Interspeech 2018
2-6 September 2018, Hyderabad

Articulatory-to-speech conversion using bi-directional long short-term memory

Fumiaki Taguchi[1], *Tokihiko Kaburagi*[2]

[1]Graduate School of Design, Kyushu University Shiobaru 4–9–1, Minami-ku, Fukuoka-shi, Fukuoka, 815–8540 Japan
[2]Faculty of Design, Kyushu University Shiobaru 4–9–1, Minami-ku, Fukuoka-shi, Fukuoka, 815–8540 Japan

taguchi.f.664@s.kyushu-u.ac.jp, kabu@design.kyushu-u.ac.jp

Abstract

Methods for synthesizing speech sounds from the motion of articulatory organs can be used to produce substitute speech for people who have undergone laryngectomy. To achieve this goal, feature parameters representing the spectral envelope of speech, directly related to the acoustic characteristics of the vocal tract, has been estimated from articulatory movements. Within this framework, speech can be synthesized by driving the filter obtained from a spectral envelope with noise signals. In the current study, we examined an alternative method that generates speech sounds directly from the motion pattern of articulatory organs based on the implicit relationships between articulatory movements and the source signal of speech. These implicit relationships were estimated by considering that articulatory movements are involved in phonological representations of speech that are also related to sound source information such as the temporal pattern of pitch and voiced/unvoiced flag. We developed a method for simultaneously estimating the spectral envelope and sound source parameters from articulatory data obtained with an electromagnetic articulography (EMA) sensor. Furthermore, objective evaluation of estimated speech parameters and subjective evaluation of the word error rate were performed to examine the effectiveness of our method.

Index Terms: Articulatory movement, EMA, Vocal-tract spectrum, Deep learning, Articulatory-to-acoustic mapping

1. Introduction

A range of methods of articulatory-to-acoustic mapping have been developed to estimate the acoustic parameters of speech from the movement patterns of articulatory organs [1–6]. Such methods are beneficial for acquired speech-impaired people, providing a way of producing substitute speech that can be generated directly from articulatory movements, without linguistic content. Articulatory movements have also been used to construct a brain-machine interface [4, 5].

Electromagnetic articulography (EMA), a well-known method of observing articulatory movements, can be used to track the position of multiple small marker coils attached to the articulatory organs such as the jaw, upper and lower lips, and tongue. The articulatory data recorded by EMA have a higher temporal resolution than other methods such as magnetic resonance imaging (MRI) or ultrasonic scanners [7]. Furthermore, EMA does not require complex post-processing processes to obtain movement data. The obtained articulatory data have been used to study articulatory-based speech synthesis, statistical voice quality conversion, and utterance learning of a foreign language, and have been used as additional feature parameters for speech recognition. Previous studies have demonstrated the effectiveness of using articulatory data for a variety of speech signal processing techniques [8]. Articulatory movements are particularly useful because they are less affected by pitch and voiced-unvoiced switching. In addition, articulatory movements are superior to spectral feature parameters in representing the articulatory state, even for non-stationary sounds such as plosive consonants.

However, it is difficult to recover the whole vocal-tract shape from measured EMA data. To overcome this issue, previous studies have used a parallel speech corpus in which speech signals and EMA data are simultaneously recorded to learn the statistical relationships between the articulatory and acoustic parameters of speech. A codebook storing pairs of articulatory and acoustic parameters can then be used to estimate the vocal-tract spectrum from the input articulatory data by selecting the neighboring pair samples [1]. However, in this codebook search method, the amount of calculation required for selecting the neighboring data samples increases when the size of the codebook increases. In other studies, articulatory-to-acoustic mapping has been achieved using Gaussian mixture models [2], feed-forward neural networks [3], deep neural networks [4, 5], and uni-directional long short-term memory (LSTM) [6].

In these previous studies, parameters representing the spectral envelope of speech (such as the mel-cepstrum parameter) have often been used as the acoustic features. This is because the vocal tract acts as a filter in the production of human speech and determines the spectral characteristics of speech. On the other hand, the vocal tract is deeply involved in the generation of the sound source for some consonants including plosives and fricatives. In addition, the acoustic characteristics of the vocal tract are responsible for the phonological representation of speech, and sound source information, such as the temporal pattern of pitch and voiced/unvoiced flag, is relevant to phonemic information. Therefore, we can expect that there might be implicit and indirect relationships between articulatory movements and parameters regarding the sound source of speech. Estimation of sound source information has been examined previously [3, 6], but the EMA datasets used in these studies were relatively small, and recording was as short as 30 minutes or less.

In the current study, we constructed an articulatory-to-speech conversion model that estimates not only the feature parameters representing the spectral envelope but also the parameters regarding the sound source of speech from articulatory data obtained with EMA. Our conversion model was constructed by training a bi-directional recurrent neural network (BRNN) with the mngu0 EMA dataset [9], which has a dataset of more than 1

10.21437/Interspeech.2018-999

hour. In addition, to determine the effectiveness of our model, we performed an objective evaluation of speech parameters estimated with the model, and a subjective evaluation examining the word error rate of synthetic speech.

2. Method

2.1. Recurrent Neural Networks

2.1.1. Bi-directional recurrent neural networks

We used a BRNN [10] method to consider forward and backward time series simultaneously by combining two recurrent structures. When time series data $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T)$ are given, $X_{-1} = (\boldsymbol{x}_T, ..., \boldsymbol{x}_2, \boldsymbol{x}_1)$ represents the backward series. Let us denote two recurrent structures as $\boldsymbol{R}_f$ and $\boldsymbol{R}_b$. The BRNN can then be written as follows:

$$H_f = (\boldsymbol{h}_{f1}, \boldsymbol{h}_{f2}, ..., \boldsymbol{h}_{fT}) = \boldsymbol{R}_f(X) \quad (1)$$

$$H_{b-1} = (\boldsymbol{h}_{bT}, ..., \boldsymbol{h}_{b2}, \boldsymbol{h}_{b1}) = \boldsymbol{R}_b(X_{-1}) \quad (2)$$

and

$$H = \begin{bmatrix} H_f \\ H_b \end{bmatrix} = \left(\begin{bmatrix} \boldsymbol{h}_{f1} \\ \boldsymbol{h}_{b1} \end{bmatrix}, \begin{bmatrix} \boldsymbol{h}_{f2} \\ \boldsymbol{h}_{b2} \end{bmatrix}, ... \begin{bmatrix} \boldsymbol{h}_{fT} \\ \boldsymbol{h}_{bT} \end{bmatrix} \right). \quad (3)$$

When time series data X are input into the recurrent structure $\boldsymbol{R}_f$, we obtain the output H_f, which takes into account the data ordered in the forward direction. We also obtain the output H_{b-1} by entering the time series data X_{-1} ordered in the backward direction to the recurrent structure $\boldsymbol{R}_b$. The total output H can be obtained by concatenating these two outputs to simultaneously consider the forward and backward time series. The time series data of all time samples are required for the inference due to the non-causal structure of a BRNN.

2.1.2. Long short-term memory

LSTM [11] is the most representative gated recurrent structure. This gated recurrent structure was developed to solve the problem of conventional RNN methods, which exhibit inferior performance when the length of an input data series is very long. LSTM uses three gates called the input gate $\boldsymbol{i}$, forget gate $\boldsymbol{f}$, and output gate $\boldsymbol{o}$. It also uses a hidden state vector called a cell $\boldsymbol{c}$ that can hold long-term information about input data series. LSTM performs the following calculations:

$$\begin{bmatrix} \overline{\boldsymbol{h}}_t \\ \boldsymbol{i}_t \\ \boldsymbol{f}_t \\ \boldsymbol{o}_t \end{bmatrix} = \begin{bmatrix} tanh \\ sigmoid \\ sigmoid \\ sigmoid \end{bmatrix} \left(\begin{bmatrix} W_{\overline{h}} \\ W_i \\ W_f \\ W_o \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{h}_{t-1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{b}_{\overline{h}} \\ \boldsymbol{b}_i \\ \boldsymbol{b}_f \\ \boldsymbol{b}_o \end{bmatrix} \right) \quad (4)$$

$$\boldsymbol{c}_t = \boldsymbol{i}_t \odot \overline{\boldsymbol{h}}_t + \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} \quad (5)$$

and

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot tanh(\boldsymbol{c}_t). \quad (6)$$

$\overline{\boldsymbol{h}}_t$ is the same hidden state vector used in the ordinal RNN. Eq. (5) shows that $\overline{\boldsymbol{h}}_t$ is adjusted by the input gate $\boldsymbol{i}_t$ to provide the updating value of the cell $\boldsymbol{c}$. In addition, the value of the cell for the past time instant decreases by using the forget gate $\boldsymbol{f}_t$. In other words, the value of the cell is updated by balancing short- and long-term information by using the input and forget gates. Finally, Eq. (6) shows that the output of LSTM is obtained by adjusting the updated cell value using the output gate $\boldsymbol{o}_t$.

2.1.3. Layer normalization

It is known that normalizing the hidden state vector has a good effect on the training of networks [12]. As a method to normalize the value of hidden state vector, layer normalization [13] is particularly effective for constructing RNN, since the length of input data series is variable and it is not possible to use a large batch size. Regardless of the number of mini-batches, layer normalization has the same effect.

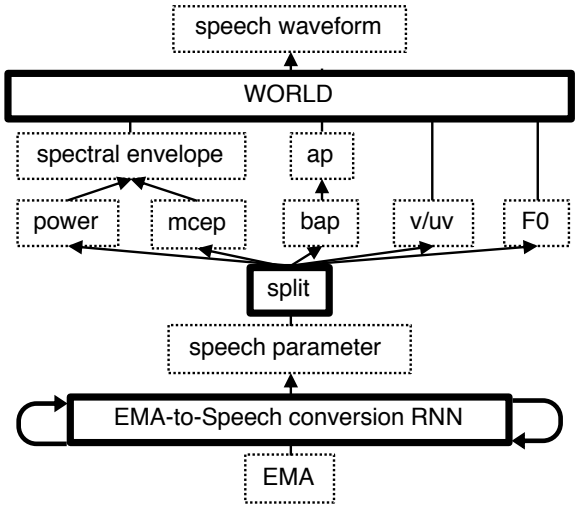


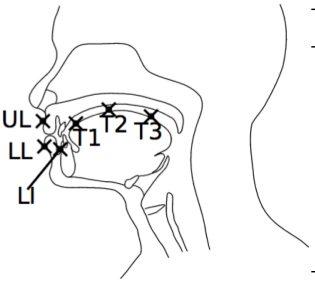
Figure 1: *Diagram of the proposed articulatory-to-speech conversion system, where "ap" means aperiodicity, "bap" means five-band averaged aperiodicity, and "v/uv" means voiced/unvoiced flag.*

2.1.4. Incremental method

The incremental method [14] is a method for training deep networks well. In this method, training of a network composed of input and output layers is first performed. When convergence is obtained, a hidden layer is added to the network, next to the input layer. Training of the new network and the addition of a hidden layer is repeated further until the desired number of hidden layers has been inserted. Note that, when a hidden layer is added, the output layer is reinitialized. Also, after the layer addition, training of the whole network should be performed, including the network layers for which the training is completed. This method is effective for training the hidden layers of a deep network.

2.2. Speech synthesis from articulatory movements

Figure 1 shows a diagram of the proposed speech synthesis method from observed articulatory movements. At the bottom of the figure, "EMA-to-speech conversion RNN" is a structure incorporating a bi-directional LSTM, which can consider the forward and backward time series of input EMA data simultaneously. When a sequence of EMA data is input, this RNN outputs a combined vector of the static and delta features of speech parameters. Speech feature parameters that have a smooth temporal trajectory can be formed from static and delta features using the maximum likelihood parameter generation (MLPG) algorithm [15]. These concatenated parameters are then separated and processed to drive a speech synthesizer called WORLD [16] (D4C Edition [17]). Finally, WORLD generates a speech signal as the output of the whole system. A cascade method has been proposed [6] in which parameter estimation is repeatedly performed to estimate other parameters, but in the current study, all the speech features were estimated through a single network.



label	location
UL	Upper lip
LL	Lower lip
LI	Lower incisor
T1	Tongue tip
T2	Tongue body
T3	Tongue dorsum

Figure 2: *Fixed positions of the receiving coils.*

RNN is trained to minimize the mean squared error with respect to the parameter vector concatenating the speech features. In our method, the conversion model is built with a single network, and can greatly reduce the labor of training and complicated ad hoc tuning of the hyper-parameters.

3. Experiment

3.1. Experimental condition

3.1.1. Articulatory-speech data pairs

In our speech synthesis method, "EMA-to-speech conversion RNN" was trained from a set of articulatory-speech data pairs. We used the mngu0 [9] dataset as an articulatory-speech parallel data corpus. This corpus was constructed by simultaneously recording speech and articulatory motion with a microphone and magnetic sensor (EMA) when a male English-speaker pronounced 1336 sentences. The position of the receiving coils attached to the articulatory organs is shown in Figure 2. The sampling frequency of EMA data was 200 Hz and that of audio signals was 16 kHz.

The articulatory data of each marker coil were represented by the two-dimensional position measured on the midsagittal plane of the speaker. We found that NaN data were included in EMA data as a result of position estimation error during data acquisition. These data were replaced with relevant values by interpolating the adjacent position data of each receiver coil over time. As the acoustic feature parameters, we used 0 to 40th-order mel-cepstrum parameters, continuous log F0 [18], voiced/unvoiced flag, and five-band averaged aperiodicity [19, 20] calculated using WORLD [16]. For mel-cepstrum parameters, we performed trajectory smoothing [21] with a low pass filter of 50 Hz. Here, the temporal adjustment of EMA data and the acoustic feature parameters were obtained so that the shift width of the analysis frame of WORLD was in accord with the sampling period of EMA data. In addition to the static features determined for each frame, their changes over time (dynamic features) were also taken into consideration as delta features for both articulatory and speech feature parameters. Each parameter was then normalized so that the mean was zero and the variance was one.

The articulatory-acoustic data pairs were divided into three subsets, and each subset was used for the training, validation of the training process, and open test of trained RNN, respectively. In the mngu0 dataset, 1336 sentences were already separated into these three subsets, 1208 for the training, 63 for the validation, and 65 for the open test. The number of data pairs was 720873 for the training, 37851 for the validation, and 39896 for the open test.

3.1.2. Training of RNN

Training of RNN was performed using articulatory-acoustic data pairs. RNN had a structure of three fully-connected layers, a layer normalization process, a sigmoid block with 128 units, two layers of bi-directional LSTM with 256 units, and an output fully-connected layer.

First, training was performed using an incremental method [14] so that the mean squared error with respect to the estimated static and delta speech feature parameters was minimized. Minimum generation error (MGE) training [22] was then performed. In this training, static features of speech parameters were generated from both the static features and dynamic features. The mean squared error was then minimized with respect to the generated static features. At each stage of the training, we used 5.0 gradient clipping and Grave ' s RMSprop [23] as an optimization method. While the training was repeated, the error for the validation dataset was calculated for each epoch, and the training was terminated when the minimum error was obtained.

4. Results and Discussion

4.1. Result of objective evaluation

The estimation error of the speech feature parameters is shown in Table 1 to compare the estimation accuracy of our method with that reported in previous studies. Here, the error for the 1st to 40th-order mel-cepstrum parameters (mcep) were evaluated in terms of mel-cepstrum distortion [dB]. The five-band averaged aperiodicity (bap), 0th-order mel-cepstrum parameter (power), and the fundamental frequency (F0) were evaluated by calculating the root mean squared error (RMSE). The voiced/unvoiced flag (v/uv) was evaluated by taking the error rate [%]. The table shows data from a previous study using the GMM-MMSE method [2], trained with a male speaker dataset included in Mocha-timit database [24] (the data set size was approximately 20 minutes). In contrast, cas DNNs [6] was trained using a subset of the mngu0 database (the dataset size was estimated as approximately 30 minutes). For cas DNNs, the estimation error of the mel-cepstrum parameters (mcep) was not shown, because the calculation procedure of the parameter values seemed to be different from the standard procedure. The results showed that the estimation error of every speech parameter was smaller in our method than in previous methods. These findings indicate that our proposed method is capable of producing speech signals from articulatory movements with better accuracy than pre-existing methods.

Figure 3 shows the estimated and target values of F0 for the sentence "Yasser Arafat understands this" included in the test subset of mngu0 database. The results revealed that the voiced/unvoiced flag and the fundamental frequency appear to be well estimated for this example.

The results of the objective evaluation revealed that the proposed method was able to estimate the speech features with better accuracy than previous methods. There are two possible reasons for this result: performance may have been enhanced by improvement of the estimation method itself, and/or the large datasets used to train the model. To determine the true cause of the improved accuracy, more detailed comparison of different estimation methods using a unified articulatory-speech dataset is required in future studies.

Table 1: *Estimated error of each feature parameter.*

	mcep [dB]	bap	power	v/uv [%]	F0 [Hz]
proposed	**4.801**	0.1260	**0.4128**	**10.32**	**10.49**
GMM-MMSE [2]	5.59	-	-	-	-
cas DNNs [6]	-	-	0.560	20.29	22.76

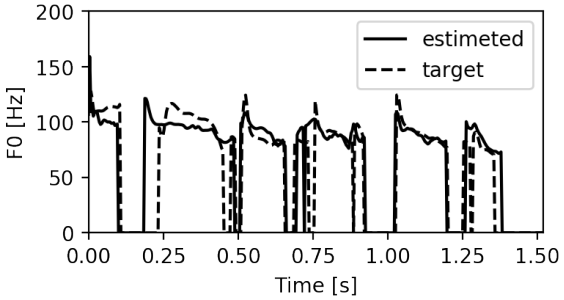


Figure 3: *The estimated and the target values of F0 for the sentence "Yasser Arafat understands this" in the mngu0 test set.*

4.2. Subjective evaluation

4.2.1. Experiment participant

Next, to perform a subjective evaluation of synthesized speech, we conducted a transcription test of speech samples generated from EMA data through Amazon Mechanical Turk (mturk). Analysis-synthesis version of each test sample (that can be regarded as the target quantity in RNN training) was also included. A sample of 10 participants took part in the test. All participants used mturk from the United States or the UK, and all declared that their native language was English. The following instructions were given to participants before they listened to the speech samples: (1) The use of earphones or headphones was recommended, (2) Speech samples were in English, and were grammatically correct, and (3) The quality of speech samples was degraded due to noise or a distortion.

4.2.2. Stimuli

Each stimulus was a speech sample synthesized from the speech feature parameters estimated from EMA data or synthesized speech where parameter values were obtained by analyzing the original speech signal (i.e., the target in RNN training). We used 65 sentences included in the test subset of the mngu0 database. Each participant transcribed these sentences once in a random order, which was synthesized from either estimated parameters or analyzed parameters. The number of the estimated version and that of the analyzed version was almost identical (i.e., 33 for estimated and 32 for analyzed, or vice versa, depending on the participant). As a result, the number of total speech samples was 130, and each sample was respectively transcribed by five participants. During the transcription, the participant listened to each sample as many times as they wanted.

4.2.3. Results

Figure 4 Shows the mean word error rate (WER) over all participants. The WER of speech sample estimated from EMA was 30.1% on average and the WER of the target value was 14.3%. In addition, the average of the mean WER difference between the target and the estimated value for each participant was 15.1%.

In the subjective evaluation, the word error rate (WER) of

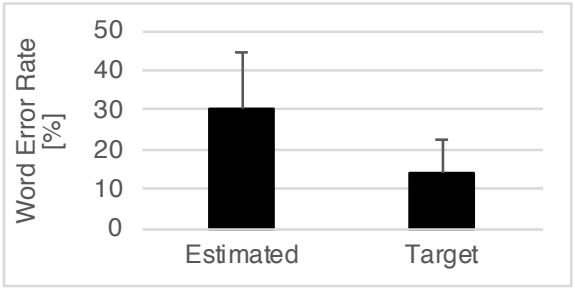


Figure 4: *Word error rate of transcription. Error bars show the standard deviation of the error rate over all participants.*

the method using a neural network [3], which was trained with a female speaker dataset included in Mocha-timit database [24] (dataset size of approximately 20 minutes), was approximately 35% to 40%, indicating that our method was capable of producing more acceptable speech signals. In our method, the difference in WER of synthesized speech between estimated parameter values and analyzed parameter values was also smaller. One possible reason for the low WER for the analyzed parameter values is that many proper nouns, such as the name of a person, were included in the speech samples. In addition, WER varied widely among participants. This may have been because the subjective evaluation was performed using mturk, and the experimental conditions, such as the type of headphones and environmental noise, were different for each participant.

5. Conclusion

The current study tested a new method for estimating feature quantities of speech, including the spectral envelope and sound source information, based on the movement trajectory of the articulatory organs. The results revealed that our proposed method was capable of producing speech from articulatory information. Using a set of articulatory-speech parallel data (i.e., the mngu0 dataset), a recurrent neural network was trained to realize the articulatory-to-speech conversion. The experimental results indicated that the proposed method can estimate speech features better than previous methods. The results of a subjective evaluation revealed that the WER was 30.1%. The intelligibility of synthesized speech was also better than that of pre-existing methods.

We plan to extend this research to compare different articulatory-to-speech conversion methods with a unified dataset in future. In addition, we plan to conduct a subjective evaluation in terms of the naturalness and speaker individuality of synthesized speech. Finally, another articulatory-to-speech conversion model will be built without using the WORLD vocoder, for Japanese.

6. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP16K00242. We thank Benjamin Knight, MSc., from Edanz Group (www.edanzediting.com/ac) for editing a draft of this manuscript.

7. References

[1] T. Kaburagi and M. Honda, "Determination of the vocal tract spectrum from the articulatory movements based on the search of an articulatory-acoustic database," in *Proc. ICSLP*, Dec. 1998, pp. 433–436.

[2] 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. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639307001495

[3] C. Kello and D. Plaut, "A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters," *The Journal of the Acoustical Society of America*, vol. 116, pp. 2354–64, 11 2004.

[4] F. Bocquelet, T. Hueber, L. Girin, P. Badin, and B. Yvert, "Robust articulatory speech synthesis using deep neural networks for bci applications," in *Proc. INTERSPEECH*, Jan. 2014, pp. 2288–2292.

[5] 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. INTERSPEECH*, Sep. 2015.

[6] 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. INTERSPEECH*, Sep. 2016, pp. 1502–1506.

[7] T. Kaburagi, K. Wakamiya, and M. Honda, "Three-dimensional electromagnetic articulography: A measurement principle," *The Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 428–443, 2005. [Online]. Available: https://doi.org/10.1121/1.1928707

[8] 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, Nov. 2015.

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

[10] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *Trans. Sig. Proc.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1109/78.650093

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[12] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, p. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[13] L. J. Ba, R. Kiros, and G. E. Hinton, "Layer normalization," *CoRR*, p. abs/1607.06450, 2016. [Online]. Available: http://arxiv.org/abs//607.06450

[14] M. Hermans and B. Schrauwen, "Training and analysing deep recurrent neural networks," in *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., 2013, pp. 190–198. [Online]. Available: http://papers.nips.cc/paper/5166-training-and-analysing-deep-recurrent-neural-networks.pdf

[15] K. Tokuda, T. Kobayashi, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *Proc. IEEE ICASSP*, 2000, pp. 1315–1318.

[16] 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. E99.D, no. 7, pp. 1877–1884, 2016.

[17] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57 – 65, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639316300413

[18] K. Yu and S. Young, "Continuous F0 modeling for HMM based statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1071–1079, July 2011.

[19] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *inProc. Int. Workshop Models Anal. Vocal Emissions Biomed. Appl.*, Sep. 2001, pp. 1–6.

[20] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with straight mixed excitation," in *Proc. INTERSPEECH*, Sep. 2006, pp. 2266–2269.

[21] S. Takamichi, K. Kobayashi, K. Tanaka, T. Toda, and S. Nakamura, "The NAIST text-to-speech system for the blizzard challenge 2015," in *Proc. Blizzard Challenge Workshop*, Sep. 2015.

[22] Z. Wu and S. King, "Improving trajectory modelling for DNN-based speech synthesis by using stacked bottleneck features and minimum generation error training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1255–1265, July 2016.

[23] A. Graves, "Generating sequences with recurrent neural networks." *CoRR*, vol. abs/1308.0850, 2013. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1308.html#Graves13

[24] A. Wrench, "The mocha-timit articulatory database," 1999. [Online]. Available: http://www.cstr.ed.ac.uk/artic/mocha.html