Correlation Analysis of Acoustic Features with Perceptual Voice Quality Similarity for Similar Speaker Selection

Yusuke Ijima, Mitsuaki Isogai, Hideyuki Mizuno

NTT Cyber Space Laboratories, NTT Corporation, Japan

{ijima.yusuke, isogai.mitsuaki, mizuno.hideyuki}@lab.ntt.co.jp

Abstract

This paper describes the correlations between various acoustic features and perceptual voice quality similarity. We focus on identifying the acoustic features that are correlated with voice quality similarity. First, a large-scale perceptual experiment using the voices of 62 speakers is conducted and perceptual similarity scores between each pair of speakers are acquired. Next, multiple linear regression analysis is carried out; it shows that five acoustic features exhibit high correlation to voice quality similarity. Last, we perform similar speaker selection based on multiple linear regression with the above features and moreover, assess its performance by classifying speakers based on the perceptual similarity. The results indicate that the combination of the five acoustic features in classifying speakers into two classes is effective in choosing speakers with similar voice quality; it reduces the error rate by about 44 % compared to using just the cepstrum.

Index Terms: perceptual similarity, voice quality, acoustic feature, speaker selection

1. Introduction

Recent research on text-to-speech synthesis has focused on generating arbitrary speaker's speech given a small amount of the target speaker's speech data. For HMM-based speech synthesis systems [1], the average-voice-based speech synthesis technique using model adaptation was proposed [2]. Given just a few minutes of speech data of the target speaker, this technique can synthesize an arbitrary speaker's speech by model transformation from the average voice model to the target speaker's model. However, it was reported that the similarity of synthesized speech to the target speaker is degraded by model conversion if the mel-cepstral distance from the average voice model is large [3]. Therefore, creating an average voice model by selecting only speakers similar to the target speaker may be more effective in synthesizing speech whose voice quality approaches that of the target speaker. In addition, using acoustic features more effective for speaker model adaptation would yield more similar synthesized speech to the target speaker by the adaptation rather than using just the cepstrum.

A variety of approaches have been proposed to analyze the relationship between speaker characteristics and acoustic features [4–6]. Studies showed that perceptual similarity is influenced by prosodic features, consisting of 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 what acoustic feature significantly influenced the human perception of voice quality [6]. Furthermore, because similarity analysis considered only a dozen speakers at most, the various voice qualities of real speakers have not been covered.

In this study, our aim is identification of the acoustic features and the vocal tract characteristics useful for the selection of perceptually similar speakers and model adaptation to a target speaker model so as to generate an arbitrary speaker's speech. While the speaker's characteristic generally consists of voice quality and prosody, our study focuses on acoustic features that impact voice quality perception. The key to finding significant acoustic features is to analyze the relationship between perceptual voice quality similarity and the individual acoustic features. Furthermore, for finding various relationships between features and perception, it is desirable to analyze the voices of many more speakers. In this paper, we 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 synthesized speech with the exactly 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. Moreover, similar speaker selection based on multiple linear regression using such features is confirmed to offer good performance.

2. 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 samples generated for the evaluation

For the subjective experiment, 62 non-professional female speakers produced a single sentence "Shoo enerugii ga sakebarete imasu" (in English "Energy savings are desired"), present in the NTT-AT Japanese multi-speaker's speech database [7]. The age of speakers ranged from 18 to 49 and each had a different dialect. The sampling frequency of the speech was 16 kHz and the quantization bit rate was 16 bits.

To analyze the relationship between the perceptual voice quality similarity and acoustic features, the prosody of speech should be separated in the evaluation. In this experiment, the synthesized speech with the prosody (F0 and phoneme duration) extracted from a speech uttered a speaker other than the chosen 62 speakers in the NTT-AT database, was employed as speech stimuli. To generate 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 [8] vocoder for

Copyright © 2011 ISCA

Table 1: *Evaluation criteria.*

Score	Description
3	very similar
2	slightly similar
1	very dissimilar

speech analysis and synthesis. The analysis frame shift was 1 ms.

2.2. 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 by 62) of the speech stimuli, and rated the similarity of presented speech pair. In order to counterbalance any effects due to the order of stimuli, the stimuli were also presented in inverse order. The rating scale is shown in Table 1. Subjects were 32 persons (14 males and 18 females) who were listening to speech stimuli for the first time. Each pair was evaluated by 8 persons. Let $s(i,j)$ be the perceptual similarity between speaker i and j averaged over the evaluation scores of 8 people. The voice quality similarity matrix component $S(i,j)$ is represented as follows.

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

This yielded the voice quality similarity matrix; $S(i,j)$ for the 62 speakers. From the results of the evaluation, we found that at least one similar speaker (in terms of voice quality) existed for any speaker, while most speakers were rated as dissimilar.

3. Regression analysis

3.1. Feature parameters

To analyze the relationship between the perceptual similarity and acoustic features, six acoustic features and the warping parameter for vocal tract length normalization (VTLN) were employed. The six acoustic features are described below.

- Low dimensional (1 to 20 dimensions) cepstral coefficients (CepL).
- High dimensional (30 to 50 dimensions) cepstral coefficients (CepH).
- Spectral slope represented as 1st cepstral coefficient (Cep1).
- Low dimensional (1 to 20 dimensions) cepstral coefficients using log spectrum from 0 kHz to 4 kHz (Cep4k).
- 1 to 20 dimensional coefficients of DCT value of aperiodic component (AP).
- Average value of aperiodic component in full band (APm).

As the acoustic distance measure of each speaker, we used the Euclidean distance of the six acoustic features of each speaker's speech. First, an acoustic feature of the synthesized 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, and the average Euclidean distance is defined as the distance between the two speakers. 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 [8].

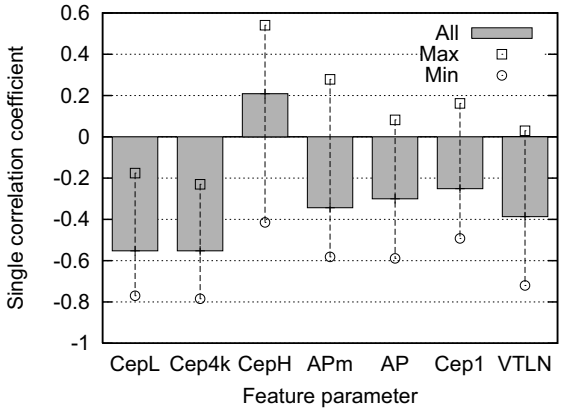

Figure 1: *Single correlation coefficients between perceptual similarity and each feature. (Error bars show maximum and minimum coefficients obtained from the 62 speakers)*

The warping parameter of VTLN was estimated from a speaker-dependent HMM for each speaker used in the subjective experiment. The feature vectors consisted of 20 cepstral coefficients not including the zeroth coefficient. We used 1-mixture 3-state left-to-right monophone HMMs. Each speaker-dependent HMM was trained using 200 sentences for each training speaker. The warping parameter, α, was estimated using 1 sentence used in the subjective experiment by selecting α with maximum likelihood from the prepared warping parameter set [9].

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

3.2. Regression analysis

In order to analyze the relationship between the perceptual similarity and feature parameters, we perform single and multiple regression analysis. In the analysis, the voice quality similarity and the feature matrix were provided except for the combination of same speaker's speech.

3.2.1. Single regression analysis

We first calculate single correlation coefficients between the perceptual similarity and each feature. These coefficients are calculated using all speakers' data and each speaker's data. Figure 1 shows single correlation coefficients for each feature. In this figure, "All" represents the coefficient using all speakers' data, and "Max" and "Min" indicate the maximum and minimum coefficients obtained from the 62 speakers. The value of "All" shows that low dimensional cepstral coefficients (CepL and Cep4k) have high correlation with perceptual similarity. It is also shown CepL and Cep4k are high correlated for all speakers from the value of "Min" and "Max". On the other hand, the values of CepH, APm and AP vary widely for each speaker. This implies that the effective features for speaker selection, other than CepL and Cep4k, are different for each speaker.

Table 2 lists the correlation coefficients between each feature. We can see that CepL has high correlation with Cep1(0.542) and VTLN(0.640). It is not desirable to utilize them simultaneously for multiple regression analysis. Other

Table 2: *Correlation coefficients between each feature.*

	Cep4k	CepH	APm	AP	Cep1	VTLN
CepL	0.448	-0.145	0.143	0.232	0.542	0.641
Cep4k		-0.140	0.107	0.376	0.108	0.290
CepH			-0.241	-0.286	0.035	-0.023
APm				0.415	0.130	0.112
AP					0.249	0.071
Cep1						0.361

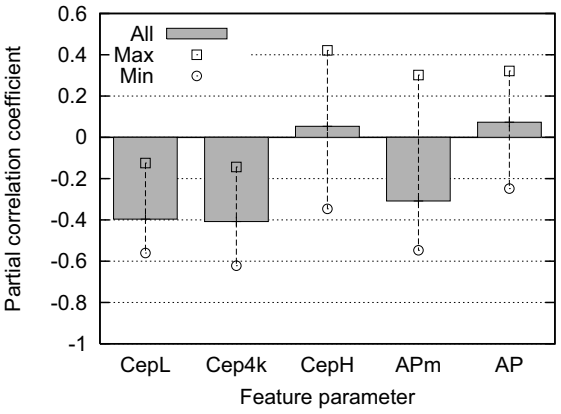

Figure 2: *Partial correlation coefficients for each feature. (Error bars show maximum and minimum coefficients obtained from the 62 speakers)*

combinations have lower correlation.

3.2.2. Multiple regression analysis

Next, we perform multiple regression analysis to investigate the effect of linearly combining multi acoustic features. CepL, Cep4k, CepH, APm and AP were utilized as the explanatory variables of the regression. Cep1 and VTLN were not used because they have high correlation with CepL. First, a multiple correlation coefficient was calculated using the above five features. We confirmed that the perceptual similarity and the estimated one were highly correlated; the multiple correlation coefficient was 0.697. This result indicates that we can use these features to estimate voice quality similarity to some extent.

We also calculate the partial correlation coefficient for each feature. The results are shown in Fig. 2. "All" indicate that three features (CepL, Cep4k and APm) have high correlation coefficients. It is also shown CepL and Cep4k are highly correlated for all speakers matching the results of Sect. **3.2.1**. On the other hand, the values of CepH, APm and AP vary widely for each speaker.

3.3. Speaker clustering using perceptual similarity

The results of the regression analysis show that some acoustic features exhibit different correlation values with the perceptual similarity for each speaker. In order to utilize these acoustic features for similar speaker selection, one useful approach is to cluster speakers considering the correlation between acoustic feature and similarity. Moreover, the relationship between acoustic features and perceptual similarity may be made clearer by analyzing each cluster obtained from the speaker clustering technique based on perceptual similarity.

In this paper, we propose a speaker clustering technique that use the perceptual voice quality similarity. We utilize the perceptual similarity matrix as the speaker vector. Let $\boldsymbol{v}_i$ be the speaker vector of speaker i. It is represented as

$$\boldsymbol{v}_i = [S(i,1), \cdots, S(i,j), \cdots, S(i,N)] \quad (2)$$

where, N represents the number of speakers in the multi speaker database. This vector representation is similar to the anchor model [10] for speaker recognition. Speaker clustering is done using the speaker vectors output by the LBG algorithm.

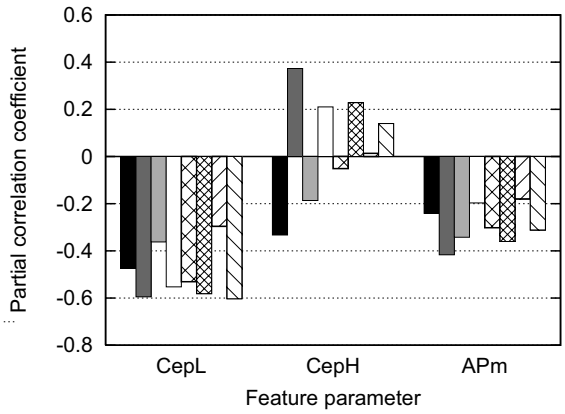

Figure 3: *Partial correlation coefficients for each cluster. (The number of clusters is 8)*

Figure 3 shows the partial correlation coefficients of each cluster for the case of 8 clusters. In this figure, bars with the same pattern indicate the coefficient of the same cluster. For instance, the black bar and white one are for cluster 1 and cluster 4, respectively. We confirmed that the clusters had different partial correlation coefficients, and that speakers with different correlativity between acoustic feature and similarity were readily classified into different classes.

4. Similar speaker selection experiment

We performed the similar speaker selection experiment to confirm the effectiveness of combining multi acoustic features and speaker clustering based on perceptual similarity.

4.1. Selection method

In order to select the perceptually similar speaker to the target speaker from the multi speaker database, we utilize the similarity as estimated by multiple regression analysis. The selection process is summarized as follows:

Step 1 Obtain the distance of each feature between the target speaker and each speaker in the multi speaker database.

Step 2 Estimate the similarity of each speaker using obtained distance in **Step 1**

Step 3 Select the speaker with the highest estimated similarity as the similar speaker.

4.2. Experimental conditions

We used the same speech data described in Sect. **??**. We utilized five acoustic features: CepL, Cep4k, CepH, APm, and AP. The number of speaker clusters was set to 1, 2, 4 and 8.

In the selection, we first select one speaker as the target speaker, and one speaker was chosen from the remaining

Table 3: *Speaker selection error rates (%) of each acoustic feature.*

Feature parameter		Error rate
single feature	CepL	29.03
	Cep4k	41.94
	APm	74.19
	AP	62.99
combination of two features	CepL+Cep4k	22.58
	CepL+APm	27.42
	CepL+AP	24.20
All		20.97

Table 4: *Speaker selection error rates (%) for different numbers of clusters.*

# of cluster	1	2	4	8
Error rate	20.97	16.13	24.2	22.58

61 speakers. From results of the subjective experiment, each speaker has at least one perceptual similar speaker. We performed a 62-fold cross-validation test. We set the classification threshold using the perceptual similarity between the target speaker and selected speaker from 2.5 to 3.0 for correct selection, and from 1.0 to 2.5 for incorrect selection. This is because the majority of subjects assessed similar subjects to be 3 (very similar) in the subjective evaluation.

We evaluate the performance of speaker selection using speaker selection error rate. The error rate was calculated by

$$error(\%) = \left(\frac{N-C}{N}\right) \times 100 \qquad (3)$$

where N and C represent the total number of selected speakers and the number of correctly selected speakers, respectively.

4.3. Experimental results

To examine the effect of combining the multi acoustic features on similar speaker selection, we first performed speaker selection using each single feature and then te combination of multi features. The number of speaker clusters was 1. In the case of a single feature, we chose the speaker with minimum distance as the similar speaker. Table 3 shows the selection error rate for each feature. In this table, the entry for "All" means the combination of five features: CepL, Cep4k, CepH, AP and APm. We can see that CepL had lower error rate than the other features in the single feature case. Moreover, "All" gave lower error rate than the other features so performance was improved by combining the multi features. The error reduction rate of "All" from CepL was about 27 %.

Next, we calculated error rates by changing the number of clusters to evaluate the performance of speaker clustering based on perceptual similarity. The feature parameter was the same as "All". Table 4 shows error rate versus the number of clusters. We can see that the performance was improved when the number of clusters was 2; the error reduction rate from CepL was about 44 %.

4.4. Discussion

We achieved the speaker selection error rate of about 20 % using the combination of five acoustic features. To improve the performance, we proposed speaker clustering based on perceptual similarity. Although the performance was improved when the number of clusters was 2, the error rates were not improved significantly. On the other hand, we performed similar speaker selection using a SVM to evaluate the performance of another classifier, and the F-measure for training data was about 0.56. This is because that the feature-space of similar speakers substantially overlaps that of dissimilar speakers. We will explore other features considering the temporal characteristics of acoustic features.

5. Conclusions

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 speakers' voices were designed and the perceptual 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 with the multiple correlation coefficient of 0.697. From experiments on similar speaker selection, we achieved the speaker selection error rate of about 20 % using the combination of these features in classifying speakers into two classes, and the error reduction rate from the cepstrum-only approach was about 44 %. In future work, we will investigate other features considering the temporal characteristics of acoustic features. Arbitrary speaker's speech synthesis based on perceptual similar speaker selection will also be confirmed.

6. 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, Sep. 2010.

[4] N. Higuchi and M. Hashimoto, "Analysis of acoustic features affecting speaker identification," in Proc. Eurospeech '95, pp.435–438, 1995.

[5] K. Amino, T. Sugawara and T. Arai, "Speaker Similarity in Human Perception and their Spectral Properties," in Proc. WESPAC IX, 2006.

[6] Y. Adachi, S. Kawamoto, S. Morishima and S. Nakamura, "Perceptual similarity measurement of speech by combination of acoustic features," in Proc. ICASSP 2008, pp.4861–4864, 2008.

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

[8] 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.

[9] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in Proc. ICASSP '96, 1:346–348, May 1996.

[10] S. R. Singer, E. Singer and J. P. Campbell, "Speaker Indexing In Large Audio Databases Using Anchor Models," in Proc. ICASSP 2001, pp.429–432, 2001.