Investigation of intra-speaker spectral parameter variation and its prediction towards improvement of spectral conversion metric

Tatsuo Inukai, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura

Graduate School of Information Science, Nara Institute of Science and Technology,
tatsuo-i@is.naist.jp, tomoki@is.naist.jp,
neubig@is.naist.jp, ssakti@is.naist.jp, s-nakamura@is.naist.jp

Abstract

In spectral conversion of statistical voice conversion (VC), distance-based measures between the converted and target spectral parameters are often used as evaluation or training criteria. However, even if the same speaker utters the same sentence, the spectral parameters vary utterance by utterance, and thus, spectral distance between utterances still exists. Moreover, the original prosodic features of input speech are often kept unchanged in some VC systems, such those that function in real-time. In such cases, prosody of converted and target speech samples are different, and these differences increases spectral distance. These potential spectral variations are not considered in the conventional evaluation/training criterion. Thus, by constructing criteria that consider this spectral difference improvements in sound quality can be expected. In this paper, we investigate intra-speaker spectral variation between utterances of the same sentence. We also propose a method for predicting this variation from prosodic parameter differences between the corresponding utterances. We conduct experimental evaluations using many speech samples of the same sentence uttered by a single speaker, with results demonstrating that the proposed method effectively predicts the intra-speaker spectral variation from the observed prosodic changes.

Index Terms: voice conversion, training/evaluation criterion, intra-speaker spectral variation, prosodic differences, prediction

1. Introduction

Statistical voice conversion (VC) is an effective technique for modifying acoustic parameters to convert non-linguistic or para-linguistic information while keeping linguistic information unchanged [1]. It was originally proposed for speaker conversion to change the voice uttered by a source speaker as if it is uttered by a specific target speaker [2]. Recent progress in VC has achieved high-quality and real-time conversion [3]. These technologies can be used in various VC applications for augmenting human-to-human speech communication, such as speaking-aid for vocally handicapped people [4, 5], silent speech interfaces [6], bandwidth extension [7], and singing voice effectors [8]. Improving VC performance has the potential to contribute greatly to practical use of these applications.

In real-time VC systems for speaker conversion, short-term speech features, such as spectral parameters, are mainly converted with little delay using complex conversion functions. On the other hand, long-term speech features, such as F_0 patterns, are fundamentally difficult to convert in real-time. Therefore, simple conversion functions, such as a global linear transform, are often used to convert F_0 values frame by frame. Consequently, performance of the real-time VC system strongly depends on spectral conversion.

As for spectral conversion, various conversion or evaluation criteria have been proposed. One of the most standard criteria is a (weighted) distance measure between converted and target spectral parameters. It is used in the most widely used VC methods, with a Gaussian mixture models (GMM) based on minimum mean square error estimation [9] or maximum-likelihood estimation [10]. Some sophisticated model training methods using it as an optimization criterion have also been proposed [11, 12]. Recently the use of not only the distance-based criterion but also other criteria have been proposed. One of them is global variance (GV), which is the second order moment of a spectral parameter trajectory [10]. It has been reported that speech quality and conversion accuracy for speaker individuality in converted speech are significantly improved by considering both the distance-based criterion and the GV. It has also been reported that mutual information is also useful [13]. The effectiveness of these criteria has also been confirmed in model training [14, 15, 16]. These results suggest that it is useful to use additional criteria rather than only the distance-based criteria, although it causes a larger distance between converted and target spectral parameters (i.e., a larger conversion error).

Previous research has not carefully investigated how much spectral distance is acceptable in VC. By considering the amount of acceptable spectral distance, it may be possible to automatically determine weight parameter controlling the balance between the distance-based criterion and additional criteria. To clarify the acceptable distance, we focus on intra-speaker spectral variation, which is the spectral distance observed when the same speaker utters the same sentence many times. It is empirically known that intra-speaker spectral variation will not go to zero between utterances. Moreover, it has been reported that larger prosodic changes cause larger spectral differences [17]. Therefore, the acceptable spectral distance possibly changes according to prosodic differences between the converted and target voices.

In this paper, we investigate intra-speaker spectral variation using many speech samples of the same sentences uttered by a single speaker. Mel-cepstral distortion [18] is used as a metric to capture the intra-speaker spectral variation. Moreover, we propose a method to predict the intra-speaker spectral variation between two utterances from their differences of various prosodic parameters. This prediction is useful to determine the acceptable spectral distortion in each utterance-pair and it has a potential to develop better training, conversion, and evaluation metrics for spectral conversion in VC.

2. Basic Procedure of VC

In the statistical VC for speaker conversion, a parallel data set consisting of utterance pairs of the source and target speakers is used to train the conversion models for individual speech parameters. As the conversion model for spectral parameters, a conditional probability density function of the target speaker's spectral parameters given the source speaker's spectral parameters is often modeled by a GMM. On the other hand, as the conversion model for F_0 parameters, the following global linear transformation is often used:

$$\log \hat{F}_0 = \frac{\sigma^{(t)}}{\sigma^{(s)}} \left(\log F_0 - \mu^{(s)} \right) + \mu^{(t)}, \tag{1}$$

where F_0 is the source speaker's F_0 value and $\hat{F}_0$ is the converted F_0 value. The conversion model parameters are $\mu^{(s)}$ and $\sigma^{(s)}$, which are mean and standard deviation values of log-scaled F_0 values of the source speaker, and $\mu^{(t)}$ and $\sigma^{(t)}$, which are those of the target speaker. Prosodic parameters, such as shape of F_0 pattern, phoneme duration, and power patterns, are kept unchanged in conversion. Note that it is also possible to convert them if real-time conversion processing is not necessary and linguistic contents are available. However, such a conversion is essentially difficult in real-time conversion processing without any linguistic contents.

In the conversion processing as mentioned above, the converted speech is generated by the converted spectral parameters and globally transformed F_0 values without any prosodic changes. Therefore, ideal converted spectral parameters will be spectral parameters of a speech sample uttered by the target speaker so that its prosody is the same as that of the source speaker. However, it is not straightforward to record such speech samples as in each utterance pair of the available parallel data the target speaker's prosody is usually different from the source speaker's prosody. Therefore, the target spectral parameters are not ideal ones. Nevertheless, in the traditional approach the spectral conversion model is basically trained so that the conversion error (i.e., the distance between the converted spectral parameters and the target spectral parameters) in the parallel data set is minimized.

3. Investigation of Intra-Speaker Spectral Parameter Variation

We investigate how much spectral parameters vary when a single speaker utters the same sentence and how much spectral parameters differ additionally by imitating prosody of other speakers.

3.1. Recording of speech samples

We recorded speech samples of the same sentence uttered by a single speaker. One Japanese male speaker uttered one sentence 200 times with his own prosody. He also uttered the same sentence while imitating prosody of other reference speakers. The number of reference speakers was 24 (12 male and 12 female). To make it easy to imitate the utterances, 1) analysis-synthesized speech samples were generated by converting F_0 values of speech samples of the reference speakers using **Eq.** (1) to make their F_0 ranges equivalent to that of the male speaker and 2) they were presented to the male speaker as reference speech samples during the recording. The male speaker recorded 8 utterances imitating each reference speaker's prosody. A total of 192 speech samples were recorded. The sampling frequency was 16 kHz.

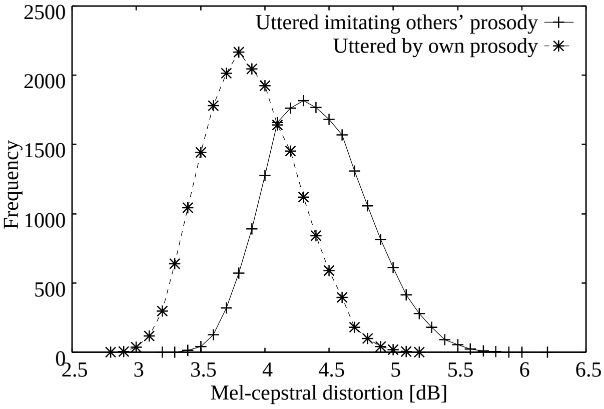


Figure 1: Frequency distribution of mel-cepstral distortion between utterances of the same sentence uttered by the same speaker.

3.2. Intra-speaker spectral parameter variation

The 1^{st} through 24^{th} mel-cepstral coefficients extracted by STRAIGHT analysis [19] were used as spectral parameters. Frame shift was 5 ms. Mel-cepstral distortion was calculated by performing dynamic time warping (DTW) in each utterance pair.

Figure 1 shows the frequency distribution of the mel-cepstral distortion for all utterance-pairs, one is a frequency distribution for speech samples with the male speaker's own prosody and the other is that for speech samples with the different speakers' prosody. We can see that even in the same sentence with the same speaker, the mel-cepstral distortion is not 0. For the speech samples with the speaker's own prosody, the mean value is 3.9 dB and the standard deviation value is 0.35 dB. On the other hand, for the speech samples with the different speakers' prosody, the spectral variation tends to be larger; its mean value is 4.4 dB and its standard deviation value is 0.38 dB.

These results suggest that 1) it is not necessary to decrease mel-cepstral distortion to 0 in VC and 2) as prosodic differences between the source and target speakers are larger, a larger mel-cepstral distortion will be acceptable.

4. Prediction of Spectral Parameter Variation

While the source and target speakers' prosody is usually different from each other in available parallel data sets, it is ideal to predict target spectral parameters in each utterance-pair when the target speaker imitates prosody of the source speaker. However, this is not straightforward to do. Although a method for predicting spectral parameter changes according to F_0 changes has been proposed [17], it still needs training data consisting of many speech samples of the same linguistic contents uttered by the target speaker with different F_0 values. It is laborious work to additionally record such a data set. Therefore, we simplify the problem to be solved. We predict spectral distortion between the original speech sample of the target speaker in the parallel data and a practically unavailable speech sample uttered by the target speaker while imitating prosody of the corresponding utterance of the source speaker. Namely, we predict not an unobserved spectral feature vector itself but its distance from an observed spectral feature vector. The predicted spectral dis-

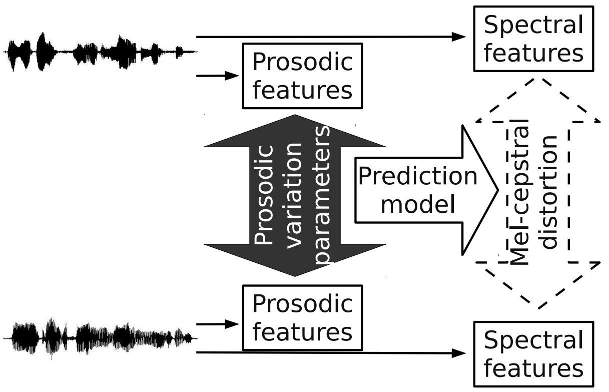


Figure 2: Prediction procedure of mel-cepstral distortion from prosodic variation parameters.

tortion is still useful as it shows acceptable spectral distortion depending on prosodic differences in each utterance-pair.

As a first step to achieve such a prediction, in this paper we propose a method for predicting the mel-cepstral distortion from prosodic variation parameters capturing prosodic differences using many speech samples of the same sentence uttered by a single speaker. **Figure 2** shows the prediction procedure. Prosodic parameters extracted from an utterance pair and the prosodic variation parameters are calculated as an explanation variable. The mel-cepstral distortion is also calculated in this utterance-pair as a target variable. Then, the mel-cepstral distortion is predicted from the prosodic variation parameters. In a practical application, the prosodic variation parameters are calculated between the source and target speech samples in each utterance pair for training or evaluation. Finally, the mel-cepstral distortion is predicted from these samples. The predicted mel-cepstral distortion is regarded as an acceptable distortion between the converted and target spectral parameters. Note that this distortion varies utterance by utterance. It is inevitable to develop a sentence/speaker-independent prediction model to make it possible to apply this prediction model in practical VC conditions.

4.1. Prediction Model

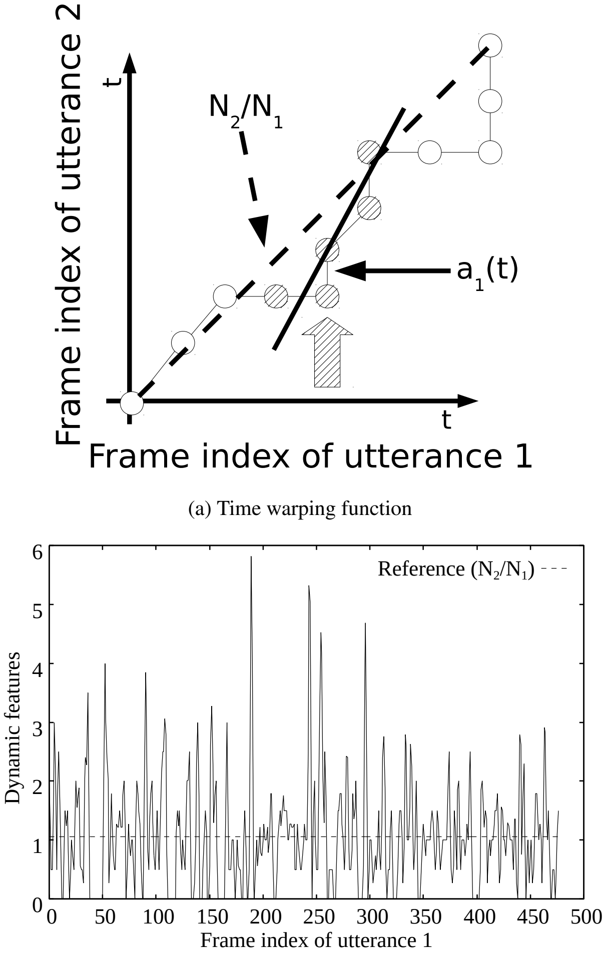
A multiple linear regression model is used to predict the mel-cepstral distortion from the prosodic variation parameters as follows:

$$\hat{m}_{i,j} = \boldsymbol{a}^\top \boldsymbol{p}_{i,j} + c, \tag{2}$$

where $\hat{m}_{i,j}$ is the predicted mel-cepstral distortion between the i^{th} utterance and the j^{th} utterance, $\boldsymbol{p}_{i,j}$ is a prosodic variation parameter vector between these utterances, $\boldsymbol{a}$ is a regression coefficient vector, and c is a bias value. The regression coefficient vector and the bias value are determined by the least square error estimation. In this paper, the mel-cepstral distortion and the prosodic variation parameters are calculated using only voice active frames, which are automatically extracted with normalized waveform power.

4.2. Prosodic Variation Parameters

Several prosodic variation parameters are used in the prediction. Duration distortion and DTW distortion capture the difference in duration. Voiced/unvoiced error rate and F_0 distortion capture the difference in F_0 patterns. Power distortion captures the difference in power patterns. These parameters take positive values and only take zeros when prosody of two speech samples is completely the same as each other.



(a) Time warping function

(b) Dynamic feature sequence of time warping function

Figure 3: Calculation of DTW distortion.

4.2.1. Duration Distortion

To capture a difference of total duration over an utterance, duration distortion is calculated as follows:

$$D_{\text{dur}} = \log N_l - \log N_s, \tag{3}$$

where N_l is the number of frames extracted from a longer utterance, and N_s is the number of frames extracted from a shorter utterance.

4.2.2. DTW Distortion

To capture the difference in local duration, DTW distortion is calculated as shown in **Figure 3**. First, temporally dynamic features of the time warping function are determined by DTW, which is given by the slope of a regression line as shown in $a_1(t)$ in **Figure 3(a)**, and calculated at each frame of each utterance. One example of the dynamic feature sequence over an utterance is shown in **Figure 3(b)**. If there is no difference in local duration, the time warping function is represented as a line and the slope at every frame is equivalent to its constant slope N_2/N_1 as shown in **Figure 3(a)**. The DTW distortion is calculated as a difference between the dynamic features and the constant slope as follows:

$$D_{\text{DTW}} = \frac{1}{2N_1}\sqrt{\sum_{t=1}^{N_1}\left(a_1(t) - \frac{N_2}{N_1}\right)^2} + \frac{1}{2N_2}\sqrt{\sum_{t=1}^{N_2}\left(a_2(t) - \frac{N_1}{N_2}\right)^2}, \quad (4)$$

where N_1 and N_2 are the number of frames of utterance 1 and utterance 2, and $a_1(t)$ and $a_2(t)$ are the dynamic feature at frame t over the utterance 1 and utterance 2. All frame pairs from frame $t-1$ to $t+1$ over the utterance 1 are used to fit a regression line to calculate $a_1(t)$. In a similar way, $a_2(t)$ is also calculated.

4.2.3. Voiced/Unvoiced Error Rate

To capture the difference of voiced/unvoiced frames, a voiced/unvoiced error rate between frames time-aligned by DTW is calculated as follows:

$$D_{\text{U/V}} = \frac{1}{N}\sum_{t=1}^{N} e(t), \quad (5)$$

where N is the number of time-aligned frame pairs, $e(t)$ is a function that returns 0 when voice/unvoiced information is the same at frame-pair t and returns 1 when they are different.

4.2.4. F_0 Distortion

To capture the difference of F_0 patterns, F_0 distortion is calculated between time-aligned frames by DTW as follows:

$$D_{F_0} = \frac{1}{N_v}\sqrt{\sum_{t=1}^{N_v}\left(\log(F_0^{(1)}(t)) - \log\left(F_0^{(2)}(t)\right)\right)^2}, \quad (6)$$

where N_v is the number of voiced frame pairs, and $F_0^{(1)}(t)$ and $F_0^{(2)}(t)$ are F_0 values of individual utterances at frame pair t.

We also calculate a maximum value ($D_{F_0}^{(\max)}$) and a minimum value ($D_{F_0}^{(\min)}$) of the absolute difference of log-scaled F_0 in each utterance pair.

4.2.5. Power Distortion

To capture the difference of power patterns, power distortion is calculated between time-aligned frames by DTW as follows:

$$D_{\text{pow}} = \frac{1}{N}\sqrt{\sum_{t=1}^{N}\left(p^{(1)}(t) - p^{(2)}(t)\right)^2}, \quad (7)$$

where N is number of frame pairs, $p^{(1)}(t)$ and $p^{(2)}(t)$ are normalized power values of individual utterances at frame pair t.

We also calculate a maximum value ($D_{\text{pow}}^{(\max)}$) and a minimum value ($D_{\text{pow}}^{(\min)}$) of the absolute difference of the normalized power in each utterance pair.

4.3. Normalization of Speaker-Dependency

The prosodic variation parameters and the mel-cepstral distortion are affected by speaker individuality. To reduce the impact of speaker dependence on these parameters, all parameters are normalized so that their mean and standard deviation values are equal to 0 and 1 in each speaker.

This normalization can be straightforwardly applied to the prosodic variation parameters (i.e., variable calculable from the input) using the parallel data in practical VC conditions. On the other hand, it is not straightforward to apply it to the mel-cepstral distortion (i.e., a target variable). Namely, the normalized mel-cepstral distortion is predicted but the unnormalized mel-cepstral distortion is hard to predict. Nevertheless, the normalized mel-cepstral distortion is still effective to improve the conventional training, conversion, and evaluation criteria because it captures additional information about the acceptable spectral distortion varying utterance by utterance.

Table 1: Prediction results of mel-cepstral distortions.

Regression model		Correlation coefficient
Speaker-dependent	Sentence-dependent	0.76
Speaker-dependent	Sentence-independent	0.75
Speaker-independent	Sentence-independent without normalization	0.64
Speaker-independent	Sentence-independent with normalization	0.72

5. Experiments

5.1. Experimental Conditions

We recorded speech data of 5 speakers (4 males, 1 female) in the same way as described in **Section 3.1**. Male 1 uttered 6 sentences 200 times, and the other speakers uttered 4 sentences 50 times. These sentences were extracted from the ATR Japanese speech database [20]. The 1^{st} through 24^{th} mel-cepstral coefficients extracted by STRAIGHT analysis [19] were used as spectral parameters. F_0 values were extracted by the F_0 estimation method of STRAIGHT analysis [21]. The sampling frequency was 16 kHz. Frame shift was 5 ms.

To evaluate prediction accuracy, we calculated a correlation coefficient between the predicted mel-cepstral distortion and the observed mel-cepstral distortion. We evaluated the following four models:

1) speaker- and sentence-dependent models: a single prediction model was trained and evaluated for each speaker and each sentence,

2) speaker-dependent and sentence-independent models: a single prediction model was trained and evaluated for each speaker using all of his/her sentences,

3) speaker- and sentence-independent models without normalization: a global prediction model was trained for all speakers using their all sentences without normalization described in **Section 4.3**,

4) speaker- and sentence-independent models with normalization: a global prediction model was trained for all speakers using their all sentences with the normalization.

In each case, five-fold cross validation was employed. All combinations of utterance-pairs of the same speaker and the same sentence were considered. In the speaker-independent model, the number of utterances of Male 1 was reduced to the same number of utterances of the other speakers. We also evaluated the effect of individual prosodic variation parameters on prediction accuracy by adding them as explanatory features one by one in the speaker- and sentence-dependent model.

Table 2: Correlation coefficients between individual prosodic difference parameters.

–	D_{DTW}	$D_{U/V}$	D_{F_0}	$D_{F_0}^{max}$	$D_{F_0}^{min}$	D_{pow}	D_{pow}^{max}	D_{pow}^{min}
D_{dur}	0.22	0.05	0.10	0.02	0.03	0.11	0.04	0.03
D_{DTW}	–	0.25	0.29	0.23	0.05	0.27	0.16	0.04
$D_{U/V}$	–	–	0.40	0.24	0.20	0.58	0.32	0.12
D_{F_0}	–	–	–	0.65	0.28	0.37	0.19	0.12
$D_{F_0}^{max}$	–	–	–	–	0.09	0.26	0.17	0.07
$D_{F_0}^{min}$	–	–	–	–	–	0.14	0.06	0.04
D_{pow}	–	–	–	–	–	–	0.63	0.20
D_{pow}^{max}	–	–	–	–	–	–	–	0.10

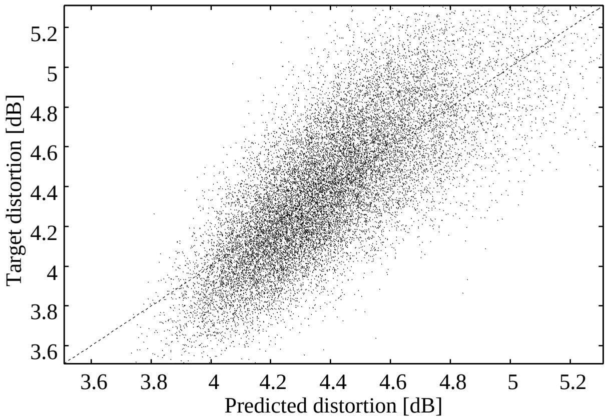


Figure 4: Scatter diagram of target mel-cepstral distortion and that predicted by sentence-dependent model (for Male 1).

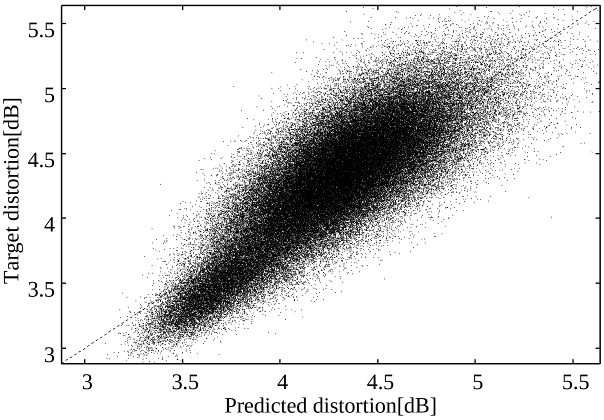


Figure 5: Scatter diagram of target mel-cepstral distortion and that predicted by sentence-independent model (for Male 1).

5.2. Experimental results

Table 1 shows results of individual prediction models using all prosodic variation parameters. The speaker- and sentence-dependent model is capable of predicting the mel-cepstral distortion with accuracy of 0.76 in correlation coefficient. A scatter plot between the predicted mel-cepstral distortion and the observed mel-cepstral distortion is shown in **Figure 4**. We can see a tendency that the prediction error is smaller as the observed mel-cepstral distortion is also smaller. The speaker-dependent and sentence-independent model does not cause any adverse effects and its correlation coefficient is 0.75 as shown in **Table 1**. Its scatter plot is shown in **Figure 5**. From this we can see that the proposed method is not sensitive to a change of sentence content. **Table 1** also shows results of the speaker-independent models with/without the normalization. If the normalization is not performed, the correlation coefficient decreases to 0.64. This result shows that the prediction model is strongly affected by the speaker differences. We can also observe this degradation in a scatter plot as shown in **Figure 6**. This degradation is alleviated by using normalization as shown in **Figure 7**.

Table 2 shows correlation coefficients between each prosodic variation parameter pairs. We can see that correlation coefficients tend to be low except for F_0 and power distortion and their maximum values. Therefore, each of the other prosodic variation parameters represents different property of prosodic differences. **Figure 8** shows changes of the correlation coefficient by adding the prosodic variation parameters to the feature set one by one. The DTW distortion has a great contribution to the prediction for all speakers. By further adding only the F_0 distortion and the power distortion, the prediction accuracy becomes almost equivalent to that achieved by using all prosodic variation parameters.

These results suggest that 1) the mel-cepstral distortion can be predicted using only three prosodic parameters (the DTW distortion, the F_0 distortion, and the power distortion) and 2) the speaker- and sentence-independent prediction model can be trained using normalization of speaker differences in each parameter.

6. Conclusions

In this paper, we investigated intra-speaker spectral variation between utterances of the same sentence. It was found that larger prosodic differences cause larger spectral variations, and acceptable spectral distortion in VC varies by prosodic variation. To predict the spectral variations caused by the prosodic differences, we proposed a prediction method using a multiple linear regression model to predict the mel-cepstral distortion from several prosodic variation parameters. The experimental results have demonstrated that 1) the mel-cepstral distortion is predicted relatively well by the proposed method (the correlation coefficient is more than 0.7), 2) the prediction model is robust against sentence differences, and 3) the prediction model is sensitive to the speaker differences but this issue is well alleviated by the parameter normalization, and 4) good prediction accuracy is achieved using only three prosodic parameters. We plan to construct training, conversion, and evaluation metrics considering the predicted spectral variation.

7. Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Number 22680016.

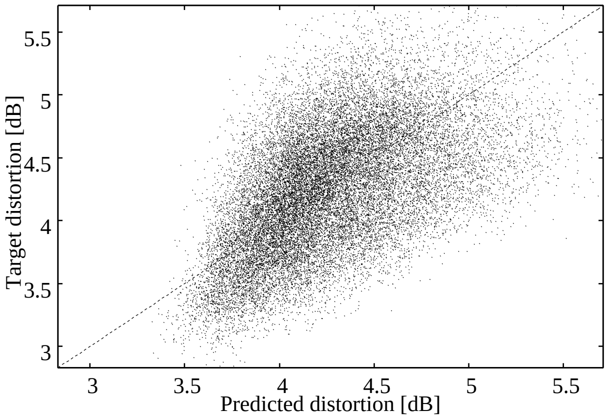


Figure 6: Scatter diagram of target mel-cepstral distortion and that predicted by speaker-independent model (for all speakers).

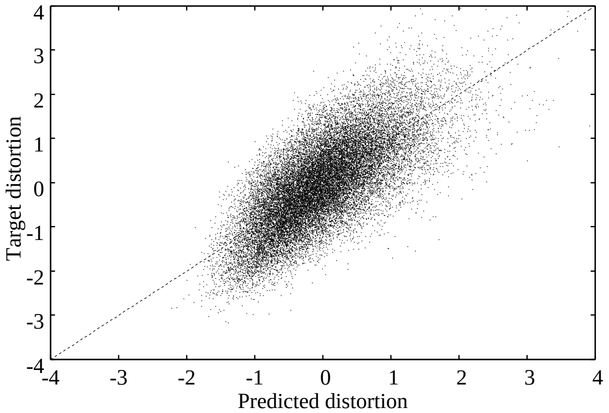


Figure 7: Scatter diagram of target mel-cepstral distortion and that predicted by normalized speaker-independent model (for all speakers).

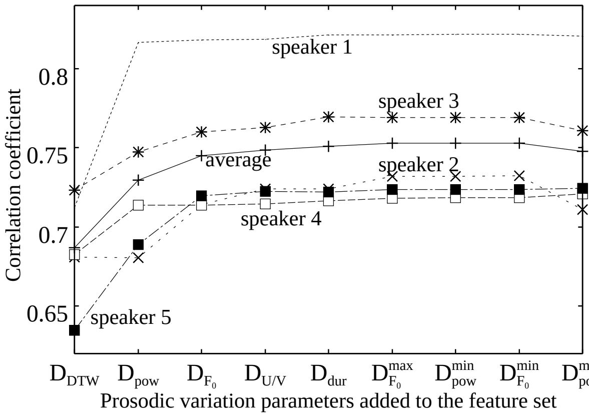


Figure 8: Correlation coefficient when adding prosodic variation parameters one-by-one.

8. References

[1] Y. Stylianou, "Voice transformation: a survey," *Proc. ICASSP*, pp. 3585–3588, Apr. 2009.

[2] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *J. Acoust. Soc. Jpn.* (E), Vol. 11, No. 2, pp. 71–76, 1990.

[3] T. Toda, T. Muramatsu, H. Banno, "Implementation of computationally efficient real-time voice conversion," *Proc. of INTERSPEECH*, Sept. 2012.

[4] Kain. A. B., Hosom. J. P., Niu. X., van Santen. J. P., Fried-Oken. M., and Staehely. J, "Improving the intelligibility of dysarthric speech," *Speech communication*, Vol. 49, No. 9, pp.743–759, 2007.

[5] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Esophageal speech enhancement based on statical voice conversion with Gaussian mixture models.", *IEEE Trans. Inf. & Syst.*, Vol. E93-D, No. 9, pp. 2472–2482, 2010.

[6] T. Toda, M. Nakagiri, K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Trans. ASLP*, vol. 20, No. 9, pp. 2505–2517, Nov. 2012.

[7] P. Jax and P. Vary. "On artificial bandwidth extension of telephone speech." *Signal Processing*, Vol. 83, pp. 1707–1719, 2003.

[8] H. Doi, T. Toda, T. Nakano, M. Goto, and S. Nakamura, "Singing Voice Conversion Method Based on Many-to-Many Eigenvoice Conversion and Training Data Generation Using a Singing-to-Singing Synthesis System." *APSIPA Annual Summit and Conference*, 2012.

[9] Y. Stylianou, O. Cappe, and E. Moulines. "Continuous probabilistic transform for voice conversion." *IEEE Trans. SAP*, Vol. 6, No. 2, pp. 131–142, 1998.

[10] T. Toda, A.W. Black, and K. Tokuda. "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory." *IEEE Trans. ASLP*, Vol. 15, No. 8, pp. 2222–2235, 2007.

[11] Yi-Jian Wu and Ren-Hua Wang, "Minimum Generation Error Training for HMM-Based Speech Synthesis." *in Proc. ICASSP*, pp. 89–92, 2006.

[12] H. Zen, Y. Nankaku, and K. Tokuda. "Continuous stochastic feature mapping based on trajectory HMMs." *IEEE Trans. ASLP*, Vol. 19, No. 2, pp. 417–430, 2011.

[13] Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Yih-Ru Wang and Sin-Horng Chen "A Study of Mutual Information for GMM-Based Spectral Conversion." *in Proc. Interspeech*, 2012.

[14] H. Benisty and D. Malah, "Voice Conversion using GMM with Enhanced Global Variance." *in Proc. Interspeech*, pp. 669–672, 2011.

[15] Zen. H, Gales. M. J F, Nankaku. Y and Tokuda. K, "Product of Experts for Statistical Parametric Speech Synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol.20, No.3, pp.794–805, 2012

[16] Hwang. H. T., Tsao. Y., Wang. H. M., Wang. Y. R., and Chen. S. H, "Exploring mutual information for GMM-based spectral conversion." *In Chinese Spoken Language Processing (ISCSLP) 2012 8th International Symposium on*, pp. 50–54, 2012.

[17] N. Minematsu and S. Nakagawa, "Analysis and modeling of spectral variations caused by F_0 changes" *Acoust. Soc. Jpn.*, Vol. 55, No. 3, pp. 165–174, 1999. (In Japanese).

[18] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment." *Communications, Computers and Signal Processing, 1993., IEEE Pacific Rim Conference on.* Vol. 1, pp. 125–128, 1993.

[19] H. Kawahara, I. Masuda-Katsuse and A. deCheveigné, "Restructuring speech representations using a pitch-adaptive time- frequency smoothing and an instantaneousfrequency-based F_0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication.*, Vol. 27, No. 3–4, pp. 187–207, 1999.

[20] Y. Sagisaka, K. Takeda, M. Abe, S. Katagiri, T. Umeda and H. Kuwahara, "A large-scale Japanese speech database.", ICSLP90, pp.1089–1092, 1990.

[21] H. Kawahara, H. Katayose, A. deCheveigné, and R.D. Pateterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity," *Proc. Eurospeech*, Vol. 99, pp.2781–2784, 1999.