Training an articulatory synthesizer with continuous acoustic data

Santitham Prom-on [1,2], *Peter Birkholz* [3], *Yi Xu* [2]

[1] Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Thailand
[2] Department of Speech, Hearing and Phonetic Sciences, University College London, United Kingdom
[3] Department for Phoniatrics, Pedaudiology and Communication Disorders, University Hospital Aachen and RWTH Aachen University, Germany

santitham@cpe.kmutt.ac.th, peterbirkholz@gmx.de, yi.xu@ucl.ac.uk

Abstract

This paper reports preliminary results of our effort to address the *acoustic-to-articulatory inversion* problem. We tested an approach that simulates speech production acquisition as a distal learning task, with acoustic signals of natural utterances in the form of MFCC as input, VocalTractLab — a 3D articulatory synthesizer controlled by target approximation models as the learner, and stochastic gradient descent as the training method. The approach was tested on a number of natural utterances, and the results were highly encouraging.

Index Terms: articulatory synthesis, embodiment constraint, target approximation, acoustic-to-articulatory inversion

1. Introduction

Speaking is one of the most complex human skills. To produce a normal speech utterance as simple as "Good morning", a person has to generate a quick succession of highly variable articulatory movements, each involving simultaneous actions of multiple articulators [1,2], and all of them coordinated in such a way that multiple layers of meanings are simultaneously encoded [3]. A human child, however, is able to acquire this highly intricate skill without formal instructions, and without direct observation of the articulators of the skilled speakers other than the visible ones such as the lips. The only sure input the child receives is the acoustics of the speech utterances. How, then, can a child learn to control her own articulators to produce speech in largely the same way as the model speakers? The present paper reports results of our preliminary effort to answer this question, which is also known as the *acoustic-to-articulatory inversion* problem [4-12].

Different approaches have been proposed to achieve acoustic-to-articulatory inversion [4-12]. These methods rely on either explicit mapping between acoustic and articulatory data [4-10], or optimization of articulatory synthesis model parameters [11-12]. The latter is an analysis-by-synthesis strategy, which iteratively adjusts parameters of a forward model to minimize the cost function. This strategy, with the implementation of forward models and supervised learning, has the potential of achieving the closest simulation of speech learning behavior.

In this study we tested a new approach to acoustic-to-articulatory inversion by simulating speech production acquisition as a distal learning task [13], with acoustic signals of natural utterances in the form of Mel-Frequency Cepstral Coefficients (MFCC) as sole input, VocalTractLab — a 3D articulatory synthesizer controlled by target approximation models [14-16] as the learner, and an iterative analysis-by-synthesis optimization as the training regimen. We have tested this approach on a number of natural utterances, and have seen encouraging results.

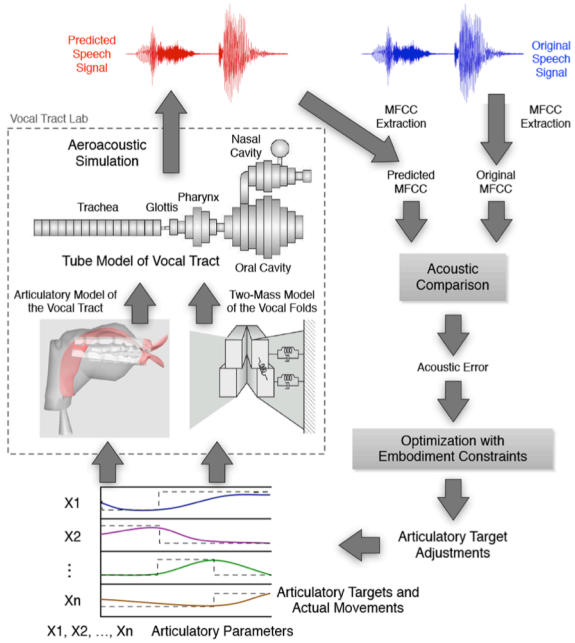

Figure 1: *Workflow diagram of the current approach.*

2. Method

Figure 1 is a schematic diagram of our method, which consists of VocalTractLab, a target approximation model, analysis-by-synthesis optimization and embodiment constraints. The following sections will go into details of each component.

2.1. The articulatory synthesizer

VocalTractLab is an articulatory synthesizer capable of generating a full range of speech sounds by controlling vocal tract shapes, aerodynamics and voice quality [14-16]. It consists of a detailed 3D model of the vocal tract that can be configured to fit the anatomy of any specific speaker and simulate growth, an advanced self-oscillating model of the vocal folds and an efficient method for the aeroacoustic simulation of the speech signal. The acoustic simulation method of the model is both stable and accurate and allows the synthesis of not only voiced sounds, but also aspiration and frication noise for fricatives and plosives.

2.2. Sequential target approximation

The control of the dynamics of VocalTractLab is based on the concept of sequential target approximation (TA), which has previously been implemented in various forms, as reviewed in [15]. TA assumes that continuous articulatory trajectories are composed of successive, non-overlapping movements, each approaching an underlying target. TA thus shares some similarities with the task dynamic model [17], but differs from it in having an explicit state transfer mechanism, which enables the simulation of extensive carryover influence, target undershoot, smooth transitions across movement boundaries, and reduced degrees of freedom due to absence of movement overlap. The TA model implemented in VocalTractLab is illustrated in Figure 2.

A key advantage of TA is that it allows the mapping of variant surface trajectories due to phonetic context, stress, speech rate, etc., to a single invariant target [2,15,18], which simplifies the problem of inverse mapping from acoustics to articulation. This different from the DIVA framework that defines targets as context-sensitive regions rather than invariant points [19, 20]. To a child trying to acquire adult-like speech, this means that for each phonetic unit a single target or a single compound target can be learned from its many context-sensitive realizations. The feasibility of this approach has been seen in our recent work on F_0 modeling [21-23]. A critical strategy in the present implementation of the TA model is an unconventional segmentation method. That is, a segmental interval is defined as the time period during which its canonical pattern is unidirectionally approached. As a result, the point where a segment best approximates its canonical pattern is marked as its offset rather than center, as shown in all the figures with segmental annotations in this paper.

2.3. Optimization via analysis-by-synthesis

The distal learning in the present study is achieved by representing surface acoustics of both natural and synthetic speech with MFCC and using the difference between the two as errors in the optimization of the articulatory targets. Each set of articulatory targets is hosted by a segmental interval, whose boundaries are *manually* defined before optimization. For each segment, articulatory targets in the form of vocal tract shapes are optimized iteratively to minimize the total sum of square errors of MFCC between original and synthesized sounds, which can be described as follows:

$$E = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(c_{ij} - \hat{c}_{ij}\right)^2 \quad (1)$$

Here, n is the number of time frames, m is the number of MFCC coefficients, and c_{ij} and $\hat{c}_{ij}$ are the jth cepstral coefficient of the ith frame in the natural and synthesized utterances, respectively.

For each segmental interval, articulatory target parameters associated with a vocal tract shape are randomly initialized. There are 23 vocal tract parameters in total [14], including HX, HY (Horz. and vert. hyoid positions), JX (Horz. jaw position), JA(Jaw angle), LP (Lip protrusion), LD (Vert. lip distance), VS (Velum shape), VO (Velum opening), TTX, TTY (Horz. and vert. tongue tip positions), TBX, TBY (Horz. and vert. tongue blade positions), TCX, TCY (Horz. and vert. tongue body positions), TRX, TRY (Horz. and vert. tongue root positions), TS1 – TS4 (Tongue side elevation at 4

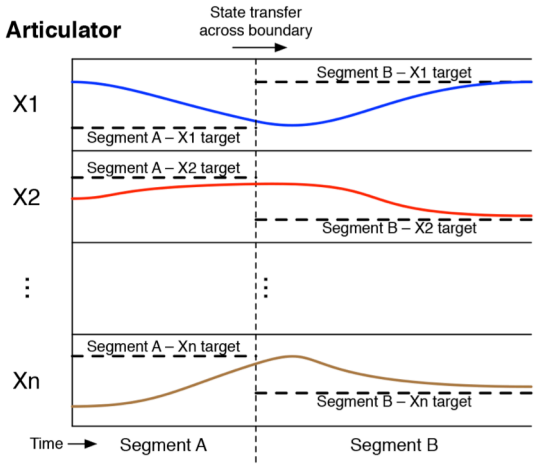

Figure 2: *Target approximation model for controlling articulatory movements.*

positions), MA1 (Min. area, tongue body), MA2 (Min. area, tongue tip), MA3 (Min. area ,teeth-lips)

For each vocal tract shape, its parameters are iteratively adjusted to minimize the error function using a stochastic gradient method. Only adjustments that result in a lower error than the previous trial are accepted. This process was repeated until either the total error converged, or an upper limit of iterations was reached.

2.4. Embodiment constraints

Some articulatory parameters may not be entirely independent of others. For examples, the tongue parameters were found to be positively correlated in articulatory movements for certain places of articulation such as alveolar, palatal and velar places [24]. This relationship suggests that there should be a constraint weakly tying these parameters together so that the changes in one parameter also affect others according to the physiological locations. This embodiment constraint will therefore make the parameter adjustment in the optimization process more realistic. In this paper, we modeled this constraint by co-adjusting the articulators located near the articulator under adjustment. For example, whenever the tongue blade parameters (TBX/TBY) were adjusted, those of tongue tip and tongue body (TTX/TTY, TCX/TCY) were also modified by a small amount (20%) relative to the main adjustment.

3. Results

3.1. Vowel coarticulation

Figure 3 shows spectrograms of a vowel sequence /a: i: u: e: o:/ produced by a male speaker at a normal speed and that of a synthetic one generated with optimized articulatory targets. Note that each vowel is annotated to terminate at a point where its target is best achieved, so that the formants in each segment move unidirectionally toward an ideal pattern. Smooth formant transitions from one vowel to another can be observed in the lower panel, just as in the upper panel. The similarity between the two spectrograms indicates a close approximation of the estimated articulatory targets (Pearson's correlation: ρ_{F1} = 0.90, ρ_{F2} = 0.96, ρ_{F3} = 0.52). The low correlation of F3 is due to the lack of F3 raising in /i/ of the synthetic utterance. The smooth synthetic formant movements are thanks to the

TA dynamics of all the articulators involved. Note, however, that the shuffling of formants in the original speech during a: and i: transition is not simulated in the synthetic speech. Such formant shuffling being speaker-specific and inaudible, as explained in some detail by Stevens [26], is not the simulation aim of our training process. Perceptual inspections by the authors indicated that the synthetic vowels sounded close to the original, except that /a:/ sounded a bit schwa-like as the low F1 also suggests. This is because its duration is the shortest and it contains less articulatory movement compared to other vowels, which possibly make it less important in terms of its contribution to the total error.

3.2. Implicit speaker normalization

Speaker normalization is generally considered as a critical step in acoustic-to-articulatory inversion [10-13]. The need for such normalization is even more obvious in children's speech acquisition, given the large child-adult differences in articulatory dimensions. Here, we tested whether the present optimization method is able to normalize speaker differences. Figure 4 shows the results of the optimization of a *male* articulatory setting based on the word "aware" spoken by a *female* American English speaker. As can be seen, the formant patterns of the two spectrograms look similar, including the convergence of F3 and F2 during the retroflex /r/ (Correlation: $\rho_{F1} = 0.72$, $\rho_{F2} = 0.95$, $\rho_{F3} = 0.85$). It should be noted that the moderate correlation in F1 is due to the lack of F1 variation in the original material. The female voice on the top not only has relatively high F_0, as indicated by the close distances between the vertical striations, but also relatively higher formant frequencies (max F2 = 2396 Hz) than that of the learned male voice (max F2 = 1832 Hz).

However, no explicit normalization strategies were used in this simulation. What seems to have enabled the learning across vocal tracts of rather different dimensions is the TA model. Despite the individual differences, formant trajectories resulting from TA bear sufficient resemblance to the original to allow the finding of optimal parameters even when the model dimensions differ. The implementation of TA as the basic dynamic control of the articulatory movements thus seems to have also helped the achievement of implicit speaker normalization. The effectiveness of the implicit normalization is also seen in the EMA comparison to be described next.

3.3. EMA comparison

Beside acoustic comparison, another way to examine the quality of the distal learning is to compare the learned articulatory shapes and trajectories with electromagnetic articulography (EMA) data of the original utterance, i.e., to see how well acoustic-to-articulatory inversion is achieved. Two four-syllable utterances of German vowel sequences, /ˌjaja'jaja/ and /ˌjOjO'jOjO/ produced by a female speaker for a previous study [25] were used as testing materials. These sequences were spoken with secondary stress on the first syllable and primary stress on the third syllable. Articulatory trajectories of multiple sensor coils were recorded by means of electromagnetic articulography at a sampling rate of 200 Hz, of which we considered three sensors on the tongue tip, tongue blade (mid), and tongue back.

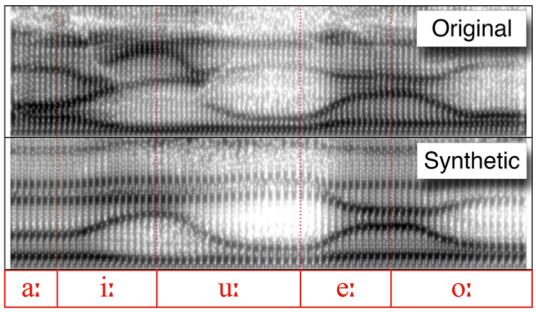

Figure 3: *Spectrograms of original and synthetic vowel sequences /a: i: u: e: o:/.*

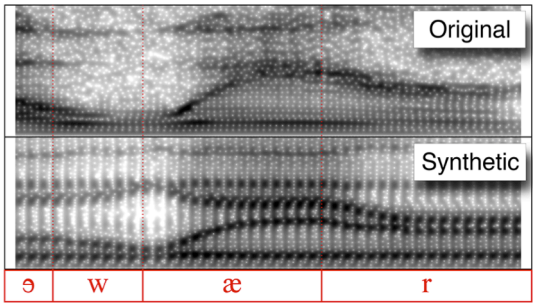

Figure 4: *Spectrograms of "aware" in American English. The synthetic one was generated by VocalTractLab with parameters learned through analysis-by-synthesis of the original, with a male articulatory setting.*

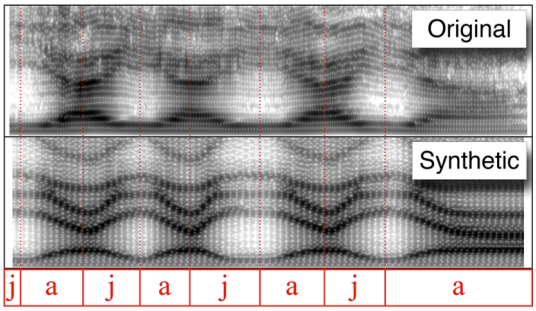

Figure 5: *Spectrograms of original and synthetic /ˌjaja'jaja/.*

In each utterance, articulatory target parameters for /j/, /a/ and /O/ were trained for all instances of these phones and used in different locations. Spectral comparison shown in Figure 5 indicates that the acoustics of the synthetic /ja/ sequence closely approximates the original (Correlation: $\rho_{F1} = 0.95$, $\rho_{F2} = 0.93$, $\rho_{F3} = 0.70$). As verifications, the learned targets were also used to synthesize the corresponding sounds in isolation to confirm that they were perceptually accurate. Figure 6 shows the learned targets for /j/ and /a/ in the first utterance, and for /j/ and /O/ in the second utterance. For comparison, the gray contours show the corresponding vocal tract shapes measured by MRI from the speaker whose vocal tract was modeled in VocalTractLab [14].

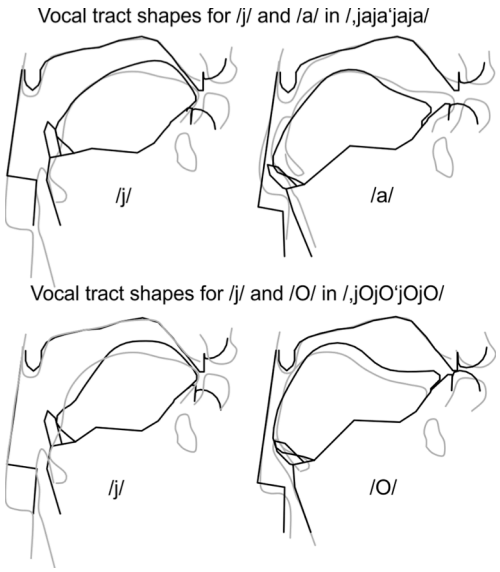

Figure 6: *The black shapes are estimated vocal tract targets for /j/ and /a/ in /ˌjajaˈjaja/, and /j/ and /O/ in /ˌjOjOˈjOjO/ For comparison, the gray contours show measured vocal tract shapes for /j/, /a/, and /O/ of the real speaker that was modeled in VocalTractLab.*

Figure 7 shows the comparison of original and synthetic EMA trajectories of the utterance /ˌjajaˈjaja/. Three EMA sensors placed along the tongue surface are compared. The EMA trajectories of the synthetic utterance were derived from movements of specific vertices of the tongue model that correspond to the EMA sensor locations of the natural speaker. Original and learned articulatory trajectories as shown in Figure 7. The root-mean-square errors (RMSE), which indicate the average distance between original and synthetic contours, of x- and y-position are (1.9 mm, 2.1 mm) for the tongue tip sensor, (3.6 mm, 2.6 mm) for the tongue mid sensor, and (4.1 mm, 4.0 mm) for the tongue back sensor. Comparing the similarity between the contours by correlation coefficients of x- and y-position shows that tongue tip positions have correlations of (0.88, 0.83), tongue mid positions have correlations of (0.83, 0.87), and tongue back position have correlation of (0.87, 0.81).

4. Discussion and Conclusions

The preliminary results of the present study are very encouraging. They have shown that it is possible to simulate speech acquisition as a distal learning process, with surface acoustics of continuous speech and predefined annotated segmental boundaries as the input, an articulatory synthesizer controlled by target approximation models as the learner, and analysis-by-synthesis optimization assisted by embodied constraints as the training regimen. The simulation tests show that underlying articulatory targets can be learned this way to generate utterances that resemble the original both acoustically, as shown in Figures 3-5, and articulatorily, as shown in Figures 6 and 7, thus largely completing the acoustic-to-articulatory inversion process with close acoustic matching.

The current results are still very preliminary, however. Though having shown implicit speaker normalization across genders, we have not yet tested the effectiveness of such normalization with a child vocal tract. Also all the consonants

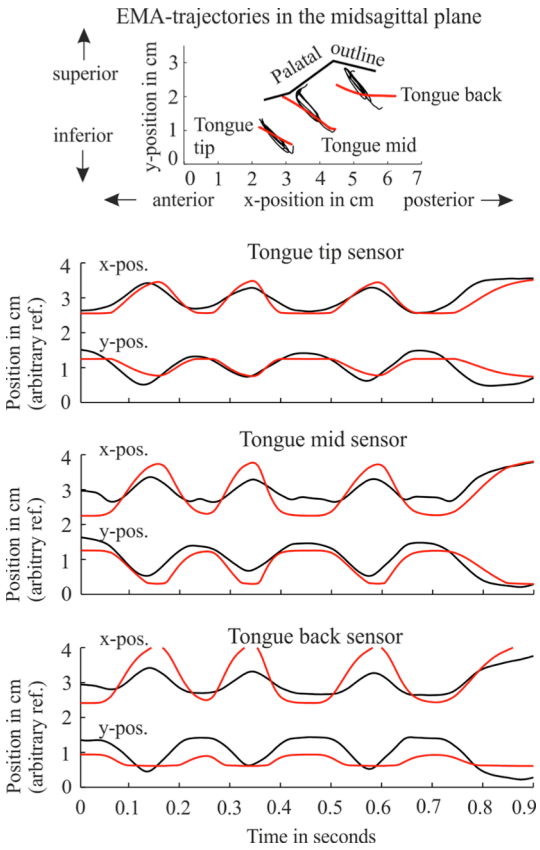

Figure 7: *EMA trajectories of three sensors on the tongue for the utterance /ˌjajaˈjaja/. Original trajectories are shown as black lines and simulated*

simulated so far are glides, which presumably have minimal gestural overlap with adjacent vowels. Strategies have yet to be developed to simulate the learning of overlapped CV gestures. Also, human learners are likely to have far more embodied constraints than the ones we have implemented in this study. A case in point is the seemingly excessively raised glottis in /j/ as compared to the measured configuration in Figure 6. Such a large discrepancy may require some additional constraints on the glottis position. Alternatively it might have been necessary to raise the glottis so far to approximate the acoustics of the natural female speaker (having a shorter vocal tract) with the male vocal tract model of the synthesizer. Finally, it should be noted that the acoustic-to-articulatory inversion in the current study is not fully complete, as the division of continuous utterances into discrete unidirectional movements is done manually. The underlying assumption is that the learning of perceptual segmentation is achieved prior to the learning of the articulatory targets. But the validity of this assumption is not fully established, and has to be addressed in future studies.

Examples of the original and synthetic sounds, and video animation of the learning progress, can be found in the supplementary material.

5. Acknowledgements

We would like to thank the Royal Academy of Engineering for support through the Newton International Fellowship Alumni follow-on funding. We are also grateful to Sascha Fagel for providing the EMA data used in this study.

6. References

[1] Mermelstein, P. “Articulatory model for the study of speech production”, J. Acoust. Soc. Am., 53(4):1070–1082, 1973..

[2] Saltzman, E. L. and Munhall, K. G., “A dynamical approach to gestural patterning in speech production”, Ecol. Psychol., 1:333-382, 1989.

[3] Xu, Y., “Speech melody as articulatorily implemented communicative functions”, Speech Commun., 46:220-251, 2005.

[4] Hofer, G., Yamagishi, J. and Shimodaira, H. “Speech-driven lip motion generation with a trajectory HMM,” in Proc. Interspeech 2008, Brisbane, Australia, pp. 2314–2317, 2008.

[5] Tamura, M., Kondo, S., Masuko, T.and Kobayashi, T. “Text-to-visual speech synthesis based on parameter generation from HMM,” in Proceedings of ICASSP 98, 1998, pp. 3745–3748.

[6] Uria, B., Renal, S. and Richmond, K. “A Deep neural network for acoustic-articulatory speech inversion”, in Proc. NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning, Sierra Nevada, Spain, 2011.

[7] Schroeter, J. and Sondhi, M. M. “Dynamic programming search of articulatory codebooks”, Proc. ICASSP 1989, Glasgow, UK, vol. 1, pp. 588–591, 1989.

[8] Ghosh, P. K. and Narayanan, S. “A generalized smoothness criterion for acoustic-to-articulatory inversion”, J. Acoust. Soc. Am., 128(4):2162-2172, 2010.

[9] Ouni, S. and Laprie, Y. “Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion”, J. Acoust. Soc. Am., 118(1):444-460, 2005.

[10] Potard, B., Laprie, Y. and Ouni, S. “Incorporation of phonetic constraints in acoustic-to-articulatory inversion”, J. Acoust. Soc. Am., 123(4):2310-2323, 2008.

[11] Panchapagesan, S. and Alwan, A. “A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the Maeda articulatory model”, J. Acoust. Soc. Am., 129(4):2144-2162, 2011.

[12] McGowan, R. “Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model test”, Speech Commun., 14:19-48, 1994.

[13] Jordan, M. I. and Rumelhart, D. E. “Forward models: Supervised learning with a distal teacher,” Cognitive Sci. 16:316-354, 1992.

[14] Birkholz, P. “Modeling consonant-vowel coarticulation for articulatory speech synthesis”, PLOS ONE, in-press.

[15] Birkholz, P., Kroger, B. J. and Neuschaefer-Rube, C. “Model-based reproduction of articulatory trajectories for consonantal-vowel sequences”, IEEE Audio, Speech and Language Process., 19(5):1422-1433, 2011.

[16] Birkholz, P., Kröger, B. J. and Neuschaefer-Rube, C. “Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis”, Proc. Interspeech 2011, 2681-2684, 2011.

[17] Saltzman, E. K. and Munhall, K. G. “A dynamical approach to gestural patterning in speech production”, Ecol. Psychol., 1:333-382.

[18] Xu, Y. and Wang, Q. E. “Pitch targets and their realization: Evidence from Mandarin Chinese”. Speech Commun., 33:319-337, 2001.

[19] Guenther, F. H. and Vladusich, T. “A neural theory of speech acquisition and production”, J. Neurolinguist., 25:402-422, 2012.

[20] Guenther, F. H. “Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production”, Psychol. Rev., 102:594-621, 1995.

[21] Prom-on, S., Xu, Y. and Thipakorn, B. “Modeling tone and intonation in Mandarin and English as a process of target approximation”, J. Acoust. Soc. Am., 125:405-424, 2009.

[22] Prom-on, S. and Xu, Y. “PENTATrainer2: A hypothesis-driven prosody modeling tool”, Proc. ExLing 2012, Athens, Greece, pp. 93-100, 2012.

[23] Prom-on, S., Liu, F. and Xu, Y. “Post-low bouncing in Mandarin Chinese: Acoustic analysis and computational modeling”, J. Acoust. Soc. Am., 132:421-432, 2012.

[24] Green, J. R. and Wang, Y.-T. “Tongue-surface movement patterns during speech and swallowing”, J. Acoust. Soc. Am., 113(5):2820-2833, 2009.

[25] Fagel, S. *Audiovisuelle Sprachsynthese*. Logos Verlag Berlin, 2004.

[26] Stevens, K. N. *Acoustic Phonetics*. MIT Press, Cambridge, MA.