11th ISCA Speech Synthesis Workshop (SSW 11)
26-28 August 2021, Budapest, Hungary

Extending Text-to-Speech Synthesis with Articulatory Movement Prediction using Ultrasound Tongue Imaging

Tamás Gábor Csapó[1,2]

[1]Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Budapest, Hungary
[2]MTA-ELTE Lendület Lingual Articulation Research Group, Budapest, Hungary
csapot@tmit.bme.hu

Abstract

In this paper, we present our first experiments in text-to-articulation prediction, using ultrasound tongue image targets. We extend a traditional (vocoder-based) DNN-TTS framework with predicting PCA-compressed ultrasound images, of which the continuous tongue motion can be reconstructed in synchrony with synthesized speech. We use the data of eight speakers, train fully connected and recurrent neural networks, and show that FC-DNNs are more suitable for the prediction of sequential data than LSTMs, in case of limited training data. Objective experiments and visualized predictions show that the proposed solution is feasible and the generated ultrasound videos are close to natural tongue movement. Articulatory movement prediction from text input can be useful for audiovisual speech synthesis or computer-assisted pronunciation training.
Index Terms: DNN-TTS, audiovisual synthesis, ultrasound

1. Introduction

Statistical parametric methods are frequently used in text-to-speech (TTS) synthesis, with two main machine learning techniques: hidden Markov-models (HMM, [1]) and deep neural networks (DNN, [2]). Recently, the focus of TTS research has moved to end-to-end solutions, applying neural vocoders (e.g. WaveNet [3] and WaveGlow [4]) and sequence-to-sequence models using attention (e.g. Tacotron2 [5]). Still, traditional (non-end-to-end, vocoder-based) DNN-TTS systems are useful in limited data scenarios, when there is few training data available, for example with speech and biosignal recordings, or in case of audiovisual speech synthesis.

1.1. Audiovisual speech synthesis

Audiovisual speech synthesis is a subfield of the more general areas of speech synthesis and computer facial animation [6]. The goal of the visible speech synthesis is typically to obtain a mask with realistic motions, not to duplicate the musculature of the face to control this mask.

The field of visual speech synthesis is fairly well established and a variety of approaches have been developed (including rule-based [7] and data-driven methods [8]). Rule-based systems include models for speech sequence planning, for muscle mechanisms and for the physical speech production apparatus. Within the biomechanical model of the vocal tract, the tongue can be represented as a finite element mesh [7] and complex biomechanical simulations are necessary to estimate the internal muscle stresses during the movement of human articulators [9]. In the context of data-driven approaches and HMM-based synthesis, there are two main categories: image-based systems are supposed to look like a video of a real person, while motion capture based systems derive features from facial points tracked over time [8]. For HMM-based audiovisual synthesis, a synchronous corpus of parametrized facial motion data and acoustic speech data is necessary. Schabus et al. showed that in combined HMM-based text-to-speech synthesis and facial animation, joint audiovisual models perform better than training separate acoustic and visual models [8].

1.2. Predicting articulatory movement from text

Another type of TTS extension is when the target is to predict articulatory motion (e.g. lip or tongue movement) and not just the face of the speaker, besides the speech output. This requires special biosignals to be recorded, which can track the movement of the articulatory organs (e.g. EMA, X-ray, vocal tract MRI, and ultrasound tongue imaging). With such a system, by giving an arbitrary input text, one is able to hear the speech and, in synchrony with it, see how to move the tongue in 3D to produce target speech sounds. This visual feedback can make a big difference for pronunciation training in L2 learning, especially when the target language contains speech sounds which are difficult to articulate.

Most earlier studies in this context were using point-tracking devices, like electromagnetic articulography (EMA) [10, 11, 12, 13, 14, 15]. Ling and his colleagues proposed a HMM-based text-to-articulatory movement prediction system, i.e. which can synthesize the speaker's mouth from text [10]. Here, the durations were not modeled, but in a subsequent study, they also investigated the timing aspects and analyzed the critical articulators [11]. Wei et al. used DNNs for the text-to-EMA prediction and confirmed that stacked bottleneck features are effective for this purpose [12]. Steiner and his colleagues experimented similarly with text-to-EMA prediction using HMMs (with synchronous text-to-speech), and they also included a geometric 3D tongue model as the target [13]. Next, they compared HMMs and DNNs for the text-to-tongue model prediction [14]. It was found that with less than 2 hours of data, DNNs outperformed HMMs. Yu and her colleagues predicted 3D articulatory movement, from text and audio inputs, therefore combining the text-to-speech and acoustic-to-articulatory inversion fields [15]. For the machine learning approach, they used a bottleneck long-term recurrent convolutional neural network. They showed that the text information complements well the acoustic features during the prediction of EMA-based articulation. The final output of the system is speech synchronized with 3D articulatory animation, using a facial mesh model [15].

As shown above, there have been several studies investigating text-to-articulatory motion with HMMs or DNNs, but all of

10.21437/SSW.2021-2

them are using point-tracking equipment (electromagnetic articulography). Medical imaging target, like ultrasound or MRI, have not been used before in this context.

1.3. Ultrasound tongue imaging

Ultrasound tongue imaging (UTI) is a technique suitable for the acquisition of articulatory data. Phonetic research has employed 2D ultrasound for a number of years for investigating tongue movements during speech [16]. Stone summarized the typical methodology of investigating speech production using ultrasound [17]. Usually, when the subject is speaking, the ultrasound transducer is placed below the chin, resulting in midsagittal images of the tongue movement. The typical result of 2D ultrasound recordings is a series of gray-scale images in which the tongue surface contour has a greater brightness than the surrounding tissue and air. Compared to other articulatory acquisition methods (e.g. EMA, X-ray, XRMB, and vocal tract MRI), UTI has the advantage that the tongue surface is fully visible, and ultrasound can be recorded in a non-invasive way [17, 18, 19]. An ultrasound device is easy to handle and move, since it is small and light, and thus it is suitable for fieldworks, as well. Besides, it is a significantly less expensive piece of equipment than the above mentioned devices.

In our earlier studies, we have shown that ultrasound is a feasible solution for articulatory-to-acoustic mapping [18, 20] and acoustic-to-articulatory inversion [21]. However, text-to-ultrasound synthesis has not been investigated before.

1.4. Contributions of this paper

The goal of this paper is to extend DNN-TTS with articulatory movement prediction, using ultrasound images of the tongue. We show on the data of several speakers that the combined TTS and synthesized articulatory motion is feasible and can result in acceptable articulatory movement video. Text-to-articulatory movement prediction might be useful for computer-assisted pronunciation training (CAPT) applications and articulatory visualization.

2. Methods

2.1. Data

For the data, we used the UltraSuite-TaL80 database [22] (`https://ultrasuite.github.io/data/tal_corpus/`). We chose four English male (03mn, 04me, 05ms, 07me) and four female speakers (01fi, 02fe, 06fe, and 09fe). In parallel with speech, the tongue movement was recorded in midsagittal orientation using the "Micro" ultrasound system of Articulate Instruments Ltd. at 81.5 fps. Lip video was also recorded in UltraSuite-TaL80, but we did not use that information in the current study. The ultrasound data and the audio signals were synchronized using the tools provided by Articulate Instruments Ltd. Each speaker read roughly 200 sentences – the duration of the recordings was about 15 minutes, which was partitioned into training, validation and test sets in a 85-10-5 ratio.

2.2. Processing the ultrasound data

In our experiments, articulatory features estimated from the raw scanline data of the ultrasound were used as the additional target of the networks (see Fig. 1). The 64×842 pixel images were resized to 64×128 pixels using bicubic interpolation, and we calculated PCA coefficients, similarly to EigenTongues [23]. While calculating the PCA, we aimed at keeping the 70% of the variance of the original images, thus having 128 coefficients. An example for the PCA eigenvectors can be seen in Fig. 2, and the result of PCA is presented in Fig. 4. To be in synchrony with the acoustic features (frame shift of 5 ms), the ultrasound data was resampled to 200 Hz.

2.3. DNN-TTS framework

Fig. 1 illustrates the proposed approach, i.e. the combined acoustic and articulatory feature prediction using a DNN from text input. The experiments were conducted in the Merlin DNN-TTS framework [24] (`https://github.com/CSTR-Edinburgh/merlin`). Textual / phonetic parameters are first converted to a sequence of linguistic features as input (based on a decision tree). Next, neural networks are employed to predict acoustic (60-dimensional MGC, 5-dimensional BAP, and 1-dimensional LF0, with delta and delta-delta features) and articulatory features (128-dimensional ULT-PCA, with delta and delta-delta) as output for synthesizing speech, at a 5 ms frame step with the WORLD vocoder. From the predicted 128-dimensional articulatory features, the 64×128 image is reconstructed using the PCA matrix, and bicubic interpolation is applied to resize the image to 64×842 pixels, to be comparable with the original data. For visualization purposes, we transformed this raw scanline data to wedge format, which shows how the real aspect ratio of the tongue surface (for an example, see Fig. 4). The transformation was done with 'ultrasuite-tools' (`https://github.com/UltraSuite/ultrasuite-tools`)

2.3.1. FC-DNN

The DNN used here is a fully-connected feed-forward multilayer perceptron architecture (FC-DNN, six hidden layers, 1024 neurons in each). We applied tangent hyperbolic activation function, SGD optimizer, and a batch size of 256. The input features had min-max normalization, while output acoustic features had mean-variance normalization. We trained the networks for 25 epochs with a warm-up of 10 epochs, applying early stopping, and a learning rate of 0.002 after that with exponential decay. We only trained both a duration model and an acoustic model, the latter also containing the articulatory features.

2.3.2. LSTM

Recurrent networks are typically more suitable to process sequential data. Therefore, we also trained an LSTM network following the Merlin recipe (four FF layers with 1024 neurons each, and one LSTM layer with 512 neurons). To ensure a longer training with the recurrent network, we used ADAM optimizer, and a warm-up of 30 epochs with early stopping. The other parameters were the same as for the FF-DDN. We trained both duration and acoustic models.

All neural network trainings were done individually with each speakers' data, without average voice training or adaptation.

3. Experimental results

After training the above models, we synthesized sentences from the test parts of the ultrasound datasets. To measure the validation and test error, we calculated both spectral prediction error (Mel-Cepstral Distortion, MCD), and an articulatory feature

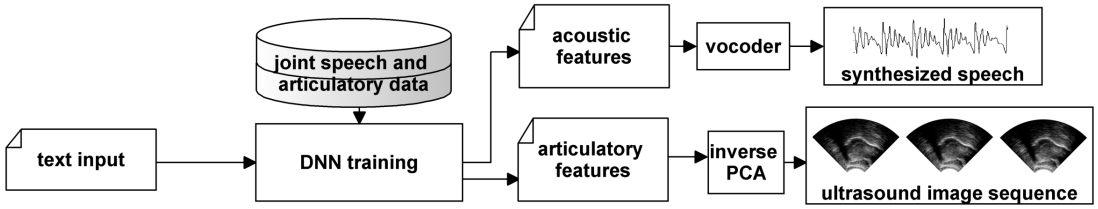

Figure 1: *Block diagram of the proposed approach.*

related error (ULT-PCA/RMSE, calculated on the normalized PCA values). We trained both duration and acoustic models, but for the error calculations, we synthesized the test sentences with their original timing. This way, warping the features in time was not necessary for calculating the error measures.

3.1. Demonstration samples

An example for the PCA eigenvectors are in Fig. 2, and the predicted articulatory feature sequence can be seen in Fig. 3 (speaker '01fi', sentence '201_aud'). As lower dimensional PCA vectors contain more information, the visualization was done in an exponential way and only 8 dimensions are plotted out of 128. Clearly, both the FC-DNN and the LSTM are following the tendencies found in the original data (e.g. in case of PCA-1, PCA-2, PCA-4), but the fine details are not modeled well. This type of oversmoothing is a frequent phenomena in statistical parametric speech synthesis. The higher dimensions (e.g. PCA-64 and PCA-128) are almost constant; i.e. they could not be modeled well with neural networks.

To visualize the individual ultrasound images, we plotted several ultrasound frames from the original videos and from the predicted ones, as a function of time, in Fig. 4. The reason why we are plotting every third frame is that for the 5 ms frame step of the Merlin toolkit, the 81.5 fps ultrasound video was interpolated to 200 Hz, and therefore, in the predicted data, roughly every 3rd frame contains visible tongue motion. In case of speaker '01fi', we can see in the top row (original ultrasound images after PCA) that there is a significant tongue movement, i.e. the tongue tip (on the right) goes higher, as the time passes. Both the predictions with the DNN and LSTM network follow the articulatory movement, but the images are smoothed – again, resulting from the statistical training. For speaker '03mn', similar tendencies can be observed: the movement of the tongue is changing its curvature as a function of time, but in the DNN-predicted and LSTM-predicted images, the tongue surface is not as clear as in the original data.

As the synthesized motion of the tongue is more visible in real-time videos, we made available several samples at `http://smartlab.tmit.bme.hu/ssw11_txt2ult`.

3.2. Objective measures

Table 1 summarizes the MCD results. The MCD values of the test sentences with the FC-DNN are between 5.8–7.0 dB (average: 6.228 dB), whereas with LSTM they are between 6.0–7.5 dB (average: 6.593 dB), indicating that the recurrent neural network was not helpful in estimating the acoustic features. The reason for this might be that we have limited articulatory-acoustic databases (roughly 200 sentences for each speaker), which is too small for training an LSTM model.

The results of the RMSE calculated on the articulatory feature are summarized in Table 2. The lowest error was achieved with the data of speaker 09fe: with FC-DNN, the test error is 2.9, while with LSTM, the test error is 3.1. The tendency is similar to the case of MCD: the LSTM network was not helpful in predicting the articulatory features, probably due to the small size of the data.

Table 1: *MCD errors on the dev/test set.*

	MCD	
Speaker	FC-DNN	LSTM
01fi	6.995 / 6.971	6.647 / 6.588
02fe	6.095 / 5.803	6.486 / 6.259
03mn	5.781 / 5.785	5.977 / 5.948
04me	5.896 / 6.024	6.318 / 6.312
05ms	6.244 / 6.256	7.235 / 7.083
06fe	5.758 / 5.582	6.444 / 6.330
07me	6.589 / 6.562	6.831 / 6.749
09fe	6.516 / 6.844	7.197 / 7.472
average	6.234 / 6.228	6.642 / 6.593

Table 2: *ULTPCA/RMSE errors on the dev/test set.*

	ULTPCA128/RMSE	
Speaker	FC-DNN	LSTM
01fi	3.292 / 3.223	3.319 / 3.208
02fe	3.533 / 3.732	3.753 / 3.904
03mn	3.147 / 3.660	3.289 / 3.680
04me	3.849 / 3.985	4.031 / 4.033
05ms	3.133 / 3.233	3.249 / 3.405
06fe	3.439 / 3.250	3.743 / 3.451
07me	3.544 / 3.595	3.498 / 3.461
09fe	3.022 / 2.864	3.234 / 3.133
average	3.370 / 3.443	3.515 / 3.534

4. Discussion and conclusions

We have shown above that text-to-ultrasound video prediction is feasible as an extension to traditional DNN-based text-to-speech synthesis, despite the relatively small amount of training data. Although the synchrony between visual and speech output is not enforced by the model, the tied acoustic and articulatory features during the DNN training ensure that the audio and visual features are in synchrony, i.e. that in the generated ultrasound videos, the tongue is moving according to the synthesized speech. To objectively check this, SyncNet, part of Wav2Lip could be applied to assess synchrony [25]. Our paper found that the joint learning of both acoustic and articulatory features has advantages, but this is not substantiated – a comparison of the joint model against two separate models remains

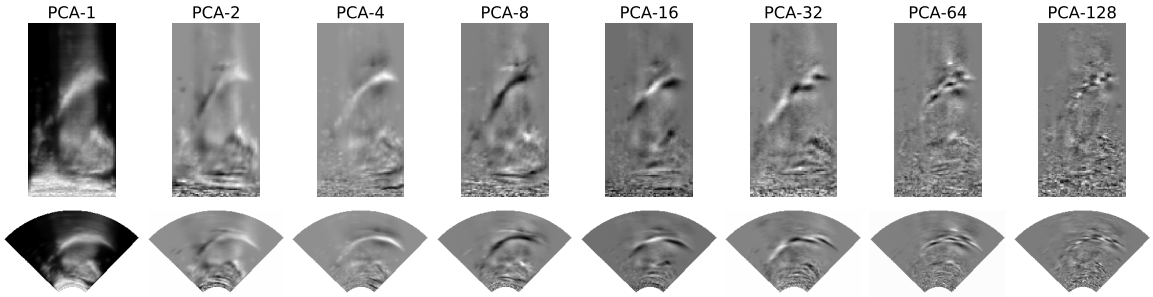

Figure 2: *PCA eigenvectors, from speaker '01fi'. Top: raw, scanline data (resized to 64×128 pixels). Bottom: wedge orientation.*

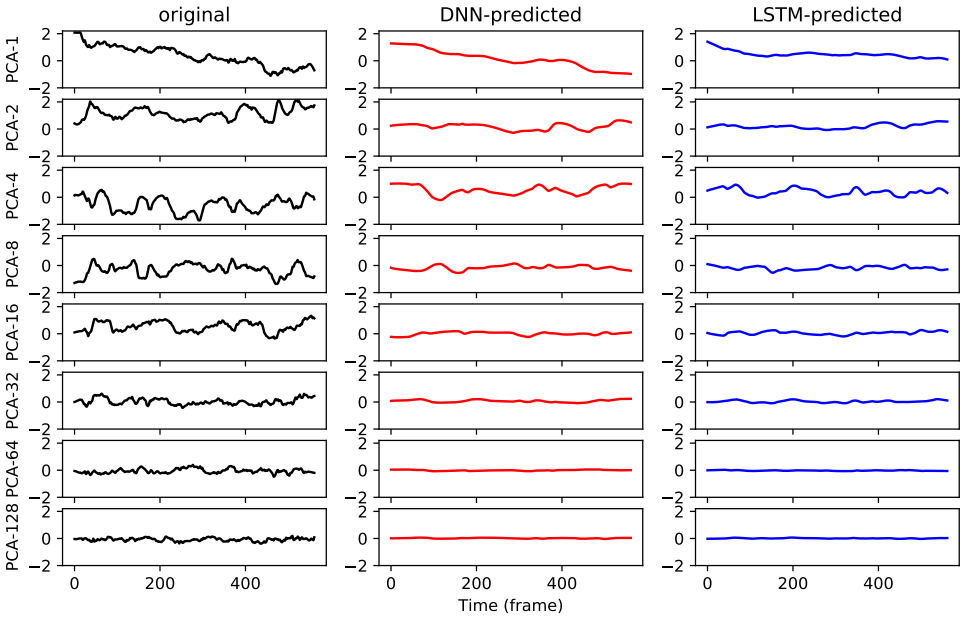

Figure 3: *Original and predicted articulatory features, from speaker '01fi'. Sentence: '"I leave it to nobody," said Shakespeare, sulkily.'*

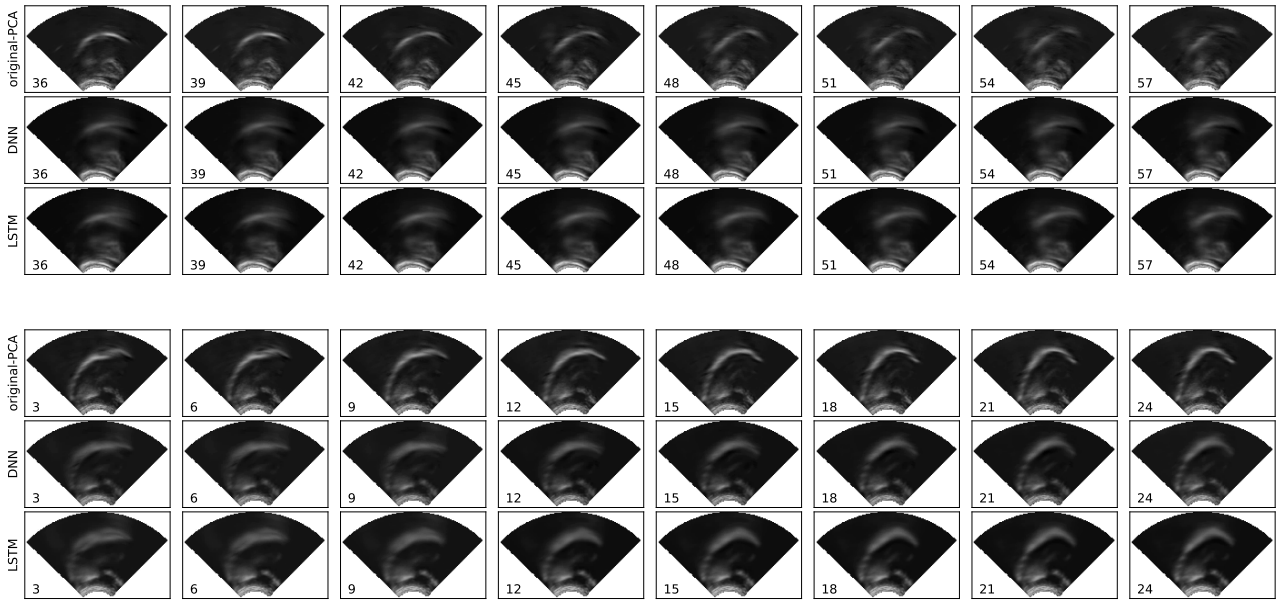

Figure 4: *Original and predicted articulatory feature sequences. Top: speaker '01fi', bottom: speaker '03mn'. The numbers at the bottom-left correspond to the frame number of the video.*

future work.

Although there have been several earlier attempts for extending text-to-speech synthesis with articulatory data, all of these studies were using EMA, being a point tracking equipment [10, 11, 12, 13, 14, 15], and containing less spatial information about the tongue than ultrasound. The advantage of ultrasound in this context is that the resulting video shows a larger portion of the tongue, compared to EMA.

Articulatory movement prediction from text input can be useful for audiovisual speech synthesis. A specific application is computer-assisted pronunciation training / computer-aided language learning [26, 27, 28], which can be beneficial for learners of second languages. With such a combined TTS and text-to-articulatory prediction system, by giving an arbitrary input text, one is able to hear the speech and, in synchrony with it, see how to move the tongue in 2D or 3D to produce target speech sounds. This visual feedback can be helpful for pronunciation training in L2 learning, especially when the target language contains speech sounds which are difficult to articulate.

In the future, we plan to investigate speaker adaptation and speaker-independent training. For this, a common articulatory space has to be defined, as the currently used PCA representation is specific for each individual speaker. Also, multi-task learning might be useful in this context: a system could potentially be pre-trained on speech-only material which is easier to acquire, and the UTI be trained only in addition. Besides, we plan to investigate the effect of the misalignments in the ultrasound transducer position [29, 30] on the text-to-ultrasound prediction results.

The code is accessible at `https://github.com/BME-SmartLab/txt2ult`.

5. Acknowledgements

The author was partly funded by the National Research, Development and Innovation Office of Hungary (FK 124584 and PD 127915 grants). We would like to thank CSTR for providing the Merlin toolkit and the UltraSuite-TaL articulatory database.

6. References

[1] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, and A. Black, "The HMM-based speech synthesis system version 2.0," in *Proc. ISCA SSW6*, Bonn, Germany, 2007, pp. 294–299.

[2] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, Vancouver, Canada, 2013, pp. 7962–7966.

[3] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *CoRR*, vol. abs/1609.0, 2016.

[4] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A Flow-based Generative Network for Speech Synthesis," in *Proc. ICASSP*, Brighton, UK, 2019, pp. 3617–3621.

[5] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *Proc. ICASSP*, Calgary, Canada, 2018, pp. 4779–4783.

[6] D. W. Massaro, M. M. Cohen, M. Tabain, J. Beskow, and R. Clark, "Animated speech: research progress and applications," in *Audiovisual Speech Processing*, G. Bailly, P. Perrier, and E. Vatikiotis, Eds. Camridge, UK: Cambridge University Press, 2012, pp. 309–345.

[7] P. Perrier, ""GEPPETO": A target-based model of speech production including optimal planning and physical modeling," in *Adventures in Speech Science*, Tokyo, Japan, 2014.

[8] D. Schabus, M. Pucher, and G. Hofer, "Joint audiovisual Hidden Semi-Markov Model-based speech synthesis," *IEEE Journal on Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 336–347, 2014.

[9] I. Stavness, M. A. Nazari, F. Cormac, P. Perrier, Y. Payan, J. Lloyd, and S. Fels, "Coupled Biomechanical Modeling of the Face, Jaw, Skull, Tongue, and Hyoid Bone," in *3D Multiscale Physiological Human*, R. O. C. H. F. E. Magnenat-Thalmann Nadia, Ed. Springer London, 2014, pp. 253–274.

[10] Z.-H. Ling, K. Richmond, and J. Yamagishi, "An Analysis of HMM-based prediction of articulatory movements," *Speech Communication*, vol. 52, no. 10, pp. 834–846, oct 2010.

[11] ——, "HMM-Based Text-to-Articulatory-Movement Prediction and Analysis of Critical Articulators," in *Proc. Interspeech*, Makuhari, Japan, 2010, pp. 2194–2197.

[12] Z. Wei, Z. Wu, and L. Xie, "Predicting articulatory movement from text using deep architecture with stacked bottleneck features," in *Proc. APSIPA*, Jeju, South Korea, 2016, pp. 1–6.

[13] I. Steiner, S. Le Maguer, and A. Hewer, "Synthesis of Tongue Motion and Acoustics from Text Using a Multimodal Articulatory Database," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 25, no. 12, pp. 2351–2361, 2017.

[14] S. Le Maguer, I. Steiner, and A. Hewer, "An HMM/DNN comparison for synchronized text-to-speech and tongue motion synthesis," in *Proc. Interspeech*. Stockholm, Sweden: International Speech Communication Association, 2017, pp. 239–243.

[15] L. Yu, J. Yu, and Q. Ling, "BLTRCNN Based 3D Articulatory Movement Prediction: Learning Articulatory Synchronicity From Both Text and Audio Inputs," *IEEE Transactions on Multimedia*, vol. PP, no. c, p. 1, 2018.

[16] M. Stone, B. Sonies, T. Shawker, G. Weiss, and L. Nadel, "Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system," *Journal of Phonetics*, vol. 11, pp. 207–218, 1983.

[17] M. Stone, "A guide to analysing tongue motion from ultrasound images," *Clinical Linguistics and Phonetics*, vol. 19, no. 6-7, pp. 455–501, jan 2005.

[18] T. G. Csapó, T. Grósz, G. Gosztolya, L. Tóth, and A. Markó, "DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 3672–3676.

[19] V. Ramanarayanan, S. Tilsen, M. Proctor, J. Töger, L. Goldstein, K. S. Nayak, and S. Narayanan, "Analysis of speech production real-time MRI," *Computer Speech and Language*, vol. 52, pp. 1–22, 2018.

[20] T. G. Csapó, C. Zainkó, L. Tóth, G. Gosztolya, and A. Markó, "Ultrasound-based Articulatory-to-Acoustic Mapping with WaveGlow Speech Synthesis," in *Proc. Interspeech*, 2020, pp. 2727–2731.

[21] D. Porras, A. Sepúlveda-Sepúlveda, and T. G. Csapó, "DNN-based Acoustic-to-Articulatory Inversion using Ultrasound Tongue Imaging," in *International Joint Conference on Neural Networks*, Budapest, Hungary, 2019, pp. N–19 221.

[22] M. S. Ribeiro, J. Sanger, J.-X. X. Zhang, A. Eshky, A. Wrench, K. Richmond, and S. Renals, "TaL: a synchronised multi-speaker corpus of ultrasound tongue imaging, audio, and lip videos," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China, 2021, pp. 1109–1116.

[23] T. Hueber, G. Aversano, G. Chollet, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel, and M. Stone, "Eigentongue feature extraction for an ultrasound-based silent speech interface," in *Proc. ICASSP*, Honolulu, HI, USA, 2007, pp. 1245–1248.

[24] Z. Wu, O. Watts, and S. King, "Merlin: An Open Source Neural Network Speech Synthesis System," in *9th ISCA Speech Synthesis Workshop*. Sunnyvale, CA, USA: ISCA, sep 2016, pp. 202–207.

[25] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild," in *Proceedings of the 28th ACM International Conference on Multimedia*, New York, NY, USA, 2020, pp. 484–492.

[26] W. Katz, T. Campbell, J. Wang, E. Farrar, J. Eubanks, A. Balasubramanian, B. Prabhakaran, and R. Rennaker, "Opti-speech: A real-time, 3D visual feedback system for speech training," in *Proc. Interspeech*, Singapore, Singapore, 2014, pp. 1174–1178.

[27] D. Jones, "Development of Kinematic Templates for Automatic Pronunciation Assessment Using Acoustic-to-Articulatory Inversion," *Master's Thesis*, jul 2017.

[28] C. Agarwal and P. Chakraborty, "A review of tools and techniques for computer aided pronunciation training (CAPT) in English," *Education and Information Technologies*, vol. 24, no. 6, pp. 3731–3743, nov 2019.

[29] T. G. Csapó and K. Xu, "Quantification of Transducer Misalignment in Ultrasound Tongue Imaging," in *Proc. Interspeech*, online, 2020, pp. 3735–3739.

[30] T. G. Csapó, K. Xu, A. Deme, T. E. Gráczi, and A. Markó, "Transducer Misalignment in Ultrasound Tongue Imaging," in *12th International Seminar on Speech Production*, 2020.