# The Correspondences between the Perception of the Speaker Individualities Contained in Speech Sounds and Their Acoustic Properties

*Kanae Amino*[†]*, Tsutomu Sugawara*[†]*, and Takayuki Arai*[‡]

[†] Faculty of Foreign Studies, [‡] Faculty of Science and Technology, Sophia University,
7-1 Kioi-cho, Chiyoda-ku, Tokyo, 102-8554, Japan
{amino-k, sugawara, arai}@sophia.ac.jp

## Abstract

This study investigates the correspondences between the differences among the phones in human speaker identification and their acoustic properties. In the speaker identification test, the Japanese CV syllables excerpted from the carrier sentences were used as the stimuli. As pointed out in the previous studies, the stimuli containing the nasal sounds were significantly effective for the identification of the speakers, compared to other stimuli containing only the oral sounds. In the acoustic analyses, we analysed the spectral properties of the stimuli in order to explain these differences in the perception test, and we found that the cepstral distances among the speakers were significantly larger in the nasal sounds than in the oral sounds. Also, there were correspondences between the rankings of the consonants in the identification test and in the cepstral distances.

## 1. Introduction

Human beings have the ability to reliably recognise speakers by speech sounds alone. This implies that the speech sounds convey not only linguistic or phonological information, but also non-linguistic information, including the speaker's individualities [1].

The individuality contained in speech sounds, which is the characteristic auditory colouring of a given speaker's voice [2], has some acoustic correlates and can be measured as acoustic parameters [3-4]. Finding these acoustic parameters enables us to improve the techniques such as automatic speech recognition and automatic speaker recognition by eliminating and by extracting those parameters, respectively [1]. Studying the acoustic correlates that reflect individualities can also contribute to linguistic phonetics. If the individual information can be more clearly outlined against the phonological information, then it follows that the phonological information can also be defined. A more specified definition of each phoneme in a language will lead to a more clean-cut model of the sound patterns of the language, and also the acoustic properties of each phoneme will be explicitly explained [5].

One way to detect the acoustic correlates of the speakers' individualities is to make an experiment where the differential effects of the phonological contents of the speech materials are examined, and to find out the phonemes by which the listeners identify the speakers most accurately [6]. It can be said that the acoustic parameters those phonemes have in common must reflect the speakers' individualities the best.

According to the previous studies that carried out this kind of experiment, it is reported that the identification rate is the highest when the voiced sonorants are given as the stimuli [7-8], especially when the nasal sounds are given [5, 9]. This tendency was confirmed both for the speech uttered in isolation and for the excerpted speech [5, 9]. As to the acoustic properties of the voiced sonorants, some studies have pointed out that the individualities are reflected in the spectral peaks of the vowels and the nasals [8-10].

In this study, we investigate the differences in the behaviour of the various Japanese sounds in a human speaker identification test, and attempted to explain the differences seen in the test by conducting spectral analyses.

## 2. Speaker Identification Test

### 2.1. Methods

*2.1.1. Selection of speakers and subjects*

In human speaker identification tests, the selection of speakers and subjects is the most difficult and important task. The number of both the speakers and the subjects must be large enough, otherwise the identification task becomes too easy or too difficult, or the reliability of the experiment may not be acceptable. Moreover, the subject group must be homogeneous. Especially when the listeners are familiar to the speakers, it is difficult to gain an equable familiarity among them [9, 11].

Taking these things into account, ten male speakers and five male subjects were selected. All of them were undergraduate students at Sophia University, and they all live in the same dormitory. Before the experiment, a questionnaire was given to the subjects, and it was affirmed that they knew all of the speakers very well and that they are in contact with them in daily life. Their native language was Japanese, and none of them had hearing impairment.

*2.1.2. Speech materials*

The speech materials of this study must represent a wide range of speech sounds of Japanese so as to judge what kinds of sounds indicate the individualities the most. At the same time, the perception tests should not be so long that the subjects get tired or lose their concentration on the tests. In this study, we selected 9 Japanese consonants articulated near the alveolar ridge, six oral sounds and three nasals, and we tested these consonants in monosyllables excerpted from the recorded sentences shown in Table 1.

As seen from Table 1, the initial words, the names of fictional political parties, whose structures are “aCaCaCa”, where “a” stands for /a/ and “C” stands for a Japanese consonant, are carried in the sentence “/… to: o ʃiʒi ʃimasɯ/ (I support the … party.).” These “aCaCaCa” are non-sense words in actual Japanese. The reason for using the names of the parties is that the suffix “-to (- party)” forms the compound words that do not have falls in accent in Japanese [12]. Therefore “aCaCaCa-to” is uttered with a relatively stable accent pattern after the third mora. Furthermore, the reason for using only one vowel /a/ is to make the experiment simple. Notably, the Japanese /a/ was the most effective vowel for speaker identification in previous experiments [5, 10, 13].

All the recordings were held in a soundproof room. The data were recorded using a microphone (SONY, ECM-MS957) and a DAT recorder (SONY, TCD-D8), and were saved at a sampling rate of 48kHz with 16-bit resolution.

In order to make the stimulus syllables, the fourth morae of the recorded sentences were excerpted. This task was executed by using the software Cool Edit Ver.96 (Syntrillium Software Corporation). The morae in question were excerpted based on the waveform, and they were cut out to be of its longest duration. The excerpted syllables were as follows: /ta/, /da/, /sa/, /za/, /ɾa/, /ja/, /ma/, /na/ and /ɲa/.

There were 5 tokens for each type of consonant, so the total number of the stimuli was 450 (10 speakers, 9 consonants, 5 tokens). These materials were randomly ordered, and edited with 3 seconds of silence between two samples. A ready signal and a 500ms white noise were inserted before each sample, in order to avoid unwanted mistakes and to degrade the auditory memory of the preceding stimulus [14], respectively. The stimuli were again recorded onto DAT at the same sampling rate and resolution as the recordings.

### 2.1.3. Test procedures

The identification tests were conducted in the same soundproof room as the recording sessions. The stimuli were presented on the DAT player (SONY, TCD-D8), and the subjects listened to them through binaural earphones (SONY, MDR-Z400) at a comfortable loudness level.

The subjects were informed of the names of the ten speakers beforehand, and they were told to write the name of the speaker on the answer sheets for each stimulus. They took breaks after every 150 trials, and the total test time was about 40 minutes.

Table 1: *Recorded sentences. Combinations of various consonants and the vowel /a/ were read in the carrier sentence.*

| /aCaCaCa/ | Carrier sentence |
|---|---|
| /a.ta.ta.ta/ | /to: o ʃiʒi ʃimasɯ/ |
| /a.da.da.da/ | |
| /a.sa.sa.sa/ | |
| /a.za.za.za/ | |
| /a.ɾa.ɾa.ɾa/ | |
| /a.ja.ja.ja/ | |
| /a.ma.ma.ma/ | |
| /a.na.na.na/ | |
| /a.ɲa.ɲa.ɲa/ | |

Table 2: *Identification results for each stimulus type. The number of the correct answers (centre column) and the percent correct (right column) are shown. The number of samples for each type (the denominator) is 250.*

| Stimulus Type | Number of Correct Answers | 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 |

## 2.2. Results

The results of the identification test are summarised for each consonant type and are shown in Table 2.

Just as the results in the previous experiments [5, 9], the nasals are the most effective sounds for the identification of the speakers, followed by the fricatives and the oral stops. Moreover, in the pairs of /ta/-/da/ and /sa/-/za/, the tendency was seen that the voiced sounds obtained higher scores than the voiceless counterparts. This tendency was also reported in the previous studies [5, 9, 13].

In the statistical analyses, the differences among the consonants were not significant in ANOVA. In *t*-test, the difference between the nasal and the oral sounds was significant ($p$ = 0.0044). There were no other pairs that differed significantly in *t*-test: for example, the pairs like oral stops-fricatives ($p$ = 0.25), obstruents-sonorants ($p$ = 0.15), and voiced-voiceless ($p$ = 0.36).

## 3. Spectral Analyses

### 3.1. Analysis Methods

The stimuli used in this study all have the same CV-structure where the V is controlled to be /a/. Then the differences in the results of the identification test should come from the consonant parts or the gliding parts to the following vowel. In order to find out the acoustic characteristics that contribute to the differences among the stimuli, we selected 6 stimulus types and analysed the spectral properties of the consonant parts. The selected consonants are /t/, /d/, /s/, /z/, /m/ and /n/.

First, the consonant parts of the target sounds were excerpted for 30ms from the stimulus syllables. The bases for the excerption are shown in Table 3. After the excerption, the cepstral distances were computed for each consonant, in the round robin manner of the 50 square matrices, 10 speakers and 5 samples for each consonant. The zero-th coefficient was excluded in the computation.

Then the average values of the intra- and the inter-speaker distances were counted. The average values for each consonant are shown in Figure 1. When evaluating an acoustic property that is thought to indicate the individuality, an effective one should have a small intra-speaker variation and at the same time a large inter-speaker variation. The ratios of the inter-speaker distances to the intra-speaker distances of all speakers were calculated for each consonant (shown in Figure 2). Larger ratio values reflect greater individualities.

### 3.2. Results of the Analyses

As seen from Figures 1 and 2, the inter-speaker distances and the ratios of the inter- and the intra-speaker distances are the largest in the nasal consonants, and then the fricatives and the oral stops follow them. In the statistical analyses, there were significant differences in the ratios among the consonants in ANOVA ($F$ (59, 5) = 19.42, $p < 0.0001$). In the results of the post-hoc test, the ratios of the two nasal consonants /m/ and /n/ were significantly larger than any of the other consonants. No other pairs or groups were significantly different.

## 4. Discussion

The rankings of the consonants in the results of the identification test and of the spectral analyses are shown in Table 4.

From Table 4, we can see the correspondences between the results of the identification test and those of the acoustic analyses. They correspond not only in that the nasals ranked above the orals, also in the rankings of the manners of articulation. The rankings were in the order of the nasals, the fricatives and the oral stops, in both results.

Table 3: *The bases for excerption of the consonant parts from the stimuli. The target consonants are /t/ & /d/ (oral stops), /s/ & /z/ (fricatives), and /m/ & /n/ (nasals), and the excerption frame was 30ms long.*

| | |
|---|---|
| Oral Stops | The interval just before the gliding part to the following vowel, including the release phase. |
| Fricatives | The stable frication parts before the gliding part to the following vowel. |
| Nasals | The interval of the nasal murmur, before the gliding part to the following vowel. |

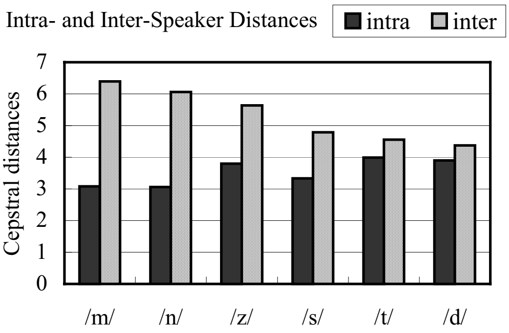

Figure 1: *The average values of the intra- and the inter-speaker cepstral distances for each consonant.*

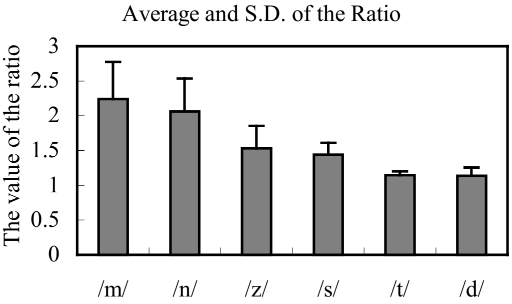

Figure 2: *The ratios of the inter-speaker cepstral distances to the intra-speaker distances. The average and the standard deviation for ten speakers are shown.*

Table 4: *The rankings of the consonants in the results of the identification test and of the spectral analyses. The ranking of the percent correct (centre column) and of the distance ratios (right column) are shown.*

| Ranking | Identification Rate | Distance-Ratios |
|---|---|---|
| 1 | /n/ | /m/ |
| 2 | /m/, /z/ | /n/ |
| 3 | | /z/ |
| 4 | /s/ | /s/ |
| 5 | /d/ | /t/ |
| 6 | /t/ | /d/ |

## 5. Summaries and Conclusion

In this study, the differences among the Japanese consonants in human speaker identification were investigated. Ten male speakers were identified by five familiar listeners, and the identification rate was the highest when the stimuli containing the nasal sounds were presented. In order to explain these differences, the spectral properties of six of the stimulus consonants were analysed. The cepstral distances of the ten speakers were calculated and it was found out that the inter-speaker distances were the largest in the nasals. These results suggest that the individualities of the speakers are reflected in the spectral information, especially in the spectra of the nasals; and the listeners perceive them when they recognise the speakers. The rankings of the consonants in the identification test and in the cepstral distances correspond not only in that the nasals had greater inter-speaker variations, but also in the orders of the manners of articulation, i.e. the nasals, the fricatives and the oral stops.

One reason that the properties of the nasal sounds are speaker-dependent is that the shapes of the resonators involved in the articulations of these sounds are considerably different for individuals [13]. Also the shapes of these resonators cannot be changed at will. This means that the properties of the nasal sounds also rarely change.

The next task will be to devise ways to put these individual characteristics to practical use. As to the automatic speaker recognition, it is reported that the recognition efficiency was improved by considering the individualities in the oro-nasal coupling and of the piriform fossa, and by using the weighted linear scale spectral features [16]. However, the acoustic properties of the nasal sounds are inevitably degraded by flu or other diseases in the supra-laryngeal part, and the study of the influences of these factors will be one of our future tasks.

## 6. Acknowledgement

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

## 7. References

[1] Niimi, Y., *Speech Recognition*, Sakai, T. (ed.), Kyoritsu Shuppan Publishing Company, Tokyo, 1979 (in Japanese).

[2] Laver, J., *The Phonetic Description of Voice Quality,* Cambridge University Press, Cambridge, 1980.

[3] Fant, G., *Acoustic Theory of Speech Production*, The Hague, Mouton, 1960.

[4] Kuwabara, H. and Sagisaka, Y., "Acoustic Characteristics of Speaker Individuality: Control and Conversion", *Speech Communications*, Vol.16: 165-173, 1995.

[5] Amino, K., "The Characteristics of the Japanese Phonemes in Speaker Identification", *Proc. of Sophia Univ. Linguistic Society*, Vol.18: 32-43, 2003 (in Japanese).

[6] O'Shaughnessy, D., *Speech Communications –Humans and Machine-*, 2nd ed., Addison-Wesley Publishing Company, New York, 2000.

[7] Matsui, T., Pollack, I., and Furui, S., "Perception of Voice Individuality Using Syllables in Continuous Speech", *Proc. of the 1993 Autumn Meeting of Acoust. Soc. Jpn.*: 379-380, 1993 (in Japanese).

[8] Sambur, M.R., "Selection of Acoustic Features for Speaker Identification", *Proc. IEEE Trans. ASSP.*, 23(2): 176-182, 1975.

[9] Amino, K., "Properties of the Japanese Phonemes in Aural Speaker Identification", *Technical Report of IEICE*, SP2004-37: 49-54, 2004 (in Japanese).

[10] Kitamura, T., and Akagi, M., "Speaker Individualities in Speech Spectral Envelopes", *J. Acoust. Soc. Jpn. (E)*, 16(5): 283-289, 1995.

[11] Bricker, P.D., and Pruzansky, S., "Speaker Recognition", in Contemporary Issues in Experimental Phonetics, Lass, N. (ed.), 295-325, Academic Press, New York, 1976.

[12] Kindaichi, H. and Akinaga, K. (ed.), *Meikai Nihongo Accent Jiten (Meikai Japanese Accent Dictionary)*, 2nd ed., Sanseido, Tokyo, 1981 (in Japanese).

[13] Nishio, T., "Can We Recognise People by Their Voices?", *Gengo-Seikatsu*, 158: 36-42, 1964 (in Japanese).

[14] Repp, B., Healy, A., and Crowder, R., "Categories and Context in the Perception of Isolated Steady-State Vowels", *J. of Exp. Psychol.: Human Perception and Performance*, Vol.5 (1): 129-145, 1979.

[15] Dang, J. and Honda, K., "Acoustic Characteristics of the Human Paranasal Sinuses Derived from Transmission Characteristic Measurement and Morphological Observation", *J. Acoust. Soc. Am.*, Vol.100 (5): 3374-3383, 1996.

[16] Yoshihara, Y., Lu, X., and Dang, J., "Speaker Identification Using Weighted Linear Scale Spectral Feature", *Proc. of the 2005 Spring Meeting of Acoust. Soc. Jpn.*: 15-16, 2005 (in Japanese).