# Generating Gestural Timing From EMA Data Using Articulatory Resynthesis

Ingmar Steiner[a,b] and Korin Richmond[a]

[a]Centre for Speech Technology Research, University of Edinburgh,
[b]Institute of Phonetics, Saarland University
E-mail: ingmar.steiner@ed.ac.uk, korin@cstr.ed.ac.uk

## Abstract

*As part of ongoing work to integrate an articulatory synthesizer into a modular TTS platform, a method is presented which allows gestural timings to be generated automatically from EMA data. Further work is outlined which will adapt the vocal tract model and phoneset to English using new articulatory data, and use statistical trajectory models.*

## 1 Background

Articulatory speech synthesis presents a phonetically intuitive alternative to other synthesis techniques. Unlike approaches such as HMM-based synthesis, there is a clear separation of source (glottis) and filter (vocal tract), avoiding many of the problems of estimating the vocal tract transfer function.

The articulatory synthesizer VocalTractLab[1] (VTL) [1, 3] uses a configurable, 3-D geometric model of the human vocal tract to synthesize high-quality speech. It is controlled by a gestural score, akin to those used in articulatory phonology [5], containing several "autosegmental" tiers on which gestures are arranged sequentially over time (Figure 1). These gestures determine the movement of a set of *control points* embedded in the vocal tract model, which in turn control its shape.

VTL does not include any text-to-speech (TTS) capabilities, and synthesis is controlled through the gestural score interface. This presents the user with the responsibility of hand-crafting suitable gestural scores, a significant task requiring expert knowledge and patience.

The main challenge in creating the gestural score lies in determining the timing of the gestures on each tier. Using the durations of acoustic segments is not a satisfactory solution, since these segments are mostly the *result* of articulatory movements that begin earlier, but how much earlier is difficult to predict, and is influenced by phonetic context.

An earlier attempt at combining VTL with the TTS platform BOSS[2] produced intellegible, but not quite natural-sounding, results [4]. This prototype used hand-written "phasing rules" [5] to convert the segmental durations predicted by the durational component of BOSS to gestural timings.

A different approach is presented in the following sections and uses human articulatory data to generate gestural timings automatically.

## 2 Data-driven articulatory resynthesis

The fundamental assumption is that gestural scores based on real human speech will result in a higher degree of naturalness in synthesis. The task, therefore, is to generate discrete gestural timings from real speech.

It is possible to capture the motion of points on the surfaces of the 3-D vocal tract model, an analysis similar to Electromagnetic Articulography (EMA). Comparing the resulting "virtual" EMA (VEMA) trajectories to actual EMA data is very straightforward and much faster than using VTL to synthesize audio and recovering gestural timings from the result by measuring distances in the acoustic domain, apart from problems arising from the nonlinear relationship between articulation and acoustics.

The procedure for each utterance in a corpus of EMA data is therefore to generate a gestural score for which the VEMA trajectories synthesized by VTL closely match the original EMA data. The set of gestural scores obtained using this analysis-by-synthesis method can then be used to train statistical

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

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

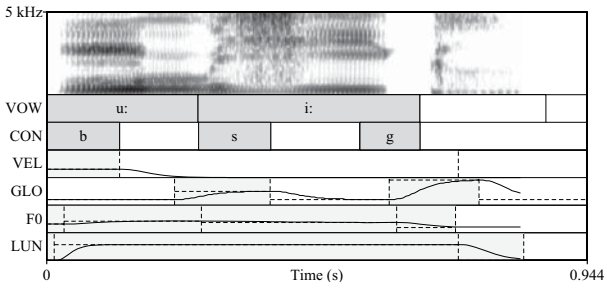


Figure 1: *Synthesis result and gestural score for the utterance Musik [mu:'zik]. The [m] is produced with bilabial occlusion (the **b** gesture) and lowered* VEL*um. The **s** and **g** gestures, combined with the* GLO*ttal gestures, produce [z] and [k], respectively.*

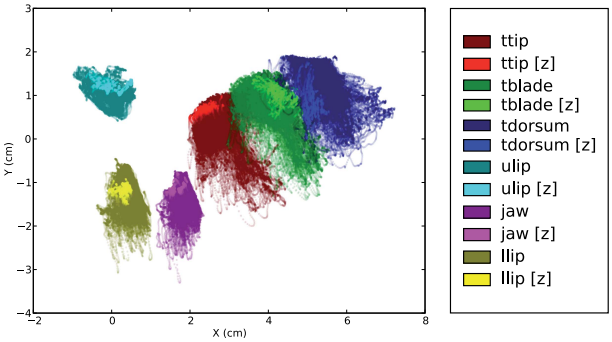


Figure 2: *Scatterplot of all datapoints in the EMA corpus, and those of [z] segments only. Note the distribution of datapoints for* ttip [z] *and* jaw [z].


Figure 3: *Finite state automaton for the consonantal gestures of the utterance [zazazaza].*

models to predict gestural scores for unseen utterances.

### 2.1 Articulatory data

To test the approach outlined above, an existing corpus of 271 two-second nonsense utterances of repetitive CV syllables was used [6]. It contains EMA data of a female native speaker of German, recorded on a Carstens AG100 Articulograph at 200 Hz sample rate, with coils attached to the upper and lower lip, jaw, and tongue tip, blade, and dorsum, as well as simultaneous audio (Figure 2).

### 2.2 Brute force baseline

As a proof of concept, one utterance ([zazazaza]) was selected from the EMA corpus, and all possible gestural scores for this simple utterance were generated. To make this a finite set, the 2 s duration was split into discrete frames, with gestural boundaries occurring only at frame boundaries.

The number of possible gestural scores $n_{f,g}$ for $f$ frames and $g$ gestures is given by the recursive function

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

For the sample utterance, the number of consonantal gestures, padded with "silent", as well as leading and trailing "neutral" gestures (corresponding to the speaker's rest position), is 12 (Figure 3). At 10 frames per second (fps), there are 20 frames, and therefore 19 possible times at which the 11 gestural boundaries can be positioned. This leads to 75,582 distinct gestural scores.[3]

For each of these resulting gestural scores, the VEMA trajectories were synthesized[4] and the root-mean-square error computed on the normalized trajectories, weighted by relevance for the phone represented by the current gesture. The relevance was determined by the ratio of phone-specific to overall EMA datapoints (Figure 2). While this already works well, integrating the approach taken by [7] is expected to further improve the results.

The gestural timings and tongue tip height trajectory of the optimal gestural score are shown in Figure 4.

### 2.3 Viterbi search

While the results of the brute force approach are promising, the vast amount of CPU time it requires renders it utterly impractical. Therefore, a Dynamic Programming algorithm was implemented to generate the gestural timings more efficiently. The consonantal gestures required for the sample utterance can be represented as Finite State Automaton (Figure 3), which can be expanded into a transition network.

Finding the optimal path through the network

[3] The number of permutations of gestural timings grows exponentially with the framerate; e.g. 15 fps yields 34,597,290 gestural scores, 20 fps yields 1,676,056,044, etc.

[4] VEMA synthesis time was ∼35 min. on 112 2.66 GHz CPUs in parallel.

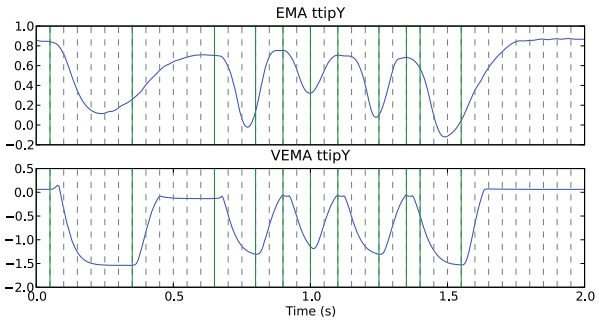


Figure 4: *Original and resynthesized ttipY trajectories for [zazazaza] at 20 fps. The dashed and solid lines represent frame and gesture boundaries.*

given an error metric based on the EMA and VEMA data is essentially a *forced alignment* problem, solved by using a Viterbi search.[5] This returns the number of frames in each gesture, and hence, the gestural timings.[6] Both the error metric used as the cost function and the search result were the same as for the brute-force search (Figure 4).

### 2.4 F0 gestures

The pitch contour from the original utterance was resynthesized by extracting, smoothing, and interpolating the original pitch using Praat.[7] The inflection point times of the resulting contour were used as F0 gesture boundaries, while the gesture parameters were derived from the slope of the tangent lines at these points.

The resulting F0 trajectory closely matches the original smoothed, interpolated pitch contour. The effort parameter was kept constant, and while an optimization of this parameter would result in an even closer match between the original and resynthesized contours, this was given a low priority.[8]

### 2.5 Gestural score assembly

Using the timing of the consonantal and F0 gestures, the remaining tiers in the gestural score were populated with gestures in a few simple steps:

*Vowel* gestures were synchronized with the consonantal gestures based on syllable structure.

[5]VTL's gestural model is a nonanticipative system.

[6]Search time (20 fps) was ∼5 min. on one 1.86 GHz CPU.

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

[8]The optimization function has no analytic solution, and the required iterative approach was (at this point) deemed too expensive for the modest improvement expected.

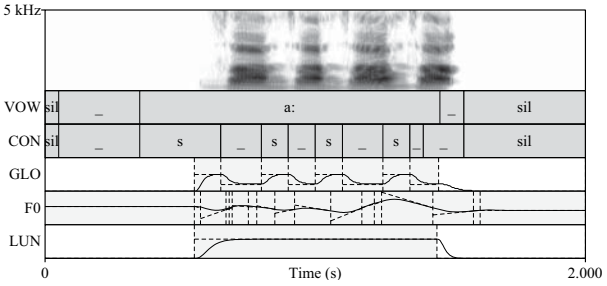


Figure 5: *Synthesis result and gestural score generated from EMA and F0 data for the resynthesis of [zazazaza].*

*Glottal* gestures were derived from consonantal and vowel gestures. Their values were set depending on whether they occurred during a vowel, fricative, or silence.

*Velic* gestures were synchronized with consonantal gestures in the same way as glottal gestures.

*Pulmonic* gestures were derived both from oral and F0 gestures.

The resulting gestural score is shown in Figure 5.

## 3 Further work

This study forms part of ongoing work to adapt VTL to an English speaker and to allow it to be used as the synthesis engine in a modular TTS system. Several aspects of this are outlined here.

### 3.1 Unified articulatory datasets

The approach presented in this paper uses EMA data from a female speaker to model the dynamics of speech and map them to VTL. However, the vocal tract configuration and phoneset available with VTL are based on vocal tract MRI data of a male speaker of German [2]. The speakers in these two datasets obviously have different vocal tracts, but no attempt was made to normalize for these differences.

A new articulatory database is currently being prepared for publication, and contains a variety of instrumentation techniques to provide articulatory data for a single 34-year old male native speaker of English. This data includes a large corpus (1,263 phonetically balanced English utterances) of 3-D EMA data recorded on a Carstens AG500 (coils on upper and lower lip, three along tongue, and on velum)[9] with simultaneous audio, as well as vocal

[9]Thanks to Phil Hoole et al. at IPS Munich

Figure 6: *Midsagittal slice of volumetric MR images for [a:]; 3-D bust prototype with digital dental cast.*

tract MRI of sustained English phones and dynamic VCV transitions.[10] Samples of the MRI data are shown in Figure 6.

These corpora will be used to build a new "voice" for VTL, consisting of vocal tract anatomy, phoneset, and a set of gestural scores resynthesizing the EMA data using the method described in the previous section. It is expected that using the same speaker for all of these components will further improve the results produced by data-driven synthesizer control.

Additionally, configuring VTL to an English speaker will provide a phoneset suitable for the synthesis of English utterances, which is currently impossible without heavy modification of VTL's original German phoneset.

### 3.2 TTS integration

The gestural timings obtained through the resynthesis method described here could be used as training data for statistical models, such as Classification and Regression Trees (CARTs) or HMM-based synthesis (HTS).[11] Controlling VTL with these models would allow the synthesis of unseen utterances.

In this way, VTL could be used as the waveform synthesis engine at the back-end of a modular TTS platform such as Festival[12] or MARY[13], both of which already contain all of the required components to convert orthographic input into sequences of phones and predict the accompanying prosodic information. Additionally, the architecture of both of these TTS platforms allows the use of CARTs or HTS synthesis techniques.

[10] Thanks to Ian Marshall et al. at SBIRC Edinburgh

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

[12] `http://www.cstr.ed.ac.uk/projects/festival/`

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

## 4 Conclusion

We have introduced a method to automatically generate gestural timings from articulatory data, and presented intermediate results demonstrating both its practicality and satisfactory results. We are now well-placed to fit the vocal tract model to single-speaker articulatory data and eventually bridge the gap between the articulatory synthesizer and TTS applications.

## References

[1] P. Birkholz. *3D-Artikulatorische Sprachsynthese*. Logos, Berlin, Germany, 2006. PhD thesis.

[2] P. Birkholz and B. J. Kröger. Vocal tract model adaptation using magnetic resonance imaging. In *7th International Seminar on Speech Production*, pages 493–500, Ubatuba, Brazil, December 2006.

[3] P. Birkholz and B. J. Kröger. A gesture-based concept for speech movement control in articulatory speech synthesis. In A. Esposito, M. Faundez-Zanuy, E. Keller, and M. Marinaro, editors, *Verbal and Nonverbal Communication Behaviours*, volume 4775 of *Lecture Notes in Computer Science*, pages 174–189. Springer, March 2007.

[4] P. Birkholz, I. Steiner, and S. Breuer. Control concepts for articulatory speech synthesis. In P. Wagner, J. Abresch, S. Breuer, and W. Hess, editors, *Sixth ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)*, pages 5–10, Bonn, Germany, August 2007. ISCA.

[5] C. P. Browman and L. M. Goldstein. Articulatory phonology: An overview. *Phonetica*, 49:155–180, 1992.

[6] S. Fagel and C. Clemens. An articulation model for audiovisual speech synthesis: Determination, adjustment, evaluation. *Speech Communication*, 44(1-4):141–154, October 2004.

[7] V. D. Singampalli and P. J. Jackson. Statistical identication of critical, dependent and redundant articulators. In *Interspeech*, pages 70–73, Antwerp, Belgium, August 2007. ISCA.

*This work was funded under the Edinburgh Speech Science and Technology (EdSST) programme.*