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

Objective Evaluation Using Association Between Dimensions Within Spectral Features for Statistical Parametric Speech Synthesis

Yusuke Ijima[1], *Taichi Asami*[1], *Hideyuki Mizuno*[2]

[1]NTT Media Intelligence Laboratories, NTT Corporation.
[2]Tokyo University of Science, Suwa.

Abstract

This paper presents a novel objective evaluation technique for statistical parametric speech synthesis. One of its novel features is that it focuses on the association between dimensions within the spectral features. We first use a maximal information coefficient to analyze the relationship between subjective scores and associations of spectral features obtained from natural and various types of synthesized speech. The analysis results indicate that the scores improve as the association becomes weaker. We then describe the proposed objective evaluation technique, which uses a voice conversion method to detect the associations within spectral features. We perform subjective and objective experiments to investigate the relationship between subjective scores and objective scores. The proposed objective scores are compared to the mel-cepstral distortion. The results indicate that our objective scores achieve dramatically higher correlation to subjective scores than the mel-cepstral distortion.
Index Terms: Statistical parametric speech synthesis, objective evaluation, spectral features, maximal information coefficient

1. Introduction

Objective evaluation of synthesized speech is one of the most important issues in the field of statistical parametric speech synthesis (SPSS) [1]. To evaluate synthesized speech quality, subjective evaluations are generally used. However, since subjective evaluations entail high cost, objective evaluations are also used widely to evaluate synthesized speech. Therefore, identifying an objective evaluation index that is highly correlated with subjective scores leads to efficient research on SPSS. If such an evaluation index could be defined, it could also be used as objective functions in SPSS. The objective evaluation index that has generally been used is parameter generation errors between natural speech and synthesized speech, e.g., mel-cepstral distortion and RMS errors of F0 and phoneme duration.

Recently, two main approaches have been used in order to improve synthesized speech quality. The first is to introduce a sophisticated machine learning technique for reducing the parameter generation errors between natural and synthesized speech. Techniques introduced for this purpose include the use of Gaussian process regression [2], deep neural networks [3], and deep recurrent neural networks [4]. The other approach is the use of speech parameter generation methods that take into account differences between the properties of natural speech and synthesized speech, such as global variance (GV) [5], local variance (LV) [6], and modulation spectrum (MS) [7]. Although these approaches effectively improve synthesized speech quality, the conventional objective evaluation index, i.e., parameter generation error, cannot necessarily be associated with the subjective scores obtained from them. This is because it is generally known that the parameter generation algorithms that take GV into account increase not only subjective scores but the mel-cepstral distortion. In this study, our aim is to identify a novel objective evaluation index that could be associated with the subjective scores obtained from these methods.

The key idea of our technique is to focus on "association" between dimensions within spectral features. In general, for natural speech, none of these associations exist since the dimensions of spectral features are mutually independent. On the other hand, in SPSS, spectral features are generated from statistical models such as HMM, which consist of a finite number of model parameters. Therefore, there would be certain associations between dimensions within the spectral features of synthesized speech. Although the analysis of degradation factors in HMM-based speech synthesis [8] implied that the association between dimensions within spectral features is one of the degradation factors, the relationship between the association and the subjective score was not clear because no quantitative analysis regarding the association was done.

In this paper, we first quantitatively analyze the relationship with the subjective score and the association between dimensions within spectral features. We then describe our objective evaluation method that takes into account the difference between the association of synthesized speech and that of natural speech by detecting the association from spectral parameter sequences. To detect the association between dimensions within spectral features, we utilized a voice conversion technique. In the proposed technique, we first divide each of the training spectral features into two spectral features. Then, we train voice conversion models to convert the divided spectral features into each other. These models can capture the association between dimensions. When the evaluation spectral features are obtained, the spectral features are converted using the trained voice conversion models. Finally, we obtained the estimation error as the evaluation index by calculating the spectral distance between the converted evaluation spectral features and the input evaluation spectral features.

2. Evaluation dataset

2.1. Speech data

We used speech data uttered by Japanese professional narrators, one male and one female. The male speaker uttered 779 sentences and the female speaker uttered 612 sentences. The sampling frequency of the speech was 22.05 kHz and the quantization bit rate was 16 bits. All speech samples were manually labeled with the phoneme segmentations and the accentual information.

For comparison with natural speech, we also used synthesized speech generated from four types of speech synthe-

http://dx.doi.org/10.21437/Interspeech.2016-584

Table 1: MOS scores obtained from the subjective evaluation.

speaker	natural speech	HMM	HMM (GV)	HMM (GV+MS)	NN
male	4.45	1.99	2.77	2.77	2.84
female	4.31	1.78	2.86	2.78	3.10

sis techniques, i.e., HMM-based speech synthesis [9], HMM-based speech synthesis taking global variance into account [5], HMM-based speech synthesis taking global variance and modulation spectrum into account [7], and neural network (NN)-based speech synthesis [3]. The training data comprised 679 of the 779 sentences uttered by the male speaker and 512 of the 612 sentences uttered by the female speaker. For the HMM-based speech synthesis, we used a five-state left-to-right hidden semi-Markov model with no skip topology. The output distribution in each state was a single Gaussian density function, and the covariance matrices were assumed to be diagonal. The control parameter of the model size was set to $\alpha = 1$. For the NN-based speech synthesis, the linguistic features were converted into 300 dimensional vectors for each frame. We set the number of hidden layers at 3 and the number of units per layer at 256. When synthesizing speech parameters by using the NN, the output parameters were modified by using global variance-based post filter [10]. We used STRAIGHT analysis [11] for speech feature extraction. The analysis frame shift was 5 ms. The spectral envelope was converted to mel-cepstral coefficients using a recursion formula. The aperiodic feature was also converted to average values for five frequency sub-bands. As a result, the feature vector was found to consist of 40 mel-cepstral coefficients including the 0th coefficient, log F0, and five-band aperiodic features with delta and delta-delta coefficients.

2.2. Subjective evaluation

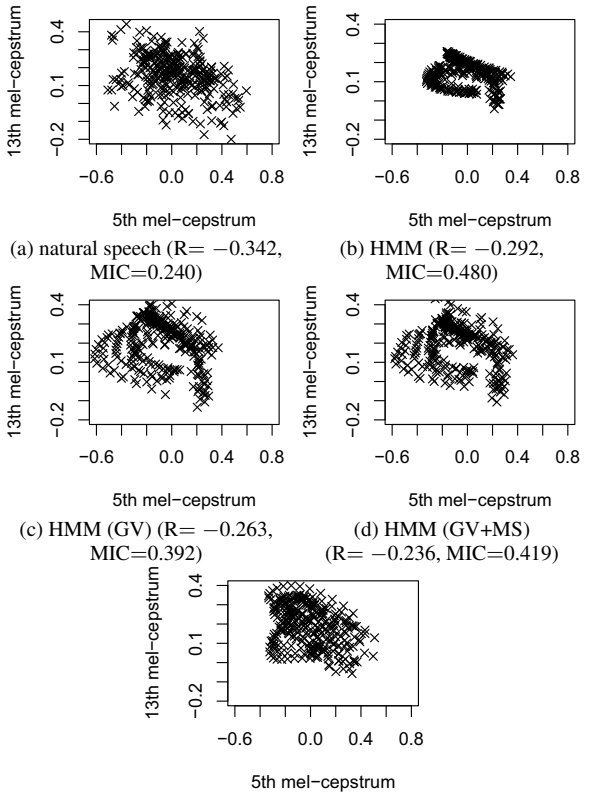
We also conducted a subjective evaluation test with respect to the naturalness of the natural and synthesized speech. For the test, we used 20 sentences for each speaker. To exclude the effects of the prosodic and excitation features (i.e., F0, aperiodic components, phoneme durations), we used these features extracted from the natural speech. Each subject evaluated each speech sample twice and rated their naturalness on a point scale ranging from 5 (very natural) to 1 (very unnatural). Table 1 shows the MOS scores obtained for 22 test subjects.

3. Association analysis of each dimension of spectral features

We first analyzed the association between dimensions within spectral features between natural and synthesized speech. However, since the association between dimensions within spectral features is changed by various factors such as phoneme context, it would be difficult to analyze the association using all the speech data. To avoid the problem, we only used the mel-cepstral coefficients having the same triphone /K-A+cl/ (17 speech segments, 312 frames) extracted from all the speech data obtained from the male speaker.

3.1. Distribution of spectral features

Figure 1 shows distributions between the 5th and 13th mel-cepstral coefficients. For the distribution of HMM (Fig. 1 (b)), there is obviously some kind of association. On the other hand, for natural speech (Fig. 1 (a)), there seems to be no such association. In comparing HMM, HMM (GV), HMM (GV+MS), and NN, we can see that the distributions of the latter three speech



(a) natural speech (R= −0.342, MIC=0.240)
(b) HMM (R= −0.292, MIC=0.480)
(c) HMM (GV) (R= −0.263, MIC=0.392)
(d) HMM (GV+MS) (R= −0.236, MIC=0.419)
(e) NN (R= −0.355, MIC=0.270)

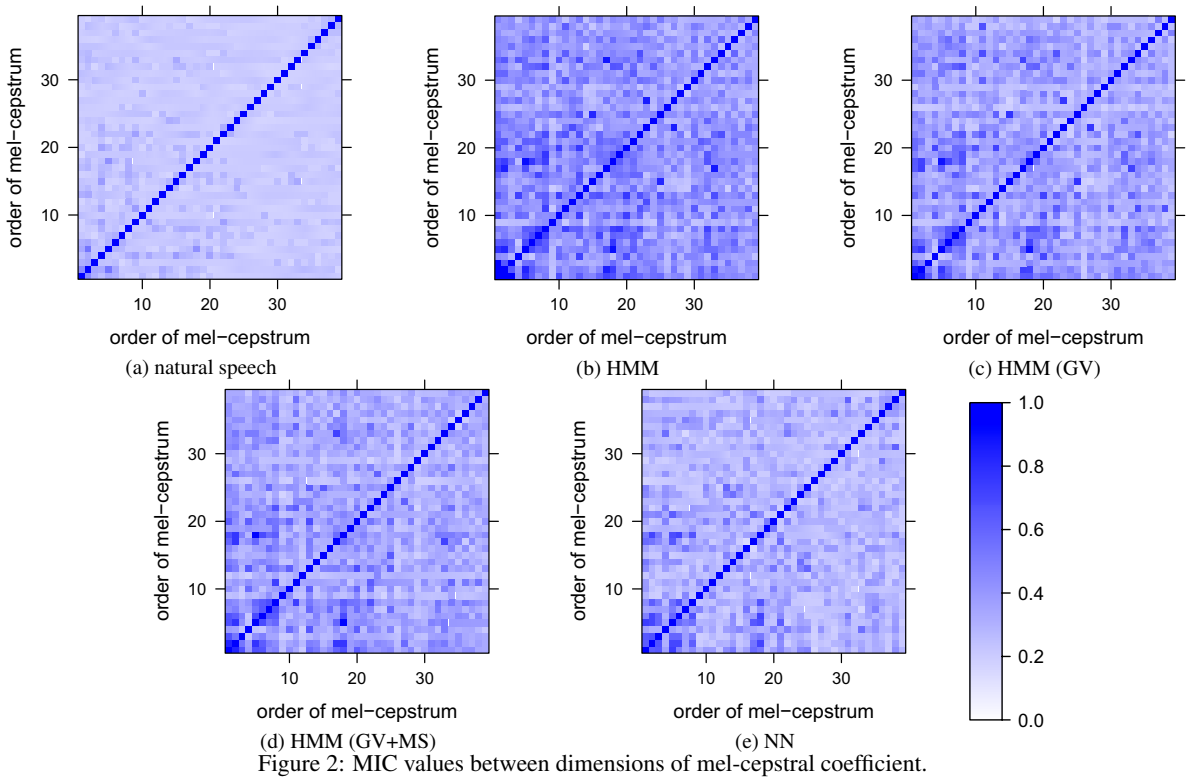
Figure 1: Distributions between 5th and 13th mel-cepstral coefficients.

instances come slightly closer to those of natural speech. From these results, it should be considered that the synthesized speech with the distribution similar to that of natural speech will produce higher subjective scores.

3.2. Association analysis using MIC

For quantitatively analyzing the association between the two variables, the correlation coefficient has been used. However, this coefficient would not be suitable for the association analysis between each dimension of spectral features. This is because it cannot detect nonlinear association such as Fig. 1. To solve this problem, we used a maximal information coefficient (MIC) [12]. The MIC makes it possible to detect nonlinear association that cannot be detected by using the correlation coefficient. Additionally, the MIC has a property similar to that of the correlation coefficient. That is, the MIC value ranges from 0 to 1, and the two variables with a strong association have a value closer to 1.

Figure 1 shows the MIC value and correlation coefficient for each distribution. A comparison with the correlation coefficients shows there are no associations with the subjective scores and the correlation coefficients. On the other hand, for the MIC values, speech with a lower MIC value has a higher subjective score. These results indicate that MIC makes it possible to capture the association. To analyze the differences in associations by the dimension of mel-cepstral coefficients, we also calculated MIC values between dimensions within mel-cepstral coefficients. Figure 2 lists the obtained MIC values for each speech. We can see that speech with a lower MIC value has a higher subjective score regardless of the dimensions of mel-cepstral coefficients. Interestingly, it can be seen that a param-



(a) natural speech (b) HMM (c) HMM (GV)

(d) HMM (GV+MS) (e) NN

Figure 2: MIC values between dimensions of mel-cepstral coefficient.

eter generation algorithm taking GV (or MS) into account decreases the MIC values. This implies that parameter generation methods that take GV (or MS) into account would be effective not only in compensating for GV (or MS) but also in compensating for the association between dimensions within spectral features. Moreover, the NN values are closest to those of natural speech. Although Fig. 2 shows only the MIC values of the triphone /K-A+cl/ uttered by the male speaker, similar tendencies were also obtained from other triphones of the female speaker. These obtained results indicate that the association between dimensions within spectral features would be effective for an objective evaluation.

4. Proposed technique

In the previous section, we confirmed that the subjective score improves with weaker association of spectral features. However, since the association will change according to various factors such as phoneme contexts, we cannot apply the MIC values directly for the objective evaluation. Therefore, we propose an objective evaluation index by using the voice conversion technique to detect the association.

4.1. Overview of proposed technique

A block diagram of the proposed method is shown in Fig. 3. The overall process is summarized below.

Training part:

Step 1 Divide the training spectral features into two spectral features on the basis of the dimensions of the training spectral features.

Step 2 Train conversion models for converting the divided spectral features into each other.

Evaluation part:

Step 3 Divide the evaluation spectral features into the two spectral features in the same manner as in **Step 1**.

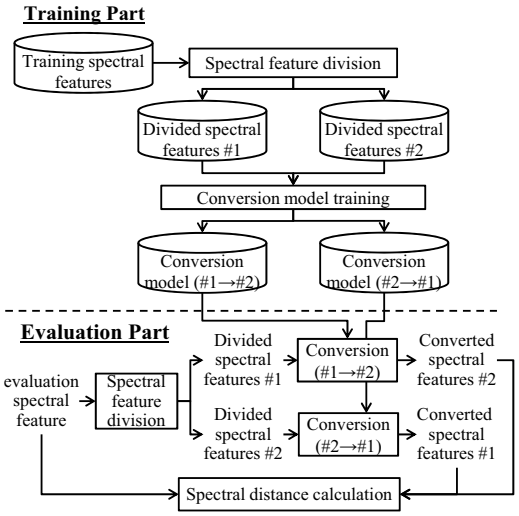


Figure 3: Block diagram of the proposed technique.

Step 4 Convert the divided evaluation spectral features into each other using the trained conversion models obtained from **Step 2**.

Step 5 Obtain the objective evaluation index by calculating the spectral distance between the converted evaluation spectral features and the evaluation spectral features.

In the proposed technique, we convert the divided spectral features into each other to detect the association within spectral features. If the obtained spectral distance between the converted and the evaluation spectral features is small, the association between dimensions within the spectral features would be strong because the divided spectral features can be converted accurately from other dimensions of spectral features. On the other hand, if the obtained spectral distance is large, the association between dimensions within spectral features would be

Table 2: Objective evaluation results.

	male			female		
	MOS	MCD [dB]	Proposed [dB]	MOS	MCD [dB]	Proposed [dB]
natural speech	4.45	–	4.18	4.31	–	4.79
HMM	1.99	4.35	0.52	1.78	4.95	0.40
HMM (GV)	2.77	4.74	1.25	2.86	5.47	1.31
HMM (GV+MS)	2.77	4.77	1.26	2.78	5.43	1.30
NN	2.84	4.35	1.41	3.10	4.83	1.51

weak. Details of each component, i.e., spectral features division, and spectral features conversion, are described as follows.

4.2. Division of spectral features

Let $\boldsymbol{X}$ be the spectral features with T frames and D dimensions. We divide the spectral features $\boldsymbol{X}$ into the two spectral features ($\boldsymbol{X}_1$, $\boldsymbol{X}_2$) on the basis of the dimensions of the spectral features. In this paper, we set odd-order mel-cepstral coefficients as $\boldsymbol{X}_1$, and even-order mel-cepstral coefficients as $\boldsymbol{X}_2$ on the basis of preliminary experiment results.

4.3. Conversion of spectral features

With our technique, we convert the divided spectral features into each other, i.e., conversion from $\boldsymbol{X}_1$ to $\boldsymbol{X}_2$, and conversion from $\boldsymbol{X}_2$ to $\boldsymbol{X}_1$. This problem can be considered as similar to that occurring with the voice conversion. Therefore, as the spectral conversion model, we can use voice conversion techniques that have been proposed such as vector quantization (VQ) [13], Gaussian mixture models (GMMs) [14], artificial neural networks (ANNs) [15], and deep bidirectional long short term memory recurrent neural networks (BLSTM-RNNs) [16].

In the evaluation part, the converted divided spectral features ($\hat{\boldsymbol{X}}_1, \hat{\boldsymbol{X}}_2$) are obtained by converting from the divided evaluation spectral features using the trained conversion models. Finally, as the objective evaluation index, we obtain the spectral distance between the evaluation spectral features $\boldsymbol{X}$ and the converted evaluation spectral features $\hat{\boldsymbol{X}}$. Here, $\hat{\boldsymbol{X}}$ is obtained by combining the converted spectral features ($\hat{\boldsymbol{X}}_1, \hat{\boldsymbol{X}}_2$). As the spectral distance, we can use mel-cepstral distortion, RMS error, and so on.

5. Experiments

5.1. Experimental conditions

We used the same speech data and the MOS regarding naturalness described in Sect. 2. As the spectral conversion model, we adopted the NN [15] to take the sequentiality of the mel-cepstral coefficients into account. The same sentences used for training of the TTS models were used as the training data of NN for the spectral feature conversion. For training of the NN for synthesized speech, the closed spectral features generated from each TTS model were used. Twenty sentences not used for the training were used for the evaluation. We set the number of hidden layers at 2 and the number of units per layer at 128. As the input vector, we used 11 consecutive (center, 5 previous, and 5 succeeding) frames to take the sequentiality of the mel-cepstral coefficients into account. The delta and delta-delta coefficients were not used. Mel-cepstral distortion was used as the spectral distance of the proposed technique. To compare our proposed objective scores with the conventional objective scores, we also calculated the mel-cepstral distortion between the mel-cepstral coefficients of natural speech and those of each synthesized speech.

5.2. Experimental results

Table 2 shows MOS, mel-cepstral distortions, and the proposed objective evaluation indexes for each speaker. We can see that the proposed evaluation indexes obtained from HMM are smallest, and those of natural speech are largest. This indicates that speech with a higher proposed evaluation index has a higher subjective score.. Furthermore, these evaluation indexes are highly correlated with subjective scores. The obtained correlation coefficients are 0.988 (the male speaker) and 0.944 (the female speaker) respectively. In contrast, it can be seen that subjective scores are not necessarily associated with the mel-cepstral distortion, especially those of HMM (GV) and HMM (GV+MS). These obtained results indicate that our objective evaluation index is more effective than the mel-cepstral distortion. In particular, our evaluation index can evaluate techniques having different tendencies of mel-cepstral distortion, such as HMM (GV), HMM (GV+MS), and NN.

5.3. Relation to prior work

Another technique that has been proposed is the objective evaluation method based on Kullback-Leibler (KL) divergence between GMMs of natural and synthesized speech [17]. With this method, however, since spectral features of natural speech are modeled by GMMs, the association between dimensions within spectral features similar to those of synthesized speech would be lost. Therefore, scores obtained from the KL divergence-based method would not be correlated with subjective scores obtained from HMM (GV) and HMM (GV+MS) as well as the mel-cepstral distortion. The perceptual evaluation of speech quality (PESQ) [18] are also highly correlated with subjective scores [19, 20]. With this method, however, the reference speech is required in order to estimate subjective scores. On the other hand, the reference speech is not required in our proposed method, although our proposed method has to train the voice conversion model.

6. Conclusions

In this paper, we presented a novel objective evaluation technique using the association between dimensions within the spectral features for statistical parametric speech synthesis. We analyzed the association between dimensions within the spectral features using a maximal information coefficient (MIC). The analysis results indicated that speech with higher naturalness has weaker associations. Then we described an objective evaluation technique based on a voice conversion technique to detect the associations for each speech. The obtained results we obtained with the method indicate that it is more effective than mel-cepstral distortion. In future work, we will explore objective evaluation indexes combining spectral features and other features such as F0 and phoneme duration since we have used only spectral features. We also plan to compare the performance of the proposed technique with that of the other methods such as PESQ [18].

7. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] T. Koriyama, T. Nose, and T. Kobayashi, "Statistical parametric speech synthesis based on gaussian process regression," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 173–183, 2014.

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

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

[5] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. and Syst.*, vol. E90-D(5), pp. 816–824, 2007.

[6] T. Nose, V. Chunwijitra, and T. Kobayashi, "A parameter generation algorithm using local variance for HMM-based speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 221–228, 2014.

[7] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A postfilter to modify the modulation spectrum in hmm-based speech synthesis," in *ICASSP 2014*. IEEE, 2014, pp. 290–294.

[8] M. Zhang, J. Tao, H. Jia, and X. Wang, "Improving HMM based speech synthesis by reducing over-smoothing problems," in *ISCSLP'08*. IEEE, 2008, pp. 1–4.

[9] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-markov model-based speech synthesis system," *IEICE Trans. Inf. and Syst.*, vol. E90-D(5), pp. 825–834, May 2007.

[10] H. Silén, E. Helander, J. Nurminen, and M. Gabbouj, "Ways to implement global variance in statistical speech synthesis." in *INTERSPEECH 2012*, 2012.

[11] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[12] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[13] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *ICASSP-88*. IEEE, 1988, pp. 655–658.

[14] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[15] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 954–964, 2010.

[16] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *ICASSP 2015*. IEEE, 2015, pp. 4869–4873.

[17] C. Do, M. Evrard, A. Leman, C. d'Alessandro, A. Rilliard, and J. Crebouw, "Objective evaluation of HMM-based speech synthesis system using Kullback-Leibler divergence," in *INTERSPEECH 2014*, 2014, pp. 2952–2956.

[18] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP 2001*, vol. 2, 2001, pp. 749–752.

[19] M. Cernak and M. Rusko, "An evaluation of synthetic speech using the pesq measure," in *Proc. European Congress on Acoustics*, 2005, pp. 2725–2728.

[20] D.-Y. Huang, "Prediction of perceived sound quality of synthetic speech," in *Proc. APSIPA*, 2011.