Acoust. Sci. & Tech. **27**, 4 (2006)

ACOUSTICAL LETTER

Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties

Kanae Amino[1,*], Tsutomu Sugawara[2] and Takayuki Arai[1]

[1]*Department of Electrical and Electronics Engineering, Sophia University, 7–1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan*
[2]*Department of Foreign Studies, Sophia University, 7–1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan*

(*Received 2 November 2005, Accepted for publication 26 December 2005*)

Keywords: Nasals, Speaker identification, Individuality, Cepstral distance
PACS number: 43.71.Bp [doi:10.1250/ast.27.233]

1. Introduction

Speech sounds convey not only linguistic or phonological information, but also nonlinguistic information, including the speakers' individualities [1]. It is known that the availability of the speech contents used for speaker identification differs depending on the types of sounds they contain, and it is reported that voiced sonorants, such as vowels and nasals, are most effective for speaker identification by both humans [2–4] and machines [5].

The speaker's individuality contained in speech sounds should have some acoustic correlations and their properties can be measured as acoustic parameters [6]. In this study, we conducted a human speaker identification test, and investigated the differences in the effectiveness of using various Japanese sounds in identifying the speakers. We also analysed the stimuli used in the experiment in order to explain these differences in terms of acoustical distances.

2. Experiment

This experiment basically followed the methodology used in the previous study [4]. One of the differences was that more speakers contributed to the study (10 compared with 3), and another was that fewer types of stimuli were used in the test (9 compared with 15). The reason for increasing the number of the speakers is that the speaker ensemble must be large enough to obtain reliable data [4,7]. At the same time, the increase in the number of the speakers leads to an increase in the test time, but this may make the subjects lose their concentration on the test. This is why we used only 9 types of stimuli this time.

Ten male speakers and 5 male subjects participated in the experiment. They were all native speakers of Japanese and had normal hearing. They all lived in the same dormitory for more than 4 years, and they had daily contact with each other.

The recording sessions were conducted in a soundproof room, using a digital audiotape (DAT) recorder (SONY TCD-D8). The data were sampled at 48 kHz with 16-bit resolution. Nine Japanese consonants were recorded in the carrier sentences that say " 'aCaCaCa' too o shiji shimasu (['aCaCaCa' toː o ʃiʒi ʃimasɯ̥], I support the 'aCaCaCa' party)." In the word 'aCaCaCa,' which is the name of a fictional political party, 'a' stands for the Japanese vowel /a/, and 'C' stands for one of the following consonants: /t/ /d/ /s/ /z/ /ɾ/ /j/ /m/ /n/ and /ɲ/. The reason for using only one vowel, /a/, is to make the experiment simple, and the reason for using the names of political parties is that the suffix '-to (-party)' forms compound words that are uttered with a relatively stable pitch pattern after the third morae [8].

After the recording, the fourth morae of 'aCaCaCa' were excerpted from the recorded sentences in order to generate the test stimuli. The excerption was performed manually on the basis of the waveforms. The speakers repeated each type of sentence 10 times, and 5 of the sentences that were uttered clearly were selected and used in the perception test. The total number of stimuli was 450, i.e., corresponding to 10 speakers, 9 consonants and 5 tokens.

The perception tests were carried out in the same soundproof room as the recording sessions. The subjects were the 5 male undergraduate students mentioned above, and they were informed of the names of the 10 speakers beforehand. The stimuli were presented in random order, and the subjects were told to write the name of the speaker on the answer sheets for each stimulus.

The number of judgments for each consonant was 250, i.e., corresponding to 10 speakers, 5 tokens, and 5 subjects. The results of the tests are summarised for each consonant and are shown in Table 1. They are listed in the order of their ranking based on scores. As in the results of the previous experiments [4], the nasals ranked above the oral sounds. No remarkable tendencies are observed among the oral sounds, except that the voiced sounds ranked higher than the voiceless counterparts in the pairs of /ta/–/da/ and /sa/–/za/. These results agreed with those in the previous studies.

In the statistical analyses, the pairs such as stops-fricatives, plosives-fricatives, and voiceless-voiced sounds were compared by a *t*-test, but none of them were significantly different. In the case of the nasal-oral pair, we could see a tendency that the nasals gained higher scores than the orals ($p = 0.057$), although the difference was not significant.

3. Acoustic analysis

In order to explain the differences among the stimuli in the perception test acoustically, we evaluated them in terms of the spectral distances. As seen in Table 1, all types of stimuli

*e-mail: amino-k@sophia.ac.jp

Table 1 Speaker identification results for each monosyllable.

Stimuli	No. of correct answers (/250)	Percent correct
/na/	215	86.0
/ɲa/	214	85.6
/ma/ /za/	202	80.8
/sa/	197	78.8
/ja/	196	78.4
/da/	195	78.0
/ɾa/	186	74.4
/ta/	184	73.6

Table 2 Criteria for excerption of consonant parts from stimuli.

Oral Stops	Interval to release phase, before gliding part to following vowel
Fricatives	Stable frication parts before gliding part to following vowel
Nasals	Interval of nasal murmur, before gliding part to following vowel

gained scores above the chance level in the perception test; however, there were considerable, although not significant, differences among them. The stimuli used in this study all have the same CV-structure, where V is restricted to be the Japanese vowel /a/. This implies that these differences among the stimuli are derived from the differences in the acoustical properties of the consonants, so this time we decided to analyse the spectral properties of the consonant parts in order to find out how they differ in their contribution to the speakers' individualities.

We selected 6 types of consonants out of 9 stimulus types. The selected consonants are /t/, /d/, /s/, /z/, /m/ and /n/. The remaining three types, /ɾ/ /j/ and /ɲ/, are realised as momentary or gliding sounds in Japanese. The target of the analysis here was the spectra of the definable consonant parts; therefore, these three were omitted this time. The consonant parts of the target sounds were excerpted manually for 30 ms from the stimulus syllables. The criteria for the excerption are shown in Table 2. A computer program called Praat [9] was used for both editing and analysis.

After the excerption, the cepstral distances of pairs of tokens for each consonant type were computed, in a round robin manner for a 50 by 50 square matrix, i.e., corresponding to 10 speakers and 5 samples for each consonant type. Then the average values of the intra- and inter-speaker distances, and the ratios of the interspeaker distances to the intraspeaker distances of all speakers were calculated for each consonant. The calculation is based on the concept of the F-ratio [1], which we usually use as a baseline when looking for an effective acoustic parameter that indicates the speakers' individualities, and larger ratios reflect higher degrees of speaker uniqueness.

The values of the inter- and intra-speaker distances for each consonant type and the average values of the distance

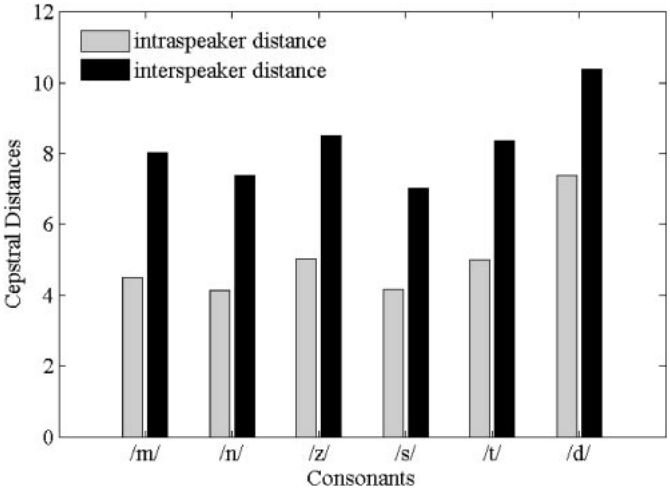

Fig. 1 Values of inter- and intra-speaker distances for each consonant.

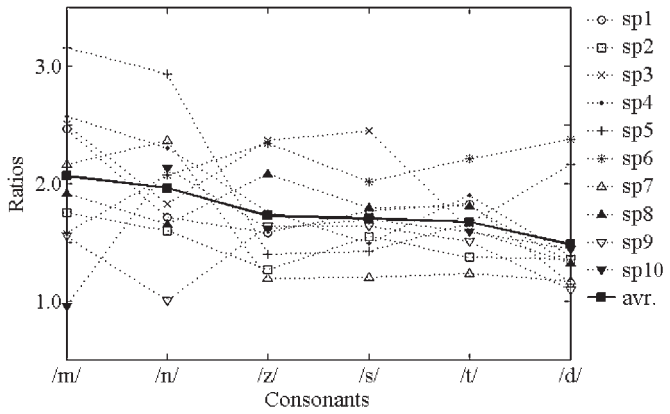

Fig. 2 Average values of inter- and intra-speaker distance ratios.

ratios for each consonant are shown in Figs. 1 and 2, respectively. As seen in Fig. 2, the interspeaker distances and the ratios of the inter- and the intra-speaker distances are largest in the nasal consonants, /m/ and /n/, and smallest in the stops, /t/ and /d/. These differences were not significant in the statistical analyses ($p = 0.075$), but the oral-nasal pair showed a significant difference in the t-test ($p = 0.0038$).

4. Discussion

The score-based rankings of the consonants in the results of the identification tests are shown in Table 3, together with those of the spectral analyses. Although there were no statistically significant correlations, the rankings correspond to each other in that they were in the order of the nasals, the fricatives and the oral stops.

5. Summary

In this study, differences among Japanese consonants in speaker identification by listening were investigated and these differences were explained by spectral analyses. Ten speakers were identified by 5 subjects who were familiar with them in the perception tests, and the identification rates were highest when nasal sounds were presented as the stimuli. Then the

Table 3 Rankings of consonants in results of identification test and of spectral analyses.

Ranking	Identification rate (Section 2)	Distance ratio (Section 3)
1	/n/	/m/
2	/m/, /z/	/n/
3		/z/
4	/s/	/s/
5	/d/	/t/
6	/t/	/d/

cepstral distances of 6 consonants of the stimuli were calculated, and it was found that interspeaker distances were largest in the nasals. These results suggest that the speakers' individualities are reflected more in the spectra of the nasal sounds than in those of the oral sounds and that the listeners perceive these individualities when they identify the speakers.

One reason that the properties of the nasal sounds are more speaker-dependent is that the shapes of the resonators involved in the articulations of these sounds differ considerably among individuals. In addition, the shapes of these resonators cannot be changed at will. This means that the properties of the nasal sounds rarely change. As pointed out in another study [10], the next task will be to inspect the morphology of the nasal cavity.

Another future task will be to devise ways to put these individualities into practical use. As to automatic speaker recognition, it is reported that the recognition rate was improved by considering the individualities in the oro-nasal coupling and by using the weighted linear scale spectral properties [11]. Also, we recommend the use of utterances containing nasals rather than other consonants for the purpose of speaker verification. However, the properties of the nasal sounds are thought to be affected greatly by head colds and other supralaryngeal diseases, and these factors must be studied as well in the future.

Acknowledgement

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

References

[1] Y. Niimi, *Speech Recognition* (Kyoritsu Shuppan, Tokyo, 1979), pp. 206–237.
[2] T. Matsui, I. Pollack and S. Furui, "Perception of voice individuality using syllables in continuous speech," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 379–380 (1993).
[3] M. R. Sambur, "Selection of acoustic features for speaker identification," *Proc. IEEE Trans. Acoust. Speech Signal Process.*, **23**, 176–182 (1975).
[4] K. Amino, "Properties of the Japanese phonemes in aural speaker identification," *Tech. Rep. IEICE*, **37**, 49–54 (2004).
[5] S. Nakagawa and T. Sakai, "Feature analysis of Japanese phonetic spectra and considerations on speech recognition and speaker identification," *J. Acoust. Soc. Jpn.* (*E*), **35**, 111–117 (1979).
[6] G. Fant, *Acoustic Theory of Speech Production* (Mouton, The Hague, 1960).
[7] P. D. Bricker and S. Pruzansky, "Speaker recognition," in *Contemporary Issues in Experimental Phonetics*, N. Lass, Ed. (Academic Press, New York, 1976), pp. 295–325.
[8] H. Kindaichi and K. Akiyama, Eds., *Meikai Japanese Accent Dictionary*, 2nd ed. (Sanseido, Tokyo, 1981).
[9] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*, Ver. 4.3.14 (Computer program), retrieved from http://www.praat.org/ (2005).
[10] T. Kitamura and K. Honda, "Unchanged parts in the vocal tract during the utterance of the vowels," *Proc. Gen. Meet. Phonet. Soc. Jpn.*, pp. 105–110 (2003).
[11] Y. Yoshihara, X. Lu and J. Dang, "Speaker identification using weighted linear scale spectral feature," *Proc. Spring Meet. Acoust. Soc. Jpn.*, pp. 15–16 (2005).