INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Articulatory-to-Acoustic Conversion with Cascaded Prediction of Spectral and Excitation Features Using Neural Networks

Zheng-Chen Liu, Zhen-Hua Ling, Li-Rong Dai

National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, P.R.China

liuzhch@mail.ustc.edu.cn, {zhling, lrdai}@ustc.edu.cn

Abstract

This paper presents an articulatory-to-acoustic conversion method using electromagnetic midsagittal articulography (EMA) measurements as input features. Neural networks, including feed-forward deep neural networks (DNNs) and recurrent neural networks (RNNs) with long short-term term memory (LSTM) cells, are adopted to map EMA features towards not only spectral features (i.e. mel-cepstra) but also excitation features (i.e. power, U/V flag and F0). Then speech waveforms are reconstructed using the predicted spectral and excitation features. A cascaded prediction strategy is proposed to utilize the predicted spectral features as auxiliary input to boost the prediction of excitation features. Experimental results show that LSTM-RNN models can achieve better objective and subjective performance in articulatory-to-spectral conversion than DNNs and Gaussian mixture models (GMMs). The strategy of cascaded prediction can increase the accuracy of excitation feature prediction and the neural network-based methods also outperform the GMM-based approach when predicting power features.

Index Terms: articulatory-to-acoustic conversion, deep neural network, recurrent neural network, Gaussian mixture model

1. Introduction

Speech is originated from articulatory movements which involve the systematic motions of a series of apparatus such as tongue, jaw, velum, etc. Therefore, articulatory features and acoustic features are inherently related. Similar to acoustic-to-articulatory inversion mapping [1], the conversion from articulatory features to acoustic features is also useful in many applications. In speech synthesis, the characteristics of the synthetic speech can be conveniently controlled by manipulating articulatory features [2]. In silent speech interface (SSI) [3], the articulatory-to-acoustic conversion can be used as an alternative to faciliate conversation in high-background-noise environments, or as an aid for the speech-handicapped, such as the laryngectomees.

To capture the movements of articulators, different techniques have been proposed [4–6]. Electromagnetic midsagittal articulography (EMA) [7] is one of them and has been exploited in many studies. EMA data is recorded by a set of sensors glued on articulators. It has quite high temporal resolution and is able to track the motions of the main articulators accurately. In previous work, EMA data has been widely utilized in speech recognition [8], acoustic-articulatory inversion mapping [9] and text-to-speech synthesis [2]. In this paper, EMA data is adopted as input for articulatory-to-acoustic conversion.

Various methods have been proposed to convert articulatory features towards acoustic features and they can be grouped into two main categories: physical-model-based approach and data-driven approach [10]. In the physical-model-based approach, physical models are constructed to approximate the structure of vocal tract and the speech production mechanism. Then speech signals can be generated by controlling excitation and vocal tract using articulatory measurements [11]. The data-driven approach learns the mapping relationship between articulatory and acoustic features from training data using statistical models. This approach developed rapidly in the last decade and is the focus of this paper. Different statistical models have been utilized in articulatory-to-acoustic conversion, such as Gaussian mixture models (GMM) [12], hidden Markov models (HMM) [6, 13], and deep neural networks (DNN) [10]. These work can obtain quite accurate spectral trajectories from articulatory parameters. However, the mapping towards excitation features has not been comprehensively investigated. On the other hand, recurrent neural networks (RNN) with long short-term memory (LSTM) cells [14] have been successfully applied to some speech sequence generation tasks, such as text-to-speech synthesis [15] and acoustic-to-articulatory inversion mapping [16]. Compared with feed-forward neural networks, RNNs provide better ability of processing sequential data by using cyclical connections among hidden nodes. Therefore, it is worthwhile to investigate the performance of using LSTM-RNNs in the articulatory-to-acoustic conversion task.

The contribution of this paper is two-fold. First, LSTM-RNNs are introduced to map EMA features towards spectral parameters. Experimental results show that this approach can achieve better objective and subjective performance than GMM and DNN-based methods. Second, the feasibility of reconstructing excitation features, including power, U/V flag and F0, from EMA features is explored. A cascaded prediction strategy is proposed, in which the predicted spectra are used as input to help the excitation prediction. Finally a complete process flow of EMA-to-waveform transformation can be achieved.

The rest of this paper is organized as follows. Section 2 provides a brief overview of the models used in our work. Section 3 introduces our proposed methods in detail. Experimental results and conclusions are given in Section 4 and 5 respectively.

2. Previous Work

2.1. GMM-based articulatory-to-acoustic conversion

Consider a sequence of input articulatory feature vectors $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T]$ and a parallel sequence of output acoustic

This work was partially funded by the National Natural Science Foundation of China (Grant No. 61273032) and National Key Technology Support Program (2014BAK15B05).

http://dx.doi.org/10.21437/Interspeech.2016-715

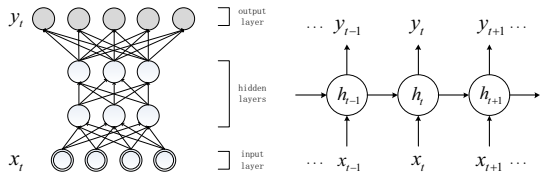

Figure 1: *Illustrations of a DNN with two hidden layers (a) and an RNN with one hidden layer (b).*

feature vectors $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T]$, where T is the number of frames. The joint distribution of the articulatory and acoustic features modeled by a GMM is

$$p(\boldsymbol{z}_t|\boldsymbol{\Theta}) = p(\boldsymbol{x}_t, \boldsymbol{y}_t|\boldsymbol{\Theta}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{z}_t, \boldsymbol{\mu}_m^{\boldsymbol{z}}, \boldsymbol{\Sigma}_m^{\boldsymbol{z}}),$$
$$\boldsymbol{z}_t = \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{y}_t \end{bmatrix}, \boldsymbol{\mu}_m^{\boldsymbol{z}} = \begin{bmatrix} \boldsymbol{\mu}_m^{\boldsymbol{x}} \\ \boldsymbol{\mu}_m^{\boldsymbol{y}} \end{bmatrix}, \boldsymbol{\Sigma}_m^{\boldsymbol{z}} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{\boldsymbol{xx}} & \boldsymbol{\Sigma}_m^{\boldsymbol{xy}} \\ \boldsymbol{\Sigma}_m^{\boldsymbol{yx}} & \boldsymbol{\Sigma}_m^{\boldsymbol{yy}} \end{bmatrix}, \quad (1)$$

where $\boldsymbol{\Theta}$ denotes the parameter set of the GMM which can be estimated from training data using EM algorithm under maximum likelihood criterion, $\mathcal{N}(\cdot, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, M is the number of mixture components, and α_m means the weight of the m-th component. At mapping stage, the distribution of acoustic features given input articulatory feature sequence can be derived from (1). Then, the converted acoustic features can be estimated using minimum mean-square error (MMSE) criterion or by maximum likelihood estimation (MLE) [17].

2.2. DNN-based articulatory-to-acoustic conversion

A DNN is a multi-layer perceptron with several hidden layers between the input layer and the output layer, as illustrated in Fig.1(a). In a DNN, units in one layer are fully connected to the units in the layer above, but there are no connections among units in the same layer. DNNs can be trained in a two-stage strategy: the pre-training stage and the fine-tuning stage [18]. For regression tasks, the fine-tuning stage adopts back-propagation (BP) algorithm to minimize the mean square error (MSE) on training set. DNNs have been applied to articulatory-to-acoustic conversion in [10], where the input vector $\boldsymbol{x}_t$ includes EMA features, frame energy, F0 and nasality at each frame, and the output vector $\boldsymbol{y}_t$ denotes MFCC features calculated from STRAIGHT spectrum [19]. Experimental results in [10] show that DNNs can achieve better performance in predicting spectral features than GMMs.

2.3. RNNs with LSTM units

RNNs are artificial neural networks with connections between units forming a directed cycle. A typical structure of RNNs is shown in Fig.1(b) where there are cyclical connections among hidden units at the same layer. Unlike feed-forward neural networks, RNNs can make better use of the context information of the input sequence, which makes it more powerful when classifying or generating sequential data. In an RNN with one hidden layer, given an input sequence $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T]$, the sequences of hidden vectors $\boldsymbol{h} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_T]$ and output vectors $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T]$ can be calculated as

$$\begin{aligned} \boldsymbol{h}_t &= \mathcal{H}(\boldsymbol{W}_{xh}\boldsymbol{x}_t + \boldsymbol{W}_{hh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_h), \\ \boldsymbol{y}_t &= \boldsymbol{W}_{hy}\boldsymbol{h}_t, \end{aligned} \quad (2)$$

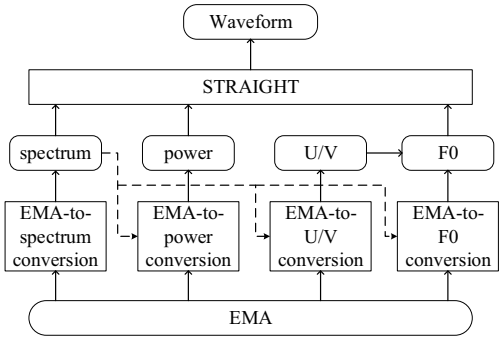

Figure 2: *Diagram of the proposed articulatory-to-acoustic conversion method. The dashed lines indicate the cascaded prediction strategy.*

where $t \in [1, T]$, T is the number of frames, $\mathcal{H}$ is the activation function, $\boldsymbol{W}$ and $\boldsymbol{b}$ denote weight matrices and bias vectors respectively. A deep RNN can be built up by stacking multiple recurrent hidden layers one on top of another. Usually, the back-propagation through time (BPTT) algorithm, which is a generalization of the BP for feed-forward networks, is used to train an RNN [20]. Traditional RNNs have the vanishing (and exploding) gradient problem. Thus, the long short-term memory (LSTM) architecture has been proposed to deal with this issue [14]. An LSTM unit is a complex hidden unit. It has three gates, namely input gates, output gates and forget gates, which determine when to remember the input, when to output the value and whether to remember or forget the value respectively. The detailed definition of the activation function in an LSTM unit used in this paper can be found in [21]. Recently, LSTM-RNNs have been successfully applied to speech generation tasks such as text-to-speech synthesis [15] and visual speech synthesis [22].

3. Proposed Method

Different from previous work [10, 17] which only predicts spectral features from EMA data, this paper aims at reconstructing speech waveforms using only EMA input. Therefore, not only spectral features but also excitation features, including power, U/V flag, and F0 values, need to be generated by articulatory-to-acoustic conversion. The schematic diagram of the proposed method is shown in Fig. 2. STRAIGHT [19] is adopted as the vocoder for acoustic feature extraction and speech waveform reconstruction in this paper. The spectral features at each frame are represented by the mel-cepstral coefficients (MCC) derived from STRAIGHT spectrum excluding the 0-th order. The power at each frame is calculated as the 0-th order of MCCs. The U/V flags and F0 values are also extracted by STRAIGHT analysis. Considering the different properties of spectral and excitation features, these four acoustic features are converted from EMA input using separate models in our method.

For EMA-to-spectrum conversion, three types of models, i.e. GMMs, DNNs, and LSTM-RNNs are implemented and compared. These models are built following the methods introduced in Section 2, where $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ denotes the input EMA feature vector and output MCC feature vector at the t-th frame respectively. Both features contain dynamic components and parameter generation algorithm with dynamic features [23] is adopted at mapping stage to generate smooth MCC trajectories. The model configurations, such as the number of mixtures for GMMs, the depths and hidden unit numbers for

DNNs and LSTM-RNNs, are tuned using a validation set in experiments.

Compared with EMA-to-spectrum conversion, to predict excitation features, such as powers and F0 contours, from EMA data is more difficult because EMA data describes the characteristic of vocal tract and its relationship with vocal cord excitation is not explicit. However, researchers have observed the influence of F0 on vowel articulation [24]. Therefore, it is worthwhile to investigate if state-of-the-art machine learning algorithms can model such influence and further predict excitation features from EMA observations. The three types of models for EMA-to-spectrum conversion are also applied to the prediction of excitation features. Considering that U/V flags are binary and it may be inappropriate to predict U/V flags using GMMs, only DNNs and RNNs are adopted to predict U/V flags and F0 values in our experiments. The neural networks for predicting U/V flag conversion are trained for classification instead of regression under cross-entropy (CE) criterion. When training the neural networks for F0 prediction, F0 interpolation at unvoiced frames is conducted first using an exponential decay function [25]. At conversion time, continuous F0 contours and binary U/V flags are mapped from input EMA features simultaneously. Then, the F0 values at the frames predicted to be unvoiced are discarded to form final F0 contours with unvoiced segments.

Limited by the number of sensors and the difficulty of data acquisition, EMA features are always low-dimensional and can only describe the articulatory configurations during pronouciation roughly. In order to utilize more information that describes the characteristics of vocal tract for EMA-to-excitation conversion, a cascaded prediction strategy is proposed and is shown as dashed lines in Fig. 2. In this strategy, the predicted spectral features are concatenated with EMA features to form the input vectors for predicting powers, U/V flags and F0 values. Although the predicted spectral features may be inaccurate, they are expected to carry on more detailed vocal tract information than EMA measurements and thus to boost the performance of predicting excitation features.

4. Experiments

4.1. Experimental setup

The MNGU0 database [26] was used in our experiments. This database consisted of 1263 British English utterances from one male speaker with parallel acoustic and EMA recordings. EMA features were captured from 6 sensors located at tongue dorsum, tongue body, tongue lip, jaw, upper lip and lower lip with a sampling frequency of 200Hz. For each sensor, the coordinates on the front-to-back axis and the bottom-to-top axis (relative to viewing the speakers head from the front) were used, making a total of 12 static EMA features at each frame. The waveforms were in 16 kHz PCM format with 16 bit precision. Spectral features (i.e. 1-st to 40-th orders of MCCs), powers (i.e. 0-th order of MCCs), U/V flags and logarithmic F0 values were extracted from the waveform using STRAIGHT vocoder [19]. As introduced in Section 3, the F0 values at unvoiced frames were interpolated by an exponential decay function [25]. The first derivatives of EMA, spectral, power, and F0 features were employed as dynamic features. Considering that the available EMA data for one speaker was usually small, 100 utterances were sampled from the database to build the training set for our experiments. The validation set and test set contained 13 and 20 utterances respectively.

Table 1: The average MCDs (dB) on test set of using GMMs, DNNs, and LSTM-RNNs for EMA-to-spectrum conversion.

Method	GMM	DNN	LSTM-RNN
MCD	3.61	3.43	**3.09**

Table 2: Preference scores (%) on intelligibility (Int.) and naturalness (Nat.) between speech generated using GMMs, DNNs, and LSTM-RNNs for EMA-to-spectrum conversion, where N/P stands for "no preference" and p means the p-value of t-test between two systems.

	GMM	DNN	LSTM-RNN	N/P	p
Int.	16.50	**42.50**	-	41.00	$<$0.001
	4.17	-	**78.63**	17.20	$<$0.001
	-	3.67	**76.79**	19.53	$<$0.001
Nat.	19.70	**41.45**	-	38.85	$<$0.001
	8.43	-	**65.83**	25.74	$<$0.001
	-	2.50	**82.19**	15.31	$<$0.001

4.2. EMA-to-spectrum conversion

The performance of EMA-to-spectrum conversion using GMMs, DNNs and LSTM-RNNs was investigated by experiments. The mel-cepstral distortion (MCD) in dB [17] between the ground truth and the predicted mel-cepstra was adopted as the objective evaluation measure. In this paper, MCDs were calculated only on non-silent frames. When training GMMs, full covariance matrices were used. The number of mixtures was set to $\{2, 4, 8, 16, 32, 64, 128\}$ respectively and 32 was found to be optimum according to their MCDs on the validation set. When training DNNs, the learning rate was fixed at 0.0001 and the momentum was set to be 0.9. The DNN architecture was tuned to be two hidden layers and 4096 hidden nodes each layer using the validation set. When training LSTM-RNNs, the iteration number was set to 30 epochs. The learning rate was initialized to be 0.001, and was halved each epoch since the 15-th epoch. The model architecture was also tuned using the validation set and finally an LSTM-RNN with two hidden layers and 1024 nodes per layer was adopted.

The average MCDs on test set of using GMMs, DNNs and LSTM-RNNs were summarized in Table 1. From this table, we can see that the two neural-network-based methods achieved lower MCDs than GMMs. This confirms the effectiveness of using neural networks with deep structures in describing the non-linear and complex mapping relationship between EMA and spectral features. The LSTM-RNN-based method performed better than the DNN-based one due to its advantage of taking the input history into account.

Furthermore, three groups of ABX preference tests were conducted on the crowdsourcing platform of Amazon Mechanical Turk (`https://www.mturk.com`) to investigate the subjective performance of EMA-to-spectrum conversion using different models. In each preference test, speech waveforms of twenty utterances in the test set were reconstructed using the spectral features predicted by two different models.[1] Each pairs of generated speech were evaluated in random order by at least twenty-five English native listeners. The listeners were asked to judge which utterance sounded more intelligible and which one was more natural. The results of the preference tests were shown in Table 2. We can see that these subjective

[1]Examples of reconstructed speech waveforms in our experiments can be found at `http://home.ustc.edu.cn/~liuzhch/IS2016/demo.html`.

Table 3: The RMSEs on test set of using GMMs, DNNs, and LSTM-RNNs for EMA-to-power conversion, where "RNN" stand for "LSTM-RNN" , the prefix "cas" in method names denotes using the proposed cascaded prediction strategy, and "DNN11" means using a context window of 11 frames to form input features when training DNNs.

Method	GMM	DNN	DNN11	RNN
RMSE	1.035	0.803	0.666	**0.612**
Method	casGMM	casDNN	casDNN11	casRNN
RMSE	0.786	0.635	0.632	**0.560**

Table 4: The error rates (%) on test set of using DNNs and LSTM-RNNs for EMA-to-U/V flag conversion. The method names are the same as the ones used in Table 3.

Method	DNN	DNN11	RNN
Error Rate	26.15	23.84	**23.40**
Method	casDNN	casDNN11	casRNN
Error Rate	21.98	**20.29**	21.49

evaluation results are consistent with the objective ones shown in Table 1. Using DNNs and LSTM-RNNs for EMA-to-spectrum conversion obtained significantly higher preference scores on intelligibility and naturalness than using GMMs. LSTM-RNNs achieved the best subjective performance among these three models.

4.3. EMA-to-excitation conversion

The objective evaluation results of EMA-to-power conversion using different models were shown in Table 3. Here, the model architectures were also tuned on validation set. The GMM had 32 mixtures. The DNN with two hidden layers and 64 nodes per layer and the LSTM-RNN with two hidden layers and 16 nodes per layer were adopted. The spectral features for cascaded prediction were generated using the LSTM-RNN model shown in Table 1 and 2. From Table 3, we can see that DNNs achieved better performance than GMMs when predicting power features. The contextual information can be utilized to further improve the prediction accuracy by either using a wider input window or adopting a recurrent model structure. The proposed cascaded prediction strategy was effective for both GMMs and neural networks. The method using LSTM-RNNs and cascaded prediction obtained the lowest RMSE of power prediction. However, when listening to the reconstructed speech waveforms, we found that the accuracy of power prediction at some silent frames still needs to be improved.

The objective evaluation results of EMA-to-U/V flag conversion and EMA-to-F0 conversion were shown in Table 4 and 5 respectively. The DNN with one hidden layer and 64 hidden nodes and the LSTM-RNN with two hidden layers and 32 nodes per layer were adopted for EMA-to-U/V flag conversion. For EMA-to-F0 conversion, we adopted the DNN and LSTM-RNN models both with two hidden layers and 64 nodes per layer. As seen in these two tables, when the cascaded prediction strategy was not applied, the LSTM-RNN-based conversion achieved the best performance. DNNs benefited more from the cascaded prediction than LSTM-RNNs for predicting either U/V flags or F0s. The DNNs with cascaded prediction and an 11-frame input window achieved the best results in both tables.

In order to evaluate the subjective performance of EMA-to-excitation conversion and the overall quality of EMA-to-waveform transformation, four systems were built for listening tests, including

Table 5: The RMSEs (Hz) on test set of using DNNs and LSTM-RNNs for EMA-to-F0 conversion. The method names are the same as the ones used in Table 3.

Method	DNN	DNN11	RNN
RMSE	25.72	24.49	**23.18**
Method	casDNN	casDNN11	casRNN
RMSE	23.74	**22.76**	23.02

Table 6: Preference scores (%) on naturalness between speech generated using the four systems introduced in Section 4.3

CvAll	*CvSpe*	*HTS100*	*HTS1000*	N/P	p
34.12	40.22	-	-	25.67	0.0735
77.27	-	13.79	-	8.94	$<$0.001
18.85	-	-	**66.42**	14.73	$<$0.001

- *CvAll*: EMA-to-waveform transformation, where spectral features were predicted by the LSTM-RNN model in Table 1, powers were predicted by the casRNN model in Table 3, and U/V flags and F0s were predicted by the casDNN11 models in Table 4 and 5;
- *CvSpe*: EMA-to-waveform transformation, where spectral features were predicted by the LSTM-RNN model in Table 1, and the other features were natural;
- *HTS100*: HMM-based text-to-speech synthesis [27] using the same training set of 100 utterances as EMA-to-waveform transformation;
- *HTS1000*: HMM-based text-to-speech synthesis using 1000 utterances in the MNGU0 database as training set.

Three groups of ABX preference tests were conducted on the crowdsourcing platform of Amazon Mechanical Turk (`https://www.mturk.com`) to compare the naturalness of speech produced by these four systems. Each test adopted twenty sentences in the test set and employed thirty English native listeners. The results were summarized in Table 6. Comparing the *CvAll* and *CvSpe* systems, we can see there was no significant difference between using the excitation features predicted from EMA data by neural networks and using natural excitation features in terms of the naturalness of reconstructed speech waveforms. In the preference tests of comparing the *CvAll* system with the two speech synthesis systems, the power of silent frames predicted by the *CvAll* system was fixed to be zero to ignore the power prediction error at non-speech frames. Under this condition, the output of the *CvAll* system was better than the HMM-based parametric speech synthesis system built using the same database and worse than the synthesis system built using a larger training set as shown in Table 6.

5. Conclusions

This paper has proposed an articulatory-to-acoustic conversion method using neural networks. In this method, spectral features and excitation features are predicted in a cascaded way from EMA input. Experimental results show that LSTM-RNNs achieve significantly better performance than GMMs and DNNs in EMA-to-spectrum and EMA-to-power conversion. The effectiveness of the cascaded prediction strategy is also demonstrated by experiments. To utilize other neural network models (such as bidirectional LSTM-RNNs) and more articulatory input (such as fMRI images) for articulatory-to-acoustic conversion will be tasks of our future work.

6. References

[1] 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.

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

[3] 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.

[4] T. Schultz and M. Wand, "Modeling coarticulation in EMG-based continuous speech recognition," *Speech Communication*, vol. 52, no. 4, pp. 341–353, 2010.

[5] S. Narayanan, E. Bresch, P. K. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim, A. C. Lammert, M. I. Proctor, V. Ramanarayanan, Y. Zhu *et al.*, "A multimodal real-time MRI articulatory corpus for speech research." in *INTERSPEECH*, 2011, pp. 837–840.

[6] T. Hueber and G. Bailly, "Statistical conversion of silent articulation into audible speech using full-covariance HMM," *Computer Speech & Language*, vol. 36, pp. 274–293, 2016.

[7] 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.

[8] A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," 2000.

[9] 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, 2010.

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

[11] P. Mermelstein, "Articulatory model for the study of speech production," *The Journal of the Acoustical Society of America*, vol. 53, no. 4, pp. 1070–1082, 1973.

[12] T. Toda, A. W. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[13] K. Nakamura, T. Toda, Y. Nankaku, and K. Tokuda, "On the use of phonetic information for mapping from articulatory movements to vocal tract spectrum," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I–I.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] 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.

[16] P. Zhu, L. Xie, and Y. Chen, "Articulatory movement prediction using deep bidirectional long short-term memory based recurrent neural networks and word/phone embeddings," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

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

[18] G. E. Hinton, "A practical guide to training restricted boltzmann machines," in *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 599–619.

[19] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 2. IEEE, 1997, pp. 1303–1306.

[20] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," DTIC Document, Tech. Rep., 1985.

[21] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM recurrent networks," *The Journal of Machine Learning Research*, vol. 3, pp. 115–143, 2003.

[22] B. Fan, L. Wang, F. K. Soong, and L. Xie, "Photo-real talking head with deep bidirectional LSTM," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4884–4888.

[23] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1315–1318.

[24] K. Honda, "Relationship between pitch control and vowel articulation," *Haskins Laboratories Status Report on Speech Research, SR*, vol. 73, pp. 269–82, 1983.

[25] C. J. Chen, R. A. Gopinath, M. D. Monkowski, M. A. Picheny, and K. Shen, "New methods in continuous mandarin speech recognition." in *Eurospeech*, 1997.

[26] K. Richmond, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus." in *Interspeech*, 2011, pp. 1505–1508.

[27] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.