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
SP2005-181 (2006-03)

Effects of the Syllable Structure on Perceptual Speaker Identification

Kanae AMINO †, Tsutomu SUGAWARA †, and Takayuki ARAI ‡

† Faculty of Foreign Studies, Sophia University
‡ Faculty of Science and Technology, Sophia University
7-1 Kioi-cho, Chiyoda-ku, Tokyo, 102-8554 Japan

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

Abstract In speaker identification by listening, the identification rates vary depending on the speech contents presented to the subjects. It is reported that the nasals are more effective than the oral sounds for identifying speakers. The present study investigates the availability of the nasal sounds in terms of syllable structures. The results showed that the coda nasals are highly effective, though onset consonants are also important. As to the place of articulation, alveolar consonants in onset positions were more effective than bilabials, and the nasals were better than their oral counterparts were.

Key words Nasals, Individuality, Speaker Identification, Syllable Structure

1. Introduction

It is manifest in everyday situations that human beings are able to recognise speakers by speech sounds alone. This is because speech sounds convey not only linguistic information but also other classes of information, including information about the speaker's identity, social background and so on [1, 2].

The term speaker individuality, or voice quality in some studies, is defined as the 'characteristic auditory colouring of a given speaker's voice' [3], and this quality is thought to be responsible for human identification of a speaker or a group of speakers [4]. In another study [5], this term is used as referring to 'a quasi-permanent quality running through all the sound that issues from a speaker's mouth.' This implies that speaker individuality is present all the time that the speaker is talking.

The characteristics that are specific to a speaker derive from his/her physiological properties, i.e. the length or the thickness of the vocal folds, the length or the volume of the vocal tract, etc., or from the learned habits, such as speaking style, speaking rate or dialects [6-8]. The latter can be extended to the modality of the utterance and to articulatory disorder in a broad sense.

Listeners exploit these characteristics for identifying the speaker and this process is necessary for successful speech communication [9, 10]. For instance, speaker information is used to gauge communicative settings. However, linguistic information, or the speech intelligibility, is of primary importance in human communication, and the speaker information is secondary to it. Indeed, the study on speech individuality has deepened only recently, compared to the study on linguistic information of the speech sounds [1, 6].

The motivations of research on speaker individuality have been based on the forensic purposes or on practical considerations. The use of speech materials in court cases has a relatively long history since 1660 [11], though it is still controversial and is being discussed actively in the forensic field. In automatic speaker recognition, where a decision-making process about the speaker identity is carried out by machine, the features that indicate speaker individuality are extracted. On the contrary, in automatic speech recognition, where speech sounds are translated into texts automatically, those speaker-dependent features are eliminated in order to pull out the abstracted elements of the speech sounds [12].

Research on the perceptual speaker identification, or speaker identification by human, has been oriented more or less to theoretical purposes. In order to select the effective texts for a speaker recognition system, researchers will find it useful to perform a speaker recognition experiment by human perception, and to use the speech contents by which the listeners identified the speakers most accurately [13]. In one study on human communication, it is reported that information about a speaker is processed separately from the recovery of linguistic content,

though these two kinds of information interact with each other [10]. In addition, it is pointed out that listeners use linguistic information, or the contents of utterances, in order to identify the speakers, and vice versa [10, 14]. This means that the study on human perception of speaker identity can contribute to a better understanding of human cognition.

This present study is concerned with the effects of linguistic structures on perceptual speaker identification.

2. Linguistic Information and Speaker Identity

2.1 Differential effects of speech sounds

Although, as mentioned above, speaker characteristics are present all the time during an utterance, research shows that there are differences among the speech sounds in the relative effectiveness for identifying the speakers. This proves that variations in the physiological properties of different speakers may be reflected in isolated utterances of different speech sounds [15].

Table 1 is a summary of the previous studies where these differential effects of speech sounds were investigated. Most of them reported that nasals and vowels of the language in question were the most effective sounds for identifying the speakers.

2.2 Speaker individuality and the nasals

In our previous experiments [16-18], also introduced in Table 1, the speakers were identified by familiar listeners using various Japanese monosyllables. The monosyllables presented to the subjects here all had the structure of CV. Fifteen kinds of consonants were used as the onset consonant in [16, 17], and nine coronal consonants in [18]. The nucleus vowel was always controlled to be /ú/ in order to make the experiments simple.

The results of the perception tests in [17, 18] are shown in Table 2. This is the list of the best five monosyllables in each speaker group. It is found that the stimuli containing nasal sounds gained the highest scores, with exception of the female speaker group in [17]. Voiced coronal obstruents ranked also in higher standings.

2.3 Problems

The effectiveness of the nasals has been reported in the previous researches. However, Japanese has another type of nasal, i.e. the coda nasal, and this has not been examined yet. Moreover, the stimuli used in the experiments above had an onset consonant and a nucleus vowel and therefore the effects of the vowel part or the transition to the following vowel were not inspected.

In this study, we carried out a perceptual speaker identification experiment in order to investigate the effects of the syllable structures and the contributions of the transitions to the identification.

3. Experiment

3.1 Recordings

3.1.1 Speakers

In selecting the speakers in a perceptual speaker identification test, one must ensure that age, gender and accent are consistent among the speakers [15].

Eight male students in the age range 22-25 (average 23.1) served as the speakers in this experiment. All of them speak Tokyo Japanese in daily conversation and had normal hearing.

3.1.2 Speech materials

The recording sessions were held in a soundproof room, using a DAT (digital audiotape) recorder (SONY TCD-D8) and a microphone (SONY ECM-MS957). The speech data were recorded on a tape at a sampling rate of 48 kHz with 16-bit resolution.

The recorded materials are Japanese non-sense monosyllables of various structure types. In order to see how the syllable structures and coda nasals work in the identification of the speakers, the materials covered the following structure types: V, VV, VN, CV, CVV and CVN. This variety of structures enables us to know the influence of the onset consonants, syllable weight and the coda nasals. The speakers read out each kind of material seven times and five of these were selected and used as the stimuli.

In order to examine the contribution of the consonant -to-vowel transitions, we prepared two more structures, –V and –VC, which were cut out from recorded CV and CVC. These two types were edited manually on the computer, using the software Praat [26]. The onset consonants were cut off just before the visible transitions of the second formant of the following vowel began on spectrograms. Thus, the stimuli –V and –VC contained the transition parts to the nucleus vowel. We will indicate it by '–'.

Table 1. List of studies on differential effects of speech sounds in speaker identification

Identification by machine

Research	No. of speakers*	No. of listeners**	Speech materials (language)	Effective sounds
Sambur [19]	11, M	—	Sentences (English)	Vowels, nasals (/ù/ /ʌ̀/), stridents (/ː/ /ə/)
Nakagawa and Sakai [20]	10, M	—	Various kinds of VCV words (Japanese)	/æ/ /ù/ /ù/ /ʌ̀/ /□/

Identification by human perception

Research	No. of speakers*	No. of listeners**	Speech materials (language)	Effective sounds
Nishio [21]	5×2, M and F	31, familiar	Sentences, phrases, isolated syllables (Japanese)	Sentences, phrases, /ú/
Ramishvili [22]	6, M	?, familiar	Isolated phonemes (Russian)	Vowels except /□/, voiced consonants
Bricker and Pruzansky [15]	10, M	16, familiar	Excerpted vowels (English)	/ú/
Stevens et al. [23]	8, M	6, naïve	Isolated words (English)	Front stressed vowels
Matsui et al. [24]	8, M	11, familiar	Excerpted CVC syllables (Japanese)	Depends on the speakers
Kitamura and Akagi [25]	5, M	8, familiar	Isolated vowels (Japanese)	/ú/
Amino [16]	3, F	14, familiar	Isolated vowels, isolated monosyllables (Japanese)	/ú/, nasals
Amino [17]	3×2, M and F	18, familiar	Excerpted monosyllables (Japanese)	Nasals, voiced coronal consonants
Amino et al. [18]	10, M	5, familiar	Excerpted monosyllables (Japanese)	Nasals

* M, F: male and female speakers, respectively.

** Familiar, unknown: whether the listeners were familiar with or unknown to the speakers.

Table 2. Best 5 stimuli and their identification rates in the previous experiments*

3 male speakers [17]	3 female speakers [17]	10 male speakers [18]
/ʌ̀ú/ (94.4%)	/□ú/ (97.8%)	/ʌú/ (86.0%)
/ːú/ (92.8%)	/ʌú/ (96.7%)	/ʌ̀□ú/ (85.6%)
/ʌ̀□ú/ /□ú/ /ɚ́ú/ (90.3%)	/□əú/ /ɚ́ú/ (95.0%)	/ùú/ /□ú/ (80.8%)
	/îú/ (93.9%)	/ːú/ (78.8%)

*The numbers of the samples (the denominators) are 195 and 180 in [17], for male and female speakers, respectively, and 250 in [18].

3.2 Perception test

3.2.1 Subjects

Eight (two male and six female) students who belong to the same research group at university as the speakers participated in the perception test. They had spent at least one year with the speakers and knew all of the speakers very well.

The mean age was 23.1 years old and they were all native speakers of Japanese. None of them had any known hearing impairment.

3.2.2 Procedures

The perception test was held in the same soundproof room as the recording sessions. The stimuli used in perception test are shown in Table 3.

The subjects listened to the sample files of each speaker first, and practised the task by use of these samples. These files are different from test samples, and the subjects listened to them and practised only once.

After the practice, test sessions followed. All the sessions were performed on a computer. The subjects listened to a test sample, identified the speaker, and then answered by clicking on a rectangle with the name of the speaker to whom s/he thought the speech belonged.

The total number of the test stimuli was 920, i.e. corresponding to 8 speakers, 23 stimulus types and 5 different samples for each type. The total test time was about an hour, and the subjects took breaks after every 230 trials.

4. Results

The results of the perception test are summarised according to the syllable structures in Figure 1 and to the onset consonants in Figure 2.

Figure 1 shows that the structures with an onset consonant (shown by black bars) gained higher scores than onsetless structures (grey and striped bars). It also tells us that there is a tendency that the heavier syllables obtained better scores except in /VN/. Coda nasals also seem to be effective for the identification in /CVN/ and /–VN/. As to the influence of the transition, we cannot tell many things only from the results of this study, but the scores of the edited syllables, /–V/ and /–VN/, did not reach those of the structures with an onset.

Table 3. List of stimuli used in perception test

Syllable structure	Stimuli
V	/ú/
VV	/úú/
VN	/úĂ/
CV	/ĺú/ /ɔ́ú/ /ùú/ /ìú/
CVV	/ĺúú/ /ɔ́úú/ /ùúú/ /ìúú/
CVN	/ĺúĂ/ /ɔ́úĂ/ /ùúĂ/ /ìúĂ/
–V	/(ĺ)-ú/ /(ɔ́)-ú/ /(ù)-ú/ /(ì)-ú/*
–VC	/(ĺ)-úĂ/ /(ɔ́)-úĂ/ /(ù)-úĂ/ /(ì)-úĂ/*

* The consonants in ( ) are cut off manually.

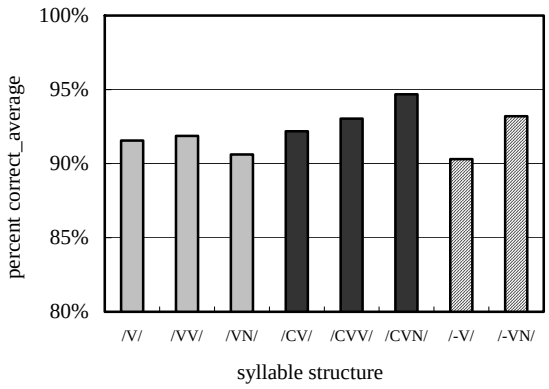

Fig. 1 Percentage of correct speaker identification (as to the syllable structure)

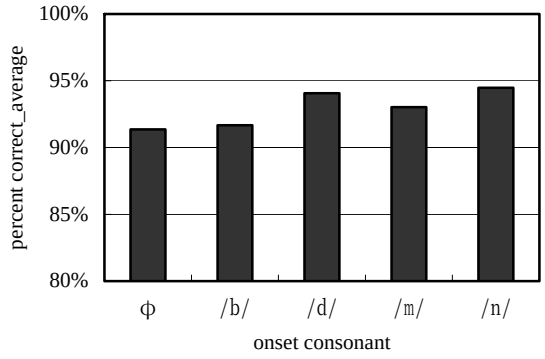

Fig. 2 Percentage of correct speaker identification (as to the onset consonant)

It is affirmed again in Figure 2 that onset consonants are important. The data here do not include the results of the edited structures. The letter φ indicates the onsetless syllables, /V/, /VV/, and /VN/. The score of these onsetless syllables were worst of all other structures, though it still gained more than 90 % correct identification.

One can also see in Figure 2 that the alveolar consonants in the onset position were more effective than the bilabial consonants in the test. Nasal consonants, /n/ and /m/, were better than their oral counterparts, /d/ and /b/, respectively.

5. Discussion

The major conclusions from this study are as follows:

I. Onset consonants are important for speaker identification.

II. Alveolar consonants convey more individuality than bilabials.

III. Nasals are effective for speaker identification both in onset and coda positions.

Onset consonants

The structures with transition or the onsetless structures in this study gained no higher identification rates. This suggests that the differential effects in the onset consonants come from the consonant parts.

Alveolars are better than bilabials

This is what was seen in our previous experiment, too [17]. Japanese has three places of articulation in oral and nasal stops, i.e. bilabial, alveolar and velar. Alveolar sounds have the largest range of possible articulation of these three, as the phonology of Japanese does not require any contrasts in place feature in the coronal area as to the stop sounds. This may lead to inter-speaker differences in articulation of alveolar sounds.

Onset and coda nasals

The properties of the nasal sounds are speaker-dependent, because the shapes of the resonators involved in the articulations of these sounds are considerably different for individuals [27]. In addition, the shapes of these resonators cannot be changed voluntarily. This means that the properties of nasals rarely change.

The nucleus vowel in the structure that has a nasal sound in onset or coda, or both, position(s) is nasalised to some extent. This nasalisation process occurs especially in the structure with a coda nasal, and the nasalised vowels are predicted to contain more individuality than non-nasalised vowels.

Coda nasal /N/ in the word-final position has been said to be articulated at the uvula, but recent work [28] reports that the place of articulation of /N/ differs among speakers, and this sound is not always uvular. This probably explains the results in this study, too.

The final goal of this study is to delimit the speaker individuality carried in speech signals and to understand the interaction between human perception of the speaker individuality and the linguistic information.

Our future task will be to look into the acoustic characteristics of the stimuli used in this study, and to show quantitative data for coronal onsets and coda nasals. We must also test on different kinds of vowels, in order to examine the effects of coarticulation. Speaker identification experiments with reversed speech may also be useful for revealing the properties of human perception.

6. Acknowledgment

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] I. Pollack, J. M. Pickett, and W. H. Sumby, “On the Identification of Speakers by Voice,” JASA, Vol.26, No.3, pp.403-406, 1954.

[2] P. Ladefoged and D. Broadbent, “Information Conveyed by Vowels,” JASA, Vol.29, No.1, pp.98-104, 1957.

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

[4] M. Ball and J. Rahilly, Phonetics, Arnold, London, 1999.

[5] D. Abercrombie, Elements of General Phonetics, Edinburgh University Press, Edinburgh, 1967.

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

[7] S. Furui, “Key Issues in Voice Individuality,” J. Acoust. Soc. Jpn., Vol.51, No.11, pp.876-881, 1995.

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

[9] J. Kreiman, D. Van Lacker, and B. Gerratt, “Perception of Voice Quality,” Chap.14 in The Handbook of Speech Perception, D. Pisoni and R. Remez (ed.), pp.338-362, Blackwell Publishing, Oxford, 2005.

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

[11] H. Hollien, The Acoustics of Crime, Plenum, New York, 1990.

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

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

[14] 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.

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

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

[17] K. Amino, "Properties of the Japanese Phonemes in Aural Speaker Identification," Tech. Rep. IEICE, Sp2004-37, pp.49-54, 2004.

[18] 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.

[19] M. Sambur, "Selection of Acoustic Features for Speaker Identification," IEEE Trans. ASSP, Vol. 23, No.2, pp.176-182, 1975.

[20] S. Nakagawa and T. Sakai, "Feature Analyses of Japanese Phonetic Spectra and Consideration on Speech Recognition and Speaker Identification," J. Acoust. Soc. Jpn., Vol.35, No.3, pp.111-117, 1979.

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

[22] G. Ramishvili, "Automatic Voice Recognition," Engineering Cybernetics, Vol.5, pp.84-90, 1966.

[23] K. Stevens, C. Williams, J. Carbonell, and B. Woods, "Speaker Authentication and Identification: A Comparison of Spectrographic and Auditory Presentations of Speech Material," JASA, Vol.44, No.6, pp.1596-1607, 1968.

[24] 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.

[25] T. Kitamura and M. Akagi, "Speaker Individualities in Speech Spectral Envelopes," J. Acoust. Soc. Jpn. (E), Vol.16, no.5, pp.283-289, 1995.

[26] P. Boersma and D. Weenik, Praat: Doing Phonetics by Computer, Ver.4.3.14 (Computer Program), retrieved from http://www.praat.org/ 2005.

[27] J. Dang and K. Honda, "Acoustic Characteristics of the Human Paranasal Sinuses Derived from Transmission Characteristic Measurement and Morphological Observation," JASA, Vol.100, No.5, pp.3374-3383, 1996.

[28] M. Hashi, A. Sugawara, T. Miura, S. Daimon, Y. Takakura, and R. Hayashi, "Articulatory Variability of Japanese Moraic-Nasal," Proc. of the 2005 Autumn Meet. Acoust. Soc. Jpn., pp.411-412, 2005.