotemporal visualization of the tongue surface using ultrasound and kriging (SURFACES)". *Clinical Linguistics & Phonetics*, **19**(6/7):529–544. doi:10.1080/02699200500113632.

Perkell, Joseph S. (1969) *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study*. Cambridge, MA, USA: MIT Press. ISBN 978-0-262-66170-6.

Perkell, Joseph S. (1974) *A physiologically-oriented model of tongue activity in speech production*. Ph.D. thesis, MIT.

Perkell, Joseph S. (1982) "Advances in the use of alternating magnetic fields to track articulatory movements". *Journal of the Acoustical Society of America*, **71**(S1):S32–S33. doi:10.1121/1.2019337.

Perkell, Joseph S. & Marc H. Cohen (1985) "Design and construction of an alternating magnetic field system for transducing articulatory movements in the midsagittal plane". *Journal of the Acoustical Society of America*, **77**(S1):S99. doi:10.1121/1.2022630.

Perkell, Joseph S. & Marc H. Cohen (1986) *An Alternating Magnetic Field System for Tracking Multiple Speech Articulatory Movements in the Midsagittal Plane*. Technical Report 512, MIT Research Laboratory of Electronics. doi:1721.1/4231.

Perkell, Joseph S., Marc H. Cohen, Mario A. Svirsky, Melanie L. Matthies, Iñaki Garabieta & Michel T. T. Jackson (1992) "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements". *Journal of the Acoustical Society of America*, **92**(6):3078–3096. doi:10.1121/1.404204.

Perkell, Joseph S. & David K. Oka (1980) "Use of an alternating magnetic field device to track midsagittal plane movements of multiple points inside the vocal tract". *Journal of the Acoustical Society of America*, **67**(S1):S92. doi:10.1121/1.2018483.

Perrier, Pascal, Louis-Jean Boë & Rudolph Sock (1992) "Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract cast". *Journal of Speech and Hearing Research*, **35**(1):53–67.

Portele, Thomas, Walter Sendelmeier & Wolfgang Hess (1990) "HADIFIX: A system for German speech synthesis based on demisyllables, diphones and suffixes". In: Gérard Bailly & Christian Benoît (eds.), *ESCA Workshop on Speech Synthesis (SSW1)*, pp. 161–164. Autrans, France: ESCA. URL `http://www.isca-speech.org/archive/ssw1/ssw1_161.html`.

Prince, Alan & Paul Smolensky (1993) *Optimality Theory: Constraint Interaction in Generative Grammar*. Technical Report RuCCS-TR-2, Rutgers University Center for Cognitive Science, New Brunswick, NJ, USA. URL `http://roa.rutgers.edu/files/537-0802/537-0802-PRINCE-0-0.PDF`. ROA 537-0802.

Radke, John C. (1984) "Kinesiograph sensor array alignment system". URL `http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=4459109`.

Riordan, R. D., M. Khonsari, J. Jeffries, G. F. Maskell & P. G. Cook (2004) "Pineapple juice as a negative oral contrast agent in magnetic resonance cholangiopancreatography: a preliminary evaluation". *British Journal of Radiology*, **77**:991–999. doi:10.1259/bjr/36674326.

Rokkaku, Mitsuhiro, Kiyoshi Hashimoto, Satoshi Imaizumi, Seiji Niimi & Shigeru Kiritani (1986) "Measurements of the three-dimensional shape of the vocal tract based on the magnetic resonance imaging technique". *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics*, **20**:47–54. URL `http://www.umin.ac.jp/memorial/rilp-tokyo/R20/R20_047.pdf`. Faculty of Medicine, University of Tokyo.

Rubin, Philip E., Thomas Baer & Paul Mermelstein (1981) "An articulatory synthesizer for perceptual research". *Journal of the Acoustical Society of America*, **70**(2):321–328. doi:10.1121/1.386780.

Rubin, Philip E., Elliot L. Saltzman, Louis M. Goldstein, Richard McGowan, Mark K. Tiede & Catherine P. Browman (1996) "CASY and extensions to the task-dynamic model". In: *1st ESCA Tutorial and Research Workshop on Speech Production Modeling*, pp. 125–128. Autrans, France: ISCA. URL `http://www.isca-speech.org/archive/spm_96/sps6_125.html`.

Sagisaka, Yoshinori (1988) "Speech synthesis by rule using an optimal selection of non-uniform synthesis units". In: *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pp. 679 – 682. New York, NY, USA: IEEE. doi:10.1109/ICASSP.1988.196677.

Sakoe, Hiroaki & Seibi Chiba (1978) "Dynamic programming algorithm optimization for spoken word recognition". *IEEE Transactions on Acoustics, Speech and Signal Processing*, **26**(1):43–49. URL `http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1163055`.

Saltzman, Elliot L. (2003) "Temporal aspects of articulatory control". In: *Eurospeech*, pp. 2565–2568. Geneva, Switzerland: ISCA. URL `http://www.isca-speech.org/archive/eurospeech_2003/e03_2565.html`.

Saltzman, Elliot L. & Kevin G. Munhall (1989) "A dynamical approach to gestural patterning in speech production". *Ecological Psychology*, **1**(4):333–382. doi:10.1207/s15326969eco0104_2.

Schaeffler, Sonja, James M. Scobbie & Ineke Mennen (2008) "An evaluation of inter-speech postures for the study of language-specific articulatory settings". In: Rudolph Sock, Susanne Fuchs & Yves Laprie (eds.), *8th International Seminar on Speech Production*, pp. 121–124. Strasbourg, France: LORIA. URL `http://issp2008.loria.fr/Proceedings/PDF/issp2008-24.pdf`.

Schenck, John F. (2000) "Safety of strong, static magnetic fields". *Journal of Magnetic Resonance Imaging*, **12**(1):2–19. doi:10.1002/1522-2586(200007)12:1<2::AID-JMRI2>3.0.CO;2-V.

Schiel, Florian (1999) "Automatic phonetic transcription of non-prompted speech". In: John J. Ohala, Yoko Hasegawa, Manjari Ohala, Daniel Granville & Ashlee C. Bailey (eds.), *14th International Congress of Phonetic Sciences*, volume 1, pp. 607–610. San Francisco, CA, USA.

Schiel, Florian (2004) "MAUS goes iterative". In: *International Conference on Language Resources and Evaluation*, pp. 1015–1018. Lisbon, Portugal: ELRA.

Schönle, Paul W., Klaus Gräbe, Peter Wenig, Jörg Höhne, Jörg Schrader & Bastian Conrad (1987) "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract". *Brain and Language*, **31**(1):26–35. doi:10.1016/0093-934X(87)90058-7.

Schönle, Paul W., Peter Wenig, Jörg Schrader, Klaus Gräbe, E. Bröckmann & Bastian Conrad (1983) "Ein elektromagnetisches Verfahren zur simultanen Registrierung von Bewegungen im Bereich des Lippen-, Unterkiefer- und Zungensystems". *Biomedizinsche Technik*, **28**(11):263–267. doi:10.1515/bmte.1983.28.11.263.

Schröder, Marc & Jürgen Trouvain (2003) "The German text-to-speech synthesis system MARY: A tool for research, development and teaching". *International Journal of Speech Technology*, **6**(4):365–377. doi:10.1023/A:1025708916924.

Scobbie, James M., Alan A. Wrench & Marietta van der Linden (2008) "Head-probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement". In: Rudolph Sock, Susanne Fuchs & Yves Laprie (eds.), *8th International Seminar on Speech Production*, pp. 373–376. Strasbourg, France: LORIA. URL `http://issp2008.loria.fr/Proceedings/PDF/issp2008-87.pdf`.

Scott de Martinville, Édouard-Léon (1857) "Fixation graphique de la voix". URL `http://www.firstsounds.org/publications/articles/Phonautographic-Manuscripts.pdf`.

Shadle, Christine H. & Robert I. Damper (2001) "Prospects for articulatory synthesis: A position paper". In: *Fourth ISCA Tutorial and Research Workshop on Speech Synthesis (SSW4)*, p. 121–126. Perthshire, Scotland: ISCA. URL `http://www.isca-speech.org/archive/ssw4/ssw4_116.html`.

Shadle, Christine H., Mohammad Mohammad, John N. Carter & Philip J.B. Jackson (1999) "Multi-planar Dynamic Magnetic Resonance Imaging: New tools for speech research". In: John J. Ohala, Yoko Hasegawa, Manjari Ohala, Daniel Granville & Ashlee C. Bailey (eds.), *14th International Congress of Phonetic Sciences*, volume 1, pp. 623–626. San Francisco, CA, USA.

Shannon, Claude E. (1949) "Communication in the presence of noise". *Proceedings of the IRE*, **37**(1):10–21.

Singampalli, Veena D. (2009) *READ ME Mocha-Dansa V 1.0.* URL `http://personal.ee.surrey.ac.uk/Personal/P.Jackson/Dansa/Mocha/readme_Mocha-Dansa_1-0.pdf`.

Sondhi, Man Mohan (1986) "Resonances of a bent vocal tract". *Journal of the Acoustical Society of America*, **79**(4):1113–1116. doi:10.1121/1.393383.

Sondhi, Man Mohan & Jürgen Schröter (1987) "A hybrid time-frequency domain articulatory speech synthesizer". *IEEE Transactions on Acoustics, Speech and Signal Processing*, **35**(7):955–967. URL `http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1165240`.

Sonies, Barbara C., Thomas H. Shawker, Thomas E. Hall, Lynn H. Gerber & Stephen B. Leighton (1981) "Ultrasonic visualization of tongue motion during speech". *Journal of the Acoustical Society of America*, **70**(3):683–686. doi:10.1121/1.386930.

Sonoda, Yorinobu (1974) "Observation of tongue movements employing magnetometer sensor". *IEEE Transactions on Magnetics*, **10**(3):954–957.

Sonoda, Yorinobu & Kohichi Ogata (1992) "Improvements of magnetometer sensing system for monitoring tongue point movements during speech". In: *2nd International Conference on Spoken Language Processing*, pp. 843–846. Banff, AB, Canada: ISCA. URL `http://www.isca-speech.org/archive/icslp_1992/i92_0843.html`.

Sonoda, Yorinobu & Kohichi Ogata (1993) "Articulatory measuring system using magnetometer sensors for tongue point movements during speech". *IEEE Transactions on Magnetics*, **29**(6):3337–3339. doi:10.1109/20.281168.

Sonoda, Yorinobu & Satoshi Wanishi (1982) "New optical method for recording lip and jaw movements". *Journal of the Acoustical Society of America*, **72**(3):700–704. doi:10.1121/1.388250.

Stevens, Kenneth N., S. Kasowski & Gunnar Fant (1953) "An electrical analog of the vocal tract". *Journal of the Acoustical Society of America*, **25**(4):734–742. doi:10.1121/1.1907169.

Stöber, Karlheinz, Thomas Portele, Petra Wagner & Wolfgang Hess (1999) "Synthesis by word concatenation". In: *Eurospeech*, volume 2, pp. 619–622. Budapest, Hungary: ISCA. URL `http://www.isca-speech.org/archive/eurospeech_1999/e99_0619.html`.

Stone, Maureen (2005) "A guide to analysing tongue motion from ultrasound images". *Clinical Linguistics & Phonetics*, **19**(6/7):455–501. doi:10.1080/02699200500113558.

Stone, Maureen, Ulla Crouse & Marty Sutton (2002) "Exploring the effects of gravity on tongue motion using ultrasound image sequences". *Journal of the Acoustical Society of America*, **111**(5):2476–2477.

Stone, Maureen & Edward P. Davis (1995) "A head and transducer support system for making ultrasound images of tongue/jaw movement". *Journal of the Acoustical Society of America*, **98**(6):3107–3112. doi:10.1121/1.413799.

Stone, Maureen, Edward P. Davis, Andrew S. Douglas, Moriel S. NessAiver, Rao Gullapalli, William S. Levine & Andrew Lundberg (2001) "Modeling the motion of the internal tongue from tagged cine-MRI images". *Journal of the Acoustical Society of America*, **109**(6):2974–2982. doi:10.1121/1.1344163.

Stone, Maureen, G. Stock, Kevin Bunin, Kausum Kumar, Melissa A. Epstein, Chandra Kambhamettu, Min Li, Vijay Parthasarathy & Jerry L. Prince (2007) "Comparison of speech production in upright and supine position". *Journal of the Acoustical Society of America*, **122**(1):532–541. doi:10.1121/1.2715659.

Story, Brad H., Ingo R. Titze & Eric A. Hoffman (1996) "Vocal tract area functions from magnetic resonance imaging". *Journal of the Acoustical Society of America*, **100**(1):537–554. doi:10.1121/1.415960.

Story, Brad H., Ingo R. Titze & Eric A. Hoffman (1998) "Vocal tract area functions for an adult female speaker based on volumetric imaging". *Journal of the Acoustical Society of America*, **104**(1):471–487. doi:10.1121/1.423298.

Sundberg, Johan (1969) "Articulatory differences between spoken and sung vovels in singers". *KTH Department for Speech, Music and Hearing Quarterly Progress and Status Report*, **10**(1):33–46. URL `http://www.speech.kth.se/prod/publications/files/qpsr/1969/1969_10_1_033-046.pdf`.

Sussman, Harvey M. & Karl U. Smith (1970) "Transducer for measuring mandibular movements". *Journal of the Acoustical Society of America*, **48**(4A):857–858. doi:10.1121/1.1912215.

Taylor, Paul (2009) *Text-to-Speech Synthesis*. Cambridge University Press. ISBN 978-0-521-89927-7.

Thimm, Georg L. & Jürgen Luettin (1999) "Extraction of articulators in X-ray image sequences". In: *Eurospeech*, pp. 157–160. Budapest, Hungary: ISCA. URL `http://www.isca-speech.org/archive/eurospeech_1999/e99_0157.html`.

Tiede, Mark K., Shinobu Masaki & Eric Vatikiotis-Bateson (2000) "Contrasts in speech articulation observed in sitting and supine conditions". In: Phil Hoole (ed.), *5th Seminar on Speech Production*, p. 25–28. Kloster Seeon, Germany. URL `http://www.phonetik.uni-muenchen.de/institut/veranstaltungen/SPS5/abstracts/60_abs.html`.

Tiede, Mark K., Shinobu Masaki, Masahiko Wakumoto & Eric Vatikiotis-Bateson (1997) "Magnetometer observation of articulation in sitting and supine conditions". *Journal of the Acoustical Society of America*, **102**(5):3166. doi:10.1121/1.420773.

Tiede, Mark K. & Eric Vatikiotis-Bateson (1994) "Extracting articulator movement parameters from a videodisc-based cineradiographic database". In: *3rd International Conference on Spoken Language Processing*, pp. 45–48. Yokohama, Japan: ISCA. URL `http://www.isca-speech.org/archive/icslp_1994/i94_0045.html`.

Tillmann, Hans G., Andreas Zierdt & Phil Hoole (1996) "Towards a three-dimensional articulographic system". *Journal of the Acoustical Society of America*, **100**(4):2662. doi:10.1121/1.417464.

Titze, Ingo R. (1984) "Parameterization of the glottal area, glottal flow, and vocal fold contact area". *Journal of the Acoustical Society of America*, **75**(2):570–580. doi:10.1121/1.390530.

Tokuda, Keiichi, Heiga Zen & Alan W. Black (2002) "An HMM-based speech synthesis system applied to English". In: *IEEE Workshop on Speech Synthesis*, pp. 227– 230. Santa Monica, CA, USA: IEEE. doi:10.1109/WSS.2002.1224415.

Tom, Kenneth, Ingo R. Titze, Eric A. Hoffman & Brad H. Story (2001) "Three-dimensional vocal tract imaging and formant structure: Varying vocal register, pitch, and loudness". *Journal of the Acoustical Society of America*, **109**(2):742–747. doi:10.1121/1.1332380.

Tuller, Betty, Shuyong Shao & J. A. Scott Kelso (1990) "An evaluation of an alternating magnetic field device for monitoring tongue movements". *Journal of the Acoustical Society of America*, **88**(2):674–679. doi:10.1121/1.399771.

Umeda, Noriko & Ryunen Teranishi (1965) "Phonemic feature and vocal feature: Synthesis of speech sounds, using an acoustic model of vocal tract". *Journal of the Acoustical Society of Japan*, **22**(4):195–203.

Viterbi, Andrew J. (1967) "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". *IEEE Transactions on Information Theory*, **13**(2):260–269.

Wakumoto, Masahiko, Shinobu Masaki, Jianwu Dang, Kiyoshi Honda, Yasuhiro Shimada, Ichiro Fujimoto & Yuji Nakamura (1997) "Visualization of dental crown shape in an MRI-based speech production study". In: *13th International Conference on Oral and Maxillofacial Surgery*, pp. 189–190. doi:10.1016/S0901-5027(97)81405-1.

Webster, Arthur G. (1919) "Acoustical impedance and the theory of horns and of the phonograph". *Proceedings of the National Academy of Sciences*, **5**(7):275–282. URL `http://www.pnas.org/content/5/7/275.full.pdf`.

Weibel, Erich S. (1955) "On Webster's horn equation". *Journal of the Acoustical Society of America*, **27**(4):726–727. doi:10.1121/1.1908007.

Weibel, Ewald R. (1963) *Morphometry of the Human Lung*. Berlin, Germany: Springer.

Weismer, Gary & Kate Bunton (1999) "Influences of pellet markers on speech production behavior: Acoustical and perceptual measures". *Journal of the Acoustical Society of America*, **105**(5):2882–2894. doi:10.1121/1.426902.

Wells, John C. (1997) "SAMPA computer readable phonetic alphabet". In: Dafydd Gibbon, Roger K. Moore & Richard Winski (eds.), *Handbook of Standards and Resources for Spoken Language Systems*, volume IV, section B. New York: Mouton de Gruyter. URL `http://www.phon.ucl.ac.uk/home/sampa/`.

Westbury, John R. (1991) "The significance and measurement of head position during speech production experiments using the x-ray microbeam system". *Journal of the Acoustical Society of America*, **89**(4):1782–1791. doi:10.1121/1.401012.

Westbury, John R. (1994) *X-Ray Microbeam Speech Production Database User's Handbook*. University of Wisconsin, Madison, WI, USA. URL `https://files.nyu.edu/ag63/public/fhs_atelier/ubdbman.pdf`. Version 1.0.

Whalen, Douglas H. (1990) "Intrinsic velar height in supine vowels". *Journal of the Acoustical Society of America*, **88**(S1):S54. doi:10.1121/1.2029052.

Whalen, Douglas H. (2003) “Articulatory synthesis: Advances and prospects”. In: *15th International Congress of Phonetic Sciences*, pp. 175–178. Barcelona, Spain.

Whalen, Douglas H., Khalil Iskarous, Mark K. Tiede, David J. Ostry, Heike Lehnert-LeHouillier, Eric Vatikiotis-Bateson & Donald S. Hailey (2005) “The Haskins Optically Corrected Ultrasound System (HOCUS)”. *Journal of Speech, Language, and Hearing Research*, **48**(3):543–553. doi:10.1044/1092-4388(2005/037).

Wilhelms-Tricarico, Reiner (1995) “Physiological modeling of speech production: Methods for modeling soft-tissue articulators”. *Journal of the Acoustical Society of America*, **97**(5):3085–3098. doi:10.1121/1.411871.

Wilhelms-Tricarico, Reiner (2000) “Development of a tongue and mouth floor model for normalization and biomechanical modelling”. In: Phil Hoole (ed.), *5th Seminar on Speech Production*, pp. 141–148. Kloster Seeon, Germany. URL `http://www.phonetik.uni-muenchen.de/institut/veranstaltungen/SPS5/abstracts/73_abs.html`.

Wilson, Ian L. (2006) *Articulatory Settings of French and English Monolingual and Bilingual Speakers*. Ph.D. thesis, University of British Columbia. URL `http://www.u-aizu.ac.jp/~wilson/Wilson2006PhD.pdf`.

Wrench, Alan A. (2000) “A multi-channel/multi-speaker articulatory database for continuous speech recognition research”. *PHONUS*, **5**:1–14. URL `http://www.coli.uni-saarland.de/Phonetics/Research/PHONUS_research_reports/Phonus5/Wrench_PHONUS5.pdf`.

Wrench, Alan A. & William J. Hardcastle (2000) “A multichannel articulatory speech database and its application for automatic speech recognition”. In: Phil Hoole (ed.), *5th Seminar on Speech Production*, pp. 305–308. Kloster Seeon, Germany. URL `http://www.phonetik.uni-muenchen.de/institut/veranstaltungen/SPS5/abstracts/03_abs.html`.

Wrench, Alan A., Alan D. McIntosh & William J. Hardcastle (1996) “Optopalatograph (OPG): A new apparatus for speech production analysis”. In: *4th International Conference on Spoken Language Processing*, pp. 1589–1592. Philadelphia, PA, USA: ISCA. URL `http://www.isca-speech.org/archive/icslp_1996/i96_1589.html`.

Wrench, Alan A., Alan D. McIntosh & William J. Hardcastle (1997) “Optopalatograph: Development of a device for measuring tongue movement in 3D”. In: G. Kokkinakis, N. Fakotakis & E. Dermatas (eds.), *Eurospeech*, pp. 1055–1058. Rhodes, Greece: ESCA. URL `http://www.isca-speech.org/archive/eurospeech_1997/e97_1055.html`.

Wrench, Alan A., Alan D. McIntosh, Colin Watson & William J. Hardcastle (1998) “Optopalatograph: Real-time feedback of tongue movement in 3D”. In: *5th International Conference on Spoken Language Processing*. Sydney, Australia: ISCA. URL `http://www.isca-speech.org/archive/icslp_1998/i98_1117.html`.

Wright, Graham A. (1997) “Magnetic resonance imaging”. *IEEE Signal Processing Magazine*, **14**(1):56–66. doi:10.1109/79.560324.

Xu, Yi & Fang Liu (2006) “Tonal alignment, syllable structure and coarticulation: Toward an integrated model”. *Italian Journal of Linguistics*, **18**(1):125–159. URL `http://alphalinguistica.sns.it/RdL/18.1/Xu&Liu.pdf`.

Xu, Yi & Q. Emily Wang (2001) “Pitch targets and their realization: Evidence from Mandarin Chinese”. *Speech Communication*, **33**(4):319–337. doi:10.1016/S0167-6393(00)00063-7.

Young, Robert W. (1939) “Terminology for logarithmic frequency units”. *Journal of the Acoustical Society of America*, **11**(1):134–139. doi:10.1121/1.1916017.

Young, Steve, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying (Andrew) Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev & Phil Woodland (2009) *The HTK Book*, HTK version 3.4 edition. URL `http://htk.eng.cam.ac.uk/docs/docs.shtml`.

Yunusova, Yana, Jordan R. Green & Antje Mefferd (2009) "Accuracy assessment for AG500, Electromagnetic Articulograph". *Journal of Speech, Language, and Hearing Research*, **52**(2):547–555. doi:10.1044/1092-4388(2008/07-0218).

Zacks, Jeff & Timothy R. Thomas (1994) "A new neural network for articulatory speech recognition and its application to vowel identification". *Computer Speech and Language*, **8**(3):189–209. doi:10.1006/csla.1994.1009.

Zen, Heiga, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W. Black & Keiichi Tokuda (2007a) "The HMM-based speech synthesis system (HTS) version 2.0". In: Petra Wagner, Julia Abresch, Stefan Breuer & Wolfgang Hess (eds.), *Sixth ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)*, pp. 294–299. Bonn, Germany: ISCA. URL `http://www.isca-speech.org/archive/ssw6/ssw6_294.html`.

Zen, Heiga, Keiichi Tokuda & Alan W. Black (2009) "Statistical parametric speech synthesis". *Speech Communication*, **51**(11):1039–1064. doi:10.1016/j.specom.2009.04.004.

Zen, Heiga, Keiichi Tokuda & Tadashi Kitamura (2007b) "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences". *Computer Speech and Language*, **21**(1):153–173. doi:10.1016/j.csl.2006.01.002.

Zhang, Le (2009) *Modelling Speech Dynamics with Trajectory-HMMs*. Ph.D. thesis, University of Edinburgh. URL `http://hdl.handle.net/1842/3213`.

Zhang, Le & Steve Renals (2008) "Acoustic-articulatory modeling with the trajectory HMM". *IEEE Signal Processing Letters*, **15**:245–248. doi:10.1109/LSP.2008.917004.

Zierdt, Andreas (1993) "Problems of electromagnetic position transduction for a three-dimensional articulographic measurement system". *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München*, **31**:137–142.

Zierdt, Andreas (2007) "EMA and the crux of calibration". In: Jürgen Trouvain & William J. Barry (eds.), *XVIth International Congress of Phonetic Sciences*, pp. 593–596. Saarbrücken, Germany. URL `http://www.icphs2007.de/conference/Papers/1511/1511.pdf`.

Zierdt, Andreas, Phil Hoole, Masaaki Honda, Tokihiko Kaburagi & Hans G. Tillmann (2000) "Extracting tongues from moving heads". In: Phil Hoole (ed.), *5th Seminar on Speech Production*, pp. 313–316. Kloster Seeon, Germany. URL `http://www.phonetik.uni-muenchen.de/institut/veranstaltungen/SPS5/abstracts/05_abs.html`.

Zierdt, Andreas, Phil Hoole & Hans G. Tillmann (1999) "Development of a system for three-dimensional fleshpoint measurement of speech movements". In: John J. Ohala, Yoko Hasegawa, Manjari Ohala, Daniel Granville & Ashlee C. Bailey (eds.), *14th International Congress of Phonetic Sciences*, volume 1, pp. 73–75. San Francisco, CA, USA.

Zitová, Barbara & Jan Flusser (2003) "Image registration methods: a survey". *Image and Vision Computing*, **21**(11):977–1000. doi:10.1016/S0262-8856(03)00137-9.