PAPER

Similar Speaker Selection Technique Based on Distance Metric Learning Using Highly Correlated Acoustic Features with Perceptual Voice Quality Similarity*

Yusuke IJIMA[†a)] ***and*** **Hideyuki MIZUNO**[†]**,** ***Members***

SUMMARY This paper analyzes the correlation between various acoustic features and perceptual voice quality similarity, and proposes a perceptually similar speaker selection technique based on distance metric learning. To analyze the relationship between acoustic features and voice quality similarity, we first conduct a large-scale subjective experiment using the voices of 62 female speakers and perceptual voice quality similarity scores between all pairs of speakers are acquired. Next, multiple linear regression analysis is carried out; it shows that four acoustic features are highly correlated to voice quality similarity. The proposed speaker selection technique first trains a transform matrix based on distance metric learning using the perceptual voice quality similarity acquired in the subjective experiment. Given an input speech, acoustic features of the input speech are transformed using the 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 of the proposed technique by comparing it to speaker selection without feature space transformation. The results indicate that transformation based on distance metric learning reduces the error rate by 53.9%.

key words: *voice quality, perceptual similarity, acoustic feature, speaker selection, distance metric learning*

1. Introduction

Recent research on text-to-speech synthesis has focused on generating arbitrary speech given only a small amount of the target speaker's speech data. The average-voice-based speech synthesis technique using model adaptation was proposed [3] for hidden Markov model (HMM)-based speech synthesis systems [4]. Given only a few minutes of speech data of the target speaker, this technique can transform an average voice model to the target speaker's model and then synthesize arbitrary speech. However, it was reported that the similarity of synthesized speech to the target speaker's speech is degraded by model transmission formation if the acoustic feature distance from the average voice model is large [5]. One useful approach to alleviating this problem is to select only speakers whose speech is similar to that of the target speaker when creating the average voice model. If the similar speakers are carefully selected, this approach is effective in synthesizing speech whose voice quality approaches that of the target speaker [6]. Furthermore, in the cross-lingual speaker adaptation technique [7], if a similar speaker whose language differs from that of the target speaker can be chosen, a model training technique based on similar speaker selection also may be effective in synthesizing different language speech where the voice quality is similar to that of the target speaker. Although the speaker characteristics generally consist of voice quality and prosody, our study focuses on voice quality because it sounds obvious that the differences of F0 and duration among speakers give the impression of prosodic similarity, while it is unclear which spectral feature impacts voice quality similarity. Therefore, determining the correlation between various acoustic features and perceptual voice quality similarity is essential for similar speaker selection. Furthermore, speaker selection taking account only of voice quality may be effective for cross-lingual adaptation that can make use only of global prosodic information.

In the field of automatic speech recognition, a variety of approaches have been proposed to train the acoustic model of the target speaker based on similar speaker selection [8], [9]. These techniques employ acoustic feature similarities such as likelihood of Gaussian mixture models (GMMs) [10]. However, even if two speakers have similar acoustic features' distributions, their voice quality is not necessarily perceptually similar. In order to enhance the effectiveness of speaker selection, we need to identify perceptual similar speakers. To do this we rely on two key components: (1) identification of acoustic features that have high correlation with perceptual voice quality similarity, (2) a speaker selection technique that takes into account the similarity of the perceived voice quality, not merely acoustic similarity.

To realize the first goal, a variety of approaches have been proposed to analyze the relationship between speaker characteristics and acoustic features [11]–[13]. Studies have shown that perceptual similarity is associated with prosodic features, consisting of fundamental frequency (F0) and phoneme duration, and acoustic features, consisting of cepstral coefficients and the aperiodic component. Because voice quality and prosody are evaluated simultaneously in subjective experiments, it was not clear which acoustic feature significantly influenced the human perception of voice quality [13]. Furthermore, because published similar-

Manuscript received June 5, 2014.
Manuscript revised September 5, 2014.
Manuscript publicized October 15, 2014.

[†]The authors are with NTT Media Intelligence Laboratories, NTT Corporation, Yokosuka-shi, 239–0847 Japan.

[*]A part of this paper was presented at INTERSPEECH 2011 [1] and at INTERSPEECH 2012 [2].

a) E-mail: ijima.yusuke@lab.ntt.co.jp

DOI: 10.1587/transinf.2014EDP7183

ity analyses considered only a dozen speakers at most, the range in voice qualities covered remains inadequate. For identifying the various relationships between acoustic features and perception in depth, it is essential to analyze the voices of many speakers.

Regarding the second goal, even if highly correlated acoustic features with perceptual voice quality similarity are found, it is not appropriate to simply use the Euclidean distance of each acoustic feature. Multiple regression analysis is widely used since it can weight each feature, but "distance metric learning [14]" (DML) is more effective since it can take side information into account. Many studies on DML have demonstrated its usefulness in applications such as image retrieval [15], music retrieval [16], and sentence retrieval [17]. This technique can realize speaker selection if the side information is set properly. We used the perceptual voice quality similarity obtained from a subjective experiment as the side information. In addition, DML has also been used for feature space transformation in a number of studies. For instance, [15] used transformation of the original image space for image retrieval. In this paper, since the perceptual voice quality similarity is used as the side information, DML can be considered to be transformation from acoustic feature space to perceptual voice quality similarity space.

In this study, our aims are to identify the acoustic features useful for the selection of perceptually similar speakers and to propose a speaker selection technique based on DML. We first conduct a large-scale subjective experiment using 62 female speakers to identify perceptual voice quality similarity. In the experiment, to exclude the influence of prosody, we use speech modified so as to exhibit exactly the same prosody (F0 and phoneme duration). Several acoustic features highly correlated to perceptual voice quality similarity are found by regression analysis of the results of the subjective experiment. In the proposed similar speaker selection technique, the transform matrix is first trained on the basis of DML to convert the acoustic feature space. Given a speech sample, the acoustic features of the sample are transformed using a trained transform matrix. Then, a similar speaker is chosen on the basis of Euclidean distances on the transformed acoustic feature space. To evaluate the proposed technique's performance, we compare it, in experiments, to speaker selection on an acoustic feature space without transformation. The results thus obtained demonstrated the technique's effectiveness.

This paper is organized as follows. Section 2 overviews the speech database used and the subjective experiment. Section 3 presents the correlation analysis conducted to link acoustic features to perceptual voice quality similarity; the key acoustic features as regards voice quality similarity are introduced. The proposed speaker selection technique, based on distance metric learning, and the results of similar speaker selection experiments are described in Sect. 4 and Sect. 5, respectively. Section 6 summarizes this paper.

2. Speech Database and Subjective Experiment

We first conducted a subjective experiment to evaluate voice quality similarity between many speakers. Speech stimuli and details of the subjective evaluation are described below.

2.1 Speech Database

We used the speech data of 62 female speakers included in the NTT-AT Japanese multi-speaker's speech database [18]. The sampling frequency of the speech was 16 kHz and the quantization bit rate was 16 bits. This database contains about 200 phonetically balanced sentences for each speaker. The speakers' ages ranged from 18 to 49.

2.2 Speech Samples Generated for the Evaluation

For the subjective experiment, we used a single sentence, "Shoo enerugii ga sakebarete imasu", (in English "Energy savings are desired") spoken by 62 non-professional female speakers included in the NTT-AT database.

To analyze the relationship between perceptual voice quality similarity and acoustic features, this evaluation removed the parameter of the prosody of speech. In this experiment, prosody modified speech with the prosody (F0 and phoneme duration) extracted from a speech uttered by a speaker other than the chosen 62 speakers in the NTT-AT database, was employed as the speech stimuli. To generate the speech stimuli with target prosody, original acoustic features (spectrum and aperiodic component) of each speech were linearly interpolated according to target duration and the F0 was modified to match the target F0. The interpolation was executed within each manually segmented phoneme boundary. We used the STRAIGHT [19] vocoder for speech analysis and synthesis. The analysis frame shift was 1 ms.

2.3 Subjective Experiment for Evaluation of Perceptual Voice Quality Similarity

A subjective experiment using the 62 speech stimuli was carried out. Subjects heard 3844 pairs (62×62) of speech stimuli, and rated the similarity of the presented speech pair. In order to counter the bias created by the order of stimuli presentation, the stimuli were also presented in inverse order. The rating scale is shown in Table 1. The subjects were 32 people (14 males and 18 females) who were listening to the speech stimuli for the first time. Each pair was evaluated by eight persons. Let $s(i, j)$ be the perceptual voice quality

Table 1 Evaluation criteria.

Score	Description
3	very similar
2	slightly similar
1	dissimilar

similarity between speaker i and j averaged over the evaluation scores of eight people. The voice quality similarity matrix component $Sim(i, j)$ is represented as follows.

$$Sim(i, j) = \begin{cases} \frac{s(i,j)+s(j,i)}{2} & (i \neq j) \\ s(i, j) & (i = j) \end{cases} \quad (1)$$

This yielded the voice quality similarity matrix, $Sim(i, j)$, for the 62 speakers.

Since each speech stimulus was evaluated by several different people in the subjective evaluation, the obtained similarity matrix might have been affected by differences in the listeners' evaluation criteria. However, it would be difficult to avoid this problem by performing a larger scale experiment because of its cost. Furthermore, the huge amount of evaluations by the same people that would be obtained in such an experiment might result in a lack of consistency in the evaluations. Therefore, each speech stimulus was evaluated by eight people we consider a minimum of subjects for subjective evaluation.

3. Regression Analysis between Perceptual Voice Quality Similarity and Acoustic Features

3.1 Acoustic Features

In analyzing the relationship between the perceptual voice quality similarity and acoustic features, we focused on ten acoustic features as described below.

- Low dimensional (1 to 12 dimensions) cepstral coefficients (CepL).
- High dimensional (13 to 24 dimensions) cepstral coefficients (CepH).
- Low dimensional (1 to 12 dimensions) cepstral coefficients using log spectrum from 0 kHz to 4 kHz (Cep4k).
- 1 to 12 dimensional coefficients of DCT value of aperiodic component (AP).
- Average value of aperiodic component in full band (APm).
- Ratio of the power in each sub-band to the power in full band (PR1–PR5).

  PR of i-th sub-band PRi is represented as follows.

  $$PRi = \frac{mean(spec_i)}{mean(spec_{full})}. \quad (2)$$

where, $spec_i$ and $spec_{full}$ represent respectively the spectrum in i-th band and the spectrum in full band ($0 - 8$ kHz). In this study, the spectrum was divided into 5 sub-bands ($0 - 1$, $1 - 2$, $2 - 4$, $4 - 6$, and $6 - 8$ kHz), using a spectral division method similar to that used for the aperiodic component in HMM-based speech synthesis.

Although many auditory features that take human perception into consideration, such as the perceptual linear predictive (PLP) feature [20], have been proposed, the goal of this study is not only achieving similar speaker selection but also applying it to speech synthesis. For this reason, we chose to use the acoustic features generally used in speech synthesis, i.e., cepstrum coefficients and an aperiodic component. We also used simple acoustic features such as power ratios (PR1–PR5) since simple features can be converted easily when synthesizing speech. In addition, previous studies, such as [21], showed the cepstrum features of speech, especially for high order cepstra, are affected by prosodic features. However, since the purpose of this paper is to identify acoustic features that have high correlation with perceptual voice quality similarity, we did not consider pitch information so as to exclude the effect of prosodic features. Furthermore, it should be noted that we used the cepstrum obtained from a lower-band rather than a higher-band log spectrum. This is because we believe that voice quality similarity would be more affected by the rough characteristics of the higher-band spectrum than affected by the detailed spectral shape of it. Since we used spectrum power ratio (PR1–PR5), we were able to take into account the rough characteristics of the higher-band spectrum in the following analysis.

As the acoustic distance measure of each speaker, we used the Euclidean distance of these acoustic features for each speaker's speech. First, an acoustic feature of the prosody modified speech used in the subjective experiment was extracted by STRAIGHT in every frame. Second, the Euclidean distance between the acoustic feature of one speaker and that of another speaker's speech was calculated in the frame; the average Euclidean distance is defined as the distance between the two speakers. The analysis frame shift was 1 ms. Because voice quality characteristics are chiefly presented by voiced phonemes rather than unvoiced phonemes, the distance was calculated using only voiced frames as detected by TEMPO [19].

As a result, the distance matrix of each acoustic feature was obtained as well as the voice quality similarity matrix.

In order to analyze the relationship between the perceptual voice quality similarity and acoustic features, we performed single and multiple regression analysis. In all analyses, the voice quality similarity and the distance matrix were provided except for the combination of same speaker's speech.

3.2 Regression Analysis

3.2.1 Single Regression Analysis

We first calculate single correlation coefficients between the perceptual voice quality similarity and each acoustic feature. Figure 1 shows single correlation coefficients for each acoustic feature. In this figure, because we calculated correlation coefficients between the distance of acoustic features and the perceptual voice quality similarity, acoustic features with negative correlation coefficient have high correlation with the perceptual voice quality similarity. The value of correlation coefficients shows that most acoustic features are correlated with perceptual voice quality similarity to some extent except for CepH and PR5. In particular, the four acoustic features CepL, Cep4k, PR1, and PR3 are correlated

Table 2 Correlation coefficients between all acoustic features.

	Cep4k	CepH	APm	AP	PR1	PR2	PR3	PR4	PR5
CepL	0.448	−0.145	0.143	0.232	0.221	0.177	0.235	0.205	0.169
Cep4k		−0.140	0.107	0.376	0.363	0.331	0.304	0.238	0.176
CepH			−0.241	−0.286	−0.293	−0.118	−0.338	−0.286	−0.254
APm				0.415	0.381	0.106	0.492	0.118	−0.003
AP					0.413	0.372	0.346	0.242	0.177
PR1						0.402	**0.799**	**0.638**	0.489
PR2							0.153	0.022	0.063
PR3								0.356	0.257
PR4									0.501

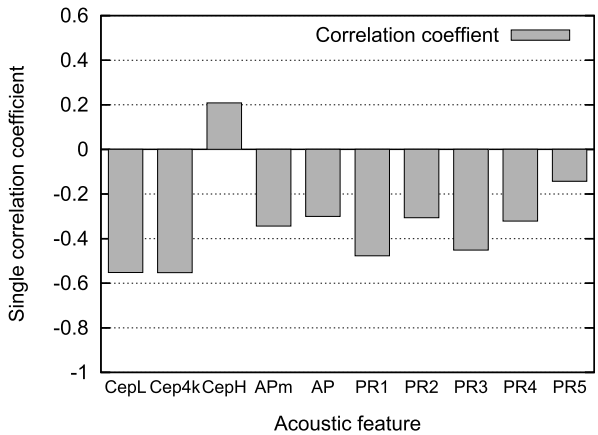

Fig. 1 Single correlation coefficients between perceptual voice quality similarity and each acoustic feature.

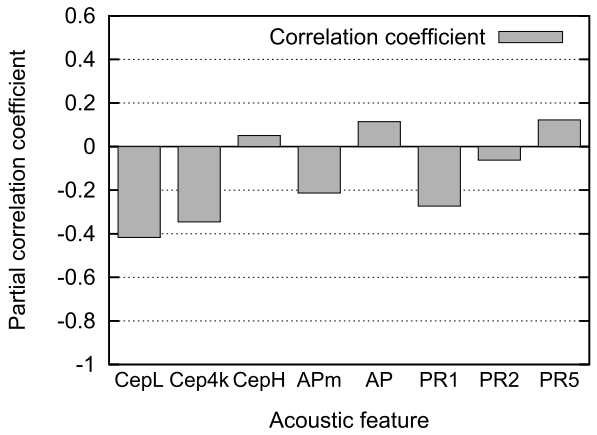

Fig. 2 Partial correlation coefficients for each acoustic feature.

with perceptual voice quality similarity because the values of the correlation coefficients are around −0.5.

Table 2 lists the correlation coefficients between each acoustic feature. We can see that PR1 has high correlation with PR3(0.799) and PR4(0.638). It is not desirable to utilize these acoustic features simultaneously for multiple regression analysis because doing so may cause multicollinearity. Other combinations have lower correlation.

3.2.2 Multiple Regression Analysis

Next, we perform multiple regression analysis to investigate the effect of linearly combining multiple acoustic features. Eight acoustic features, i.e., CepL, Cep4k, CepH, APm, AP, PR1, PR2, and PR5, were utilized as the explanatory variables of the regression. To avoid multicollinearity, PR3 and PR4 were not used since we confirmed from Table 2 they have high correlation with PR1. To obtain precise results from multiple regression analysis, it is necessary to avoid the use of these acoustic features simultaneously. We therefore used PR1, which has the highest single correlation coefficient. First, a multiple correlation coefficient was calculated using the above eight acoustic features. We confirmed that the perceptual voice quality similarity and the estimated one were highly correlated; the multiple correlation coefficient was "0.741". This result indicates that we can use these acoustic features to estimate voice quality similarity to some extent.

We also calculate the partial correlation coefficient for each acoustic feature. The results are shown in Fig. 2. The partial correlation coefficient values indicate that four acoustic features (CepL, Cep4k, APm, and PR1) have high correlation coefficients, which matches the results of Sect. **3.2.1**. On the other hand, the other four acoustic features, i.e., CepH, AP, PR2, and PR5, have low correlation coefficients.

Furthermore, in order to confirm the impact on similar speaker selection for each acoustic feature, we investigated the Bayesian information criterion (BIC) values for each combination of acoustic features. Table 3 shows the BIC values. In this table, each column shows the combinations of acoustic features which have the minimum BIC value when changing the number of acoustic features. From this table, we can see that the BIC values decrease as the number of acoustic features increase. This implies that these eight acoustic features are effective for similar speaker selection to some extent. However, since the BIC value reductions are different according to each acoustic feature, the impact on similar speaker selection is considered to be large in the order corresponding to Cep4k, CepL, PR1, and APm. This result is consistent with the result of multiple regression analysis.

From these results, these four acoustic features, i.e., CepL, Cep4k, APm, and PR1, are considered to be acoustic features highly correlated with perceptual voice quality similarity. Thus, in the following speaker selection experiments, we used these four acoustic features. Although the other acoustic features, i.e., CepH, AP, PR2, and PR5,

Table 3 Bayesian information criterion values for each combination of acoustic features.

# of acoustic feature	combination of acoustic features	BIC value
1	(1) Cep4k	3418.2
2	(2) (1)+CepL	2776.1
3	(3) (2)+PR1	2323.4
4	(4) (3)+APm	2157.4
5	(5) (4)+CepH	2051.1
6	(6) (5)+PR5	1962.9
7	(7) (6)+AP	1900.6
8	(8) (7)+PR2	1895.4

seem to be effective for speaker selection from the BIC values, we did not use these features since they correlated poorly with the perceptual voice quality similarity. We also calculated multiple correlation coefficients by using three (CepL+Cep4k+PR1) and four (CepL+Cep4k+PR1+APm) selected features. The multiple correlation coefficients obtained were 0.704 and 0.720, respectively.

4. Speaker Selection Technique Based on Distance Metric Learning

Next, we use distance metric learning (DML) to propose a similar speaker selection technique using the obtained acoustic features. An overview of our proposed selection system and its details are given below.

4.1 Overview of Proposed Similar Speaker Selection

A block diagram of the proposed method is shown in Fig. 3. In the proposed technique, we first employ distance metric learning to train a transform matrix using training data with speaker class. When an input utterance is given, the input utterance vector, described in Sect. **4.4**, extracted from the input utterance is transformed using the trained transform matrix. After that, k-nearest neighbor (kNN) [22] classifier-based speaker selection is performed by calculating the Euclidean distance between the transformed input utterance vector and the transformed utterance vectors extracted from all training data. The overall speaker selection process is summarized below.

Training part:

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

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

Step 3 Perform DML (RCA) to obtain the transform matrix $\boldsymbol{A}$ and transformed utterance vectors (training vectors) using the speaker classes of training data and dimension-reduced training utterance vectors.

Selection part:

Step 4 Extract an input utterance vector from the input speech.

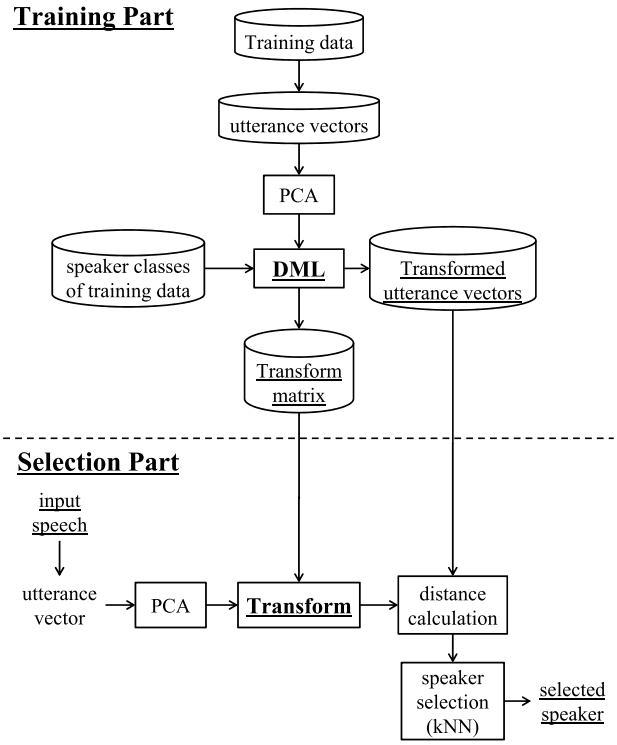

Fig. 3 A block diagram of the speaker selection system based on distance metric learning.

Step 5 Perform PCA using the extracted input 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 utterance vectors obtained from **Step 3**.

Step 8 Select one speaker as the most similar speaker, i.e., the speaker having the most frequent vectors among the k nearest neighbor vectors.

Because utterance vectors (described in Sect. **4.4**) are generally highly dimensional vectors, it is necessary to reduce the number of dimensions of the training vector to avoid the curse of dimensionality. For this reason, we perform PCA to achieve simple dimension reduction in **Step 2** and **Step 5**. After the dimension reduction, DML is performed using the dimension-reduced training vectors.

Details of each component, i.e., distance metric learning, the utterance vector, the speaker class, and kNN classifier-based speaker selection, are described as follows.

4.2 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

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{M} (\boldsymbol{x}_i - \boldsymbol{x}_j) \quad (3)$$

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 to $\boldsymbol{M} = \boldsymbol{A}^T\boldsymbol{A}$. This reduces Eq. (3) to

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \| \boldsymbol{A}(\boldsymbol{x}_i - \boldsymbol{x}_j) \|_2 \tag{4}$$

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

In this study, in order to avoid the sparse data problem, we used Relevant Component Analysis (RCA) [23], a well known supervised distance metric learning method. Although a number of DML techniques, such as Neighborhood Component Analysis (NCA) [24] and Large Margin Nearest Neighbour (LMNN) [25], have been proposed to train a more precise transform matrix $\boldsymbol{A}$, these techniques generally require much training data. However, since our proposed technique requires the perceptual voice quality similarity obtained from the subjective evaluation, we cannot collect sufficient training data for such DML techniques. Therefore, we used RCA to train the transform matrix because this technique is simple and effective.

4.2.1 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 global linear matrix $\boldsymbol{M}$ to minimize the distance between the 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:

$$\hat{\boldsymbol{C}} = \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{N_j}(\boldsymbol{x}_{ji} - \boldsymbol{\mu}_j)(\boldsymbol{x}_{ji} - \boldsymbol{\mu}_j)^T \tag{5}$$

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; 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.

4.3 Speaker Class Using Perceptual Voice Quality Similarity

To set the speaker class for each speaker, we adopt a speaker clustering technique based on perceptual voice quality similarity. We utilize the perceptual voice quality similarity matrix as the speaker vector obtained from Sect. **2.3**. Let $\boldsymbol{v}_i$ be the speaker vector of speaker i. It is represented as

$$\boldsymbol{v}_i = [Sim(i, 1), \cdots, Sim(i, j), \cdots, Sim(i, N_s)] \tag{6}$$

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_s to 62. Speaker clustering is done by applying the k-means algorithm to the speaker vectors.

4.4 Utterance Vector

We utilized the GMM supervector [26] as the utterance vector to realize 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 for mixture i and dimension j. The GMM supervector $\boldsymbol{m}$ is represented as

$$\boldsymbol{m} = [\mu_{11}, \cdots, \mu_{ij}, \cdots, \mu_{ML}] \tag{7}$$

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

To advance the field of speaker recognition, i-vector [27] was proposed to improve the performance of text-independent speaker recognition. Because the purpose of this paper is to confirm the effectiveness of applying distance metric learning to similar speaker selection, we used the GMM supervector only in the following experiments.

4.5 kNN Classifier-Based Speaker Selection

The k-nearest neighbor (kNN) classifier is the simplest classifier of all machine learning algorithms in pattern recognition. Because it is an effective and simple technique, it is used in various research fields. This paper also uses this technique for speaker selection.

Given an input utterance vector and all training utterance vectors described in Sect. **4.4**, the Euclidean distances between an input vector and all training vectors are calculated. Next, the k training vectors that have the smallest distance from the input vector are chosen. Finally, the speaker that yields the greatest number of selected k training vectors is selected as the similar speaker.

5. Experiments

5.1 Experimental Conditions

In the following experiments, we used the speech data of 62 female speakers as described in Sect. **2.1**. We used the perceptual voice quality similarity between all speaker pairs (62×62) as determined by the subjective experiment as described in Sect. **2.3**. Thirty sentences uttered by 61 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.

We utilized the four acoustic features with the highest correlation with the perceptual voice quality similarity as identified in Sect. **3.2.2**, i.e., CepL, Cep4k, APm, and PR1. These acoustic features were extracted using STRAIGHT [19]. The analysis frame shift was 5 ms. Although the frame shift was 1 ms in Sect. **2** and **3**, we changed it to 5 ms because a 1 ms frame shift is generally too short for GMM supervectors. The following experiments were performed using only voiced frames as in Sect. **3.2**.

A speaker-independent GMM was trained from all speech data uttered by the 62 female speakers (12400 utterances = 62 speakers × 200 utterances) to extract the GMM supervector. We set 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 in Sect. **2.3**. Let $sel(utt_{ij})$ be the speaker identified by the speaker selection technique using utt_{ij}, which represents the j-th utterance uttered by speaker i. The average similarity is expressed as

$$\frac{1}{N_{eval}} \sum_{i=1}^{S} \sum_{j=1}^{U} Sim(i, sel(utt_{ij})) \tag{8}$$

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; $Sim(i, sel(utt_{ij}))$ represents the perceptual voice quality similarity between input speaker i and the selected speaker from utt_{ij}.

5.2 Acoustic Feature Performance

To compare acoustic feature performances, we first performed speaker selection by changing the acoustic features. In this experiment, we did not use RCA to perform distance metric learning. In the proposed speaker selection methods, the optimal selection parameters (i.e., the number of GMM mixtures and the number of PCA dimensions) differ for each combination of acoustic features. Therefore, to set optimal parameters for each combination, we performed a preliminary experiment by changing these parameters. In the experiment, we set the number of GMM mixtures at 32, 64, and 128 and the number of dimensions of PCA from 10 to 40. From the obtained results, we respectively set the number of GMM mixtures for four combinations (i.e., CepL, CepL+Cep4k, CepL+Cep4k+PR1, and CepL+Cep4k+PR1+APm) at 32, 64, 64, and 128, and the number of PCA dimensions at 31, 30, 28 and 27.

Table 4 shows the average similarity obtained for each acoustic feature. We can see that the average similarity increased by adding Cep4k and PR1. On the other hand, the selection performance hardly changed at all when APm was added. This is because the utterance vectors we used fail to make allowance for the temporal characteristics of acoustic features. In Sect. **3**, we used speech with exactly the same prosody (F0 and phoneme duration) to exclude the effect of the prosody. In this section, however, we used GMM supervector, which cannot represent temporal characteristics because it represents only the average characteristics of the whole utterance.

5.3 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 suggested by the previous experiment, we used three acoustic features, i.e., CepL, Cep4k, and PR1. Figure 4 shows the average similarity for each speaker class. We can see that the average similarity for each acoustic feature was improved by distance metric learning using RCA.

However, it can be seen that for the case of two speaker classes, the average similarity decreased, but the change was slight, if at all, for four or more classes. This is because RCA fails to take into account the complexity according to the number of speaker classes. In general, when the number of speaker classes is increased, a transform matrix that can process the details of the acoustic feature space is required. However, RCA can train only a global transform matrix, and so cannot take account of the complexity created by the increase in the number of speaker classes.

Table 4 Average similarity for each acoustic feature.

Acoustic feature	Average similarity
CepL	2.35
CepL+Cep4k	2.43
CepL+Cep4k+PR1	2.44
CepL+Cep4k+PR1+APm	2.41

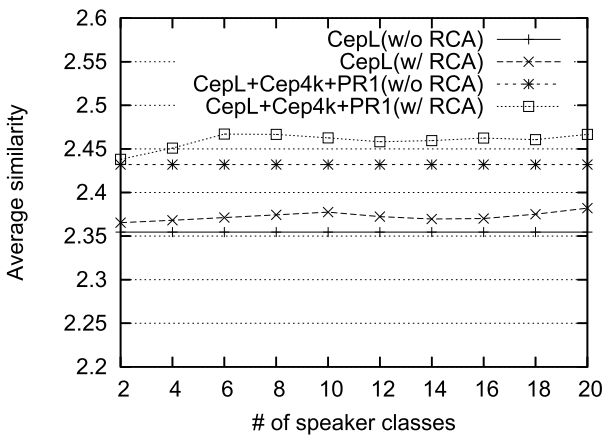

Fig. 4 Average similarity versus the number of speaker classes.

5.4 Overall Performance

We investigated the overall speaker selection performance by combining acoustic features and distance metric learning. Figure 5 shows the histogram of the similarity between the selected speaker and the input speaker. The number of speaker classes was set to 8 from the results of Sect. **5.3**. 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 rate, we found it was reduced by 53.9%, i.e., from 20.22% to 9.31%. A paired t-test we performed confirmed that the difference between the two methods is statistically significant at the 1% level. This indicates that the proposed method can significantly reduce the speaker selection error rates.

5.5 Comparison with Speaker Recognition Technique

Finally, we compared our proposed technique's speaker selection performance with that of a conventional speaker recognition technique based on GMM [10]. To obtain each speaker's GMM, we performed MAP adaptation from a speaker-independent GMM. As the speaker-independent GMM, we used the same model used in the proposed technique (described in Sect. **5.1**). We used CepL+Cep4k+PR1 as the acoustic feature, and 30 sentences uttered by each speaker were used for MAP adaptation.

Table 5 shows the average similarity results obtained from the experiment. These results confirmed that the two methods (the proposed technique without RCA and the GMM-based one) have comparable performance. In addition, applying feature space transformation using RCA confirmed that the average similarity obtained in doing so is higher than that of the GMM-based method. From these results, we confirmed the effectiveness of the proposed technique compared with the conventional speaker recognition technique.

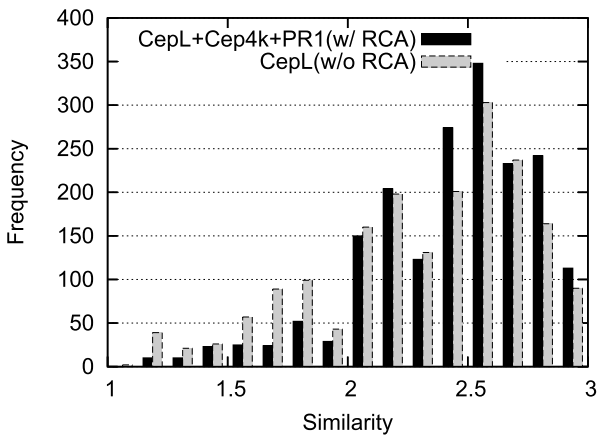

Fig. 5 Histogram of the similarity between the selected speaker and the input speaker.

Table 5 Performance comparison with GMM-based speaker recognition.

Method	Average similarity
GMM	2.43
Proposed (w/o RCA)	2.44
Proposed (w/ RCA)	2.47

6. Conclusion

In this paper, we analyzed the relationship between the perceptual voice quality similarity and various acoustic features for perceptually similar speaker selection. First, perceptual experiments using 62 female speakers' voices were designed and the perceptual voice quality similarity matrix between each speaker was determined. The results of multiple regression analysis showed that low dimensional cepstrum coefficient, low dimensional cepstrum coefficient under 4 kHz and the aperiodic component had high correlation to perceptual voice quality similarity; the multiple correlation coefficient was "0.741". Furthermore, we have presented a new speaker selection technique that takes perceptual voice quality similarity into account in the selection process. This technique utilizes distance metric learning to transform the acoustic feature space into the perceptual voice quality similarity space. Experiments showed that the proposed technique improves speaker selection performance. In particular, the proposed technique can significantly reduce the speaker selection error rates.

In future work, we will investigate other distance metric learning techniques, other speaker classes, and other utterance vectors to improve the technique's speaker selection performance. Although we have selected acoustic features using speaker selection by regression analysis, a unified approach to feature selection (i.e., [28]) will also be performed to select acoustic features matching our selection method based on distance metric learning.

References

[1] Y. Ijima, M. Isogai, and H. Mizuno, "Correlation analysis of acoustic features with perceptual voice quality similarity for similar speaker selection," INTERSPEECH 2011, pp.2237–2240, 2011.
[2] Y. Ijima, M. Isogai, and H. Mizuno, "Similar speaker selection technique based on distance metric learning with perceptual voice quality similarity," INTERSPEECH 2012, 2012.
[3] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," IEICE Trans. Inf. & Syst., vol.E90-D, no.2, pp.533–543, Feb. 2007.
[4] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-markov model-based speech synthesis system," IEICE Trans. Inf. & Syst., vol.E90-D, no.5, pp.825–834, May 2007.
[5] J. Yamagishi, O. Watts, S. King, and B. Usabaev, "Roles of the average voice in speaker-adaptive HMM-based speech synthesis," INTERSPEECH 2010, pp.418–421, Sept. 2010.
[6] R. Dall, M. Veaux, J. Yamagishi, and S. King, "Analysis of speaker clustering strategies for HMM-based speech synthesis," INTERSPEECH 2012, 2012.
[7] Y.J. Wu, S. King, and K. Tokuda, "Cross-lingual speaker adaptation

for HMM-based speech synthesis," ISCSLP2008, pp.1–4, 2008.
[8] S. Yoshizawa, A. Baba, K. Matsunami, Y. Mera, M. Yamada, and K. Shikano, "Unsupervised speaker adaptation based on sufficient HMM statistics of selected speakers," ICASSP 2001, pp.341–344, May 2001.
[9] C. Huang, T. Chen, and E. Chang, "Speaker selection training for large vocabulary continuous speech recognition," ICASSP 2002, pp.609–612, May 2002.
[10] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech Communication, vol.17, no.1-2, pp.91–108, Aug. 1995.
[11] N. Higuchi and M. Hashimoto, "Analysis of acoustic features affecting speaker identification," Eurospeech-95, pp.435–438, 1995.
[12] K. Amino, T. Sugawara, and T. Arai, "Speaker similarity in human perception and their spectral properties," WESPAC IX, 2006.
[13] Y. Adachi, S. Kawamoto, S. Morishima, and S. Nakamura, "Perceptual similarity measurement of speech by combination of acoustic features," ICASSP 2008, pp.4861–4864, 2008.
[14] L. Yang, "An overview of distance metric learning," http://www.cs.cmu.edu/~liuy/dist_overview.pdf, 2007.
[15] H. Chang and D.Y. Yeung, "Kernel-based distance metric learning for content-based image retrieval," Image Vision Comput., vol.25, no.5, pp.695–703, May 2007.
[16] M. Slaney, K. Weinberger, and W. White, "Learning a metric for music similarity," ISMIR 2008, pp.313–316, Sept. 2008.
[17] 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.
[18] NTT-AT, "Japanese speech database (in Japanese)." http://www.ntt-at.co.jp/product/denwa_j
[19] 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.
[20] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," The Journal of the Acoustic Society of America, vol.87, pp.1738–1752, 1990.
[21] N. Minematsu, K. Tsuda, and K. Hirose, "Quantitative analysis of f0-induced variations of cepstrum coefficients," ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding, pp.113–117, 2001.
[22] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inf. Theory, vol.IT-13, no.1, pp.21–27, 1967.
[23] N.S. A. Bar-Hillel, T. Hertz, and D. Weinshall, "Learning distance functions using equivalence relations," ICML 2003, pp.11–18, 2003.
[24] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," NIPS, pp.513–520, 2005.
[25] K. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbor classification," NIPS, pp.1473–1480, 2006.
[26] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," IEEE Signal Process. Lett., vol.13, no.5, pp.308–311, May 2006.
[27] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," INTERSPEECH 2009, pp.1559–1562, 2009.
[28] H. Liu and H. Motoda, Feature selection for knowledge discovery and data mining, Springer, 1998.

Yusuke Ijima received the B.E. degree in electric and electronics engineering from National Institution for Academic Degrees and University Evaluation by graduation from Yatsushiro National College of Technology, Japan, in 2007, and the M.E. degree in information processing from Tokyo Institute of Technology, Japan, in 2009. He joined NTT Cyber Space Laboratories (currently Media Intelligence Laboratories) in 2009, where he engaged in the research and development of speech synthesis. He is a member of the Acoustical Society of Japan(ASJ), the International Speech Communication Association(ISCA) and the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan.

Hideyuki Mizuno received the B.E. and M.E. degrees from Nagoya University, Japan, in 1986 and 1988, respectively. He received the Dr. Eng. degree in systems and information engineering from University of Tsukuba, Japan, in 2006. He joined NTT Human Interface Laboratories (currently Media Intelligence Laboratories) in 1988, where he engaged in the research and development of speech synthesis and voice quality conversion. He received the Technical Development Award of the Acoustical Society of Japan in 1998. He is a member of the ASJ and the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan.