GESTURE-BASED ARTICULATORY TEXT TO SPEECH SYNTHESIS*

Benjamin Weitz[1,2], *Ingmar Steiner*[2,3], *Peter Birkholz*[4]

[1]*SemVox GmbH*
[2]*Saarland University*
[3]*DFKI GmbH*
[4]*Technische Universität Dresden*
bweitz@coli.uni-saarland.de

Abstract: We present work carried out to extend the text to speech (TTS) platform MaryTTS with a back-end that serves as an interface to the articulatory synthesizer VocalTractLab (VTL). New processing modules were developed to *(a)* convert the linguistic and acoustic parameters predicted from orthographic text into a gestural score, and *(b)* synthesize it to audio using the VTL software library. We also describe an evaluation of the resulting gesture-based articulatory TTS, using articulatory and acoustic speech data.

1 Introduction

Articulatory synthesis, the simulation of speech production using a model of the human vocal tract, has seen steady advancements over the past decades. However, it is still a complex task to design the control structures which are required to drive the dynamics of the vocal tract model, that in turn determine the evolution of its shape over time. Depending on the nature of the articulatory synthesizer front-end, these control structures may take the form of a *gestural score*, which arranges the relative timing of high-level "macros", setting vocal tract target configurations that correspond to the desired speech sounds [2].

While the ability to fine-tune parameters of the speech simulation process is a unique advantage of articulatory synthesis, it would nevertheless be valuable to generate gestural scores from an underspecified, text-based input representation. However, only limited work has been done to integrate these concepts with speech technology applications such as text to speech (TTS) synthesis [3].

In this paper, we present work carried out to extend the TTS platform MaryTTS[1] [4] with a back-end that serves as an interface to the articulatory synthesizer VocalTractLab (VTL)[2] [5]. New processing modules were developed to *(a)* convert the linguistic and acoustic parameters predicted from orthographic text into a gestural score, and *(b)* synthesize it to audio using the VTL software library.

The remainder of this paper is structured as follows: Section 2 gives a brief overview of the VTL synthesizer and the MaryTTS platform, and describes how the two were integrated with each other. In Section 3, we provide details of several experiments designed to evaluate the accuracy of the articulatory synthesis in the articulatory and acoustic domains. Finally, we conclude with a summary and outline future work.

*This paper is based on unpublished work by Weitz [1].

[1] `http://mary.dfki.de` and `https://github.com/marytts/marytts`

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

2 Methods

2.1 VocalTractLab

The VocalTractLab (VTL) articulatory synthesizer comprises three main components:
(a) a geometric vocal tract model,
(b) a gestural control model, and
(c) an acoustic model.
The geometric model consists of 3D meshes representing the oral and pharyngeal cavities, the tongue surface, teeth, and lips. The shape of this vocal tract model was adapted to fit the anatomy of a male native speaker of German, using magnetic resonance imaging (MRI) [6].

The control model is based on articulatory phonology [7], in particular its concept of gestures on multiple independent tiers. In VTL, these tiers include the lungs, glottis, and F0 (for direct control of air pressure, phonation, and fundamental frequency, respectively), a vocalic tier (for vowels), and several tiers for consonantal constriction formation (tongue body, tongue tip, lips, and velum). Each gesture on one of these tiers is characterized by its onset and offset times, and its target, which can be a numeric value (such as air pressure in Pa or F0 in semitones), or a symbol representing a predefined setting for the vocal tract model's low-level control parameters. The shape of the vocal tract itself is defined at each point in time by a combination of these control parameters, such as the tongue body center coordinate and radius in the midsagittal plane, angle of jaw aperture, lip protrusion, etc. [for details, cf. 8].

The acoustic model is based on a branched tube model, and uses the simulated glottal waveform and a noise generator, combined with the vocal tract transfer function calculated from the shape of the vocal tract model at each point in time, to generate an acoustic waveform.

To synthesize an utterance, the user of VTL is first required to provide a gestural score, which specifies the required gestures on all tiers, before letting the synthesizer simulate the resulting audio; however, creating such a gestural score, and timing the gestures correctly, is far from trivial, and mistakes can result in unintelligible output, or audio which does not even resemble speech.

2.2 MaryTTS

MaryTTS is an open-source, multilingual TTS platform implemented in Java. It is designed to process data, represented by XML documents, using a sequence of modules, each of which in turn enriches the document, generating audio in the final, Synthesis module. Several of these modules are responsible for normalizing the input text, determining the pronunciation using a lexicon and rules, and predicting segment durations and intonation contours using statistical models.

In order to generate audio, the MaryTTS user needs to specify an available "voice", which is tied to a specific language, and which is configured with acoustic models for prosody prediction. Only after the target sequence of phonetic segments, along with their durations and target F0 values, is determined, can these acoustic parameters be passed to the Synthesis module for waveform generation, which in turn relies on one of several available synthesizers, including diphone concatenation or unit selection, or hidden Markov model (HMM)-based synthesis.

2.3 Articulatory TTS pipeline

We chose to use MaryTTS as the "front-end" to predict phonetic-acoustic parameters from text and generate a gestural score from them, and VTL as the synthesizer back-end, from which the resulting audio waveform is then retrieved and presented to the user.

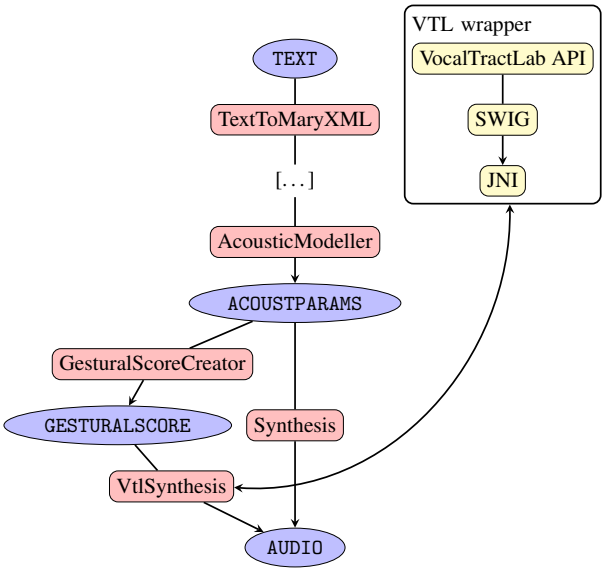

Figure 1 – Architecture of the extended MaryTTS processing pipeline with a native wrapper, and custom modules and datatypes. MaryTTS modules are shown as red rectangles, datatypes as blue ellipses.

This involved developing new MaryTTS components, and wrapping the VTL API with a Java interface (see Figure 1). Instead of the conventional module processing pipeline, where the Synthesis module requires input in the form of MaryXML data with acoustic parameters (`ACOUSTPARAMS`) to produce `AUDIO` output data, we need to first generate a gestural score in VTL format from the acoustic parameters, and then process it using the VTL API, before returning it as output to the MaryTTS synthesis request. For this reason, the new MaryTTS data type `GESTURALSCORE` was defined, and two new modules were implemented, a GesturalScoreCreator (which converts `ACOUSTPARAMS` input to `GESTURALSCORE` output), and a new VtlSynthesis module (which takes the `GESTURALSCORE` input, sends it to the VTL API, waits for the resulting audio samples, and finally outputs them as `AUDIO`).

In order to wrap VTL, which is written in C++ and provided as a native library, and expose its public API to MaryTTS, we rely the Java Native Interface (JNI) via automatically generated Java bindings using the Simplified Wrapper and Interface Generator (SWIG).[3]

While the phone-level pronunciation and symbolic prosody can be generated using the German language components provided by MaryTTS, the acoustic parameters in the form of segment durations and F0 target contours also needed to be predicted. For this purpose, we used an HMM-based voice trained on the PAVOQUE corpus [9] (neutral subset).

3 Experiments

At the heart of our gestural score generation is the prediction of gesture duration and alignment. In order to evaluate different approaches to this challenge, we designed several experiments, including a phasing rule based approach, as well as data-driven techniques in the articulatory and acoustic domains.

[3] `http://www.swig.org/`

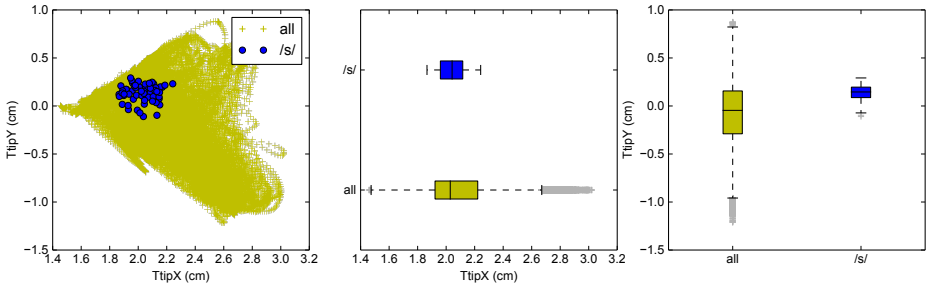

Figure 2 – Relevance plots for the distribution of tongue tip EMA coil positions for all vs. [s] frames. The compactness of the distributions illustrates the relevance of the tongue tip for the production of this apical sibilant.

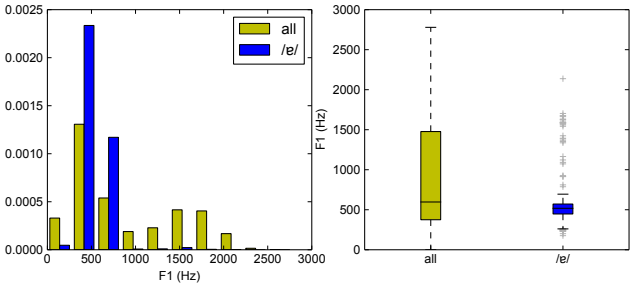

Figure 3 – Relevance plots for the distribution of the first formant for all vs. [ɐ] frames. The compactness of the distributions underscores the strong localization of this vowel in the formant space.

3.1 Phasing rules

In order to evaluate a baseline of manually specified rules for gesture durations and alignment, we implemented a version of the phasing rules described by Kröger [10]. Unfortunately, the resulting synthetic speech was far from intelligible, due in part to an inability to model syllables with complex structure.

3.2 Data

Turning to a data-driven approach, we used a corpus of electromagnetic articulography (EMA) data recorded from the same speaker whose MRI scans had been used to configure the vocal tract model [6]. The EMA corpus contained 172 German sentences designed to study German vowel articulation, and was kindly provided by Jörg Dreyer [for details, see 11]. This data provides intra-oral motion capture, with EMA coils tracking the lower lip, jaw, and tongue tip, blade, dorsum, and back, in addition to acoustic recordings (sampled at 16 kHz).

The EMA data was first smoothed, and then automatically segmented at the phone and syllable level, using a combination of WebMAUS [12] and MaryTTS. The data was then used in a resynthesis approach to generate a set of gestural scores for training a statistical model to predict gestural scores for unseen utterances. Different cost functions were evaluated for comparing the resynthesized version to their natural counterparts.

3.3 Optimization using RMSE

In order to optimize an initially generated gestural score, for each gesture in the score, a search space was sampled by synthesizing the preceding gesture with different durations (in 10 ms

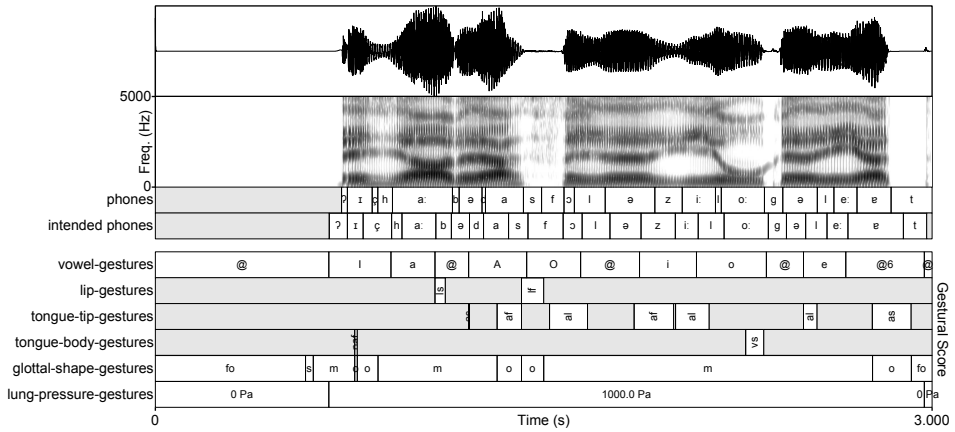

Figure 4 – Resynthesis of the utterance *Ich habe das volle Silo geleert*, with actual and intended phones and generated gestural score, optimized using the RMSE cost function.

steps, up to a maximum of 200 ms). The best gesture duration was determined by evaluating a cost function, based on the reference recording. The distance metric was a RMSE with the following features:

- EMA coil positions, as well as their first and second derivatives
- voicing
- harmonics to noise ratio (HNR)
- frequency and bandwidth of the first three formants, as well as their first and second derivatives
- 12 mel-frequency cepstral coefficients (MFCCs), as well as their first and second derivatives

The features were weighted based on their relevance for the production of each phone, using the ratio of the variances for that phone vs. all values of that feature (after normalizing). This heuristic allowed us to bootstrap a weighting matrix with minimal supervision. Examples of the feature distribution analysis for relevance weighting are shown in Figure 2 and Figure 3. An example utterance is shown in Figure 4.

Overall however, this approach produced unsatisfactory results. In the articulatory domain, we relied on "virtual" EMA trajectories obtained by tracking mesh vertices on the articulators of the geometric vocal tract model; however, the dynamics of these trajectories did not appear to resemble those of the natural motion capture data.

3.4 Optimization using phoneme classifier

Reformulating the comparison of natural and resynthesized audio as a pronunciation evaluation problem, we explored an alternative cost function using a log-likelihood phoneme classifier. Our classifier used Gaussian mixture models (GMMs), trained using the smacpy library[4] in a bag-of-frames model [13], with MFCCs and EMA trajectories and their derivatives as features.

The performance of the phoneme classifier can be seen in Table 1, and an example utterance resynthesized using a cost function based on phoneme classification is shown in Figure 5.

[4] `https://github.com/danstowell/smacpy`

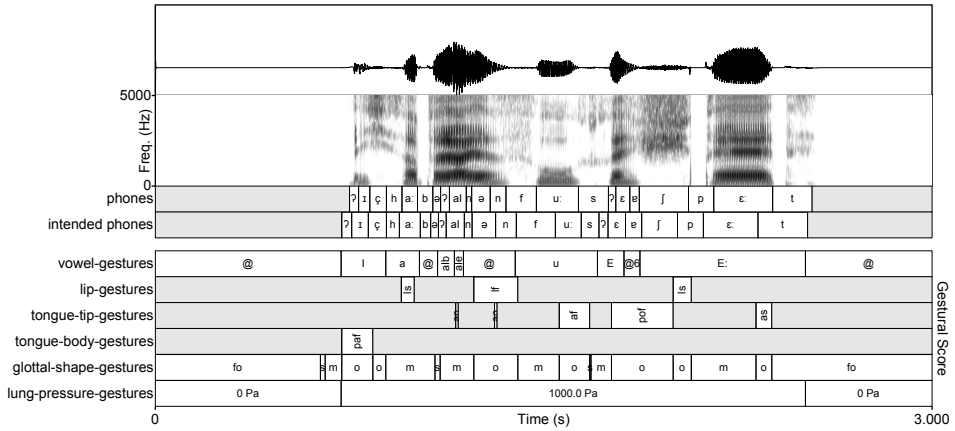

Figure 5 – Resynthesis of the utterance *Ich habe einen Fuß erspäht*, with actual and intended phones and generated gestural score, optimized using the phoneme classification log-likelihood cost function.

3.5 Optimization using log-spectral distance

In view of the problems in using VTL to generate naturalistic "virtual" EMA trajectories for comparison in the articulatory domain, we decided to explore a gestural score optimization using only acoustic features. In this case, the cost function used the log-spectral distance (LSD) as a distance metric, similar to the approach of Nam et al. [14].

In an informal evaluation, we determined that the best results could be obtained using a window length of 15 ms. An example utterance is shown in Figure 6.

We resynthesized the entire corpus, optimizing the gestural scores using the LSD-based cost function, and then trained a classification and regression tree (CART) for each tier in the gestural score. These CARTs were then provided as resources to the GesturalScoreCreator module introduced in Section 2.3, where they are used to predict gesture durations based on input phones and predicted acoustic durations. However, the performance of the CARTs is limited by the small amount of training data.

4 Conclusion

We have presented an experimental system for gesture-based articulatory TTS synthesis, extending the MaryTTS platform with new components which integrate the VTL articulatory synthesizer. The gestural scores required as input for VTL are generated automatically, using

	MFCC window size				
features	5 ms	10 ms	15 ms	20 ms	25 ms
MFCCs	67.98	70.94	73.69	70.89	68.31
MFCCs + Δ features	72.04	81.72	87.69	87.69	86.70
MFCCs + EMA	88.16	89.45	89.80	88.71	87.69
MFCCs + Δ features + EMA	89.42	91.34	91.89	91.67	90.46
EMA	86.13				

Table 1 – Performance of the phoneme classifier. Scores represent average accuracy (in %) over a 10-fold cross-validation. Note that for window sizes other than 5 ms, the EMA data had to be resampled to match the MFCC frame rate.

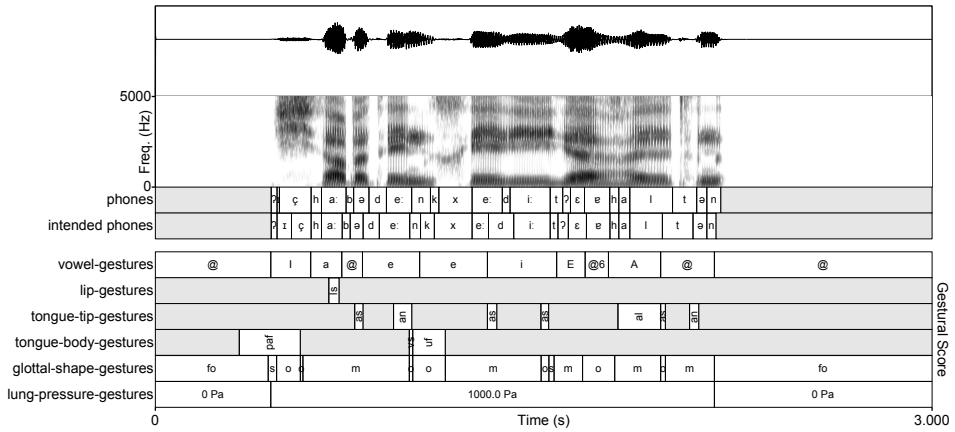

Figure 6 – Resynthesis of the utterance *Ich habe den Kredit erhalten*, with actual and intended phones and generated gestural score, optimized using the LSD cost function.

CARTs trained on a multimodal speech corpus recorded from the same speaker whose anatomy was used to adapt the VTL vocal tract model.

We also presented several experiments to optimize generated gestural scores in a data-driven approach. While a comparison in the articulatory domain is intuitive and theoretically more efficient, in practice it proved challenging, due to inherent differences in the natural and synthetic EMA trajectories. Therefore, we achieved the highest accuracy after optimizing with a purely acoustic LSD metric.

This work paves the way towards articulatory TTS synthesis, and the corresponding software is available online,[5] under an open source license. Further work can focus on aspects such as using larger corpora to train the models for gesture duration prediction. It would also be interesting to revisit evaluating in the articulatory domain, but this would require modification of the generation of "virtual" EMA trajectories in VTL's geometric vocal tract model. Alternatively, we could explore adapting the approach taken by Steiner et al. [15] to sidestep the gestural model completely, and generate VTL control parameters directly from text using state-of-the-art statistical parametric synthesis techniques.

References

[1] WEITZ, B.: *Gesture-Based Articulatory Text-to-Speech Synthesis*. Master's thesis, Saarland University, 2014. Unpublished.

[2] BIRKHOLZ, P.: *Modeling consonant-vowel coarticulation for articulatory speech synthesis*. *PLoS ONE*, 8(4), 2013. doi:10.1371/journal.pone.0060603.

[3] BIRKHOLZ, P., I. STEINER, and S. BREUER: *Control concepts for articulatory speech synthesis*. In *6th ISCA Workshop on Speech Synthesis (SSW)*. Bonn, Germany, 2007. URL `http://www.isca-speech.org/archive_open/ssw6/ssw6_005.html`.

[4] SCHRÖDER, M. and J. TROUVAIN: *The German text-to-speech synthesis system MARY: A tool for research, development and teaching*. *International Journal of Speech Technology*, 6(4), pp. 365–377, 2003. doi:10.1023/A:1025708916924.

[5] `https://github.com/marytts/marytts-vocaltractlab`

[5] BIRKHOLZ, P.: *3D-Artikulatorische Sprachsynthese*. Logos Verlag, Berlin, Germany, 2005.

[6] BIRKHOLZ, P. and B. KRÖGER: *Vocal tract model adaptation using magnetic resonance imaging*. In *7th International Seminar on Speech Production (ISSP)*, pp. 493–500. Ubatuba, Brazil, 2006.

[7] BROWMAN, C. P. and L. GOLDSTEIN: *Articulatory phonology: An overview*. *Phonetica*, 49(3-4), pp. 155–180, 1992. doi:10.1159/000261913.

[8] BIRKHOLZ, P.: *VocalTractLab 2.1 User Manual*, 2013.

[9] STEINER, I., M. SCHRÖDER, and A. KLEPP: *The PAVOQUE corpus as a resource for analysis and synthesis of expressive speech*. In *Phonetik & Phonologie 9*, pp. 83–84. Zurich, Switzerland, 2013.

[10] KRÖGER, B. J.: *Ein phonetisches Modell der Sprachproduktion*. Niemeyer, 1998.

[11] STEINER, I.: *Observations on the dynamic control of an articulatory synthesizer using speech production data*. Ph.D. thesis, Saarland University, 2010. `urn:nbn:de:bsz:291-scidok-32243`.

[12] KISLER, T., F. SCHIEL, and H. SLOETJES: *Signal processing via web services: the use case WebMAUS*. In *Digital Humanities*, pp. 30–34. Hamburg, Germany, 2012.

[13] GIANNOULIS, D., D. STOWELL, E. BENETOS, M. ROSSIGNOL, M. LAGRANGE, and M. D. PLUMBLEY: *A database and challenge for acoustic scene classification and event detection*. In *21st European Signal Processing Conference (EUSIPCO)*. Marrakech, Morocco, 2013. URL `http://www.eurasip.org/Proceedings/Eusipco/Eusipco2013/papers/1569738295.pdf`.

[14] NAM, H., V. MITRA, M. TIEDE, M. HASEGAWA-JOHNSON, C. ESPY-WILSON, E. SALTZMAN, and L. GOLDSTEIN: *A procedure for estimating gestural scores from speech acoustics*. *Journal of the Acoustical Society of America*, 132(6), pp. 3980–3989, 2012. doi:10.1121/1.4763545.

[15] STEINER, I., S. LE MAGUER, and A. HEWER: *Synthesis of tongue motion and acoustics from text using a multimodal articulatory database*. Submitted. `arXiv:1612.09352`.