Control Concepts for Articulatory Speech Synthesis

Peter Birkholz[1], *Ingmar Steiner*[2], *Stefan Breuer*[3]

[1]Institute for Computer Science, University of Rostock, Germany
[2]Department of Computational Linguistics and Phonetics, Saarland University, Germany
[3]Institute of Communication Sciences (IfK), University of Bonn, Germany
piet@informatik.uni-rostock.de, steiner@coli.uni-saarland.de, breuer@ifk.uni-bonn.de

Abstract

We present two concepts for the generation of gestural scores to control an articulatory speech synthesizer. Gestural scores are the common input to the synthesizer and constitute an organized pattern of articulatory gestures. The first concept generates the gestures for an utterance using the phonetic transcriptions, phone durations, and intonation commands predicted by the Bonn Open Synthesis System (BOSS) from an arbitrary input text. This concept extends the synthesizer to a text-to-speech synthesis system. The idea of the second concept is to use timing information extracted from Electromagnetic Articulography signals to generate the articulatory gestures. Therefore, it is a concept for the re-synthesis of natural utterances. Finally, application prospects for the presented synthesizer are discussed.

1. Introduction

Articulatory speech synthesis is the most rigorous way of synthesizing speech, as it constitutes a simulation of the mechanisms underlying real speech production. Compared to other approaches in speech synthesis, it has the potential to synthesize speech with any voice and in any language with the most natural quality. Further advantages of articulatory speech synthesis are discussed by Shadle and Damper [17]. However, despite its potential, it is still a difficult task to actually achieve an average speech quality for one specific voice and language with an articulatory speech synthesizer. The problem are the high demands on the models for the various aspects of speech production. One of these aspects is the generation of speech movements, i.e., the control of the model articulators. In this paper, we present (i) a novel control model based on articulatory gestures and (ii) propose two concepts for the high-level prediction of the gestural parameters. The control model was implemented as part of an articulatory speech synthesizer based on a 3D model of the vocal tract and a comprehensive aeroacoustic simulation method [3, 4, 5]. The goal of the proposed high-level concepts is to specify the articulatory gestures in the form of a *gestural score* needed for the generation of the speech movements from different sources of input.

The idea of the first concept is to generate speech from text using the open source software platform BOSS (Bonn Open Synthesis System) [8]. BOSS was originally developed as a unit-selection speech synthesis system comprising modules for phonetic transcription, phone duration prediction, intonation generation, and the actual unit-selection step. In this study, we present a way to transform the output of the modules for phonetic transcription and phone duration prediction into the gestural score for the articulatory synthesizer.

The idea of the second concept is to use timing information extracted from Electromagnetic Articulography (EMA) signals to create the artificial gestural scores. Since EMA signals reflect the articulatory movements of real speakers, this is a concept for the *re*synthesis of speech. In other words, the second concept is an attempt to copy the speech of a speaker recorded by an EMA device, primarily with respect to gestural timing.

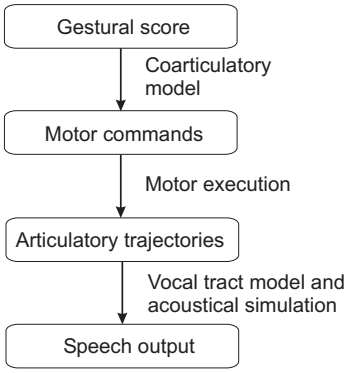

Figure 1: Flow diagram of the articulatory synthesizer.

The speech generation chain of the articulatory synthesizer is depicted in Figure 1. As mentioned above, the input to the synthesizer is a gestural score. It can be regarded as a representation of the intended utterance in terms of gestures for the glottal and the supraglottal articulators. As in the framework of articulatory phonology by Browman and Goldstein [10] and the gestural control model by Kröger [15], we regard gestures as characterizations of discrete articulatory events that unfold during speech production in terms of goal-oriented articulatory movements. However, the actual characterization of these events differs from the aforementioned approaches and will be discussed later. After a gestural score has been specified, it is transformed into sequences of *motor commands* – one sequence for each parameter of the glottis and the vocal tract model. The execution of the motor commands, i.e. the generation of the actual articulatory trajectories, is simulated by means of third order linear systems. These systems were designed to produce smooth movements similar to those observed in EMA signals. The movements are directly generated in terms of time-varying parameter values for the vocal tract and the glottis. They determine the shape of the vocal tract and the state of the glottis which are the input to the aeroacoustic simulation generating the speech output.

This article is organized as follows. In Section 2, the components in Figure 1 will be described in more detail, in particular the models for the vocal tract and the glottis, the specification

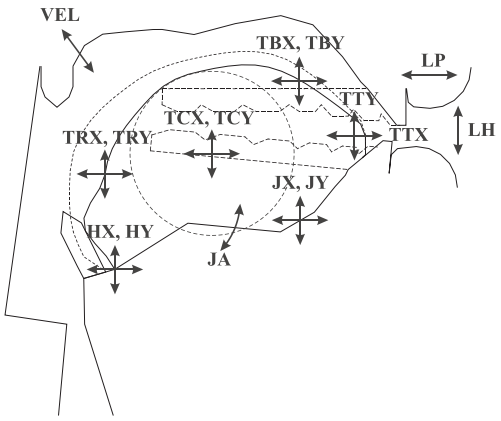

Figure 2: Schematic overview of the parameters of the vocal tract model and the articulatory structures that they control.

of gestural scores, and their transformation into speech movements. Section 3 presents the concepts for the high level control of the synthesizer, i.e. the generation of gestural scores from text using BOSS on one hand, and from timing information extracted from EMA tracks on the other hand. In Section 3.3 we discuss application prospects for the presented synthesizer. Conclusions are drawn in Section 4.

2. Articulatory speech synthesizer

2.1. Models for the vocal tract and the glottis

Vocal tract model. The vocal tract model of the synthesizer is a three-dimensional wire frame representation of the surfaces of the articulators and the vocal tract walls of a male speaker [3, 4]. The shape and position of all movable stuctures is a function of 23 adjustable parameters. Figure 2 shows the midsagittal section of the 3D vocal tract model along with the most important parameters. The arrows indicate how the corresponding parameters influence the articulation. Most of these parameters come in pairs and define the position of certain structures directly in Cartesian coordinates in a fixed frame of reference. For example, the point defined by the parameters (TCX, TCY) specifies the position of the tongue body (represented by a circle), (TTX, TTY) defines the position of the tongue tip, and (JX, JY) the position of the jaw. Therefore, the temporal change of these parameters should be comparable to the movement of pellets glued to the tongue or mandible in real articulations, as measured by EMA devices. The parameter values that best represent the ideal articulatory target shapes for German vowels and consonants have recently been determined by means of magnetic resonance images (MRI) [4]. The articulatory targets for consonants represent the vocal tract shape at the time of the maximum constriction, uttered without a specific phonetic context. However, it is well known that the actual articulatory realization of consonants strongly depends on the phonetic context. Only a few articulators (or parts of them) are really involved in the formation of the consonantal constriction while others are subject to coarticulation with adjacent phones. For example, the [g] in [igi] is realized differently from the [g] in [ugu]. In both cases, the tongue body is raised to make a palatal closure, but it is clearly more anterior in the context of the front vowel [i] than in the context of the back vowel [u]. In our synthesizer, such coarticulatory differences are handled by means

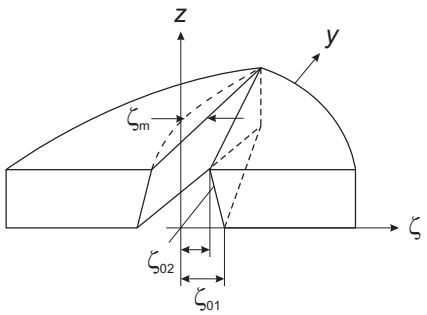

Figure 3: Model for the glottis based on Titze [19].

of a dominance model. This model specifies a dominance value or "degree of importance" for each vocal tract parameter of each consonant. A high dominance means that a certain parameter is important for the formation of the consonantal constriction, and a low dominance value means that it is not important and therefore subject to coarticulation. In the above example for the consonant [g], the parameter TCY for the height of the tongue body has a high dominance, but TCX for its horizontal position a low dominance. The actual target parameter value $x_{c|v}[i]$ of a parameter i for a consonant c in the context of a vowel v at the moment of maximum constriction/closure is expressed as

$$x_{c|v}[i] = x_v[i] + w_c[i] \cdot (x_c[i] - x_v[i]), \qquad (1)$$

where $w_c[i]$ is the weight (dominance) for parameter i, and $x_c[i]$ and $x_v[i]$ are the parameter values of the ideal targets for the consonant and vowel. The optimal dominance values for all parameters of all consonants have been determined in a previous study [4]. It was also shown that this simple dominance model is capable of reproducing the major coarticulatory differences in the realization of consonants.

Vocal fold model. For the voiced excitation of the synthesizer, we implemented a parametric model of the glottal geometry based on the proposal by Titze [19]. A schematic representation of the model is shown in Figure 3. The vocal fold parameters are the degree of abduction at the posterior end of the folds at the lower and upper edge (ζ_{01} and ζ_{02}), the fundamental frequency F_0, the phase difference between the upper and lower edge, and the subglottal pressure. Based on these parameters, the model generates the time-varying cross-sectional areas at the glottal inlet and outlet opening. We extended Titze's original model to account for a smooth diminishment of the oscillation amplitude with increasing abduction [2] and for a parametrization of glottal leakage similar to [11].

Combination of the models. The geometric models of the vocal folds and the vocal tract are transformed into a combined area function. This area function, supplemented with the area functions of the subglottal system and the nasal cavity, serve as input to a time domain simulation of the flow and acoustics in the vocal system, producing the actual speech output [2, 1].

2.2. From gestural scores to speech movements

The intermediate representation layer for an utterance in the synthesizer is a gestural score. It defines an utterance in terms of an organized pattern of articulatory gestures. The specification and execution of these gestures differs, however, from previously proposed gestural control concepts (e.g., Browman and Goldstein [10], and Kröger [15]).

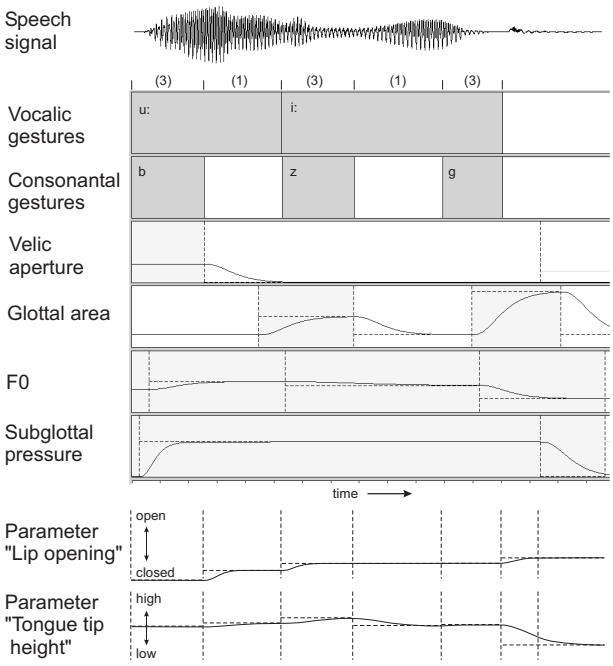

Figure 4: Gestural score for the utterance [mu:zi:k] with the generated speech waveform (top) and the resulting targets and their execution for two of the vocal tract parameters (bottom).

Figure 4 shows a gestural score for the utterance [mu:zi:k]. This example will illustrate the following explanations of the model. We differentiate between six types of gestures. Each row in Figure 4 contains the gestures of one type. The gestures in the first two rows are *vocalic* and *consonantal* gestures. Together with the *velic* gestures (third row) they determine the parameters of the vocal tract model, i.e., the supralaryngeal articulation. The gestures in the remaining rows control the glottal rest area (degree of abduction), the F_0, and the subglottal pressure. They determine the parameters of the model of the vocal folds, i.e., the laryngeal articulation. Each gesture has a certain temporal activation interval (defined by the vertical boundary lines) and is associated with a target for one or more vocal tract parameters or laryngeal parameters.

Let us first turn towards the supraglottal articulation. In Figure 4, the first vocalic gesture is associated with the target configuration for the vowel [u:], and the second one is associated with the vowel [i:]. The fixed target configurations were determined a priori for each vowel, as discussed in Section 2.1. The consonantal gestures in Figure 4 are associated with the consonants [b], [z] and [g]. We must point out that the target configuration for consonants with the same place of articulation are represented by only one configuration for each group. The groups {[b],[p],[m]}, {[d],[t],[n]}, and {[g],[k],[ŋ]} are represented by the target configurations for [b], [d], and [g], respectively. The voiceless plosives and the nasals are assumed to differ from the voiced plosives only in the state of the velum and the glottal area, which can be controlled individually in the gestural scores. Also the supraglottal articulation of voiced and voiceless fricatives with the same place of articulation is represented by only the voiced cognates. In Figure 4, the intervals for [b], [z], and [g] overlap with the intervals for the vowels [u:] and [i:]. This means that these consonants are coarticulated with the corresponding vowels. All vocalic and consonantal gestures are associated with an articulatory effort parameter. This effort translates into the transition speed towards the associated targets during the execution of the gestures.

But how are the vocalic and consonantal gestures executed, i.e., how are they transformed into the time-varying vocal tract parameter functions? First, a sequence of motor commands is generated for each parameter. In the context of this control model, a motor command is defined as target value for a vocal tract parameter within a defined time interval. Below the gestural score in Figure 4, these sequences of target values are shown for the lip opening LH and the vertical tongue tip position TTY by means of horizontal dashed lines. An individual motor command is generated for each combination of a vocalic and a consonantal gesture. The motor command boundaries are indicated by vertical dotted lines. The actual target value associated with a motor command for a vocal tract parameter depends on the underlying gestures. We differentiate between three cases: (1) The target value is that for a vowel. (2) The target is that for an isolated consonant. (3) The target is that for a consonant coarticulated with a vowel calculated according to Equation (1).

In Figure 4, we have only the cases (1) and (3), which are marked accordingly on top of the gestural score. In this way, a sequence of motor commands is calculated for each vocal tract parameter. The only exception is the parameter for the velic aperture, which is controlled separately by the velic gestures. These gestures directly correspond to the motor commands for the parameter *VEL* (cf. Figure 2).

The execution of the motor commands is modeled by means of a critically damped dynamical third order linear system with the transfer function

$$H(s) = 1/(1+\tau s)^3, \tag{2}$$

where s is the complex frequency and τ is a time constant to be described later. The input to the system is the sequence of targets for a certain parameter. The system's output is the time dependent function value for that parameter. For the parameters LH and TTY, the resulting functions are drawn as solid lines below the gestural score in Figure 4. Note that the systems behave in such a way that the vocal tract parameters successively approximate the target values associated with the motor commands. In other words, they implement the original articulatory gestures as goal-oriented movements. The parameter τ in Equation (2) is a measure for the speed of target approximation. A small value for τ corresponds to a fast movement, and vice versa. The τ parameters for the individual motor commands are derived from the articulatory effort parameters for the vocalic and consonantal gestures. Therefore, τ can vary for adjacent motor commands.

As stated before, the parameter for the velic aperture of the vocal tract model is controlled independently from the other supraglottal parameters by means of velic gestures. The velic gestures directly define the target positions for motor commands, which are executed in the same way as described above. Similarily, the gestural targets for the glottal rest area, F_0, and the subglottal pressure defined in the gestural score are directly mapped on motor commands for the corresponding parameters of the model of the vocal folds.

A more detailed description of the gestural control model and the underlying ideas can be found in [6].

3. High level control concepts

3.1. Bonn Open Synthesis System (BOSS)

The Bonn Open Synthesis System (BOSS) [8] is a developer framework for the design of unit selection speech synthesis applications in C++. Its main goal is to relieve researchers in the field of speech synthesis of the need to implement their own systems from scratch. It is available under the GPL open source license from the IfK website [9]. BOSS is designed to be used as a client/server application over a network. Most of the symbolic preprocessing, the selection of units and their concatenation and manipulation are performed by the server while the client software is responsible for text normalization and tokenization and for encoding this information into the XML vocabulary understood by the server. By this choice of design, BOSS can be flexibly employed for either CTS or TTS, depending on what type of client is used. The core class of the BOSS server, also called the module scheduler, processes the client-generated information sentence by sentence. Required modules are loaded dynamically upon initialization of the scheduler class. The names and calling order of module libraries are defined in a configuration file, so that a developer who wishes to adapt BOSS to a new language or application is not required to change the source code of the server software. For the application described in this paper, we used the German transcription module, the CART [7] duration prediction module and the Fujisaki-based [13] intonation module delivered with the BOSS distribution. In summary, these modules provide the phonetic transcription (structured into syllables and phones) of a German input text with a duration specification for each phone, and Fujisaki-based intonation commands for each syllable. In the following, we will discuss a proposal how to translate this information into a gestural score for the articulatory synthesizer.

The major problem in this context is to translate the phone durations given by BOSS into activation intervals of the gestures, especially of the vocalic, consonantal, velic and glottal gestures. BOSS predicts the phone durations corresponding to the conventional way of phone segmentation, i.e. the beginning and the end of phones is associated with striking landmarks in the auditory signal or the spectrogram. In this sense, the consonant [t], for example, starts where the acoustic signal energy suddenly drops due to the apico-alveolar closure and ends after the aspiration phase following the release of the closure. In general, these acoustical landmarks can be assigned to special *articulatory* events that are also reflected in the gestural scores. Furthermore, each class of phones exhibits typical patterns of temporal coordination of the involved articulatory gestures, such as the coordination between the constriction forming gesture (consonantal gesture) and the glottal abduction gesture for voiceless plosives. These patterns are sometimes called "phasing rules" [10, 15]. The phasing rules, together with the associations between acoustical landmarks and time instants in the gestural scores allow to calculate phone durations from gestural scores, and vice versa, to create gestural constellations for phones of a given class and with a given duration.

Figure 5 illustrates the phasing rules and the correspondence between gestural constellations and the resulting speech waveform for plosives, fricatives, and nasals. The consonants in these examples were embedded into the context [iːCaː]. First of all, the consonantal gestures were always aligned to be coarticulated with the vowel of the second syllable, according to Xu [20]. The time intervals of consonantal closure (or critical constriction in the case of [s]) are marked by vertical dashed lines. Typically, these intervals start 30–60 ms after the onset of

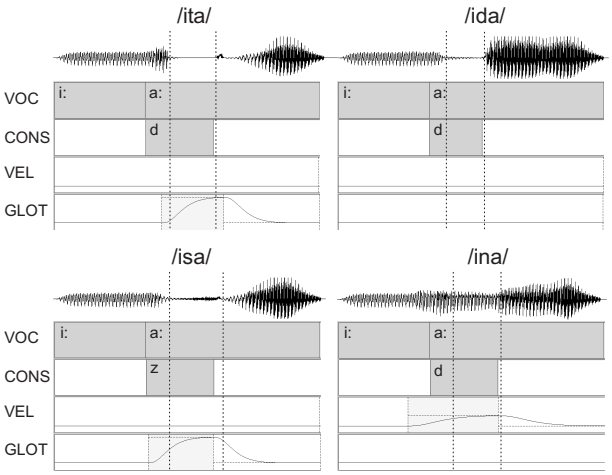

Figure 5: Gestural constellations for voiced and voiceless plosives, voiceless fricatives, and nasals in the context [iːCaː]. *VOC*=vocalic gestures, *CONS*=consonantal gestures, *VEL*=velic gestures, and *GLOT*=glottal gestures. The vertical dotted lines indicate the beginnings and ends of the consonantal closure/constriction intervals. The gestures for subglottal pressure and F_0 are not shown.

the consonantal gestures. This is the time the constriction forming articulators need to reach their target positions. The ends of the constriction/closure intervals are typically very shortly after the offset of the consonantal gestures, where the articulators start moving towards their targets for the following vowels. For [iːdaː], [iːsaː], and [iːnaː] (and the corresponding classes of consonants), the constriction intervals directly correspond to the phone durations according to the BOSS predictions. However, for voiceless aspirated plosives as the [t] in [iːtaː], BOSS does not predict the constriction duration, but the duration from the onset of the closure to the end of the burst and aspiration phase. In the gestural score, this end point is roughly where the glottal aperture is reduced to 50% of its maximal area.

The velic and glottal gestures in Figure 5 illustrate appropriate phasing rules for the different classes of consonants. Voiced plosives need neither a velic nor a glottal gesture. For voiceless aspirated plosives, glottal abduction should approximately start at the beginning of the closure interval [18]. To get a fair amount of aspiration, glottal adduction should start approximately by the end of the oral closure interval. For voiceless fricatives, the glottal gesture should start and end roughly simultaneously with the consonantal gesture to produce good synthetic results. Nasals need a lowering of the velum by means of a velic gesture. Preliminary synthesis results suggest that the onset and offset of the velic aperture is not very critical. For [iːnaː] in Figure 5, we made the velic gesture start shortly before the corresponding consonantal gesture and end simultaneously with it. Similar rules can easily be established for voiced fricatives, laterals, glottal consonants, and the generation of consonant clusters. The duration of vowels and diphthongs is determined by the borders of the adjacent consonants.

This section was mainly meant to illustrate basic ideas for the rule-based creation of gestural scores from a given phonetic transcription and phone durations. A quantitative implementation of these rules is in progress, and first speech examples will be presented at the conference. To improve the naturalness of the synthetic utterances, a prototypical transformation from

BOSS intonation commands to gestures for F_0 control will also be implemented.

3.2. Speech resynthesis based on EMA data

The duration of predicted parameters (both segmental and suprasegmental) using conventional TTS "preprocessing" is based on observations of acoustic landmarks in speech. In articulatory synthesis, we must predict the movements of the articulators which cause these landmarks, after a certain delay. To analyze and directly implement this delay in an articulatory synthesizer, we must first study the actual movements of the articulators during speech production. One possibility of doing this is through Electromagnetic Articulography (EMA).

For the analysis of articulatory parameters during actual speech production, we were given access to two EMA corpora ([12], [14]). The first of these contains recordings of a female German speaker uttering /CVCVCVCV/ sequences, with all combinations of a set of 9 consonants and 15 vowels of German, in two conditions (EMA sensors: jaw, lower and upper lip, tongue tip, blade and dorsum). The second corpus consists of recordings of 7 German speakers (1 female, 6 male) uttering /CVC/ syllables embedded in a carrier phrase, with all combinations of 3 consonants and 14 vowels, in two conditions, as well as reading a list of 108 German sentences (EMA sensors: jaw, lower lip, tongue tip, blade, dorsum, and back).

The aim of an intermediate study is to resynthesize the utterances of the recorded speakers, comparing the trajectories of the articulatory parameters. Since the virtual vocal tract is modeled upon that of one speaker and the natural data obtained from another, a direct comparison of raw articulator movements does not make sense. Rather, the *timing* of the simulated EMA trajectories produced by the synthesizer is modeled on the temporal structure of articulatory gestures performed by the original speaker, and thereby indirectly on his speech rhythm.

While it could in theory be possible to directly transfer the EMA trajectories to the virtual articulators (normalized for differences in anatomy) and produce similar, if not identical utterances, such a low-level approach is not the goal of an articulatory synthesizer with high-level control mechanisms. Rather, the purpose of this resynthesis is twofold: to test the parametric fidelity of the synthesizer; and to analyze the observed delay from gestural onsets to the acoustic landmarks traditionally regarded as the beginning of the corresponding segment in the synthesis output.

For a preliminary comparison of natural and synthetic articulatory trajectories, the word *Methanol* [meta'no:l] was resynthesized, using EMA parameters of one of the male speakers. The resynthesis process involved two steps: first, identifying intervals in which the relevant EMA trajectories approached the respective target values; and second, providing this timing information to the synthesizer in the form of a gestural score. Additionally, the F_0 contour was extracted from the acoustic signal and included in the gestural score in a smoothed form. The resulting synthesis output is presented alongside the original recording in Figure 6. The relative height and arrangement over time of the peaks and valleys in these curves displays an encouraging similarity. One should keep in mind that our aim was not to produce an exact copy of the trajectories, but to combine the gestural targets of the virtual vocal tract with timing derived from EMA data, creating the desired perceptual impression.

In addition to gestural timing, it is conceivable to extract measures of articulatory effort from the EMA trajectories and include these in the gestural score, since the synthesizer allows fine control over this parameter.

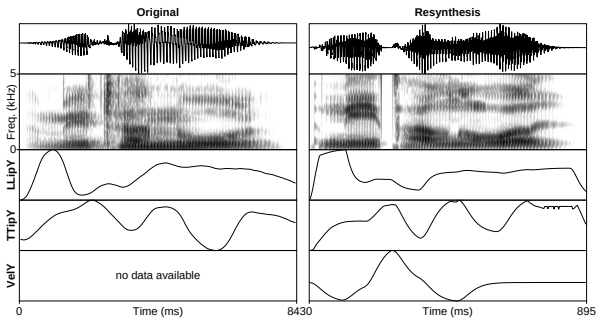

Figure 6: Gestural constellations for original (left) and resynthesized (right) version of the word *Methanol* [meta'no:l]. Below the spectrograms are the normalized trajectories of the parameters corresponding to height of the lower lip (*LLipY*), tongue tip (*TTipY*), and velum (*VelY*).

3.3. Application prospects

Combining high-level articulatory control with natural-sounding synthesis breaks out of the widely-accepted compromise that naturalness and parametric flexibility are inversely correlated in speech synthesis and cannot both be satisfied at once. This opens up many new opportunities for a variety of applications for the presented system. A few immediate prospects are outlined below, but listing all the possibilities would be well outside the scope of this paper.

Considerable naturalness can already be achieved with unit-selection and similar synthesis approaches (especially in a limited domain), but at the cost of prosodic control. In fact, many unit-selection synthesis platforms currently choose to abandon explicit prosody modeling altogether and therefore lack control over parameters such as F_0. Those that do allow F_0 target specification (either through the unit selection algorithm itself or subsequent signal manipulation) may introduce significant artifacts in an unpredictable way, depending on whether or not suitable units can be found in the unit-selection corpus.

Expressive speech synthesis. One possible area of application for an articulatory synthesizer with full flexibility and high naturalness is of course expressive (a.k.a. "emotional") speech synthesis (cf. [16] for a detailed survey). This expanding field of speech synthesis relies heavily on flexible control over prosodic and/or paralinguistic parameters, mainly F_0, but also voice quality, among others. For this reason, expressive speech synthesis has largely been unable to make use of the progress in unit-selection approaches, being forced to rely instead on less natural-sounding, but more flexible diphone concatenation or formant synthesis.

Certain other relevant parameters, such as voice register, articulatory effort, lip spreading, etc. can only be controlled with elaborate effort, if at all, using the synthesis methods mentioned above. The system presented here, however, is ideally suited to such tasks and can be extended to provide high-level control over precisely such parameters.

Multilingual speech synthesis. With a certain amount of adjustment, the presented system could easily be adapted to new languages, the phoneset being, after all, a set of gestural "macros". The resulting synthesis output would be in the

same voice as long as the vocal tract characteristics remain unchanged. This would allow true multilingual synthesis without depending on necessarily distinct native speaker recordings.

Voice morphing. On the other hand, vocal tract characteristics could be deliberately modified to create a different voice. This allows control over gender, age, timbre, as well as a multitude of other extralinguistic parameters. Since all synthesis output is rendered to an acoustic signal only once, no degradation of quality occurs, as is inevitable with voice morphing techniques and similar signal processing. The presented system provides full control over numerous physiological properties of the synthesis voice, permitting finely detailed voice design for e.g. artificial agents in dialog systems.

Prosody research. Phonetic research in prosody would benefit greatly from an instrument allowing at leisure the synthesis of natural-sounding, prosodically fully-flexible speech. This would provide the means to e.g. implement and test autosegmental phonological models, generate high-quality stimuli for experiments, and much more. Currently, many synthetic stimuli created for prosody experiments suffer from limited naturalness, depending on the synthesis technique used to produce them, for the same reasons as outlined above under *expressive speech synthesis*. Whereas in a (commercial) TTS system, intelligibility takes precedence over naturalness, in prosodic experiments, a lack of naturalness may distract test subjects and affect their responses, skewing the results of the study.

Nevertheless, it must be acknowledged that the computational complexity of articulatory synthesis as implemented in the presented system currently prevents synthesis in realtime on an average desktop PC. It is our belief, however, that realtime synthesis will become realistic in the very near future, owing to advances in processing power as well as code optimization.

4. Conclusions

We have presented two concepts for the high-level control of an articulatory speech synthesizer. First, we outlined rules for the transformation of phonetic transcriptions and phone durations predicted by the Bonn Open Synthesis System (BOSS) into gestural scores, extending the synthesizer to a text-to-speech system. Second, we demonstrated the generation of gestural scores based on EMA signals. Our preliminary results suggest that both ways lead to well intelligible synthetic speech.

For future research, it is conceivable to train BOSS to directly predict gestural parameters, e.g. gestural durations, instead of phone durations in the conventional sense, as it currently does. This would considerably simplify the rules for the generation of gestural scores, but would require a corresponding segmentation of the original EMA data.

5. Acknowledgments

This research was partially funded by the German Research Foundation (DFG) with the grant JA 1476/1-1. We would like to thank Sascha Fagel and Phil Hoole for making their EMA data available to us.

6. References

[1] P. Birkholz and D. Jackèl, "Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system," in *Interspeech 2004-ICSLP*, Jeju, Korea, pp. 1125–1128, 2004.

[2] P. Birkholz, "3D-Artikulatorische Sprachsynthese," Ph.D. dissertation, University of Rostock, 2005.

[3] P. Birkholz, D. Jackèl, and B. J. Kröger, "Construction and control of a three-dimensional vocal tract model," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, Toulouse, France, pp. 873–876, 2006.

[4] P. Birkholz and B. J. Kröger, "Vocal tract model adaptation using magnetic resonance imaging," in *7th International Seminar on Speech Production (ISSP'06)*, Ubatuba, Brazil, pp. 493–500, 2006.

[5] P. Birkholz, D. Jackèl, and B. J. Kröger, "Simulation of losses due to turbulence in the time-varying vocal system," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1218–1226, 2007.

[6] P. Birkholz, "Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets," *submitted to Interspeech 2007 - Eurospeech*, Antwerp, Belgium, 2007.

[7] L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees*, Wadsworth International, Belmont, CA, 1984.

[8] S. Breuer, P. Wagner, J. Abresch, J. Bröggelwirth, H. Rohde and K. Stöber *Bonn Open Synthesis System (BOSS) 3 Documentation and User Manual,* `http://www.ikp.uni-bonn.de/boss/BOSS_Documentation.pdf`, 2005.

[9] `http://www.ikp.uni-bonn.de/boss`

[10] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, pp. 155–180, 1992.

[11] B. Cranen and J. Schroeter, "Modeling a leaky glottis," *Journal of Phonetics*, vol. 23, pp. 165–177, 1995.

[12] S. Fagel, *Audiovisuelle Sprachsynthese: Systementwicklung und -bewertung*. Logos Verlag, Berlin, 2004

[13] H. Mixdorff, "Intonation Patterns of German - Model-based Quantitative Analysis and Synthesis of F_0 contours," Ph.D. dissertation, TU Dresden, 1998

[14] P. Hoole and C. Mooshammer, "Articulatory analysis of the German vowel system", In: Auer, P., Gilles, P. & Spiekermann, H. (eds.), *Silbenschnitt und Tonakzente*. Niemeyer, Tübingen, pp. 129–152, 2002.

[15] B. J. Kröger, *Ein phonetisches Modell der Sprachproduktion*. Niemeyer, Tübingen, 1998.

[16] M. Schröder, "Approaches to emotional expressivity in synthetic speech," in K. Izdebski (ed.), *Emotions in the Human Voice*, vol. 3, 2007.

[17] C. H. Shadle and R. I. Damper, "Prospects for articulatory synthesis: A position paper," in *Fourth ISCA Tutorial and Research Workshop on Speech Synthesis*, Pitlochry, Scotland, pp. 121–126, 2001.

[18] K. N. Stevens, *Acoustic Phonetics*. MIT Press, Boston, 1998.

[19] I. R. Titze, "Parameterization of the glottal area, glottal flow, and vocal fold contact area," *Journal of the Acoustical Society of America*, vol. 75, no. 2, pp. 570–580, 1984.

[20] Y. Xu and F. Liu, "Tonal alignment, syllable structure and coarticulation: Toward an integrated model," *Italian Journal of Linguistics (in press)*, 2007.