INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Integrating Articulatory Information in Deep Learning-based Text-to-Speech Synthesis

Beiming Cao[1], *Myungjong Kim*[1], *Jan van Santen*[3], *Ted Mau*[4], *Jun Wang*[1,2]

[1]Speech Disorders & Technology Lab, Department of Bioengineering
[2]Callier Center for Communication Disorders
University of Texas at Dallas, Richardson, Texas, United States
[3]Center for Spoken Language Understanding, Oregon Health & Science University
[4]Department of Otolaryngology - Head and Neck Surgery
University of Texas Southwestern Medical Center, Dallas, Texas, United States

{beiming.cao, myungjong.kim, wangjun}@utdallas.edu;
vansantj@ohsu.edu;ted.mau@utsouthwestern.edu

Abstract

Articulatory information has been shown to be effective in improving the performance of hidden Markov model (HMM)-based text-to-speech (TTS) synthesis. Recently, deep learning-based TTS has outperformed HMM-based approaches. However, articulatory information has rarely been integrated in deep learning-based TTS. This paper investigated the effectiveness of integrating articulatory movement data to deep learning-based TTS. The integration of articulatory information was achieved in two ways: (1) direct integration, where articulatory and acoustic features were the output of a deep neural network (DNN), and (2) direct integration plus forward-mapping, where the output articulatory features were mapped to acoustic features by an additional DNN; These forward-mapped acoustic features were then combined with the output acoustic features to produce the final acoustic features. Articulatory (tongue and lip) and acoustic data collected from male and female speakers were used in the experiment. Both objective measures and subjective judgment by human listeners showed the approaches integrated articulatory information outperformed the baseline approach (without using articulatory information) in terms of naturalness and speaker voice identity (voice similarity).

Index Terms: text-to-speech synthesis, articulatory data, deep learning, deep neural network

1. Introduction

Text-to-speech (TTS) synthesis is a process of generating speech waveform from textual input [1, 2]. TTS is widely used for human-computer interaction and communication aid. Clinically, TTS is used as an assistive technology for speech communication. TTS can be adapted in book reading devices for blind people [3, 4] and augmentative & alternative communication (AAC) device for patients with neurological speech disorders [5]. In addition, speech synthesis can serve as a speech output component of a silent speech interface [6], which converts voiceless patient's articulatory movement (e.g., tongue and lip) to speech. Besides the success of speech synthesis with a standard voice output for normal people, researchers recently started working on reconstructing personalized voice of people who are losing or already lost their ability to speak [7–10], e.g., laryngectomees (individuals following a laryngectomy, surgical removal of larynx).

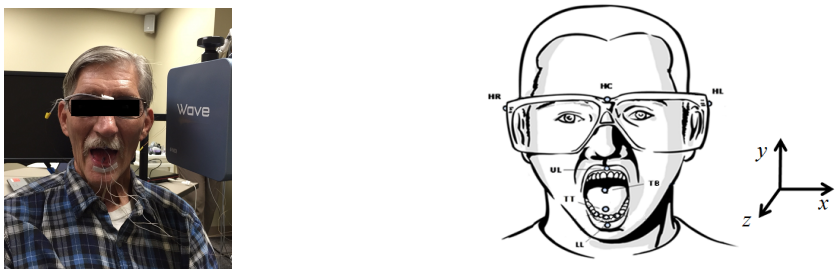
Statistical parametric speech synthesis (SPSS) [11] has the advantage of flexibility, more controllable, and small footprint which make it suitable for clinical applications. Deep learning models (e.g. deep neural network and variations) have been used for acoustic modeling in SPSS recently and outperformed traditional HMM-based approaches [11–14]. Despite the recent success of TTS, the conventional deep learning-based TTS still requires a large training data set from one speaker, which is a limit for some clinical application. It is difficult to collect large amount of high-quality audio data from patients with speech disorders (e.g., laryngectomees, who are unable to produce quality speech).

Articulatory movement information provides a description to speech utterances [15] and a compensation of insufficient training data. Articulatory movement information is expected to improve the personalization and naturalness of synthesized voice [15]. Another potential of articulatory movement data is that it can be collected almost harmlessly from people who are unable to produce speech sounds (e.g., laryngectomees). Ling and colleagues [15] proposed an approach of integrating articulatory features into HMM-based parametric speech synthesis [16], and proved the improvement in acoustic parameter prediction and speech naturalness [15]. Integration of articulatory data in deep learning-based TTS system, however, has rarely been studied. Relevant work including articulation-to-speech synthesis using DNN can be found in [17–20].

In this paper, we investigated the effectiveness of integrating articulatory movement information in deep learning-based TTS with small amount of training data. Experiments of two methods of integrating articulatory data have been conducted. One is the direct integration (DI) method, which is to train the DNN model with the output layer composed by concatenating acoustic features and articulatory features. During testing stage, only the predicted acoustic parameters are sent into vocoder to generate speech waveform. The other approach is the direct integration plus forward-mapping (DI+FM) method which uses a learning mode same to DI method, but during the testing session, the output articulatory features were forward-mapped to acoustic features by another DNN trained with same training dataset. Then the weighted average of original predicted acoustic features and forward-mapped acoustic features was used as the input of vocoder to generate speech waveform.

Speaker-dependent TTS experiments were conducted with acoustic and articulatory movement data collected from two speakers, one male and one female. The synthesized speech was measured both objectively and subjectively. The objective eval-

http://dx.doi.org/10.21437/Interspeech.2017-1762



(a) Wave System (b) Sensor Locations

Figure 1: *Articulatory (tongue and lip) motion data collection setup.*

uation is the accuracies of acoustic parameter predictions. The subjective evaluation is the average preference score in naturalness and speaker voice identity (similarity to original voice) of multiple listeners. The experimental results show that applying articulatory data to deep learning-based TTS (DI and DI+FM) outperformed the baseline approach (without using articulatory information) in both objective and subjective evaluations.

2. Articulatory Data Set

For this study, we used data from two speakers (one male and one female). No history of other speech, language, or cognitive problem was reported. Each subject repeated a sequence of 132 phrases at their habitual speaking rates twice during data recording. The phrases were selected from phrases that are frequently spoken by AAC devices users [21, 22]. An example phrase is *how are you doing?* A total of 528 phrases were recorded.

2.1. Tongue motion tracking device

The articulatory movement (and acoustic) data were collected using the Wave system (Northern Digital Inc., Waterloo, Canada) (Figure 1a). Movement of articulators was recorded by attaching four small sensors to the surface of the lips and tongue (two on the tongue and two on lips). A head sensor was attached to the middle of forehead for head correction. Tongue sensors were attached to the tongue using dental glue (Peri-Acry1 90, GluStitch). Head and lip sensors were attached with skin tape. Articulatory data from the four-sensor set tongue tip (TT, 5-10 mm to tongue apex), tongue back (TB, 20-30mm back from TT), upper lip (UL), and lower lip (LL) was used for this study [23, 24]. The locations of the five sensors are shown in Figure 1b. The sampling rate of recording was 100Hz. Participants were seated with their head next to the magnetic field generator (The blue box in Figure 1a) during data collection session. The spatial precision of movement tracking using Wave System is about 0.5mm [25]. Before the formal data collection, a three-minute training session helped participants adapt to speak with tongue sensors.

2.2. Data processing

Preprocessing was applied to the raw movement data before analysis. First, the movement of head was subtracted from the motion data of tongue and lips to obtain head-independent articulator movement. The derived 3D Cartesian coordinates system is shown in Figure 1b, in which x is left-right direction, y is vertical, and z is front-back direction. In this study, only y and z coordinates were used for analysis [26]. Second, a 5th-order Butterworth low-pass filter with a cutoff frequency of 20Hz was used to remove noise in the movement data [27]. Since the frame shift length of acoustic feature extraction is set to 5 milliseconds, all articulation data were upsampled to 200Hz.

3. Method

3.1. DNN without articulatory data: baseline approach

A conventional DNN-based TTS model includes a duration model DNN (DM-DNN) that predicts the durations of phones from input phone labels, and an acoustic model DNN (AM-DNN) that predicts an acoustic sequence from a sequence of acoustic labels [28]. In this study, a relatively smaller size DNN is used in DM-DNN. The input of DM-DNN is the full-context labels mapped to numerical features, the output is the corresponding duration (as the number of frames) of the units the labels belong to, which was limited to phone [28]. AM-DNN, on the other hand, takes the full-context labels and positional information frames as input according to the duration generated by DM-DNN and predicts acoustic parameters for vocoder to synthesized speech voice. This implementation is used for obtaining the baseline results in this study since no articulatory data was used in this model.

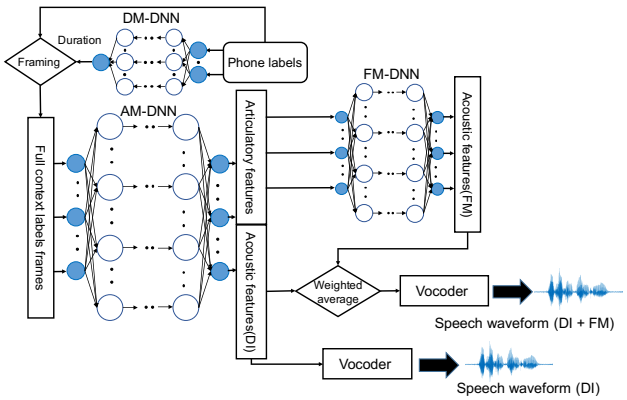


Figure 2: *Implementation of whole articulatory integration, DM-DNN denotes duration-model DNN, AM-DNN denotes acoustic model DNN, FM-DNN denotes forward-mapping DNN.*

3.2. TTS based on DNN with articulatory data

The design of our approaches for integrating articulatory information in DNN-based TTS is shown in Figure 2. In the direct integration (DI) approach, the articulatory features are used as the additional output of AM-DNN and the speech waveform is generated from the acoustic features predicted by AM-DNN. For the direct integration plus forward-mapping (DI+FM) approach, FM-DNN is additionally used, which is used for mapping the output articulatory data predicted by AM-DNN to acoustic parameters. Then, acoustic speech is synthesized by

the weighted average of acoustic features predicted by the DI approach and the forward-mapped (FM) acoustic features.

The weights for combining DI acoustic features and FM acoustic features were found when minimizing the RMSE of predicted acoustic features (DI+FM) during development. These weights are expected to represent the relationship between acoustic and articulatory features and vary between different speakers. For experiments with our dataset, the weights that minimizing the RMSE of acoustic features during development is 12:1 (DI vs FM acoustic feature) for the female speaker and 20:1 for the male speaker.

3.3. Experimental setup and measures

As mentioned, 264 phrases were recorded from each participant. Speaker dependent TTS experiments were conducted. 220 phrases were used for training, 24 phrases as development set, and 20 as testing set, there is no overlap between development and testing set. The vocoder used in this study is WORLD [29, 30].

In the preliminary experiment (use training and development set only), three different DNNs, 6×512, 4×1024 and 6×1024 (i.e., $n \times m$ denotes n hidden layers with hidden units of m) were validated for AM-DNN, the two speaker's average results of AM-DNN are shown in Table 1, which indicated that 6×1024 DNN is optimal for our experiment. The structure of FM-DNN was found by validating the performance of 6×1024 AM-DNN and FM-DNN in three smaller different sizes (Table 2) together. In addition, for each of three different FM-DNN, summation weights from 8 to 30 were verified. Eventually 4×512 DNN was found optimal in this case and as mentioned, 12 and 20 were found as optimal weight for female and male speakers, the results are shown in Table 2.

The predicted acoustic parameters include: 1-dimensional log scale of fundamental frequency (log-f0), 5-band aperiodicities (BAP) (0-1, 1-2, 2-4, 4-6, 6-8 kHz) [31], 60-dimensional mel-cepstral coefficients (MCCs), concatenated with their first and second derivatives, and 1-dimensional voiced/unvoiced label (V/UV). Articulatory feature is the 2-dimensional movement of four sensors and their first and second derivatives, which is 24-dimensional ($2 \times 4 \times 3 = 24$). Therefore, the dimension of output layer in AM-DNN without and with articulatory data is 199 ($(1 + 5 + 60) \times 3 + 1 = 199$) and 223 ($199 + 24 = 223$) respectively. All acoustic features were extracted frame by frame in a step size of 5ms. The inputs of duration model-DNN (DM-DNN) and AM-DNN are 416 dimension label features and 416-dimensional label features plus 9-dimensional positional information respectively. The 416-dimensional label features were extracted by HTS [32], including quinphone identity, part-of-speech within a syllable word and phrase, etc [33]. Other details of experiment setup are shown in Table 1. The DNN models in this study were implemented with Merlin speech synthesis toolkit [33].

Table 1: *Average performance of baseline AM-DNN of two speakers with only acoustic features as output in development session.*

Acoustic Feature	6×512	4×1024	6×1024
log-f0	0.163	0.161	0.153
BAP Distortion (dB)	1.263	1.266	1.206
MCD (dB)	5.259	5.235	5.187
V/UV Error (%)	15.90	15.58	14.54

Table 2: *Average performance of DI+FM model with FM-DNN in different sizes (AM-DNN is fixed to 6 × 1024) of two speakers in development session.*

Acoustic Feature	4×512	6×512	4×1024
log-f0	0.143	0.153	0.150
BAP Distortion (dB)	1.180	1.219	1.242
MCD (dB)	5.013	5.158	5.119
V/UV Error (%)	12.50	14.49	13.91

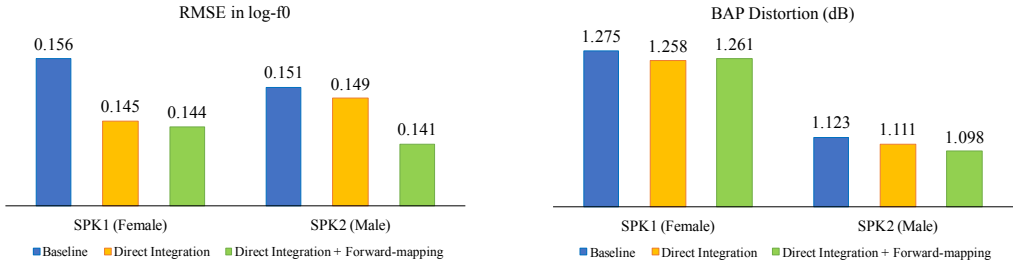
Both subjective and objective evaluation were used in the project. The objective evaluation includes root-mean-square error (RMSE) of the logarithm of fundamental frequency (log-f0), Mel-Cepstral distortion (MCD) [34], band aperiodicities (BAPs) distortion, and voiced/unvoiced (V/UV) prediction error rate. The V/UV error rate is the error rate of predicting a frame is voiced or unvoiced. The computation of MCD and BAP distortion is shown in equation (1). In equation (1), C and C^{gen} denote the original and generated voice, respectively, m is the frame step (or time), d denotes dth dimension in frame m. D is the dimension of features, which is 60 for MCD and 5 for the BAP, respectively.

$$MCD/BAP = \frac{10}{log10} \sum_{m=1}^{T} \sqrt{2 \sum_{d=1}^{D} (C_{m,d} - C_{m,d}^{gen})^2} \quad (1)$$

The subjective evaluation is the average preference scores in terms of naturalness and speaker voice identity (similarity to the original voice) of 20 synthesized phrases. The subjective evaluation was done by 10 listeners, who were all English speakers, speakers of our data were not included, the average score of two speakers are shown in Figure 4.

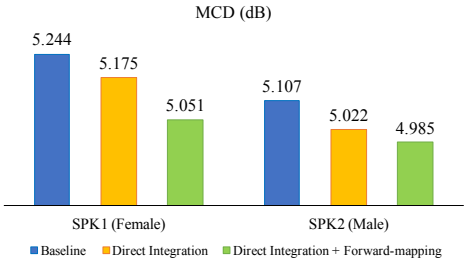
Table 3: *Experimental setup.*

Acoustic Feature	199-dim. vectors
Mel-Cepstral Coefficients (MCCs)	(60-dim. vectors) + Δ + ΔΔ (180-dim.)
Band Aperiodicities (BAPs)	(5-dim. vectors) + Δ + ΔΔ(15-dim.)
Fundamental Frequency on log scale (log-f0)	(1-dim. vectors) + Δ + ΔΔ (3-dim.)
Voiced/Unvoiced (V/UV) error	(1-dim.)
Sampling rate	16000 Hz
Windows length	25 ms
Articulatory Feature	24-dim. vectors
articulatory movement vector (8 sensors)	(8-dim. vectors) + Δ + ΔΔ (24-dim.)
Common	
Frame rate	5 ms
DNN Topology	
Duration-model DNN (DM-DNN)	
Input	416-dim. full-context label
Output layer dim.	1-dim. duration
No. of nodes each hidden layer	256
Depth	4-depth hidden layers
Learning rate	0.002
Acoustic-model DNN (AM-DNN)	
Input	416-dim. full-context label + 9-dim. position
Output layer dim.	199-dim. for baseline 223-dim. for articulatory
No. of nodes each hidden layer	1024
Depth	6-depth hidden layers
Learning rate	0.002
Forward-mapping DNN (FM-DNN)	
Input	24-dim. articulatory
Output layer dim.	199-dim
No. of nodes each hidden layer	512
Depth	4-depth hidden layers
Learning rate	0.0015
Vocoder	WORLD

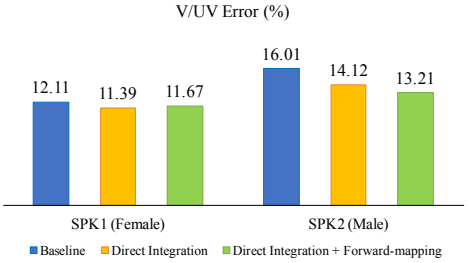


(a) RMSE in log-f0

(b) BAP Distortion



(c) MCD Distortion



(d) Voiced/Unvoiced Error

Figure 3: *Results of objective evaluations.*

4. Results & Discussion

The synthesized speech by the three methods (baseline, DI, and DI+FM) was measured both objectively (Figure 3) and subjectively (Figure 4). Lower objective measures indicated a better performance.

As shown in Figure 3, RMSE of log-f0, distortion of BAP and MCCs, and V/UV error rate were all decreased by integrating articulatory features in either of two integration approaches. The results showed that integrating articulatory movement information in DNN-based TTS improved the quality of synthesized speech. When comparing the direct integration (DI) and the DI+FM approaches, results were not consistent in all measures. In the DI+FM approach, RMSE of log-f0 and MCD in all experiments were further decreased, except BAP and V/UV for the female speaker (SPK1).

Conventionally, the accuracy of acoustic features such as f0 and V/UV are not thought to have certain relationship with articulators' movement. In this study both of them were improved by adding articulatory features. These findings provide evidence for speech science that extracted f0 and V/UV flag from the synthesis speech could be reflected by articulatory information to some extent. This could possibly be explained that tongue body back movement/position may slightly affect vibration of vocal folds and then improve the speech output quality in terms of pitch and voiced/unvoiced errors.

Subjective evaluation (Figure 4) also indicated that adding articulatory information (in either integration approach) increased the naturalness and similarity significantly. For the comparison between the two integration approaches, the DI approach slightly outperformed the DI+FM approach in naturalness, while the DI+FM approach generated voice that sounds more similar to speaker's original voices.

As mentioned previously, the training data size is generally limited in clinical applications. This study demonstrated that another type of information (although also of limited size) could benefit the TTS quality (naturalness and speaker voice identity). There is a logistic difficulty in collecting articulatory data [26]. Fortunately, acoustic-to-articulatory inverse mapping is available and could be used to generate pseudo-articulatory data in practical applications (e.g., [35–37]).

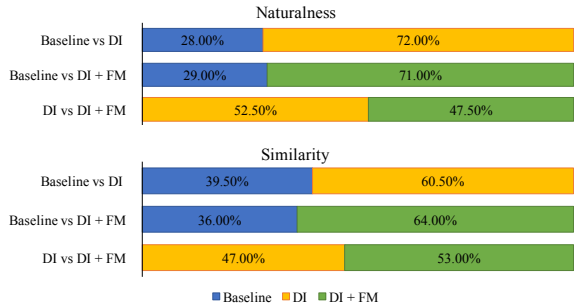


Figure 4: *Results of subjective evaluations.*

5. Conclusion & Future Work

This study demonstrated the effectiveness of integrating articulatory movement information into deep learning-based TTS with a small training data set. Both of the proposed integration approaches: direct integrating (DI) and direct integrating plus forward-mapping approach (DI+FM) outperformed the baseline approach (DNN without articulatory information) in terms of both objective and subjective measures (naturalness and speaker voice identity). These findings proved the effectiveness of applying articulatory data to deep learning-based TTS. Future work includes verifying the conclusions above with large amount of data and other approaches to merge two types of acoustic features (DI and FM).

6. Acknowledgement

This work was supported by the National Institutes of Health (NIH) under award number R03 DC013990 and by the American Speech-Language-Hearing Foundation through a New Century Scholar Research Grant. We thank Kristin Teplansky, Katie Purdum, and the volunteering participants.

7. References

[1] J. P. H. van Santen, J. P. Olive, R. W. Sproat, and J. Hirschberg, Eds., *Progress in Speech Synthesis*. New York, NY, USA: Springer-Verlag New York, Inc., 1997.

[2] X. Huang, A. Acero, H. Hon, Y. Ju, J. Liu, S. Meredith, and M. Plumpe, "Recent improvements on Microsoft's trainable text-to-speech system-Whistler," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, Munich, Germany, 1997, pp. 959–962.

[3] T. Portele and J. Kramer, "Adapting a TTS system to a reading machine for the blind," in *Proc. IEEE International Conference on Spoken Language Processing*, vol. 1, 1996, pp. 184–187.

[4] A. Reddy, N. Pratap *et al.*, "Communication of Dumb & Blind People with TTS," *International Journal of Research in Computer and Communication*, vol. 1, no. 6, pp. 364–372, 2012.

[5] D. Beukelman and P. Mirenda, "Augmentative and alternative communication: Supporting children and adults with complex communication needs," 2005.

[6] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[7] Z. Ahmad Khan, P. Green, S. Creer, and S. Cunningham, "Reconstructing the Voice of an Individual Following Laryngectomy," *Augmentative and Alternative Communication*, vol. 27, no. 1, pp. 61–66, 2011.

[8] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction," *Acoustical Science and Technology*, vol. 33, no. 1, pp. 1–5, 2012.

[9] C. Veaux, J. Yamagishi, and S. King, "Towards personalized synthesized voices for individuals with vocal disabilities: Voice banking and reconstruction," *Workshop on Speech and Language Processing for Assistive Technologies*, vol. 107–111, 2013.

[10] ——, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference*. IEEE, 2013, pp. 1–4.

[11] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7962–7966.

[12] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Interspeech*, 2014, pp. 1964–1968.

[13] M. Coto-Jiménez and J. Goddard-Close, "Speech synthesis based on hidden Markov models and deep learning," *Research in Computing Science*, vol. 112, pp. 19–28.

[14] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, "On the training aspects of deep neural network (DNN) for parametric TTS synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3829–3833.

[15] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1171–1185, 2009.

[16] J. Yamagishi, "An introduction to HMM-based speech synthesis," *Technical Report*, 2006.

[17] S. Aryal and R. Gutierrez-Osuna, "Data driven articulatory synthesis with deep neural networks," *Computer Speech & Language*, vol. 36, pp. 260–273, 2016.

[18] F. Bocquelet, T. Hueber, L. Girin, P. Badin, and B. Yvert, "Robust articulatory speech synthesis using deep neural networks for bci applications," in *15th Annual Conference of the International Speech Communication Association (Interspeech 2014)*, 2014, pp. 2288–2292.

[19] B. H. Story and K. Bunton, "An acoustically-driven vocal tract model for stop consonant production," *Speech Communication*, pp. 1–17, 2016.

[20] P. Birkholz, L. Martin, Y. Xu, S. Scherbaum, and C. Neuschaefer-Rube, "Manipulation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis," *Computer Speech & Language*, vol. 41, pp. 116–127, 2017.

[21] J. Wang, A. Samal, and J. R. Green, "Preliminary test of a real-time, interactive silent speech interface based on electromagnetic articulograph," *ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies*, pp. 38–45, 2014.

[22] J. Wang and S. Hahm, "Speaker-independent silent speech recognition with across-speaker articulatory normalization and speaker adaptive training," in *Interspeech*, 2015, pp. 2415–2419.

[23] J. Wang, A. Samal, P. Rong, and J. R. Green, "An optimal set of flesh points on tongue and lips for speech-movement classification," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 1, pp. 15–26, 2016.

[24] J. Wang, S. Hahm, and T. Mau, "Determining an optimal set of flesh points on tongue, lips, and jaw for continuous silent speech recognition," in *6th Workshop on Speech and Language Processing for Assistive Technologies*, 2015, pp. 79–85.

[25] J. Berry, "Accuracy of the NDI wave speech research system," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 5, pp. 1295–1301, 2011.

[26] J. Wang, J. R. Green, A. Samal, and Y. Yunusova, "Articulatory distinctiveness of vowels and consonants: A data-driven approach," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 5, pp. 1539–1551, 2013.

[27] J. R. Green and Y.-T. Wang, "Tongue-surface movement patterns during speech and swallowing," *The Journal of the Acoustical Society of America*, vol. 113, no. 5, pp. 2820–2833, 2003.

[28] B. Potard, M. P. Aylett, D. A. Baude, and P. Motlicek, "Idlak Tangle: An open source Kaldi based parametric speech synthesiser based on DNN," *Interspeech 2016*, pp. 2293–2297, 2016.

[29] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[30] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

[31] H. Silen, E. Helander, and M. Gabbouj, "Prediction of Voice Aperiodicity Based on Spectral Representations in HMM Speech Synthesis," in *Interspeech*, 2011, pp. 105–108.

[32] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0." in *SSW*. Citeseer, 2007, pp. 294–299.

[33] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," *Proc. SSW, Sunnyvale, USA*, 2016.

[34] M. Shannon, H. Zen, and W. Byrne, "Autoregressive models for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 587–597, 2013.

[35] S. Hahm and J. Wang, "Parkinsons condition estimation using speech acoustic and inversely mapped articulatory data," in *Proc. of INTERSPEECH*, 2015, pp. 513–517.

[36] A. Ji, "Speaker independent acoustic-to-articulatory inversion," Ph.D. dissertation, Marquette University, 2014, Department of Electrical Engineering.

[37] A. Ji, M. T. Johnson, and J. J. Berry, "Parallel reference speaker weighting for kinematic-independent acoustic-to-articulatory inversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1865–1875, 2016.