PERCEPTUAL SIMILARITY MEASUREMENT OF SPEECH BY COMBINATION OF ACOUSTIC FEATURES

Yoshihiro Adachi[†] [‡], *Shinichi Kawamoto*[†], *Shigeo Morishima*[‡], *and Satoshi Nakamura*[†]

†ATR Spoken Language Communication Research Laboratories,
2-2-2 Keihanna, Science City, Kyoto, 619-0288 Japan
‡Science and Engineering, Waseda University,
3-4-1 Okubo Shinjuku-ku Tokyo, 169-8555 Japan
xyadachi@toki.waseda.jp, shinichi.kawamoto@atr.jp, shigeo@waseda.jp, satoshi.nakamura@atr.jp

ABSTRACT

Future cast system is a new entertainment system where participant's face is captured and rendered into the movie as an instant Computer Graphics (CG) movie star, which had been first exhibited at the 2005 World Exposition in Aichi Japan. We are working to add new functionality which enables mapping not only faces but also speech individualities to the cast. Our approach is to find a speaker with the closest speech individuality and apply voice conversion. This paper investigates acoustic features to estimate perceptual similarity of speech individuality. We propose a method linearly combined eight acoustic features related to the perception of speech individualities. The proposed method optimizes weights for the acoustic features considering perceptual similarities. We have evaluated performance of our method with Spearman's rank correlation coefficients to perceptual similarities. As the results, the experiments evidenced that the proposed method achieves a correlation coefficient of 0.66.

Index Terms— Acoustic correlators, Speaker recognition, Speech analysis

1. INTRODUCTION

Future cast system (FCS) [1] is the world's first entertainment system which enables anyone to easily participate in a prerecorded movie as an instant CG movie star. FCS can automatically perform all the processes from capturing participant's facial characteristics using a 3D range scanner for rendering them into the movie. Additionally, this system allocates a suitable role for the participants in the story and each actor begins to speak and perform in a fully CG based movie as vividly as any real actor. However, the prerecorded voice of either an actor or actress is used as a substitute for that of each participant. The substitute voice is selected depending on only each participant's gender information which is estimated based on the scanned face shape without consideration of other information such as age and voice quality. This caused some mismatch for those who perceive the voice of the character to be different from their own or the people they know. Therefore we decided to focus on selecting the similar speaker from speech database to reduce the mismatch. We propose a method to measure the perceptual similarity of speech, because it is impossible to record all speech of participants in advance, and to convert voice quality sufficiently with the present conversion technology.

Speaker recognition which deals with the similarity of speech have been researched and applied to several fields such as security field. In the security field, the similarity of speakers is generally determined based on likelihood of Gaussian Mixture Model (GMM) between speakers. However, we focus on the perceptual similarity rather than the similarity of speaker models. Additionally, the aim of speaker recognition is to perceive oneself, not to search a similar speaker. Meanwhile, as for the relation between the perceptual similarity and the acoustic distance, Amino et al. [2] proved the strong correlation between the cepstral distance and the perceptual similarity. Nagashima et al. proved the strong correlation between spectrum distances at 2 - 10 kHz and the perceptual similarity of speech in which utterance speed and intonation were controlled by speakers [3]. Because the personality of speech appears not only in voice quality but also in utterance speed or intonation, we examine the acoustic feature for the estimation of perceptual similarity between perceptual speech similarity using natural speech.

In this paper, acoustic feature means following features such as Mel Frequency Cepstral Coefficient (MFCC), STRAIGHT Cepstrums, spectrum, STRAIGHT-Ap (aperiodic component) which is an analysis parameter of STRAIGHT [4], fundamental frequency, formants and spectrum slope. All these features are related to the perceptual similarity. In this paper, we demonstrate how the combined feature estimates the perceptual similarity.

A method to represent the perceptual similarity is mentioned in section 2. In section 3, the acoustic feature to estimate the perceptual similarity is introduced. In section 4, we

explain the estimation experiment of the perceptual similarity using the acoustic feature, and show the results. In section 5, conclusions and future works are described.

2. PERCEPTUAL SIMILARITY

By sorting the speech data according to the perceptual similarity, it is clear which is the second or third similar speech to the target speech. Therefore, we can assigned several similar speech to multiple target speakers from same database with no overlaps in FCS. For that reason, the perceptual similarity is rendered with a permutation. The speech data in the database is sorted in descending order of perceptual similarity to the target speech as follows.

1. The target speech is defined as X.
2. A speech is selected randomly from our speech database and defined as A.
3. Rest of the speech data are divided into two groups (one is more similar to X than A, and the other is opposite.)
4. Same processing is applied (2 and 3) to each divided group.
5. Speech data are sorted according the perceptual similarity by repeating this process recursively.

In the process 3, the similarity between two speech data is judged from comprehensive impression not just focusing one of the acoustic features (e.g. quality of voice, intonation and speaking rate).

3. ACOUSTIC FEATURE

The acoustic feature to estimate the perceptual similarity is the combination of eight acoustic features. These acoustic features are weighted for stronger relationship between the combined feature and the perceptual similarity. Following sentences explain the eight acoustic features, how to calculate the acoustic similarity, and how to decide the weights for the acoustic features. Acoustic features are extracted except for the silent part: a pause.

3.1. Acoustic Features

3.1.1. MFCC

MFCC is one of acoustic features which is robust in noisy environments, and commonly used for not only speech recognition but also speaker recognition with GMM. In our study, MFCC is represented with the vector of 25 dimensions (12 static, 12 dynamic, 1 dynamic power).

3.1.2. STRAIGHT Cepstrums

Kitamura described that the perception of personality is influenced by the high order STRAIGHT Cepstrum and the first STRAIGHT Cepstrum which represent the fequency characteristic of vocal sound source and that gradient respectively [5]. Therefore, we decided to focus on the relation between the perceptual similarity and the high order STRAIGHT Cepstrum or first STRAIGHT Cepstrum.

Cepstrum is extracted by STRAIGHT analysis, and that of over 35 dimensions are defined as the high order STRAIGHT Cepstrum (CepH). The first STRAIGHT Cepstrum (Cep1) is the first dimension of the calculated STRAIGHT Cepstrum.

3.1.3. Spectrum

High frequency spectrum has also the strong relation with personality. Furui et al. [6][3] demonstrated the strong relation between a high frequency spectrum and personality. Therefore, we research the relation between the high frequency log spectra and the perceptual similarity. In this paper, the high spectrum means over 2.6 kHz [5] spectrum frequency.

3.1.4. STRAIGHT-Ap

Saito et al. [7] discovered the individual feature in STRAIGHT-Ap (Ap) under 2 kHz. Therefore, we focus on the relation between Ap and the perceptual similarity.

3.1.5. Fundamental Frequency

We investigated the relation between fundamental frequency (F0) and the perceptual similarity, because Hashimoto et al. [8] have proved that F0 has an effect on the personality perception. F0 is extracted using STRAIGHT-TEMPO which is a part of STRAIGHT analysis.

3.1.6. Formants, Spectrum Slope

Voice quality is a critical acoustic feature to assess the similarity of speech. Kido et al. [9] described that formants (Formant) and spectrum slope (SpecSlope) are indispensable features for representation of voice quality. In this paper, Formant means from 1st to 4th formant, and SpecSlope is a gradient from 0 kHz to 3 kHz log Spectrum.

3.2. Acoustic Similarity

3.2.1. GMM Likelihood

The aim of speaker identification is to search the target speaker. Likelihood of GMM is commonly used in speaker identification [10]. Remarkable advantages of using GMM are to be able to represent complex feature vectors, to be robust in noisy environments, and to be independent from the context of speech. Therefore, GMM is effective in speaker identification and verification [11][12].

GMM likelihood defined as the acoustic similarity is examined to reveal the relation between GMM likelihood and the perceptual similarity. 16 mixtures GMM is used in our experiments.

3.2.2. DTW Distance

We recognize the Dynamic Time Warping (DTW) distance as a measure of the acoustic speech similarity. Sakoe et al. [13] have developed the DTW distance for matching of speech signals with time warping. DTW is commonly used in a wide

range of pattern recognition because of the simplicity of the theory, the ease of implementation and a small amount of calculation.

In this study, firstly, all acoustic features are extracted per every 10 ms. Secondly, the Euclidean distances of acoustic features are calculated by every affiliated frame based on DTW, and then the mean distance is defined as the acoustic similarity calculated by DTW.

3.3. Weighting

The perceptual similarity is estimated using the combined feature that is a weighted linear coupling of the eight acoustic features when the correlation between the acoustic similarity and the perceptual similarity is strong. That is to say, the weights should be decided to have high correlation coefficient. Therefore, the weight is calculated using the steepest descent method in a way that Spearman's rank correlation coefficient between the acoustic similarity and the perceptual similarity is getting bigger. The Spearman's rank correlation coefficient ρ is represented in equation (1).

$$\rho = 1 - \frac{6\sum_{i=1}^{N}(a_i - b_i)^2}{N^3 - N} \quad (1)$$

a is the permutation according to the perceptual similarity by the subject. b is the permutation according to the acoustic similarity. Because the perceptual similarity is described in a permutation, the weights are evaluated using the rank correlation coefficient.

4. ESTIMATION EXPERIMENT OF PERCEPTUAL SIMILARITY

4.1. Experimental Procedure

We evaluate the estimation accuracy of the perceptual similarity using the combined feature. First, we examine the reproducibility of the perceptual similarity represented by the subject. The subject sorts the speech data in our database according to the perceptual similarity twice per a target speech. Spearman's rank correlation coefficient is calculated using those two permutations. We target this correlation coefficient for estimating the perceptual similarity by the acoustic feature. Next, we evaluate the estimation accuracy using the combined feature. Leave-one-out cross-validation is applied to this evaluation. The weights are optimized for eight acoustic features using the steepest descent method. Finally, we compare the rank correlation coefficient between a combined feature and its each component for evaluating accuracy of similarity estimation.

4.2. Experimental Conditions

We prepare 36 speech data uttered by 36 females for the evaluation. By limiting the gender of speakers, we can select the perceptual similar speech to the target speaker from speech data expected to have similar voice quality. The sentence of the speech is the Japanese sentence 'Arayuru genjitu o subete jibun no hoe nejimageta noda.' Average duration of the speech data is about 4 seconds. The range of female speaker's ages was from 18 to 59 years old. These speech signals are sampled with 16 kHz and quantized with 16 bit. The permutation of the perceptual similarity has been decided by one subject.

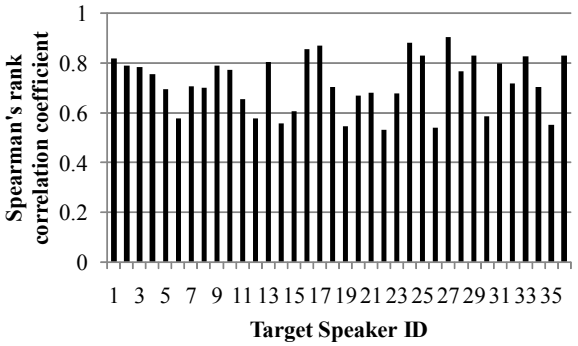

Fig. 1. The reproducibility of perceptual similarity.

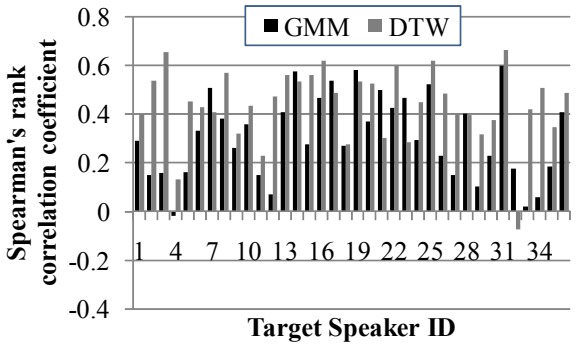

Fig. 2. The estimation accuracy of the perceptual similarity using the combined feature.

4.3. Experimental Results

The result of the reproducibility of the perceptual similarity is shown in Figure 1. The average of the rank correlation coefficients is 0.72, the standard deviation is 0.11, and the maximum is 0.90. This result indicates that the subject can sort the speech data this rank correlation coefficient, which we regard as the target of the estimation accuracy by using the acoustic feature.

The result of the estimation accuracy of the perceptual similarity using the combined feature is shown in Figure 2. The averages of the correlation coefficients are 0.29 (GMM), 0.44 (DTW), the standard deviations are 0.21 (GMM), 0.15 (DTW), and the maximums are 0.60 (GMM), 0.66 (DTW). This result shows that estimation using DTW is more reliable than GMM. Temporal structure of acoustic features is

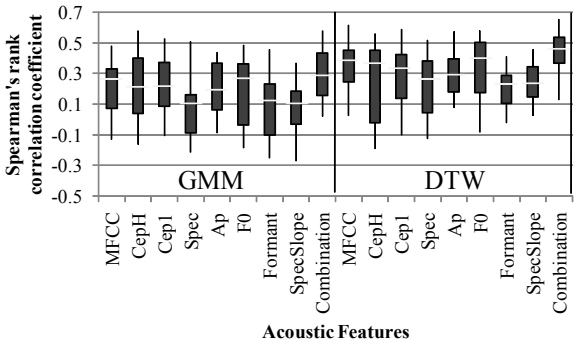

Fig. 3. The comparison of estimation accuracy between a combined feature and its each component.

one of important information. This is because that human generally identify the speech similarity considering the information. Both DTW and GMM calculate the similarity based on acoustic features: those which have temporal structure are used for DTW and others with no temporal structure are used for GMM. Therefore, the result of DTW at this experiment has evaluated the perceptual similarity effectively more than that of GMM.

The estimation accuracy of perceptual similarity with the combined feature and its each component are shown in Figure 3. The median of combined feature (combination) is higher than that of other component features around 0.07-0.23 in rank correlation coefficient using DTW.

As a result, the estimation with the combined feature is more accurate than with the component features, though it is not enough to be equal to the subject's reproducibility in Figure 1.

5. CONCLUSIONS

This paper described the combined acoustic feature to estimate the perceptual speech similarity. The experimental result with 36 speech data uttered by 36 females proved that the acoustic feature combined with 8 acoustic features was more effective than its component features.

We have researched with the perceptual similarity answered by one subject. However, the similarity of the subject is not always equal to other subjects. As future work, we need to increase the number of subjects. Moreover, we need to deal with not only female speakers, but also male speakers. Furthermore, we would apply Hidden Markov Models (HMMs) and Support Vector Machine (SVM) for the calculation of the acoustic similarity. HMMs have potential of higher resolution since it can represent temporal structures. On the other hand, SVM also can provide higher performance since it has higher discriminative abilities [14].

6. ACKNOWLEDGMENTS

This research is supported by the Special Coordination Funds for Promoting Science and Technology of Ministry of Education, Culture, Sports, Science and Technology.

7. REFERENCES

[1] S. Morishima, A. Maejima, S. Wemlera, T. Machida, and M. Takebayashi.: Future Cast System. ACM SIGGRAPH 2005 Sketch. ACM SIGGRAPH 2005 Full Conference DVD-ROM Disc 2. ISBN 1-59593-099-X.020-morishima.pdf (2005)

[2] K. Amino, T. Sugawara, and T. Arai.: Speaker Similarities in Human Perception and their Spectral Properties. Proc. of WESPAC IX, (2006)

[3] I. Nagashima, M. Takagiwa, Y. Saito, Y. Nagao, H. Murakami, M. Fukushima, and H. Yamnagwa.: An investigation of speech similarity for speaker discrimination. Acoustical Society of Japan 2003 Spring Meeting, (2003) 737–738 [in Japanese].

[4] H. Kawahara.: STRAIGHT: An extremely high-quality VOCODER for auditory and speech perception research. in Computational Models of Auditory Function (Eds. Greenberg and Slaney), IOS Press, (2001) 343–354

[5] T. Kitamura and T. Saitou.: Contribution of acoustic features of sustained vowels on perception of speaker characteristic. Acoustical Society of Japan 2007 Spring Meeting, (2007) 443–444 [in Japanese].

[6] S. Furui and M. Akagi.: Perception of voice individuality and acoustic correlates. Journal of the Acoustical Society. of Japan, vol. J66-A, (1985) 311–318

[7] T. Saitou and T. Kitamura.: Factors in /VVV/ concatenated vowels affecting perception of speaker individuality. Acoustical Society of Japan 2007 Spring Meeting, (2007) 441–442 [in Japanese].

[8] N. Higuchi and M. Hashimoto.: Analysis of acoustic features affecting speaker identification. Proc. of EUROSPEECH '95, (1995) 435–438

[9] H. Kido and H. Kasuya.: Voice quality expressions of speech utterance and their acoustic correlates. Technical report of IEICE, SP2002-95, WIT2002-35, (2002)

[10] A. Martin, M. Przybocki, G. Doddington, and D. Reynolds.: The NIST speaker recognition evaluation - overview, methodology, system, results, perspectives. Speech Communication, Vol. 31, (2000) 225–254

[11] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg.: Using Prosodic and Lexical Information for Speaker Identification. Proc. ICASSP, Vol. 1, (2002) 141–144

[12] D.A. Reynolds.: Speaker Identification and Verification using Gaussian Mixuture Speaker Models. Speech Communication , V. 17, (1995) 177–192

[13] H. Sakoe and S. Chiba.: A Dynamic Programming Algorithm Optimization for Spoken Word Recognition. IEEE Trans. on ASSP, Vol. 26, No. 27, (1978) 43–49

[14] M. Schmidt and H. Gish.: Speaker Identification via Support Vector Machines. Proc. ICASSP, (1996) 105–108