Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

Acoustic Features Affecting Speaker Identification by Imitated Voice Analysis

Mari TANAKA*, Hideki KAWAHARA** and Shigeo MORISHIMA*

*Department of Applied Physics, School of Advanced Science and Engineering,

Waseda University, Japan

**Design Information Sciences Department, Faculty of Systems Engineering,

Wakayama University, Japan

ABSTRACT

In this paper, physical correlates of perceived personal identity are investigated using imitated 16 utterances spoken by 11 mimicry speakers and 24 test subjects. Our unique strategy to use non-professional impersonators enabled to prepare test utterances with wide range of perceived similarities. Reasonably high correlations (0.46 and 0.44) in multiple regression analysis were attained by grouping subjects into three groups based on cluster analysis of the subjective test results. Without clustering, the correlation was only 0.17. Cluster analysis also revealed differences in their focusing physical correlates between three groups indicating importance of individual differences both in speakers and listeners.

1. INTRODUCTION

Personal characteristics in voice quality are focused recently, because of its efficiency to improve not only speaker recognition performance but also speech synthesis quality.

Many researchers tried to reveal relationship between voice individuality and acoustic features. Most of them used a synthetic speech as a stimulus in a subjective experiment to create a very similar impression by an interpolation of two speakers' voices using morphing technique.

Kitamura[1] modified several acoustic properties of sustained vowel /a/ uttered by 10 male speakers by morphing technique and investigated those effects on perception of closeness of speaker characteristics. An interval scale for sound quality of the stimuli was also measured in order to confirm whether the degradation of sound quality affects the results and revealed a strong positive correlation between interval scales of closeness of speaker characteristics and sound quality of the stimuli implying that sound quality might affect to the experimental results.

In a few of studies that used mimicry voices, there is not problem of sound quality. According to Laver [2], mimicry is a stereotyping process and that does not involve exactly copying the target speaker. So a few previous study explore the acoustical characteristics that a professional impersonator changes from his natural voice to imitated target voice. To get close to the target voice and to succeed with the voice imitation, the impersonator needs to change his voice and speech behavior in a number of ways. A process of imitation is very useful, especially in case of professional impersonator.

In a case study of imitated voice, Eriksson & Wretling [3] found that duration was perfectly parallel between the voice imitation and the natural rendition at word level as well as at segment level. Zetterholm [4]-[6] has carried out auditory and acoustic analysis of imitated utterances and demonstrated that a professional impersonator captures the speech style, dialect, pronunciation, and intonation.

Kitamura [7] revealed importance of the mean and dynamics of pitch frequency for imitating. However these studies are performed by the stimuli of professional impersonator's voice, so it is difficult to capture enough number of stimuli because the number of professional impersonator is limited.

In this study, we recorded imitated voices that 11 non professional mimicry speakers imitated 16 target voices and compared some imitated voices with the target voices by Dynamic Time Warping Distance and perceptual similarity score. And we analyze these similarity standards to reveal which acoustic features are very important for perception of personal characteristics.

2. ANALYSIS PROCEDURE

Physical correlates of perceptual similarity were investigated by the following three-step procedure. Firstly, multiple linear regression analysis was conducted using acoustic parameters as independent variables and the perceived similarity as the dependent variable. Then, secondly, correlation coefficients were investigated between these acoustic parameters and the perceived similarity. Thirdly, cluster analysis was conducted to group subjects into groups. Finally, multiple linear regression analysis and correlation analysis were conducted for each group.

Speech data

We used 16 target voices including two sentences spoken by 8 Mimicry speakers.

Sentence 1:

"Sasuga Tensai Programmer" by an emotional voice

Sentence 2:

"Ucyuukoukakaishi Sen-SanjyuuRokunichi" by a monotonic voice

Non-professional 11 impersonators were asked to mimic 16 target voices. They were instructed to mimic immediately after listening to each target voice sample. This procedure yielded 176 utterances sampled at 16 kHz with 16-bit resolution.

Perceptual Similarity

We examine the perceptual similarity between target voice and imitated voice spoken by mimicry speakers. It is difficult for us to represent the perceptual similarity in numeric value because the standard and the scaling of similarity have individual difference.

We represent perceptual similarity in numeric value using Mean Opinion Score (MOS). We investigated the similarity by the score of five-grade evaluation and ratios of five scores, which are shown in Table 1, by 24 subjects. We set the score closest to target person to 0.

Table 1. The score of five-grade evaluation and ratios

Score	evaluation	raito
0	The same speaker	1.01%
1	Quite similar	8.24%
2	Similar	21.22%
3	Rather not similar	31.77%
4	Not similar	37.76%

Acoustic similarity

We recognized the Dynamic Time Warping (DTW) distance as a measure of the acoustic features similarity. Sakoe et al. [8] 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. Adachi et al. [9] compared perceptual similarity with similarity of acoustic features by distance of DTW and GMM (Gaussian Mixture Model), and showed that the result of DTW at the experiment has evaluated the perceptual similarity effectively more than that of GMM.

In this study, we used DTW distance with 12 acoustic parameters between target voice and imitated voice to represent speaker similarity.

- MFCC

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

- STRAIGHT Cepstrums

Kitamura [1] described that the perception of personality is influenced by the high order STRAIGHT[10] Cepstrum and the first STRAIGHT Cepstrum which represent the detailed spectral shape and that gradient respectively.

Cepstrum is extracted by STRAIGHT analysis, and 35th and higher cepstral coefficients are defined as the high order STRAIGHT Cepstrum (CepH). The first STRAIGHT Cepstrum (Cep1) is the first coefficient of the calculated STRAIGHT Cepstrum.

- Spectrum

Higher frequency region of spectrum has also the strong relation with personality. Furui et al. [11][12] demonstrated the strong relation between a high frequency band width of spectrum and personality.

Therefore, we investigate the relation between log spectrum in a higher frequency region and perceptional similarity. In this paper, the region boundary was set to 2.6 kHz.

- STRAIGHT-Ap

Saito et al. [13] discovered the information of personality in STRAIGHT-Ap (Ap) under 2 kHz. They indicated such information in characteristic of vocal sound source. Therefore, we focus on the relation between STRAIGHT-Ap and perceptional similarity.

- Fundamental Frequency

We investigated the relation between fundamental frequency (f0) and perceptual similarity because Hashimoto et al. [14] have proved that the fundamental frequency has an effect on the personality perception. We extracted the fundamental frequency every 10 ms using STRAIGHTTEMPO which is a part of STRAIGHT analysis.

- Formants, SpectrumSlope

Voice quality is a critical acoustic feature to assess the similarity of speech. Kido et al. [15] described that formants (F1-F4) 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.

- Utterance Speed

As for the relation between utterance speed and personality perception, Francis et al proved that the utterance speed is changed depending on the personality perception [16]. Therefore we research the relation between utterance speed and perceptual similarity. In our research, the utterance speed is the average duration of a mora. A mora consists of one or zero consonants and a vowel, and is a phonetic unit similar to a syllable. Many resarcher of japanese speech used 'mora'[17].

3. EXPERIMENTAL RESULT

The multiple linear regression analysis

The result of the multiple linear regression analysis was that the standardised partial regression coefficient by 24 subjects' perceptual scores and DTW distance of 12 acoustic features by all sentences was 0.12, by sentence1 was 0.17, by sentence2 was 0.12 and showed that relationship between all subject's perceptual similarity and 13 acoustic features similarity by DTW is infirm.

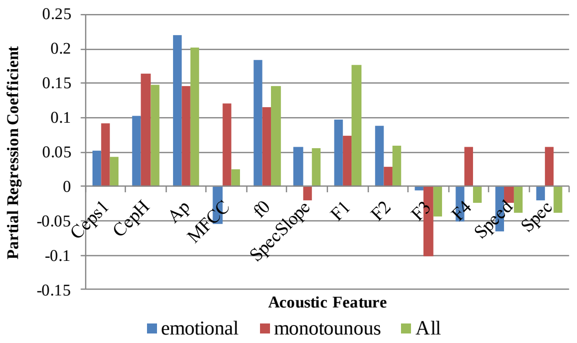


Figure 1. Coefficient of regression for each sentence

One of the causes of this result is person depended equation of perceptual similarity. The standardised partial regression coefficient by these regression equations were shown in Figure 1. This result shows that difference of sentence affects the coefficient of regression, and similarity of emotional voices was determined by pitch, Ap F1, F2, SS, and CepH, on the other hand, similarity of monotonous voices was determined by cep1, cepH, Ap, mfcc,pitch,f1,f4.

Especially cepsH, pitch, and Ap show positive correlation with perceptual similarity irrespective of sentence. On the other hand, mfcc, specSlope, F4 and spec are affected by difference of senetence.

Cluster analysis of subjectsIt

To examine personal equation of perceptual similarity and tendency first, we calculate the correlation coefficient by each 12 acoustic features and scores of perceptual similarity by each of subjects, second, cluster analysis is performed based on correlation coefficients of each subject. The tree diagram by cluster analysis of subjects is shown in Figure 2.

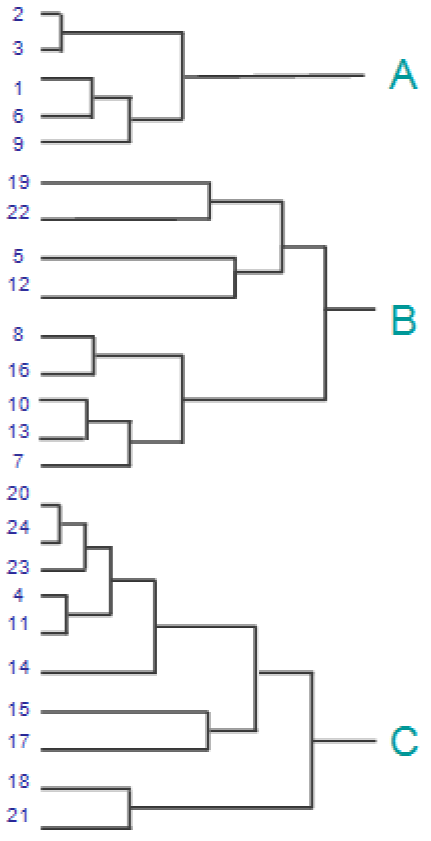
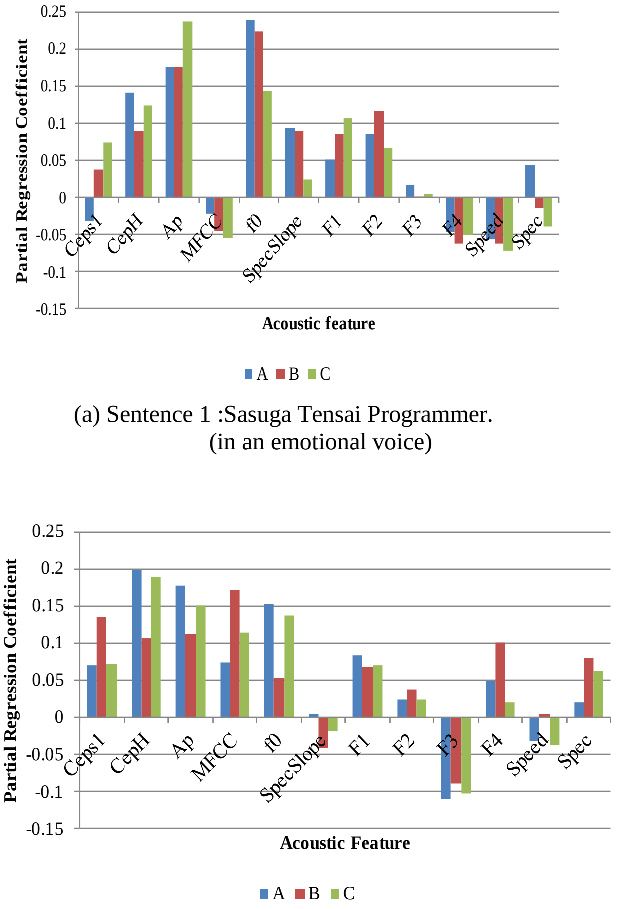


Figure 2. Tree diagram by cluster analysis of 24 subject



(a) Sentence 1 :Sasuga Tensai Programmer.
(in an emotional voice)

(b) Sentence 2 : Ucy uukoukakaishi Sen-SanjyuuRokunichi.
(in a monotonous voice)

Figure 3. Coefficient of regression for each cluster

As a result, we found three big clusters of subjects, group A, B and C, to examine difference of important acoustic features of each cluster. Because Figure 1 shows that difference of sentence affects the coefficient of regression, we calculated the coefficient of regression by sentences. The coefficients of regression by sentence 1 and sentence2 are shown in Figure 3 (a) and (b).

Table 2. Multiple correlation coefficient of quadratic equation

group	Sentence1	Sentence2
A	0.46	0.44
B	0.27	0.24
C	0.24	0.22
All	0.17	0.12

The coefficient of multiple correlation of the quadratic equation by subject’s perceptual scores and DTW distance of 12 acoustic features for every cluster is shown in Table 2.

A parallel between (a) and (b) shows that difference of clusters affects the difference of the coefficient of regression. Especially, Ceps1 and spec showed difference by every cluster.

Because the coefficient of multiple correlations calculated for every cluster independently is higher than that of all subjects, personal characteristics in perceptual similarity affect the difference of regression coefficients.

4. DISCUSSION

Of course, in this study, imitated voices were lopsided and dissimilar to target voices than voices by professional impersonator and scores of perceptual similarity leaned to 4 (Not Similar), because 11 mimicry speakers were not professional in voice imitation.

However, we succeeded to get much variety of imitated voice samples depending on the skill of mimicry speakers. So some of them are very similar to original and others are not. As a result, we got samples with a variety of similarity degrees. If we use professional, all the samples are perfectly similar and it's not suitable for our research target.

As a result, we note that uttered sentence and subjects affect relationship between acoustic feature and perceptual similarity. When utterance is emotional, pitch and CepH show heavy correlation with perceptual similarity by all subjects, but importance of Cep1 and spec differs depending on the subject. On the other hand when utterance is monotonous, personal equation of perceptual similarity is smaller than emotional, and MFCC indicates correlation too.

In a perceptual similarity measurement, we have to consider the influence of the spoken sentence and personal characteristics of subjects with focusing features and personal preference.

5. CONCLUSION

In this study, we analyzed a variety of imitated voice by several speakers and compared with subjective similarity scores to examine acoustic features concerning with perceptual similarity measurement. As a result, we found the difference depending on the personal characteristics of subject and succeeded to categorize these personal features into a few clusters to achieve higher correlation between acoustic feature vector distance and perceptual similarity scores.

By analysis for each cluster, there are very strong correlations between perceptual similarity and acoustic feature vectors in each group. As a future subject, we have to define speaker similarity measurement considering the subject focusing feature, preference and relation between target speaker and imitation speaker.

REFERENCES

1. T. Kitamura and T. Saitou., "Contribution of acoustic features of sustained vowels on perception of speaker characteristic." *Acoustical Society of Japan 2007 Spring Meeting*, 443-444(2007), [in Japanese]
2. Laver, J., "Principles of phonetics", *Cambridge: Cambridge University Press*, (1994)
3. Eriksson, A, Wretling, P., "How flexible is the human voice? – A case study of mimicry. " *In Proc. Eurospeech '97*, Vol.2. Rhodes: 1043-1046,(1997)
4. Zetterholm, E., "Impersonation: reproduction of speech," *Work-ing Papers, Dept. of Linguistics, Lnud University*, 49, 176-179(2001)
5. Zetterholm, E., "Same speaker: different voices: A study of one impersonator and some of his different imitations," *Proc. Int. Conf. Speech Sci. & Tech*., 70-75(2006)
6. Zetterholm, E., "A comparative survey of phonetic features of two impersonators," *TMH-QPSR*, 44, 129-132(2002)
7. T. Kitamura., "Acoustic Analysis of Imitated Voice Produced by a Professional Impersonator," *2008 ISCA*, 813-816(2008)
8. H. Sakoe and S. Chiba.: A Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Trans. on ASSP*, Vol. 26, No. 27, 43-49(1978)
9. Y. Adachi, S. Kawamoto, S. Morishima, and S. Nakamura., "Perceptual Similarity Measurement of Speech by Combination of Acoustic Features." *ICASSP2008*, 4861-4864(2008)
10. 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*, 343-354(2001)
11. 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*, 737-738(2003) [in Japanese].
12. S. Furui and M. Akagi., "Perception of voice individuality and acoustic correlates." *Journal of the Acoustical Society of Japan*, vol. J66-A, 311-318(1985)
13. T. Saitou and T. Kitamura., "Factors in /VVV/ concatenated vowels affecting perception of speaker individuality." *Acoustical Society of Japan 2007 Spring Meeting*, 441-442(2007) [in Japanese].
14. N. Higuchi and M. Hashimoto., "Analysis of acoustic features affecting speaker identification." *Proc. of EUROSPEECH'95*, 435-438(1995)
15. H. Kido and H. Kasuya. "Voice quality expressions of speech utterance and their acoustic correlates." *Technical report of IEICE*, SP2002-95, WIT2002-35(2002)
16. A.L. Francis and H.C. Nusbaum. "Paying attention to speaking rate." *Proc. of ICSLP 96*(1996)
17. R.Tachibana, T,Nagano, G.Kurita, M.Nishimura, and N.Babaguchi, " Automatic Labeling Using Multiple Models for Japanese . " *IEICETRANS.INF.&SYST.,VOL.E90D,NP.11 NOVEMBER 2007*,1806-1812(2007)