ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Similar Speaker Selection Technique Based on Distance Metric Learning with Perceptual Voice Quality Similarity

Yusuke Ijima, Mitsuaki Isogai, Hideyuki Mizuno

NTT Cyber Space Laboratories, NTT Corporation, Japan

Abstract

This paper describes a similar speaker selection technique based on distance metric learning. Our aim is selection of a perceptually similar speaker using acoustic features from a multi-speaker database. A novel point of the proposed technique is training a transform matrix using the perceptual voice quality similarity between many speakers obtained from a subjective evaluation to convert acoustic feature space. Given an input speech, acoustic features of the input speech are transformed using a trained transform matrix, after which speaker selection is performed based on the Euclidean distance on the transformed acoustic feature space. We perform speaker selection experiments and evaluate the performance results by comparing them with those of speaker selection on acoustic feature space without feature space transformation. The results indicate that transformation based on distance metric learning provides about 60% of the error reduction rate.

Index Terms: speaker selection, perceptual similarity, voice quality, distance metric learning

10.21437/Interspeech.2012-534

1. Introduction

Realization of text-to-speech synthesizer which can speak with any desired speakers' voice is one of the most important issues in the research area of human computer interaction systems. For HMM-based speech synthesis systems [1], the average-voice-based speech synthesis technique using model adaptation was proposed [2]. Given only a few minutes of speech data of the target speaker, this technique can synthesize an arbitrary speaker's speech by model transformation from an average voice model to the target speaker's model. However, it was reported that the similarity of synthesized speech to the target speaker's speech is degraded by model conversion if the acoustic feature distance from the average voice model is large [3]. One useful approach to alleviating this problem is creating an average voice model by selecting only speakers whose speech is similar to that of the target speaker. This approach may be an effective one in synthesizing speech whose voice quality approaches that of the target speaker.

In the field of speech recognition techniques, a variety of approaches have been proposed to train an acoustic model of the target speaker based on similar speaker selection [4, 5]. These techniques select similar speakers based on acoustic feature similarities (MFCCs) such as likelihood of GMM [6]. However, even if the distribution of MFCCs between two speakers is close, the voice quality of the two speakers is not necessarily perceptually similar. For this reason, we previously analyzed the relationship between perceptual voice quality similarity and various acoustic features and in so doing found several acoustic features that were highly correlated to perceptual voice quality similarity [7]. From a text-dependent speaker selection experiment, we showed the selection performance was improved by using these acoustic features, but we have not confirm the effectiveness of these acoustic features in the text-dependent speaker selection. For this purpose, in this paper, we used the GMM supervector [8], whose effectiveness in text-independent speaker recognition has been confirmed. Moreover performance sufficient to select a perceptually similar speaker has not yet been obtained in [7]. This is because some speakers' voices are not perceptually similar even if the Euclidean distance on acoustic feature space between two speakers is close. To overcome this problem, in this paper, we propose a similar speaker selection technique based on "distance metric learning [9]" (DML) to learn a distance metric considering the perceptual voice quality similarity.

Using given side information, DML can learn an optimal distance metric. Many studies have been devoted to DML due to its importance for many applications such as image retrieval [10], music retrieval [11], and sentence retrieval [12]. This technique can affect speaker selection if the side information can be set properly. We used the perceptual voice quality similarity obtained from the subjective experiment as the side information. This learning has been used as the feature space transformation in a number of studies. For instance, [10] used transformation of the original image space for image retrieval.

In the proposed method, the transform matrix is trained based on distance metric learning to convert acoustic feature space. Given an input speech, the acoustic features of an input speech are transformed using a trained transform matrix. Then a similar speaker is chosen based on the Euclidean distance on the transformed acoustic feature space. To evaluate the proposed technique's performance, we compared it with that of speaker selection on acoustic feature space without transform. The results we obtained in speaker selection experiments demonstrated the technique's effectiveness.

2. Speaker selection system

2.1. Distance metric learning

Let us denote a set of N vectors in d-dimensional space as $\boldsymbol{X} = \{x_i \in \mathbb{R}^d\}_{i=1}^N$, where the Mahalanobis distance between two vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is defined as

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{M} (\boldsymbol{x}_i - \boldsymbol{x}_j) \quad (1)$$

where $\boldsymbol{M}$ is a positive semi-definite matrix that satisfies valid metric properties. The goal of DML is to find an optimal Mahalanobis matrix $\boldsymbol{M}$ from the side information. We can uniquely decompose any positive semi-definite matrix as $\boldsymbol{M} = \boldsymbol{A}^T \boldsymbol{A}$. This reduces Eq. (1) to

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \| \boldsymbol{A}(\boldsymbol{x}_i - \boldsymbol{x}_j) \|_2 \quad (2)$$

the Euclidean distance after the transformation $\boldsymbol{x}_i \rightarrow \boldsymbol{A}\boldsymbol{x}_i$. Thus DML is equivalent to transformation of the vector space

using matrix $\boldsymbol{A}$.

In this study, as the first step in applying distance metric learning to speaker selection, we used Relevant Component Analysis（RCA） [13] which is one of the supervised distance metric learning methods since it is a simple and efficient technique.

2.1.1. Relevant component analysis

Given a set of vectors $\boldsymbol{X} = \{x_i\}_{i=1}^{N}$ and setting the K class for each vector, RCA trains a global linear matrix $\boldsymbol{M}$ to minimize the distance between vectors in each class. The optimal transformation by RCA is computed as $\boldsymbol{A} = \hat{\boldsymbol{C}}^{-1/2}$ and the Mahalanobis matrix is equal to the inverse of the average covariance matrix of classes, i.e., $\boldsymbol{M} = \hat{\boldsymbol{C}}^{-1}$, where $\hat{\boldsymbol{C}}$ is defined as follows:

$$\hat{\boldsymbol{C}} = \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{N_j}(\boldsymbol{x}_{ji} - \boldsymbol{\mu}_j)(\boldsymbol{x}_{ji} - \boldsymbol{\mu}_j)^T \quad (3)$$

here $\boldsymbol{\mu}_j$ denotes the mean of the j-th class and, x_{ji} denotes the i-th vector in the j-th class, and N and N_j are the total number of vectors and the total number of vectors in the j-th class respectively.

To apply RCA to speaker selection, we need to define the class and the vector. In this paper, the class and the vector are called the speaker class and the utterance vector respectively.

2.2. Speaker class using perceptual voice quality similarity

To set the speaker class for each speaker, we adopt a speaker clustering technique that uses the perceptual voice quality similarity. We utilize the perceptual similarity matrix as the speaker vector [7]. Let $\boldsymbol{v}_i$ be the speaker vector of speaker i. It is represented as

$$\boldsymbol{v}_i = [Sim(i,1), \cdots, Sim(i,j), \cdots, Sim(i,N_s)] \quad (4)$$

where, $Sim(i,j)$ represents the perceptual voice quality similarity between speakers i and j and N_s represents the number of speakers participating in the subjective experiment. In this paper, we set N as 62. Speaker clustering is done using the speaker vectors output by the k-means algorithm.

2.3. Utterance vector

We utilized the GMM supervector [8] as the utterance vector to actualize a text-independent similar speaker selection technique because its effectiveness in text-independent speaker recognition has been confirmed. The GMM supervector was created by concatenating the mean parameter of an individual GMM mixture. Given a speaker utterance, MAP adaptation is performed using a speaker-independent GMM that is trained in advance. Let μ_{ij} be the mean parameter of the adapted GMM's output distribution at mixture i and dimension j. The GMM supervector $\boldsymbol{m}$ is represented as

$$\boldsymbol{m} = [\mu_{11}, \cdots, \mu_{ij}, \cdots, \mu_{ML}] \quad (5)$$

where M and L represent respectively the number of GMM mixtures and the number of acoustic features' dimensions.

To compare the selection performance obtained with the utterance vector, we also used another vector called the "statistical vector", which was created by computing the mean, variance, and quartile of each acoustic feature.

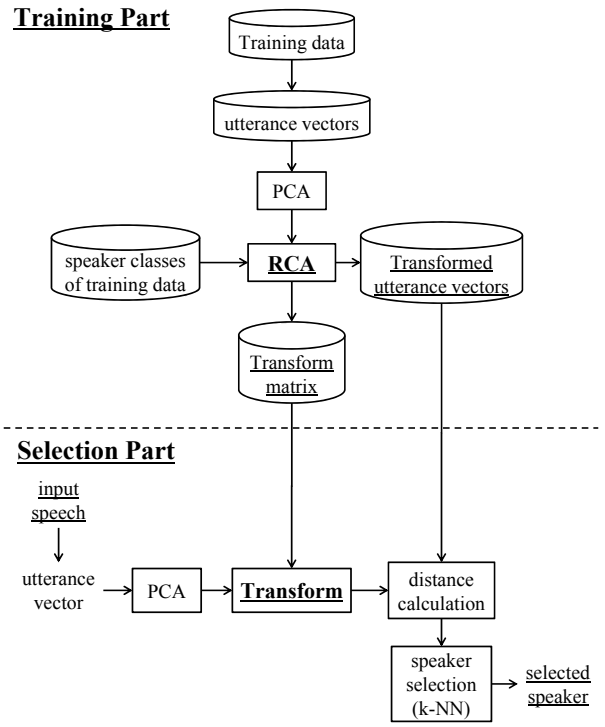

Figure 1: *A block diagram of the speaker selection system.*

2.4. Overview of proposed similar speaker selection

A block diagram of the proposed method is shown in Fig. 1. In the proposed technique, we first trained a transform matrix using training data with speaker class by RCA. When an input utterance is given, the utterance vector extracted from the input utterance is transformed using the trained transform matrix. After that, k-nearest neighbor (kNN) classifier-based speaker selection is performed by calculating the Euclidean distance between the input vector and the utterance vectors extracted from all the training data. The overall speaker selection process is summarized below.

Training part:

Step 1 Extract utterance vectors for each utterance from all training data.

Step 2 Perform PCA using the extracted utterance vectors for dimension reduction.

Step 3 Obtain the transform matrix **A** and transformed vectors (training vectors) using the speaker classes of training data and utterance vectors by RCA.

Selection part:

Step 4 Extract an utterance vector from an input speech.

Step 5 Perform PCA using the extracted utterance vector for dimension reduction.

Step 6 Transform the input utterance vector using the transform matrix **A** obtained from **Step 3**.

Step 7 Calculate the Euclidean distances between the transformed input utterance vector and the transformed training vectors obtained from **Step 3**.

Step 8 Select one speaker as a similar speaker, i.e., the speaker having the most utterances of the k nearest neighbors vector.

3. Experiments

3.1. Experimental conditions

In the following experiments, we used 62 female speakers' speech data included in the NTT-AT Japanese multi-speaker's speech database [14]. The sampling frequency of the speech was 16 kHz and the quantization bit rate was 16 bits. This database contains about 200 sentences for each speaker. The speakers' ages ranged from 18 to 49. We set the voice quality similarity between all speaker pairs (62×62) by means of a subjective experiment [7]. The rating is a 3-point scale, that is, 3 for very similar, 2 for slightly similar, and 1 for very dissimilar. The subjective experiment results made it clear that each speaker had at least one perceptually similar speaker.

Thirty sentences uttered by 61 speakers of the 62 speakers were used for the training data and 30 sentences uttered by the other speaker not included in the training data were used as the evaluation data. In the selection experiment, we first select one speaker as the evaluated speaker, and one speaker was chosen from the remaining 61 training speakers. We performed a leave-one-out cross-validation test in order to ensure the validity of the results obtained.

A speaker-independent GMM was trained from all speech data uttered by the 62 female speakers to extract the GMM supervector. We set selection parameters on the basis of preliminary experiment results, i.e., the number of GMM mixtures was set at 64, the number of dimensions in PCA at 27, and the number of nearest neighbors in the kNN classifier at 5.

To evaluate the speaker selection performance, we used average similarity. The average similarity is calculated by the perceptual voice quality similarity between the input speaker and the selected speaker obtained from the above mentioned subjective experiment [7]. Let $sel(utt_{ij})$ be the speaker selected from the speaker selection technique using utt_{ij}, which represents the j-th utterance uttered by speaker i. The average similarity is expressed as

$$\frac{1}{N_{eval}} \sum_{i=1}^{S} \sum_{j=1}^{U} Sim(i, sel(utt_{ij})) \tag{6}$$

where N_{eval}, S, and U represent respectively the number of evaluation utterances (S by U), the number of evaluated speakers, and the number of utterances per evaluated speaker, and $Sim(i, sel(utt_{ij}))$ represents the perceptual voice quality similarity between the input speaker i and the selected speaker from utt_{ij}.

3.2. Acoustic features

We utilized four acoustic features having high correlation with the voice quality similarity [7]. These acoustic features are described below.

- Low dimensional (1–12 dimensions) cepstral coefficients (cep)
- Low dimensional (1–12 dimensions) cepstral coefficients using log spectrum from 0 kHz to 4 kHz (cep4k)
- Average value of aperiodic component in full band (APm)
- Ratio of the power of 1 kHz or less to the power in full band (BSP)

These acoustic features were extracted using STRAIGHT [15]. The analysis frame shift was 5 ms. Because voice quality characteristics are chiefly presented by voiced phonemes rather than unvoiced phonemes, the following experiments were performed using only voiced frames as detected by TEMPO [15].

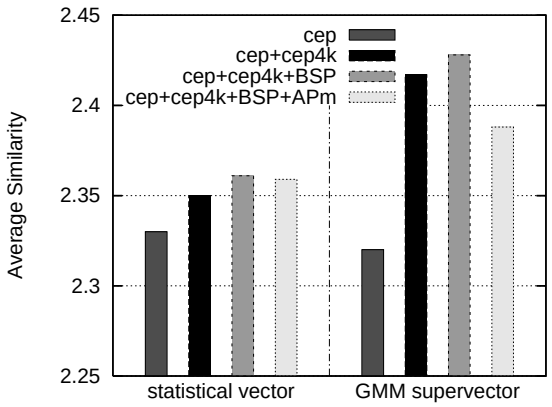

Figure 2: *Average similarity for each acoustic feature and utterance vector.*

3.3. Performance with acoustic feature and utterance vector

To compare the acoustic feature and utterance vector performances, we first performed speaker selection by changing the acoustic features and utterance vectors. In this experiment, we did not use RCA to perform distance metric learning. Figure 2 shows the average similarity obtained for each acoustic feature and utterance vector.

We can see that the average similarity increased by adding cep4k and BSP. On the other hand, the selection performance hardly changed at all when APm was added. This is because the utterance vectors we used (statistical vector and GMM supervector) cannot be considered the temporal characteristics of acoustic features. In our previous report [7], we used speech with exactly the same prosody (F0 and phoneme duration) to exclude the effect of the prosody. In this study, however, we used two types of vectors that cannot represent the temporal characteristics because they only represent the average characteristics of the whole utterance.

In addition, it can be seen that the average similarity of the GMM supervector was improved compared with that of that statistical vector in all acoustic features. This indicates that the GMM supervector is able to express the speaker characteristics better than the statistical vector. Therefore, we used the GMM supervector in the following experiments.

3.4. Performance comparison with distance metric learning

Next, we performed a speaker selection experiment by changing the number of speaker classes to investigate the effectiveness of distance metric learning in similar speaker selection. As in the previous experiment, we used two acoustic features, i.e., cep, cep4k, and BSP. Figure 3 shows the average similarity for each speaker class. We can see that the average similarity for each

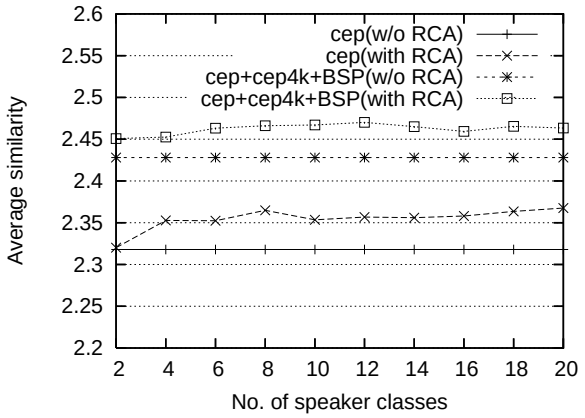

Figure 3: *Average similarity for the number of each speaker class.*

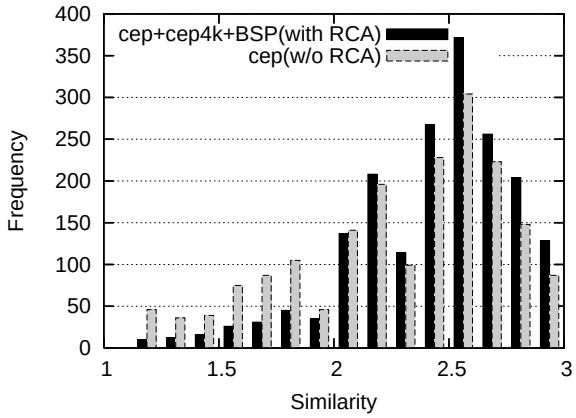

Figure 4: *Histogram of the similarity between the selected speaker and the input speaker.*

acoustic feature was improved by distance metric learning using RCA. Moreover, it can be seen that for the case of two speaker classes, the average similarity decreased, but for four or more classes the average similarity hardly changed at all. This is because the acoustic features that affect voice quality similarity differ for each speaker class. In our previous report [7], we confirmed that the speaker classes had different partial correlation coefficients and that speakers with different correlativity between acoustic feature and similarity were readily classified into different classes. However, since RCA can train a global transform matrix only, it cannot train the transform considering the difference between speaker classes of acoustic features.

Figure 4 shows the histogram of the similarity between the selected speaker and the input speaker. It can be seen that the number of speakers having low similarity decreased with distance metric learning when acoustic features were added. When we calculated the selection error rates, which is the rate at which a selected speaker has low similarity (i.e., less than 2), we found it was reduced from 23.33% to 9.41% and the error reduction rates was 59.6%. This indicates that the proposed method can significantly reduce the speaker selection error rates.

4. Conclusion

In this paper, we have presented a new speaker selection technique that takes the perceptual similarity into account in the selection process. This technique utilizes distance metric learning to transform acoustic feature space to perceptual similarity space. Speaker selection experiments showed that the proposed technique improves speaker selection performance. In particular, this technique can significantly reduce the speaker selection error rates.

In future work, we will investigate other distance metric learning techniques to improve the technique's speaker selection performance. Arbitrary speaker's speech synthesis based on perceptual similar speaker selection will also be investigated.

5. References

[1] 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, no.5, pp.825–834, May. 2007.

[2] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," IEICE Trans. Inf. and Syst. vol.E90-D, no.2, pp.533–543, Feb. 2007.

[3] J. Yamagishi, O. Watts, S. King and B. Usabaev, "Roles of the Average Voice in Speaker-adaptive HMM-based Speech Synthesis," in Proc. Interspeech 2010, pp.418–421, Sept. 2010.

[4] S. Yoshizawa, A. Baba, K. Matsunami, Y. Mera, M. Yamada and K. Shikano, "Unsupervised speaker adaptation based on sufficient HMM statistics of selected speakers," in Proc. ICASSP 2001, pp.341–344, May. 2001.

[5] C. Huang, T. Chen, E. Chang, "Speaker selection training for large vocabulary continuous speech recognition," in Proc. ICASSP 2002, pp.609–612, May. 2002.

[6] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech Communication, 17(1-2), pp.91–108, Aug. 1995.

[7] Y. Ijima, M. Isogai and H. Mizuno, "Correlation analysis of acoustic features with perceptual voice quality similarity for similar speaker selection," in Proc. INTERSPEECH 2011, pp.2237–2240, Aug. 2011.

[8] W. M. Campbell, D. E. Sturim and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification", IEEE Signal Processing Letters, vol. 13, No. 5, pp. 308–311, May, 2006.

[9] L. Yang, "An Overview of Distance Metric Learning", http://www.cs.cmu.edu/l̃iuy/dist_overview.pdf, 2007.

[10] H. Chang and D. Y. Yeung, "Kernel-based distance metric learning for content-based image retrieval", Image Vision Comput. 25, 5, pp.695–703, May 2007.

[11] M. Slaney , K. Weinberger and W. White, "Learning a metric for music similarity", in Proc. ISMIR 2008, pp.313–316, Sept. 2008

[12] D. Mochihashi, G. Kikui and K. Kita, "Learning an optimal distance metric in a linguistic vector space", Systems and Computers in Japan, pp.12–21, 2006

[13] A. Bar-Hillel, T. Hertz, N. Shental and D. Weinshall, "Learning Distance Functions Using Equivalence Relations", in Proc. ICML 2003, pp.11–18, 2003.

[14] Japanese speech database (in Japanese), http://www.ntt-at.co.jp/page.jsp?id=1793&content_id=337

[15] 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, 27, pp.187–207, 1999.