社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2007-191 (2008-3)

Perception of Speaker Identity and its Relation to the Phonological Features

Kanae AMINO and Takayuki ARAI

Department of Electrical and Electronic Engineering, Sophia University

7-1 Kioi-cho, Chiyoda-ku, Tokyo, 102-8554 Japan

E-mail: {amino-k, arai}@sophia.ac.jp

Abstract In perceptual speaker identification, it is known that the identification accuracy depends on the contents of the stimuli presented to the listeners. In our previous experiments, we found that the stimuli containing a nasal sound are effective for speaker identification by listening, and that coronal sounds are more effective than labial ones.

In this present study, we investigated the effects of the stimulus contents again, this time focusing on the vowels in CV monosyllabic stimuli. Through the experiment we obtained the following outcomes: 1) nasal sounds gained significantly higher scores than oral sounds, although the difference between the sonorants and the obstruents was not significant; 2) alveolar nasals were more effective than the labial nasal; 3) palatalisation of a consonant improved speaker identification performances; and 4) back vowels were more effective than front vowels significantly. These tendencies imply the following phonological grounds: 1) [+nas] is more important than [+son]; 2) [+nas] comes under [PLACE] as to the judgment of the speaker identity; 3) [son] may lie under [nas] in the identification of the speakers, and 4) [-ant] and [+back] are also important for speaker identification.

Key words Individuality, Perceptual Speaker Identification, Nasality, Phonological Features, Feature Geometry

1. Introduction

In daily conversations, listeners obtain many kinds of information from speech sounds besides linguistic information. One of them is the information about speakers. Speaker information conveyed by speech sounds includes speaker’s identity, information about his/her health states, and the regional and social backgrounds [1, 2].

Speech sounds differ both among individuals and within individuals. In spite of these differences, or inter- and intra-speaker differences, human beings perceive and recognise the utterances, and this fact has been of great interest to many of the speech scientists as well as to the psychologists and the linguists. Inter-speaker differences derive from two principal factors: speaker’s physiological properties and learnt properties. The former is, for example, the length and the thickness of the vocal folds, the length and the shapes of the vocal tract, etc. On the other hand, the latter is more like the speaker’s habits in articulation, or the dialects and speaking styles [3-5].

Identifying familiar people only by their voices is another ability of the human beings. This ability, also seen in other primates such as rhesus macaques (*Macaca mulatta*) [6] and Japanese macaques (*Macaca fuscata*) [7], plays an important role for a communication to be successful. Listeners exploit speaker characteristics in order to gauge the communicative settings, and also use them for understanding the contents of the utterances [8-10].

The study on perceptual speaker identification has often been oriented for forensic purposes. In order to make forensic speaker recognition more efficient, we should use the speech data selectively, i.e. select the sounds by which human listeners identify speakers most accurately [11]. Also, in order to examine the adequacy of earwitnesses’ testimonies, perceptual properties of human speaker identification need to be clearly understood.

There are some kinds of limitations reported as to human speaker identification. For example, voice disguise can degrade the speaker identification performances [12]. In Hirson and Duckworth [13], creaky voices degraded the speaker identification accuracy from 90% to 65% compared to the modal phonation. Speaking in whisper also degrades the performances [14].

Listeners cannot tell who is speaking when the speech samples are too short. However, the effects of utterance duration on identification task seem to be important only when the utterances in question are very brief [1, 15]. Consensus on the duration of the speech materials that is enough for the perception of the speaker identity is not yet obtained [15]. Correct speaker identification rates fall when the speech samples are in foreign languages [16]. This means that the perception of speaker information and that of phonological information interact with each other.

Identification and verification of familiar speakers are usually easier than the identification of previously unknown speakers [17-18]. How long a listener can remember a given speaker’s voice is another problem. There is a report that the participant could remember a voice over a period of two years [19-20].

Finally, the phonemic variations of the speech materials also affect human speaker identification. This is another evidence for the existence of the interaction between the processing of the phonological and the speaker information. These differences among the speech sounds in the relative effectiveness for identifying the speakers also mean that variations in the physiological properties of different speakers may be reflected in isolated utterances of different speech sounds [21].

Table 1 shows the list of the studies that investigated the effects of the stimulus contents on human speaker identification. Most of them reported that nasals and vowels of the language in question were the most effective sounds for identifying the speakers. Also in our previous experiments, as can be seen, the stimuli containing nasal sounds gained consistently highest scores. In our experiments, however, the monosyllables contained only /a/ as the vowel in order to make the experiments simple. This present study explores the differential effects of the consonants and the following vowels on perceptual speaker identification.

Table 1. Studies on Differential Effects of Speech Sounds in Perceptual Speaker Identification

Research	No. of speakers*	No. of listeners**	Stimuli (language)	Effective sounds
Nishio [22]	5×2, M and F	31, familiar	Sentences, phrases, isolated syllables (Japanese)	Sentences, phrases, isolated vowel /a/
Ramishvili [23]	6, M	?, familiar	Isolated phonemes (Russian)	Vowels except /u/, voiced consonants
Bricker and Pruzansky [24]	10, M	16, familiar	Excerpted vowels (English)	/a/
Stevens et al. [25]	8, M	6, unknown	Isolated words (English)	Front stressed vowels
Matsui et al. [26]	8, M	11, familiar	Excerpted CVC syllables (Japanese)	Depends on speakers
Kitamura and Akagi [27]	5, M	8, familiar	Isolated vowels (Japanese)	/a/
Amino [28]	3, F	14, familiar	Isolated vowels, isolated monosyllables (Japanese)	/a/, nasals
Amino [29]	3×2, M and F	18, familiar	Excerpted monosyllables (Japanese)	Nasals, voiced coronal consonants
Amino et al. [30]	10, M	5, familiar	Excerpted monosyllables including coronal consonants (Japanese)	Nasals
Amino et al. [31]	8, M	8, familiar	Isolated monosyllables of various syllable structures (Japanese)	Nasal onsets, coda nasals
Amino et al. [32]	10, M	16, unknown	Excerpted monosyllables including coronal consonants (Japanese)	Nasals

* M, F: male and female speakers, respectively.
** Familiar, unknown: whether the listeners were familiar with or unknown to the speakers.

Table 2. Speakers of this Experiment

Speaker ID	Age	Height [cm]	Hometown	Average F0 [Hz]*	*S.D.* of F0 [Hz]
#1	In 20s	181	Yokohama	148.9	6.7
#2	In 20s	171	Tokyo	127.0	3.9
#3	In 30s	169	Tokyo	164.7	6.5
#4	In 40s	164	Chiba	121.5	3.9

*Averaged among all the utterances used in this experiment.

Table 3. Stimulus Monosyllabes

Consonant		/a/	/e/	/i/	/o/	/ɯ/
None	ϕ	/a/	/e/	/i/	/o/	/ɯ/
Stops	/t/	/ta/	/te/	-	/to/	-
	/d/	/da/	/de/	-	/do/	-
Tap / Flap	/ɾ/	/ɾa/	/ɾe/	/ɾi/	/ɾo/	/ɾɯ/
Fricatives	/s/	/sa/	/se/	-	/so/	/sɯ/
	/z/	/za/	/ze/	-	/zo/	-
	/ʃ/	/ʃa/		/ʃi/	/ʃo/	/ʃɯ/
Affricates	/t͡ʃ/ /t͡s/	-	-	/t͡ʃi/	-	/t͡sɯ/
	/d͡ʒ/ /d͡z/	-	-	/d͡ʒi/	-	/d͡zɯ/
Nasals	/m/	/ma/	/me/	/mi/	/mo/	/mɯ/
	/n/	/na/	/ne/	/ni/	/no/	/nɯ/
	/nʲ/	/nʲa/	-	-	/nʲo/	/nʲɯ/
Approximants	/j/	/ja/	-	-	/jo/	/jɯ/
	/w/	/wa/	-	-	-	-

2. Experiment

2.1 Speech Materials and Participants

In making the stimuli for the experiment, speech materials were selected from JEIDA Japanese Common Speech Data Corpus (JEIDA-JCSD) [33]. Among the one hundred and ten monosyllables in the corpus, we used forty-eight syllables uttered by four male speakers. These four speakers were selected, because they all spoke Tokyo Japanese, and their recordings were conducted in relatively quiet environments. Information on the speakers and the stimulus monosyllables are shown in Tables 2 and 3, respectively.

Fifteen (eight male and seven female) volunteers served as the listeners in this study. They had never heard the speakers' voices before.

The mean age of the participants was 23.4 years old and they were all native speakers of Japanese. None of them had any known hearing impairments.

2.2 Procedures

The experiment was held in a sound-treated room. First, the participants listened to the sample speech of each of the speakers played on a computer through headphones (SONY MDR-Z700). The sample words were: /horʲɯ:/ (保留, suspension), /kaigʲo:/ (改行, creating a new line), and /heɴkaɴ/ (変換,

conversion).

The participants were instructed to remember the four speakers' voices, and they could listen to the samples as many times as they wanted. Then they practised the task by using these samples. Learning of the speakers and the practice trials were repeated until they reached 90% correct speaker identifications in the practice session.

Test session followed the practice session. All the tests were again performed on a computer. The participants listened to a monosyllable uttered by one of the speakers, identified the speaker, and answered by clicking on a button of the speaker ID. The total number of the test stimuli was 576, i.e. corresponding to four speakers, forty-eight monosyllables and three different tokens for each syllable. The total test time was about an hour, and the participants took breaks after every 192 trials.

3. Results

The results of this experiment were evaluated by the speaker identification accuracy (% correct) of the fifteen listeners. The identification performances are summarised according to the consonants in Table 4 and Figure 1 (a) and to the following vowels in Figure 1 (b).

We can see in Table 4 and Figure 1 that the identification accuracies by all the consonants were above the chance level (25%). Bricker and Pruzansky [21] and Clarke and Becker [34] reported that the mean correct identification rates for the identification of four unknown speakers was 58%. Compared to this report, the results we gained this time is a little better, but within the ballpark.

Among the consonants, coronal nasals /n/ and /n^{j}/ gained the highest identification scores. The voiced coronal stop /d/ followed them, and then the coronal fricatives /z/ and /ʃ/. Syllables without the onset consonant gained the lowest scores. All these tendencies are consistently seen in our previous experiments [28-32].

Results of the one-way ANOVA among all the consonants showed a significant tendency (*p* = 0.058), and the difference between nasals and non-nasals was significant in Mann-Whitney's *U*-test (*p* = 0.045). Difference between obstruents and sonorants in the same test was not significant.

Another tendency is that palatalisation of the consonants enhanced the speaker identification performances. Post-alveolar and palatalised sounds /ʃ/ and /n^{j}/ obtained higher scores than their alveolar counterparts, /s/ and /n/, though the differences were not significant.

As for the vowels, back vowels /o/, /ɯ/ and /a/ ranked higher than front vowels /i/ and /e/. In Mann-Whitney's *U*-test, the difference between back-front vowels was significant (*p* = 0.003). In open-close vowel comparison, no significant differences were observed.

4. Discussion

The major findings from the results of this study are as follows:

1) Nasal sounds gained significantly higher scores than oral sounds, although the difference between the sonorants and the obstruents was not significant.
2) Alveolar nasals were more effective than the labial nasal.
3) Palatalisation of consonants improved speaker identification performances.
4) Back vowels were more effective than front vowels significantly.

Table 4. Speaker Identification Results According to Consonants

Consonant		% Correct
None	ϕ	56.1
Stops	/t/	59.3
	/d/	64.8
Tap / Flap	/ɾ/	57.0
Fricatives	/s/	61.4
	/z/	65.0
	/ʃ/	63.9
Affricates	/t͡ʃ/ /t͡s/	62.2
	/d͡ʒ/ /d͡z/	56.9
Nasals	/m/	60.4
	/n/	65.1
	/n^{j}/	65.6
Approximants	/j/	62.4
	/w/	61.1
Average		61.6

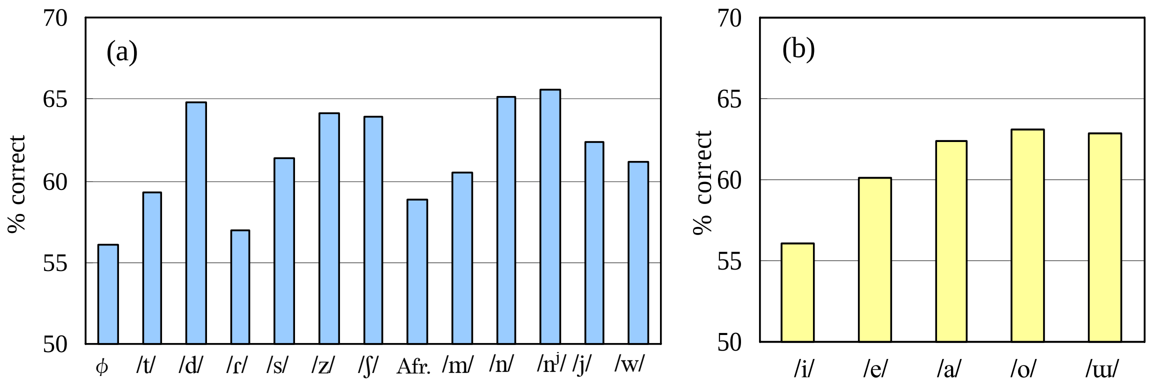

Figure 1. Speaker Identification Results According to (a) Consonants, and (b) Vowels.

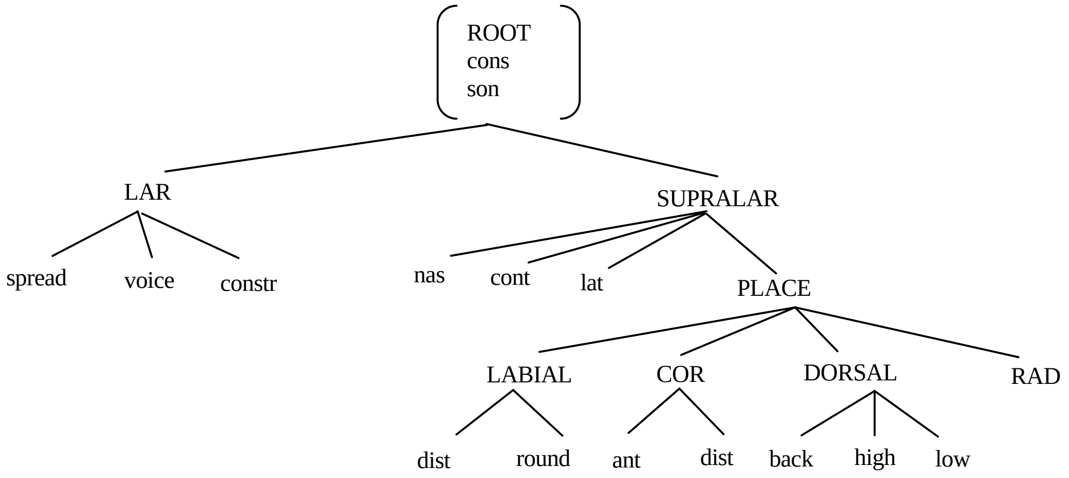

Figure 2. Consensus Feature Tree [42]

These tendencies can be grasped as followings in the framework of the phonological features:

a) [+nas] (nasal) is more important than [+son] (sonorant).
b) [+nas] comes under [PLACE] as to the judgment of the speaker identity.
c) [son] may lie under [nas] in the identification of the speakers.
d) [-ant] (anterior) and [+back] are also important for speaker identification.

The features mentioned above thus rank in the following order:

[ant], [back] > [PLACE] > [nas] > [son]

It is interesting to note that the hierarchy of the features mentioned above is just in the reversed order of the phonological feature geometry advocated by Gussenhoven and Jacobs [35] shown in Figure 2.

Phonological features are devised to reflect the elements involved in speech production, and thus they can explain various phonetic phenomena as well as language variation and language change. The dominance structures of the features in the feature tree are important for the explanation of those phenomena.

On the other hand, speaker individuality contained in speech sounds should not, and usually do not, exceed phonological information [3]. This implies that the features critical for phonological distinctions do not contribute much to speaker identity. This can account for the reason that the phonological feature hierarchy and that of speaker information are converse.

Our next task will be to think of the relationship to the acoustical properties looking into the resonance properties of the stimuli and coarticulation between the consonants and the following vowels. It will help understand in detail the production and the perception of the speaker individuality, and search for a more elegant way to describe the relation between phonological and phonetic aspects of the speech.

5. Acknowledgment

This study was supported by Sophia University Open Research Centre from MEXT.

6. References

[1] I. Pollack, J. M. Pickett, and W. H. Sumby, “On the Identification of Speakers by Voice,” J. Acoust. Soc. Am., Vol.26, pp.403-406, 1954.

[2] P. Ladefoged and D. Broadbent, “Information Conveyed by Vowels,” J. Acoust. Soc. Am., Vol.29, pp.98-104, 1957.

[3] Y. Niimi, Speech Recognition, T. Sakai (ed.), Kyoritsu Shuppan Publishing Company, Tokyo, 1979.

[4] S. Furui, “Key Issues in Voice Individuality,” J. Acoust. Soc. Jpn., Vol.51, pp.876-881, 1995.

[5] H. Traunmueller, “Modulation and Demodulation in Production, Perception, and Imitation of Speech and Bodily Gestures,” Proc. FONETIK 98, pp.40-43, 1998.

[6] D. Rendall, P. Rodman and R. Emond, “Vocal recognition of individuals and kin in free-ranging rhesus monkeys,” Anim. Behav., 51, pp. 1007-1015, 1996.

[7] N. Masataka and K. Fujita, “Vocal learning of Japanese and rhesus monkeys,” Behaviour, Vol. 109, pp. 191-199, 1989.

[8] S. Furui, Acoustic and Speech Engineering, Kindai Kagaku-sha, Tokyo, 1992.

[9] L. Nygaard, “Perceptual Integration of Linguistic and Nonlinguistic Properties of Speech,” Chap. 16 in The Handbook of Speech Perception, D. Pisoni and R. Remez (eds.), pp.390-413, Blackwell Publishing, Oxford, 2005.

[10] J. Goggin, C. Thompson, G. Strube, and L. Simental, “The Role of Language Familiarity in Voice Identification,” Memory and Cognition, Vol. 19, pp. 448-458, 1991.

[11] D. O’Shaughnessy, Speech Communications –Human and Machine–, second ed., Addison-Wesley Publishing Company, New York, 2000.

[12] H. Hollien, W. Majewski, and E. Doherty, “Perceptual Identification of Voices under Normal, Stress, and Disguise Speaking Conditions,” J. Phonetics, Vol. 10, pp.139-148, 1982.

[13] A. Hirson and M. Duckworth, “Glottal Fry and Voice Disguise: a Case Study in Forensic Phonetics,” J. Biomed. Eng., Vol. 15, pp.193-200, 1993.

[14] T.L. Orchard and A.D. Yarmey, “The Effects of Whispers, Voice-sample Duration, and Voice Distinctiveness on Criminal Speaker Identification,” Appl. Cog. Psychol., Vol. 9, pp.249-260, 1995.

[15] H. Hollien, Forensic Voice Identification, Academic Press, San Diego, 2002.

[16] C. Thompson, “A Language Effect in Voice Identification,” Appl. Cog. Psychol., Vol.1, 1979.

[17] A. Schmidt-Nielsen and K.R. Stern, “Identification on Known Voices as a Function of Familiarity and Narrow-band Coding,” J. Acoust. Soc. Am., Vol.77, pp.658-663, 1985.

[18] A. Schmidt-Nielsen and K.R. Stern, “Recognition of Previously Unfamiliar Speakers as a Function of Narrow-band Processing and Speaker Selection,” J. Acoust. Soc. Am., Vol.79, pp.1174-1177, 1986.

[19] F. McGhee, “The Reliability of the Identification of the Human Voice,” J. Gen. Psychol. Vol.17, pp.249-271, 1937.

[20] F. McGhee, “An Experimental Study of Voice Recognition,” J. Gen. Psychol., Vol. 31, pp.53-65, 1944.

[21] P. Bricker and S. Pruzansky, “Speaker Recognition,” Chap. 9 in Experimental Phonetics, N. Lass (ed.), pp.295-326, Academic Press, London, 1976.

[22] T. Nishio, “Can We Recognise People by Their Voices?” Gengo-Seikatsu, Vol.158, pp.36-42, 1964.

[23] G. Ramishvili, “Automatic Voice Recognition,” Engineering Cybernetics, Vol.5, pp.84-90, 1966.

[24] P. Bricker and S. Pruzansky, “Effects of Stimulus Content and Duration on Talker Identification,” J. Acoust. Soc. Am., Vol. 40, pp.1441-1450, 1966.

[25] 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., Vol.44, pp.1596-1607, 1968.

[26] T. Matsui, I. Pollack, and S. Furui, “Perception of Voice Individuality Using Syllables in Continuous Speech,” Proc. of the 1993 Autumn Meet. Acoust. Soc. Jpn., pp.379-380, 1993.

[27] T. Kitamura and M. Akagi, “Speaker Individualities in Speech Spectral Envelopes,” J. Acoust. Soc. Jpn. (E), Vol.16, pp.283-289, 1995.

[28] K. Amino, “The Characteristics of the Japanese Phonemes in Speaker Identification,” Proc. Sophia Univ. Linguistic Soc., Vol. 18, pp.32-43, 2003.

[29] K. Amino, “Properties of the Japanese Phonemes in Aural Speaker Identification,” IEICE Tech. Rep., Vol. Vol.104, pp.49-54, 2004.

[30] 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, pp.2025-2028, 2005.

[31] K. Amino, T. Sugawara, and T. Arai, “Effects of the Syllable Structure on Perceptual Speaker Identification,” IEICE Tech. Rep., Vol.105, pp.109-114, 2006.

[32] K. Amino, T. Arai, and T. Sugawara, “Phoneme-dependency of Accuracy Rates in Familiar and Unknown Speaker Identification,” J. Acoust. Soc. Am., Vol. 120, pp.3291, 2006.

[33] JEIDA Japanese Common Speech Data Corpus, http://www.sunrisemusic.co.jp/dataBase/fl/voicebase01_fl.html

[34] F. Clarke and R. Becker, “Comparison of Techniques for Discriminating among Talkers,” J. Speech. Hear. Res., Vol. 12, pp.747-761, 1969.

[35] C. Gussenhoven and H. Jacobs, Understanding Phonology, Hodder Arnold, London, 1998.