Speaker-Independent Mel-cepstrum Estimation from Articulator Movements Using D-vector Input

Kouichi Katsurada[1], *Korin Richmond*[2]

[1] Department of Information Sciences, Tokyo University of Science, Tokyo, Japan
[2] Center for Speech Technology Research, University of Edinburgh, Edinburgh, United Kingdom
katsurada@rs.tus.ac.jp, korin@cstr.ed.ac.uk

Abstract

We describe a speaker-independent mel-cepstrum estimation system which accepts electromagnetic articulography (EMA) data as input. The system collects speaker information with d-vectors generated from the EMA data. We have also investigated the effect of speaker independence in the input vectors given to the mel-cepstrum estimator. This is accomplished by introducing a two-stage network, where the first stage is trained to output EMA sequences that are averaged across all speakers on a per-triphone basis (and so are speaker-independent) and the second receives these as input for mel-cepstrum estimation. Experimental results show that using the d-vectors can improve the performance of mel-cepstrum estimation by 0.19 dB with regard to mel-cepstrum distortion in the closed-speaker test set. Additionally, giving triphone-averaged EMA data to a mel-cepstrum estimator is shown to improve the performance by a further 0.16 dB, which indicates that the speaker-independent input has a positive effect on mel-cepstrum estimation.

Index Terms: articulatory movement, EMA, mel-cepstrum estimation, speech synthesis

1. Introduction

The performance of speech synthesis has improved dramatically in recent years, owing to state-of-the-art neural network-based approaches. Tacotron2 [1] achieved sound quality almost indistinguishable from that of a real human using an encoder-decoder type network and the Wavenet vocoder [2]. Many systems employ an end-to-end approach which directly synthesizes voice from text (e.g. [1]). Such systems are applicable in numerous applications that require high quality voices. However, actual human articulation information for voice generation is still required in some applications that involve simulation of the mechanism of human voice production. For example, Anumanchipalli et al. synthesized speech from ECoG brain activity data, but claim their approach is only viable using estimated articulatory movements as an intermediary representation [4]. Meanwhile, Tobing et al. successfully demonstrated articulatory control over speech synthesis by modifying tongue tip height in their articulatory-to-acoustic and inversion mapping systems in order to change several vowels [3]. Such studies show that taking articulation into account can prove valuable for speech synthesis in multiple cases and can ease the simulation of more human-like variety of pronunciation.

Articulatory movements can be recorded with several types of equipment which may be roughly divided into two main categories: imaging-based and point tracking-based techniques [5]. The biggest advantage of imaging-based methods, such as X-ray cineradiography [6], magnetic resonance imaging (MRI) [7], and ultrasound [8], is that a fuller picture of articulation can be obtained. These techniques can scan the shape of the internal vocal tract with X-ray, radio wave, or ultrasound that can penetrate the human body. However, they present some difficulties for use in speech synthesis due to their loud noise, low frame rate, or danger of X-ray radiation. On the other hand, electromagnetic articulography (EMA) [9], which is perhaps the most popular point tracking-based method, is widely used in articulation-based speech synthesis because it has a high frame rate and reasonable accuracy in point tracking [5]. The purpose of this study is to provide a speaker-independent mel-cepstrum estimation system from the articulatory movements recorded by EMA.

To construct a speaker-independent mel-cepstrum estimation model, some kind of speaker information should be provided as input. In recent years, speech synthesis studies have shown that feature vectors transferred from a speaker verification network can improve the performance of multi-speaker synthesis [10, 11]. We borrow this idea in our proposed system. We show that appending d-vectors [12] generated from EMA data to the inputs given to the EMA-based mel-cepstrum estimator is helpful for estimating mel-cepstra in a speaker-independent model. Subsequently, we attempt to make the inputs given to the mel-cepstrum estimator speaker-independent by taking the average of all speakers' EMA data when the same tri-phone is spoken. Finally, we show the result of experiments conducted with a two-stage network in which the first stage network is trained to output a speaker-independent average EMA sequence that is given to the second mel-cepstrum estimation network.

2. Related work

Generating acoustic features from recorded articulatory movements using data-driven models has been studied for around two decades at least. In some of the earlier studies, Gaussian mixture models (GMM) or hidden Markov models (HMM) were employed for modeling the statistical relation between the articulatory and acoustic features [3, 13-16]. The aim of those studies was to construct a GMM (or HMM) that predicts acoustic features from articulatory movements or models the co-occurrence relation between articulatory and acoustic features. Neural networks have also been applied to this problem numerous times. As an earlier example, Kello & Plaut used simple feed-forward networks with a single hidden layer [17]. In more recent years, deep learning has gained popularity in training the mapping between articulatory movements and acoustic features, and a variety of networks

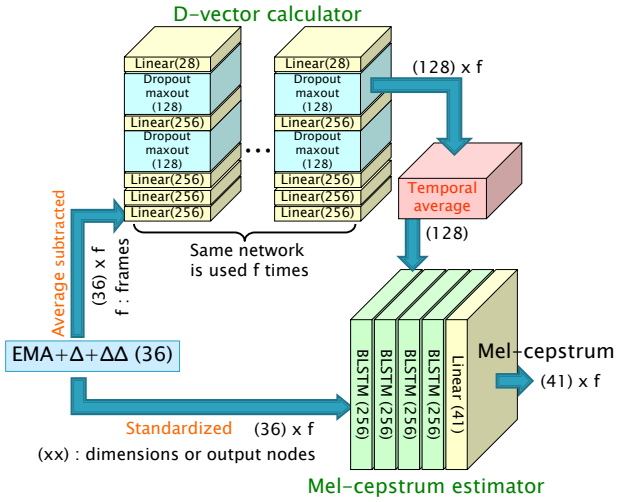

Figure 1: *Configuration of mel-cepstrum estimation system.*

have been used to handle the sequential articulatory movements [18-24]. In [19], windows of an entire sequence are input to a standard DNN. In contrast, [20] and [21] use long short-term memory (LSTM) units which have frequently been shown to be suitable for modeling sequential input and output relations. The latest research, which shows the best performance in estimating the acoustic features from articulatory movements, uses the bi-directional LSTM (BLSTM), which is an extension of LSTM that can deal with both forward and backward sequences of the input/output data [22-24]. However, these deep-learning-based studies have aimed at developing a speaker-specific acoustic feature generator. The purpose of this paper in contrast is to provide a speaker-independent mel-cepstrum estimator which can generate any of multiple speakers' acoustic features using a single model.

3. Mel-cepstrum estimation using d-vector input

3.1. Capturing speaker characteristics

Our mel-cepstrum estimator is composed of a 4-layer BLSTM with a linear output layer, as this showed the best performance in our preliminary experiments. We incorporate speaker information to the estimator in a similar way to recent multi-speaker speech synthesis research. For obtaining speaker information from EMA data, Illa et al. proposed an LSTM-based speaker identifier [25]. Although its performance is good, it requires more training data than the EMA data we have available. Other synthesis studies have used the d-vector for capturing speaker characteristics [10, 11]. D-vectors have the advantage of being obtained with small footprint data [12] and are easy to calculate by taking the average of all frames at the last intermediate vector. Thus, we use the d-vector to represent speaker information in this study. The configuration of our system is illustrated in Figure 1. The d-vector calculation network is the same as in the original paper [12].

3.2. Database preprocessing

There have emerged many databases [27-31] since MOCHA-TIMIT [26], the first widely used EMA database, was released. Some of these databases contain multiple speakers [28-31],

Table 1: *Speakers and number of sentences used in the experiments in this paper.*

	Speakers (number of sentences)		
English native female	05ENF (102)	07ENF (138)	09ENF (134)
	21ENF (69)	28ENF (134)	36ENF (114)
	37ENF (82)	40ENF (76)	
English native male	06ENM (147)	15ENM (67)	16ENM (107)
	32ENM (133)	33ENM (125)	34ENM (37)
	35ENM (119)	38ENM (50)	
Chinese-accented female	01MBF (104)	02MBF (122)	04MSF (142)
	11MBF (83)	14MSF (118)	20MBF (158)
	22MBF (146)	24MSF (141)	
Chinese-accented male	08MBM (101)	10MSM (35)	23MBM (60)
	25MSM (114)	26MSM (150)	27MSM (56)
	29MBM (135)	31MBM (51)	

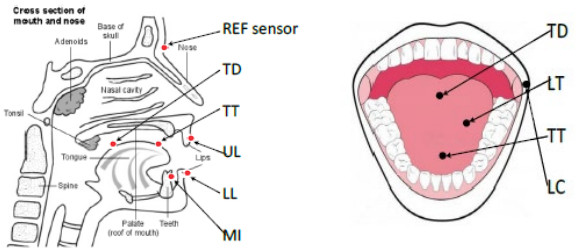

Figure 2: *Location of sensors* [30].

which is the essential requirement in our study, since we aim to develop a speaker-independent mel-cepstrum estimator. Among these databases, the EMA-MAE database [30] has the largest number of speakers, with EMA data sampled at 400 kHz and audio for 40 speakers (10 males and females for native English and Chinese-accented English respectively). Therefore, this is the data we have used in the research here.

The EMA-MAE database contains several sentences in a file. For convenience, we first divided the database into sentences, then removed those sentences with NaN (not a number) data, as well as those containing obvious mistakes (such as data where coil positions are at an unreasonable distance from the rest of the human body). Next, we removed those speakers with fewer than 10 sentences. Ultimately, we obtained 3,350 sentences from 32 speakers (8 of each group). Table 1 summarizes the speakers and their respective number of sentences as used in our experiments. We chose to use six of the articulator positions (TD, LT, TT, UL, LL and MI in Figure 2), as these are also available in other major databases [27, 29]. The positions were projected onto the midsagittal plane. The inputs given to the mel-cepstrum estimator were the z-score normalized EMA, Δ-EMA and ΔΔ-EMA sequences. Normalization was done sentence by sentence, as we have found this can notably improve estimation performance. For the d-vector calculator input, we simply subtracted the sentence average from the original EMA data, which is the same as in paper [25].

3.3. Experimental results

We conducted an eight-fold cross-validation test in which data were divided into three sets: training set, closed-speaker test set and open-speaker test set. In any one test, 28 out of 32

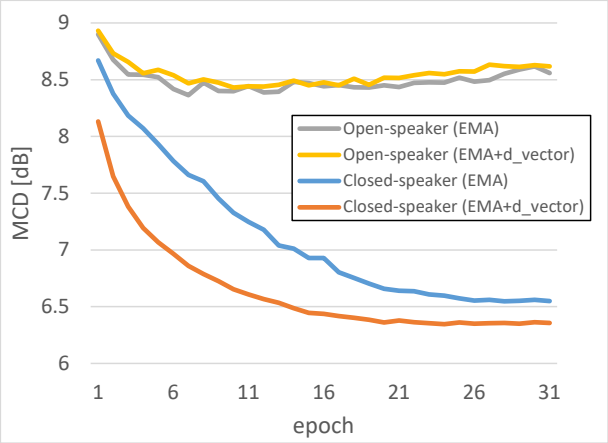

Figure 3: *Effect of appending d-vector to input in mel-cepstrum estimation.*

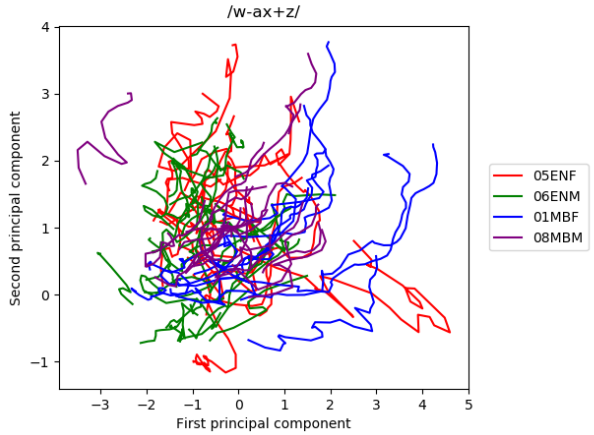

Figure 4: *Trajectory of normalized EMA data when the same tri-phone /w-ax+z/ is pronounced.*

speakers (seven English native/Chinese-accented male/female) were used as the training data and the closed-speaker test data. Of these data, 90% were used as the training data and 10% as the closed-speaker test data. The remaining four speakers' data were placed in the open-speaker (or "unseen-speaker") test set. We used SPTK-3.9 for mel-cepstrum extraction. The Adam optimizer was used and the learning rates were set at 0.0001 during training. We evaluated the effectiveness of the d-vectors in both closed-speaker and open-speaker experiments.

Figure 3 presents the results. The horizontal axis shows the number of training epochs. The vertical axis gives mel-cepstrum distortion (MCD) compared with the original speech audio. The d-vectors are observed to improve the accuracy of mel-cepstrum estimation by 0.19 dB with respect to the lowest MCD in the closed-speaker experiment. These results show that appending speaker information to the inputs given to the mel-cepstrum estimator is effective in generating accurate mel-cepstra. Although the inputs (normalized EMA, Δ-EMA and ΔΔ-EMA) implicitly include speaker identity, the d-vectors, which are the averages of all frames of data, seem to complement the speaker information. Meanwhile, the d-vectors are not found to improve results in the same way in the open-speaker experiment. We suspect this is due to having too few speakers available for calculating the d-vectors for unseen speakers. In comparable speech synthesis studies, more than

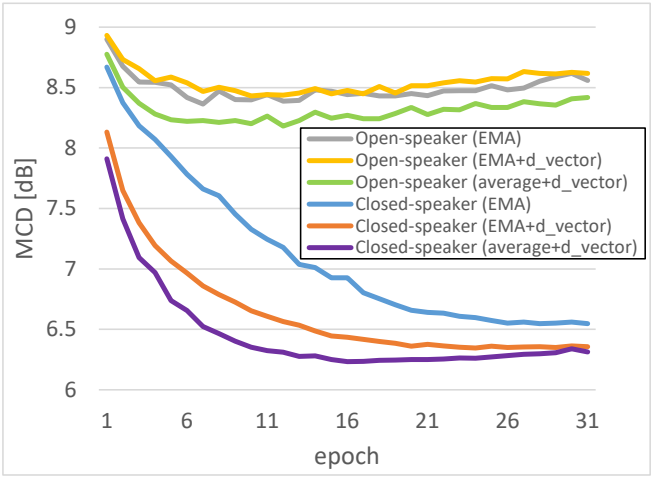

Figure 5: *Giving tri-phone average to mel-cepstrum estimator.*

10,000 speakers may be used to train the models involved in calculating these vectors [10].

4. Speaker-independent inputs and a two-stage network

4.1. Inputting mean EMA data to the estimator

Frames of EMA data encode multiple types of information, such as the phones, their context, speaker identity, and so on. Figure 4 illustrates the trajectories of the sentence-by-sentence normalized EMA when the same tri-phone /w-ax+z/ is uttered multiple times by different speakers. The axes are the first and second components obtained after applying principal component analysis (PCA). Each line represents an utterance of the tri-phone. From the figure, we can observe few common characteristics among the multiple utterances of the same tri-phone. To clarify the tri-phone features and to make the inputs speaker-independent, we calculated the all-speaker average of the normalized EMA data when the same tri-phone is spoken. For that, we performed forced alignment to the speech data using the PocketSphinx speech recognizer [32], and then applied dynamic time warping (DTW) with mep-cepstra as the local distances to obtain the temporal correspondence between the tri-phones in different sentences.

The results of providing these mean EMA trajectories to the mel-cepstrum estimator are shown in Figure 5. We find that giving average input improves the accuracy of generated mel-cepstra. This result indicates that the speaker-independent average vector can be an appropriate intermediate representation in the process of mel-cepstrum estimation. However, there is a problem for using this in a practical articulatory-to-acoustic mapping in that it is impossible to calculate averages during the process of mel-cepstrum estimation because only the EMA data corresponding to a single utterance is available. We address this limitation next.

4.2. A two-stage network and its evaluation

Instead of calculating mean trajectories, we attempted to introduce a preprocessing network that estimates the average of EMA data. Figure 6 illustrates the entire system structure that is composed of two stages: a frontend network and the mel-cepstrum estimation network. The frontend network, which is trained to output the average EMA sequence, is composed of 2 BLSTM layers with 1,000 nodes and a linear

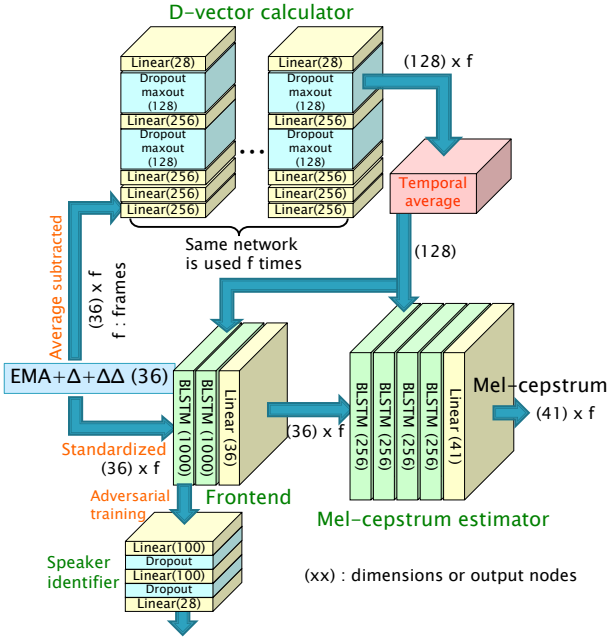

Figure 6: *Two stage network for mel-cepstrum estimation.*

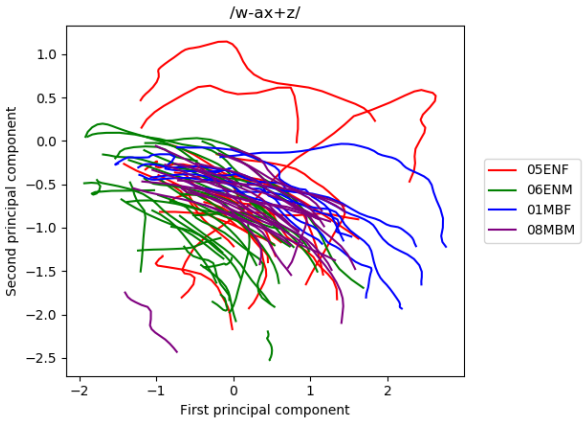

Figure 7: *Speaker dependency in frontend output.*

output layer. It is trained with the pairs of normalized EMA data and its corresponding average calculated in Section 4.1. The frontend outputs are 36 dimensional EMA, Δ-EMA and ΔΔ-EMA sequences, because estimating all three was found to improve performance of mel-cepstrum estimation in preliminary experiments. The first BLSTM layer of the frontend network is trained with domain adversarial training [33] in which the reversed gradient is back-propagated from a speaker-identification component. In this experiment, the gradients are multiplied by a small value (-0.01) to attenuate the effect of adversarial training. The speaker identification component, which estimates the speaker frame by frame, consists of an output linear layer and two hidden linear layers followed by a tanh activation function and a 50% dropout layer. The second stage is the same as the one described in Section 3.1. After being separately trained, they are connected and a fine-tuning training operation is conducted.

Figure 7 illustrates the outputs of the fine-tuned frontend network when the tri-phone data /w-ax+z/ are provided to the frontend. Although there are still differences between samples,

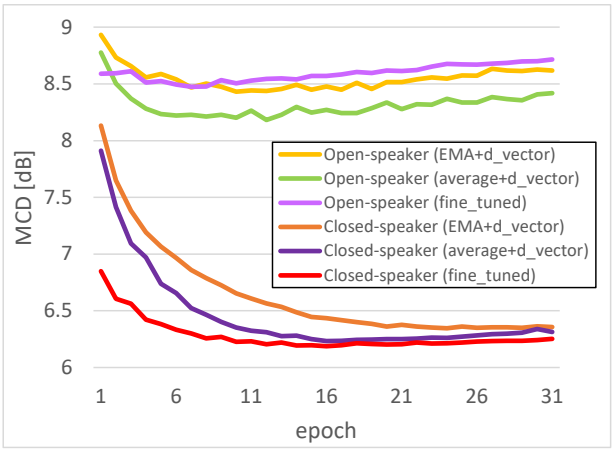

Figure 8: *Results of the two-stage network.*

the trajectories seem to have more similar shape for the same tri-phone compared to the lines in Figure 4. Figure 8 shows the result of mel-cepstrum estimation. We could obtain encouraging results in the closed-speaker experiment, where the lowest MCD is improved by 0.16 dB compared with the EMA+d-vector experiment. However, in the open-speaker experiment, the results obtained are relatively unsatisfactory. Analysis indicates this may be due to inadequate performance of average estimation in the frontend network. In fact, the mel-cepstrum estimator receiving the correct averages shows better results than that accepting the output of the frontend network. If the frontend performance can be sufficiently improved, the mel-cepstrum estimation accuracy can be expected to be correspondingly high since more accurate averages will be given to the estimator. This will be the focus of our future work.

5. Conclusions

We have presented a speaker-independent mel-cepstrum estimator from EMA data input that models speaker information using d-vectors. We have also investigated the effectiveness of giving speaker-averaged EMA data to the estimator, and constructed a two stage network in which the averages are trained to be output in the first stage network. Experimental results show that using d-vectors in mel-cepstrum estimation and training to output averages in the two-stage network can lead to improvement in the closed-speaker mel-cepstrum estimation. However, the accuracy in the open-speaker experiment did not show the same improvement. As it was observed that giving average EMA trajectory inputs to the mel-cepstrum estimator led to an improvement in accuracy in the open-speaker experiments, we intend to improve the performance in the open-speaker experiment in the future.

6. Acknowledgements

This work has been supported by a Grant-in-Aid for Scientific Research (C) 19K12024 2019 and for Scientific Research (B) 16H03211 2019 by MEXT, Japan.

7. References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in Proc. of ICASSP 2018, pp.4779–4783 (2018).

[2] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," arxiv: 1609.03499 (2016).

[3] P. L. Tobing, T. Toda, G. Neubig, S. Sakti, S. Nakamura, and A. Purwarianti, "Articulatory controllable speech modification based on statistical feature mapping with Gaussian mixture models," in Proc. of InterSpeech2014, pp.2298-2302 (2014).

[4] G. K. Anumanchipalli, J. Chartier, and E. Chang, "Speech synthesis from neural decoding of spoken sentences," Nature, vol.568, no.7753, pp.493-498 (2019).

[5] K. Richmond, Z. Ling, and J. Yamagishi, "The use of articulatory movement data in speech synthesis applications: An overview - Application of articulatory movements using machine learning algorithms -," Acoustical Science and Technology, vol.36, no.6, pp.467-477 (2015).

[6] S. Kiritani, "X-ray microbeam method for the measurement of articulatory dynamics: Technique and results," Speech Communications, vol.5, no.2, 119–140 (1986).

[7] S. G. Lingala, B. P. Sutton, M. E. Miquel, and K. S. Nayak, "Recommendations for real-time speech MRI," Journal of Magnetic.Resonance Imaging, vol.43, no.1, pp.28-44 (2015).

[8] Y. S. Akgul, C. Kambhamettu, and M. Stone, "Extraction and tracking of the tongue surface from ultrasound image sequences," in Proc. of CVPR1998, pp.298-303 (1998).

[9] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," Brain and Language vol.31, no.1, pp.26–35 (1987).

[10] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in Proc. of NISP2018, pp.4485-4495 (2018).

[11] R. Doddipatla, N. Braunschweiler, and R. Maia, "Speaker adaptation in DNN-based speech synthesis using d-vectors," in Proc. of InterSpeech2017, pp.3404–3408 (2017).

[12] E. Variani, X. Lei, E. McDermott, I. Lopez Moreno and J. Gonzalez-Dominguez, "Deep neural networks for small foot-print text-dependent speaker verification," in Proc. of ICASSP2014, pp.4052–4056 (2014).

[13] T. Toda, A.W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," Speech Communications, vol.50, no.3, pp.215–227 (2008).

[14] S. Aryal and R. Gutierrez-Osuna, "Reduction of non-native accents through statistical parametric articulatory synthesis," J. Acoust. Soc. Am., vol.137, no.1, pp.433–446 (2015).

[15] Z. H. Ling, K. Richmond, and J. Yamagishi, "Feature-space transform tying in unified acoustic-articulatory modelling for articulatory control of HMM-based speech synthesis," in Proc. of InterSpeech2011, pp.117–120 (2011).

[16] T. Hueber, G. Bailly, and B. Denby, "Continuous articulatory-to-acoustic mapping using phone-based trajectory HMM for a silent speech interface," in Proc. InterSpeech2012, Tue.P3c.01 (2012).

[17] C. Kello and D. Plaut. "A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters," J. Acoust. Soc. Am., vol.116, no.4, pp.2354-2364 (2004).

[18] F. Bocquelet, T. Hueber, L. Girin, C. Savariaux, and B. Yvert, "Real-time control of a DNN-based articulatory synthesizer for silent speech conversion: A pilot study," in Proc. InterSpeech2015, pp.2405-2409 (2015).

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

[20] Z. C. Liu, Z. H. Ling, and L. R. Dai, "Articulatory-to-acoustic conversion with cascaded prediction of spectral and excitation features using neural networks," in Proc. of InterSpeech2016, pp1502-1506 (2016).

[21] B. Cao, M. Kim, J. R. Wang, J. Santen, T. Mau, and J. Wang, "Articulation-to-speech synthesis using articulatory flesh point sensors' orientation information," in Proc. of InterSpeech2018, pp.3152-3156 (2018).

[22] Z. C. Liu, Z. H. Ling, and L. R. Dai, "Articulatory-to-acoustic conversion using BLSTM-RNNs with augmented input representation," Speech Communication, vol. 99, pp.161-172 (2018).

[23] F. Taguchi and T. Kaburagi, "Articulatory-to-speech conversion using bi-directional long short-term memory," in Proc. of InterSpeech2018, pp.2499-2503 (2018).

[24] A. Illa and P. K. Ghosh, "An investigation on speaker specific articulatory synthesis with speaker independent articulatory inversion," in Proc. of InterSpeech2019, pp.121-125 (2019).

[25] A. Illa and P. K. Ghosh, "Inferring speaker identity from articulatory motion during speech," in Proc. of MLSLP2018 (2018).

[26] A. Wrench and W.J. Hardcastle, "A multichannel articulatory database and its application for automatic speech recognition," 5th Seminar on Speech Production: Models and Data, pp.305–308 (2000).

[27] K. Richmond, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," in Proc. of IntersSeech2011, pp.1505-1508 (2011).

[28] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," Language Resources and Evaluation, vol.46, no.4, pp.1-19 (2012).

[29] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. Ghosh, A. Katsamanis, and M. Proctor, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," J. Acoust. Soc. Am., vol.136, no.3, pp.1307-1311 (2014).

[30] A. Ji1, J. J. Berry, and M. T. Johnson, "The electromagnetic articulography Mandarin accented English (EMA-MAE) corpus of acoustic and 3D articulatory kinematic data," in Proc. ICASSP2014, pp.7769-7773 (2014).

[31] J. Scobbie, A. Turk, C. Geng, S. King, R. Lickley, and K. Richmond, "The Edinburgh speech production facility DoubleTalk corpus," in Proc. of InterSpeech2013, pp.764-766 (2013).

[32] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in Proc. of ICASSP1988, pp.I-185-I-188 (1988).

[33] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand and V. Lempitsky, "Domain-adversarial training of neural networks," Journal of Machine Learning Research, vol.17, pp.1-35 (2016).