WESPAC IX 2006
The 9th Western Pacific Acoustics Conference
Seoul, Korea, June 26-28, 2006

SPEAKER SIMILARITIES IN HUMAN PERCEPTION AND THEIR SPECTRAL PROPERTIES

Kanae AMINO[1)], Tsutomu SUGAWARA[2)], and Takayuki ARAI[1)]

[1)] Department of Electrical and Electronics Engineering, [2)] Graduate Division of Foreign Studies
Sophia University, Tokyo, Japan
E-mail: {amino-k, sugawara, arai}@sophia.ac.jp

ABSTRACT

The acoustic properties of voiced sonorants are said to be speaker-dependent, and it is also reported that these sounds are effective for identifying the speakers. In our previous experiments, we found that the nasal sounds are highly effective for speaker identification by listening, and that the inter-speaker distances in the spectra were also greater in nasals than in oral sounds. This present study further analyses the spectral properties of nasals and orals in terms of perceptual and acoustical voice similarity and shows that nasal sounds may have longer interval that listeners can exploit for speaker identification.

KEYWORDS: Speaker individuality, Speech similarity, Cepstral distances, Nasals

INTRODUCTION

Human beings have the ability to reliably identify the speakers by speech sounds alone [1]. This is because speech sounds contain some speaker information as well as the linguistic information. Although they are treated as peripheral in linguistic research, speaker characteristics contained in speech sounds are important when it comes to automatic speaker recognition techniques or to speaker identification in forensic cases.

It is said that the perception of speaker identity interacts with the perception of linguistic information [2]. Thus the listeners use the linguistic contents of utterances in order to identify the speakers. Also, speaker information is important for listeners in order to obtain the framework for discriminating the phonemes or to gauge communicative settings.

Previous studies [3-5] pointed out that there are differences among speech sounds in the effectiveness for perceptual speaker identification, i.e., the accuracy of the identification varies according to the speech contents presented to the subjects. Some studies report that vowels [6, 7] and voiced consonants [3, 6, 8-10] are effective. In our previous experiments, we found that nasal sounds are more effective for speaker identification than oral sounds [8-10], and that coronal sounds are more effective than the sounds articulated at other places [10].

According to a study where various synthesised sounds were tested, the most effective acoustic parameter for perceptual speaker identification was the spectral information, and fundamental frequency and temporal structures followed it [11]. Also in our previous study, we found that the perception of the speaker identity corresponded to the spectral distances of the stimuli [9], i.e., inter-speaker distances were greater in nasal sounds and smaller in oral sounds.

In this present study, we further investigate the perceptual and spectral properties of the speech sounds in terms of voice similarity in order to explain the differential effects among the speech sounds in speaker identification.

METHODS

Speech materials. The materials used in this study were the following six monosyllables recorded in a previous experiment [9]: /da, ta, ma, na, sa, za/. These monosyllables were uttered by ten male speakers in carrier sentences, and excerpted manually from these sentences. Five tokens for each monosyllable and for each speaker were used for the analyses, thus we had fifty samples for each consonant type. All the materials were recorded onto DAT (digital audiotape) at the sampling frequency of 48 kHz with 16 bit resolution, and then down-sampled to 16 kHz for the analyses.

Analysis Frames. As was shown in the previous study [10], syllable onset is one of the most important sections for perceptual speaker identification. According to this report, we analysed three different intervals excerpted from the onset part of the monosyllables. The types of the frames and the criteria for the excerption are shown in Table 1, and Fig. 1 shows an example of the excerption. Each frame had 30 ms length.

Cepstral distances. Cepstral coefficients were calculated up to the 30th order, and the inter- and intra-speaker cepstral distances were computed for all the possible speaker-pairs, separately for each consonant type and for each frame type. Thus we obtained eighteen (six monosyllables and three frame types) square matrices of 50×50 (ten speakers and five tokens). Then the average values of intra- and inter-speaker distances were obtained and confusion matrices for cepstral distances of the ten speakers were drawn using these values.

Perceptual speech similarity. The results of the perceptual speaker identification test in a previous experiment [9] were used in this study. Five students who know all of the speakers very well served as the subjects in the experiment. The speakers and the subjects had known each other for at least four years. The percentages of the correct speaker identification for the stimuli in question are shown in Table 2. The nasals ranked above, and fricatives and oral stops followed them. In order to give an index for perceptual voice similarity, confusion matrices were generated as for the ten speakers for each consonant.

Table 1. Types of analysis frames and criteria for excerption

ID	Excerpted section	Criterion for excerption
Frame 1	Stable consonant part	Interval that covers only the true consonant part*
Frame 2	Consonant part including transition	Interval before the second formant of the following vowel gets stable
Frame 3	Vowel part including transition	Interval from the first pulse of the vowel

* The analysis targets for Frame 1 were four types of consonants that have stable consonant parts: /m/, /n/, /s/ and /z/.

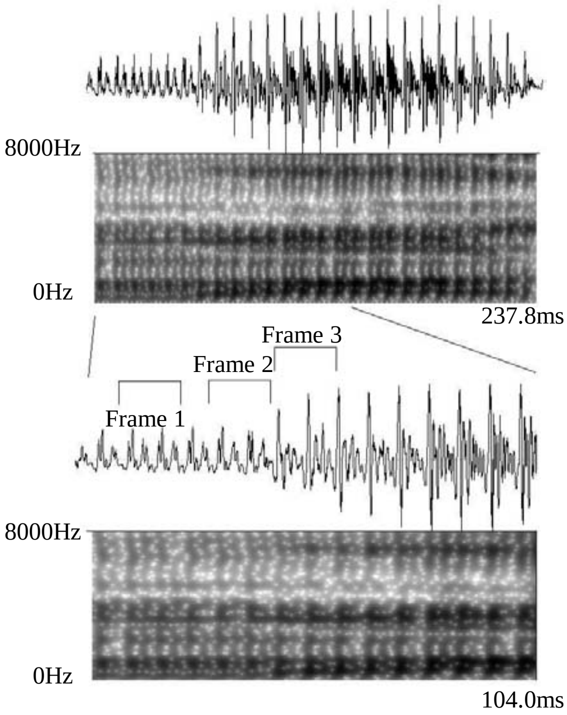

Fig. 1. Example of excerption

Table 2. Results of the perception test: percent correct for each stimulus type
The total number of each stimulus type was 250

Stimulus	Percent correct (%)
/na/	86.0
/ma/, /za/	80.8
/sa/	78.8
/da/	78.0
/ta/	73.6

RESULTS

Relationship between perception and cepstral distances. In order to investigate the relationship between perceptual voice similarity and cepstral distances, we calculated the correlation coefficients between the confusion matrices of the ten speakers in human perception and in cepstral analyses. The results are shown in Table 3. The number of the elements in a matrix was a hundred.

Table 3. Correlation coefficients between confusion matrices of perceptual speaker identification and cepstral distance matrices

Stimulus	Frame 1	Frame 2	Frame 3
/da/	N/A	-0.309	-0.635
/ta/	N/A	-0.342	-0.595
/ma/	-0.812	-0.793	-0.748
/na/	-0.788	-0.765	-0.624
/za/	-0.334	-0.334	-0.635
/sa/	-0.375	-0.375	-0.663

DISCUSSION

As seen in Table 3, the correlation coefficients of the nasals are consistently high through the three frames. On the contrary, those of the oral stops and the fricatives were high only in the third frame, in the vowel part including the transition from the preceding consonant. This leads to the following two implications:

1. In nasals, listeners use all three frames as the cue of the speaker individuality.
2. In oral stops and fricatives, listeners use only the vowel part for identifying the speakers.

In other words, the nasal sounds contain speaker information in longer spans than the oral sounds. This may explain the effectiveness of the nasals in speaker identification.

The shapes of the speech organs that are involved in the nasal resonance are said to be speaker-dependent [12], and also the shapes of these resonators cannot be changed voluntarily. This is why the acoustical properties of the nasal sounds indicate speakers' physiological characteristics. Furthermore, the fourth formant frequency and the spectral dip around 4 to 6 kHz are said to reflect the speakers' hypopharyngeal properties [13]. The monosyllables /ma/ and /na/ contain both these properties, and this also accounts for the advantage of nasals in the perception test.

SUMMARY

In order to explain the effectiveness of nasals in perceptual speaker identification, we inspected the relationship between confusion matrices of the perception test and of the cepstral distances. We observed the correlation coefficients as to three analysis frames, the stable consonant part, the consonant part including the transition to the following vowel and the vowel part including the transition from the preceding consonant. It was found that the correlation is high all through the frames in nasals, but only in the vowel part in oral sounds. These results show that the sections used by listeners for identifying the speakers may be sound-specific, and the nasals have longer intervals that indicate speaker individuality than the oral sounds do.

One of the final goals of the research is to understand the mechanisms of human cognition, and there will be many steps before we reach there. Our next task will be to test syllables with different vowels in order to see the effects of co-articulation. The analyses of the stable vowel part were also not included in this study. Moreover, inspection of the spectra is necessary in order to estimate the frequency range that reflects speaker characteristics.

ACKNOWLEDGEMENTS

This study was supported by MEXT Grant-in-Aid for Scientific Research (A) 16203041, and by Grant-in-Aid for JSPS Fellows 17-6901.

REFERENCES

1. F. Nolan, *The Phonetic Basis of Speaker Recognition* (Cambridge University Press, Cambridge, 1983)

2. L. Nygaard, “Perceptual integration of linguistic and nonlinguistic properties of speech,” Chap. 16 in D. Pisoni and R. Remez (ed.), *The Handbook of Speech Perception* (Blackwell Publishing, Oxford, 2005)

3. T. Nishio, “Can we recognise people by their voices?,” *Gengo-Seikatsu*, **158**, 36-42 (1964)

4. P. Bricker and S. Pruzansky, “Speaker Recognition,” Chapt.9 in N. Lass (ed.), *Experimental Phonetics* (Academic Press, London, 1976)

5. D. O’Shaughnessy, *Speech Communication –Human and Machine–*, 2nd ed. (Addison-Wesley Publishing Company, New York, 2000)

6. G. Ramishvili, “Automatic Voice Recognition,” *Engineering Cybernetics*, **5**, 84-90 (1966)

7. K. Stevens, C. Williams, J. Carbonell, and B. Woods, “Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material,” *J. Acoust. Soc. Am.*, **44**, No.6, 1596-1607 (1968)

8. K. Amino, “Properties of the Japanese phonemes in aural speaker identification,” *IEICE Tech. Rep.*, **104**, No.149, 49-54 (2004)

9. K. Amino, T. Sugawara, and T. Arai, “Correspondences between the perception of the speaker individualities contained in speech sounds and their acoustic properties,” *Proc. of Interspeech*, 2025-2028 (2005)

10. K. Amino, T. Sugawara, and T. Arai, "Effects of the syllable structure on perceptual speaker identification," *IEICE Tech. Rep.*, **105**, No.685, 109-114 (2006)

11. S. Furui, "Key issues in voice individuality," *J. Acoust. Soc. Jpn.*, **51**, No.11, 876-881 (1995)

12. J. Dang and K. Honda, "Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation," *J. Acoust. Soc. Am.*, **100**, No.5, 3375-3383 (1996)

13. T. Kitamura, K. Honda and H. Takemoto, "Individual variation of the hypopharyngeal cavities and its acoustic effects," *Acoust. Sci. Tech.*, **26**, No.1, 16-26 (2005)