INTERSPEECH 2009 BRIGHTON

Towards unsupervised articulatory resynthesis of German utterances using EMA data

Ingmar Steiner[1,2], *Korin Richmond*[2]

[1]Institute of Phonetics, Saarland University, Germany
[2]Centre for Speech Technology Research, University of Edinburgh, UK
steiner@coli.uni-saarland.de, korin@cstr.ed.ac.uk

Abstract

As part of ongoing research towards integrating an articulatory synthesizer into a text-to-speech (TTS) framework, a corpus of German utterances recorded with electromagnetic articulography (EMA) is resynthesized to provide training data for statistical models. The resynthesis is based on a measure of similarity between the original and resynthesized EMA trajectories, weighted by articulatory relevance. Preliminary results are discussed and future work outlined.

Index Terms: articulatory speech synthesis, copy synthesis, electromagnetic articulography

10.21437/Interspeech.2009-590

1. Background

This study presents progress towards the integration of an articulatory synthesizer [1, 2] into a modular text-to-speech (TTS) framework. The synthesizer, *VocalTractLab*[1] (VTL), uses a three-dimensional, geometric model of the vocal tract that has been configured to approximate the anatomy of a specific male speaker of German [3], and provides a set of target configurations of the vocal tract model corresponding to a German phoneset.

VTL's synthesis is controlled by means of a *gestural score* (a concept borrowed from articulatory phonology [4]), arranging appropriate articulatory gestures, which are timed in a way that produces suitable output. While the nature and sequence of gestures in the score can be generated automatically from textual input, optimizing their *timing* presents a significant problem for the use of VTL as a synthesis back-end in TTS, where manual optimization at synthesis time is out of the question.

It may be possible to predict the gestural timing automatically using statistical models, but suitable data must be available to train such models. One possible way of obtaining this training data is to generate it by resynthesizing real human utterances with VTL in such a way that the motion of the vocal tract model matches that of the human speaker as closely as possible, while producing intelligible copies of the original utterances. This requires a corpus of articulatory movement data such as that obtained using electromagnetic articulography (EMA).

The present study aims to demonstrate a method of achieving this with as little human intervention as possible, building on work presented in [5]. In the previous study, nonsense utterances with a phonetically simple structure were resynthesized. Expert knowledge was used to select a single, most relevant articulatory trajectory to be matched. However, in the method we are currently developing and describe here, the aim is to examine an optimized cost function, based on automatically derived weights, in order to cope with the increased phonetic complexity of natural utterances.

This study is also exploratory in nature and attempts to solve some aspects of the overall challenge while avoiding certain potential problems. In particular, the automatic generation of gestural scores at this stage still uses the speaker definition supplied with VTL by its author, which constrains its deployment to German; additionally, we have sought to circumnavigate the issue of vocal tract normalization between the vocal tract model and the EMA articulatory movement data by using the same speaker for both.

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

2. Method

The method described here centers on the comparison of articulatory movements in the human vocal tract and the vocal tract model. EMA requires a significant amount of effort, equipment, and expertise, tracking the movement of fleshpoints on articulators in the human vocal tract in real time during speech production. In contrast, it is trivial to sample the coordinates of specific vertices on the wire-frame mesh of the vocal tract model in VTL while a gestural score is rendered. However, the resulting trajectories are comparable, and by quantifying their similarity, it is possible to select, from a finite set of generated gestural score candidates, the one whose "virtual" EMA (VEMA) trajectories are most similar to the real EMA trajectories of a natural target utterance.

2.1. Gestural score generation

The control structures required by VTL take the form of a gestural score containing consonant and vowel gestures (see [6] for a detailed explanation). The identity of these gestures is known from textual input and lexical lookup. In a resynthesis paradigm, the total duration of the gestural score is equal to that of the previously recorded target utterance, and this duration can be split into a number of discrete frames.

Each frame represents a state of the gestural score, in which a gesture is assumed. The sequence of states required to generate the target score can therefore be generated by a finite state automaton (FSA), which, given a number of frames, can be expanded into a transition network. In this network, each row represents one gesture and each column represents one frame.

Every possible path through the network is equivalent to a distinct gestural score; the total number of scores $n_{f,g}$ is given by the recursive function

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

where f and g are the number of frames and gestures, respectively.

Finding the best gestural score is therefore essentially a forced alignment problem, and since the search space can quickly become prohibitively large, a dynamic programming technique such as a Viterbi search is invaluable to efficiently find the optimal gestural score. One example of such a path through the network is shown in Fig. 1.

2.2. Articulatory data

In order to avoid the need for vocal tract normalization, EMA data of the same speaker as that used for the speaker configuration in VTL was required. Of the numerous EMA recordings available of the speaker in question, a set of utterances was selected that contained 24 normal German utterances[2], repeated 7 times each.

This EMA corpus was recorded at the Centre for General Linguistics (ZAS), Berlin, in 2002. The data was sampled at 200 Hz on a Carstens AG100 articulograph, with measurement coils in the midsagittal plane, on the lower incisors (`LI`), lower lip (`LL`), and tongue tip (`TT`), blade (`TB`), and dorsum (`TD`), with simultaneous audio. Further details of the recording procedure can be found in [7].

2.3. Error metric

The suitability of a gestural score candidate is determined by a cost function, which calculates the difference between the VEMA trajectories synthesized by VTL using this gestural score, and the corresponding EMA trajectories in the original data. Our cost function is similar to that of [8], in that a correlation coefficient r is combined with the squared error s_e to quantify the similarity in shape of, as well as the distance between, two trajectories.

$$c = f s_e + (1 - f)(1 - r) \tag{2}$$

These metrics are balanced by a scaling factor f and weighted according to a matrix of weights, which stores the relevance of each articulatory trajectory for the production of each element in the phoneset. This allows deviation in the shape or scale of the VEMA trajectories from the corresponding original EMA trajectories to be penalized much more strongly where the trajectories represent an articulator critical for the current gesture.

2.4. Relevance of articulators

Not every articulatory movement is equally critical to the gestures required for a given phone to be produced; e.g. for the production of a [z], the `TT` trajectories are far more relevant than the `LL`. To quantify this relevance, a weight is applied to each articulatory trajectory, in each frame. While it is possible to define these weights using phonetic expert knowledge, the task of doing so is tedious and potentially error-prone, and general phonetic knowledge cannot always predict which specific strategies of speech production are actually employed by the individual speaker whose utterances are to be resynthesized.

A promising possibility is to statistically analyze the EMA data, given a phonetic transcription, and automatically identify the articulators relevant to the realization of each phone type. Such an approach, using Kullback-Leibler divergence, is detailed in [9], and was applied to the EMA data used in this study.

[2]The utterances were "normal" in the sense that they were well-formed, grammatical German sentences and did not contain nonsense syllables.

Phone	Identified articulatory trajectories					
[ə]	TTy					
[ɐ]	TDy					
[a]	TTy	TTx	TDy	LIx	LLy	
[eː]	LIy	TBy	LLy			
[g]	TBx	LIy	TTy			
[h]	TTx	TDy	TTy	LIy		
[iː]	TBx	TTy	LLy	TTx	TDy	
[l]	TBx					
[m]	LIy	TDy	TTy	LLy	TDx	
[o]	TDy	TBx	TTy	LLx	LIy	
[ʁ]	TDx	TTy	LIy	LLx	TTx	TDy
[ʊ]	TTy	TDy	TDx	LIx	LLy	
[x]	LIy					
[yː]	LIy					
[z]	TBy					

Table 1: *Sample results of applying the algorithm presented in [9] to the German EMA data. For each phone (in IPA notation) the articulatory trajectories identified as relevant are listed. In numerous cases, the identified trajectories defy phonetic knowledge, which is almost certainly due to the low suitability of this particular EMA data set for this type of analysis. Phones for which no relevant trajectories were found have been omitted here.*

In order to apply this algorithm to the EMA corpus, the acoustic data was first automatically segmented using forced alignment with HTK [10] via MAUS[3] [11].

The results of the automatic identification, listed in Table 1, are somewhat unsatisfactory. However, it is more than likely that this can be attributed to any or all of a number of possible error sources, all of which implicate the data as less than suitable, not the algorithm itself. In particular,

- the number and placement of measurement coils exclude the velum and upper lip, so a number of phones cannot be correctly described with this particular EMA data set;
- only 24 distinct utterances were used, and they are not phonetically balanced, so there is additional data sparsity;
- the phonetic segments were automatically aligned and not hand-corrected (see below);
- the articulatory trajectories were treated as independent; it is possible that the "2D" variant of the algorithm, which combines the `x` and `y` trajectories of each measurement coil, would produce clearer results.

The biggest problem seems to be the fact that the segment boundaries are not optimal. While they were visually checked for obvious alignment errors, the precision of automatic boundary placement was low. It is observed in [9] that labeling errors can affect the performance of the algorithm, and our preliminary results confirm this sensitivity.

Despite the limited applicability of the results obtained here, the automatic identification of relevant articulators remains an attractive possibility, and will be tested more thoroughly on a more suitable EMA corpus. In the mean time, as a fallback for the present study, we have used a manually assigned baseline weighting instead to work around the problems we have described.

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

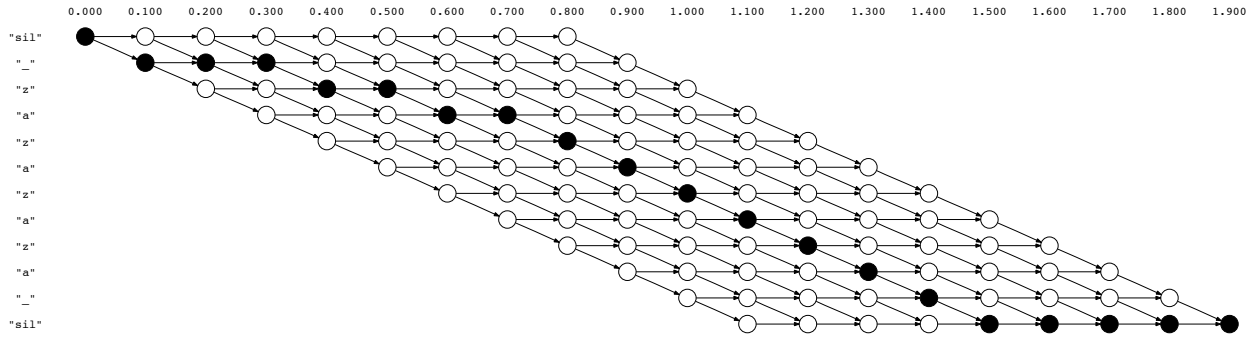

Figure 1: *Transition network for the nonsense utterance [zazazaza], illustrating the search space. The columns represent the frames of the 2 second utterance (at 10 frames per second; columns labels are frame start times), while the rows represent the sequence of gestures. The path through the shaded states is equivalent to the least costly gestural score (from [5]).*

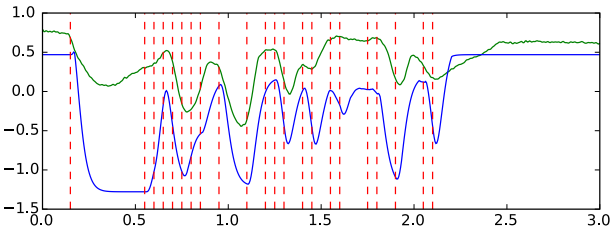

Figure 2: *Original (top) and resynthesized (bottom) `TTy` trajectories for the 3 second utterance "Ich habe Daten analysiert" ("I've analyzed data") at 20 frames per second. The vertical lines indicate gesture boundaries. The underlying gestural score was generated in such a way that the global shape of the trajectories takes precedence over the actual values.*

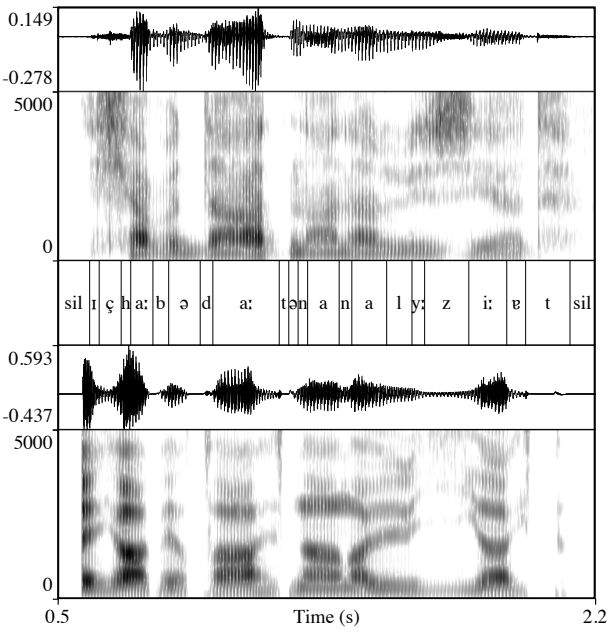

Figure 3: *Original (top) and resynthesized (bottom) utterance "Ich habe Daten analysiert" ("I've analyzed data"); the automatically aligned acoustic segmentation (in IPA) is shown in the middle.*

3. Results

Using the method described above, and presuming an appropriate cost function, an underspecified gestural score containing consonant and vowel gestures can be generated for a given utterance in the EMA corpus. As an example, one utterance illustrating the VEMA trajectories resynthesized by VTL using such a score is shown in Fig. 2. It should be noted that priority in determining the gesture boundaries was given to the overall shape of the trajectories, not the absolute values (see also Section 4.1 below).

Prior to waveform synthesis, the gestural score is enriched with F_0 gestures generated automatically by extracting the pitch contour from the original utterance (using Praat[4]), and aligned with the consonant gestures. The result of the waveform synthesis is shown in Fig. 3.

4. Discussion

We have described a method to automatically generate a gestural score from a spoken utterance for which EMA data is available and whose phone sequence is known. The method aims to find the score which most closely approximates the timing of the articulatory gestures assumed to underlie the original utterance. By performing this analysis-by-synthesis process for a suitably large number of utterances, we can build a training set of gestural scores. These may be used directly to train models (for example a CART model) to predict gestural score timings directly from text. Alternatively, by synthesizing the training set of gestural scores with VTL, a set of parameter trajectories may be generated, which encode the changing shape of the vocal tract model over time. These may be used to train statistical models to predict similar parameter trajectories for unseen utterances, for example using models similar to those used in statistical parametric speech synthesis [12].

4.1. Vocal tract geometry

By using EMA data from the same speaker as the one whose magnetic resonance imaging (MRI) data forms the basis for the vocal tract model's geometry and target configurations [3], we

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

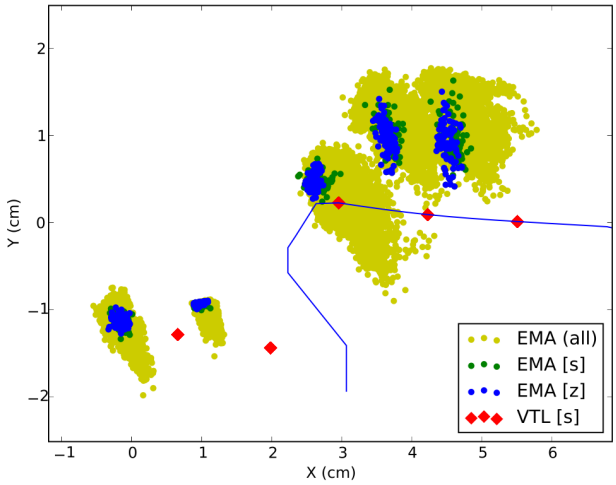

Figure 4: *Vocal tract geometry comparison between EMA data for [s] and [z] and VTL [s]. The EMA data is restricted to one midpoint sample per segment. Coils are (from left to right)* `LL`*,* `LI`*,* `TT`*,* `TB`*, and* `TD`*. Additionally, for the EMA data, all midpoints are shown in a light color to illustrate the fact that the distribution of samples for [s] and [z] is much more compact for some articulators (specifically* `TT`*,* `LI`*, and* `LL`*), which corresponds to their relevance in the production of these phones. The line represents the vocal tract model tongue contour.*

had expected to avoid the need for vocal tract normalization. However, it was found that several factors disrupt the anticipated close relationship.

The arrangement of vertices selected for tracking in the vocal tract model is not exactly the same as the placement of EMA coils on the speaker's articulators. While the flexibility of defining the virtual EMA coils could be improved, even a perfect match in a given static configuration would not automatically lead to the same trajectories during resynthesis, since the deformations of the vocal tract model surfaces and the speaker's articulators are inherently different. Even statically, the vocal tract model is a parametric representation of the speaker's real vocal tract and not intended to reproduce its shape as faithfully as a static vocal tract reconstruction from volumetric data could.

Finally, the fact that the vocal tract target configurations are based on MRI data of a supine speaker while the EMA data was recorded in an upright position leads to an additional, noticeable effect on the data for the tongue, and (to a lesser extent) the jaw. A recent study of this phenomenon is described in [13].

The consequences of these issues can be observed in Fig. 4. Despite these considerations, it may be possible to translate the geometries in a straightforward manner, an avenue yet to be explored in the context of this study.

4.2. Open issues

A number of issues await solution before the overall goal of allowing VTL to be used as an articulatory synthesis back-end in a TTS framework becomes possible. These issues are the focus of ongoing research and while they lie outside the scope of this paper, they are briefly addressed here.

EMA Data. A large EMA corpus of over 2,000 phonetically balanced English utterances has subsequently been recorded with a single male speaker using a Carstens AG500 3D articulograph.

Automatic weighting. For a large portion of this corpus, the arrangement of measurement coils includes the velum, and this data, along with the corpus design, is expected to yield improved results when analyzed using the method described by [9] for automatic identification of relevant articulators. While segmentation was also done automatically, the alignment process was performed with greater care than that mentioned above.

Vocal tract adaptation. The same speaker was also recently scanned using vocal tract MRI. The resulting volumetric data can be used to adapt the anatomy of the vocal tract model to this speaker, and midsagittal scans of dynamic speech production are intended to define the vocal tract configurations required for an English phoneset.

5. Acknowledgements

The authors would like to thank Susanne Fuchs and Jörg Dreyer for providing the EMA data, Veena Singampalli and Philip Jackson for generous help in applying their approach to this data, and Peter Birkholz for allowing extension of his VocalTractLab synthesizer.

This work was supported by Marie Curie Early Stage Training Site EdSST (MEST-CT-2005-020568).

6. References

[1] Birkholz, P., *3D-Artikulatorische Sprachsynthese*, Logos, 2006.

[2] Birkholz, P. and Kröger, B. J., "A gesture-based concept for speech movement control in articulatory speech synthesis", in A. Esposito, M. Faundez-Zanuy, E. Keller, M. Marinaro [Eds.], *Verbal and Nonverbal Communication Behaviours*, 174–189, Springer, 2007.

[3] Birkholz, P. and Kröger, B. J., "Vocal tract model adaptation using Magnetic Resonance Imaging", *Proc. 7th International Seminar on Speech Production*, 493–500, 2006.

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

[5] Steiner, I. and Richmond, K., "Generating gestural timing from EMA data using articulatory resynthesis", *Proc. 8th International Seminar on Speech Production*, 313–316, 2008.

[6] Birkholz, P., Steiner, I. and Breuer, S., "Control concepts for articulatory speech synthesis", *Proc. 6th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)*, 5–10, 2007.

[7] Fuchs, S., "Articulatory correlates of the voicing contrast in alveolar obstruent production in German", *ZAS Papers in Linguistics*, 41, 2005.

[8] Zacks, J. and Thomas, T. R., "A new neural network for articulatory speech recognition and its application to vowel identification", *Computer Speech and Language*, 8(3):189–209, 1994.

[9] Jackson, P. J. and Singampalli, V. D., "Statistical identification of articulation constraints in the production of speech", *Speech Communication*, 51(8):695–710, 2009.

[10] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P., *The HTK Book (for HTK Version 3.3)*, Cambridge University Engineering Department, 2005.

[11] Schiel, F., "Automatic phonetic transcription of non-prompted speech", *Proc. 14th International Congress of Phonetic Sciences*, 607–610, 1999.

[12] Zen, H., Tokuda, K. and Black, A. W., "Statistical parametric speech synthesis", *Speech Communication*, doi: 10.1016/j.specom.2009.04.004, 2009, in press.

[13] Stone, M., Stock, G., Bunin, K., Kumar, K., Epstein, M. A., Kambhamettu, C., Li, M., Parthasarathy, V. and Prince, J. L., "Comparison of speech production in upright and supine position", *Journal of the Acoustical Society of America*, 122(1):532–541, 2007.