Effects of the Phonological Contents on Perceptual Speaker Identification*

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

[1] Department of Electrical and Electronics Engineering
[2] Graduate Division of Foreign Studies, Sophia University,
7-1 Kioi-cho, Chiyoda-ku, 102-8554, Tokyo, Japan
{amino-k, arai, sugawara}@sophia.ac.jp

Abstract. It is known that the accuracy of perceptual speaker identification is dependent on the stimulus contents presented to the subjects. Two experiments were conducted in order to find out the effective sounds and to investigate the effects of the syllable structures on familiar speaker identification. The results showed that the nasal sounds were effective for identifying the speakers both in onset and coda positions, and coronal sounds were more effective than labial counterparts. The onset consonants were found to be important, and the identification accuracy was degraded in onsetless structures.

Keywords: Perceptual speaker identification, Familiar speaker identification, Nasal sounds, Coronal sounds.

1 Introduction

It is manifest that human beings have the innate ability to recognise speakers by speech sounds alone. This means that speech sounds convey information about the speakers as well as the linguistic contents.

Speech individualities, or speaker-specific characteristics contained in speech sounds, derive either from physiological properties of the speaker or from his/her learned habits. The former includes the length or the thickness of the vocal folds, the length or the volume of the vocal tract, and all other physical properties of a given speaker, and the latter is exemplified by speaking style, speaking rate, and social and regional dialects. The modality of an utterance and the articulatory disorders may also be included among the learned habits.

The term "speaker individuality," called "voice quality" in a broad sense in some studies, can be used to refer to "a quasi-permanent quality running through all the sound that issues from a speaker's mouth [1]," and the "characteristic auditory

* This work was originally presented at the 9th European Conference on Speech Communication and Technology Interspeech 2005 (Experiment 1) and at the International Workshop on Frontiers in Speech and Hearing Research 2006 (Experiment 2). For details see references [23] and [24].

C. Müller (Ed.): Speaker Classification II, LNAI 4441, pp. 83–92, 2007.

colouring of a given speaker's voice [2]." Ball and Rahilly [3] regard this quality as being responsible for human identification of a speaker or a group of speakers.

In forensic speech sciences, it is important to know about the relationship between speech individualities and how people perceive them. The use of speech materials in court cases has a relatively long history since 1660 [4], though there are still some controversial issues remaining, such as the correspondences between acoustic properties of the sounds and the human percepts, the limits of the human memory, and the effects of the recording conditions and transmission.

If we could find acoustic correlates of speaker individuality, they can be exploited for various fields in speech technology [5]. For example, speaker individuality is extracted and used in automatic speaker recognition and in voice conversion [6, 7]. In automatic speech recognition, on the other hand, speech individuality should be excluded. One way to find the acoustic parameters that indicate speech individuality is to conduct a speaker identification experiment by listening, and to investigate the property of human perception [8]. The factors that affect the perception would be important in defining speech individuality.

When identifying a speaker, listeners abstract speaker-specific characteristics of the utterance, and collate them with the information stored in the brain. It is reported that the processing of speech contents and that of speaker identity occur separately, though they interact with each other [9, 10]. Also, it is also pointed out that listeners use linguistic information in order to identify the speakers, and vice versa [9, 11].

One example of the interaction between the linguistic information and the speaker information is that different speech sounds are more or less effective for perceptual speaker identification [12]. This means that the accurateness of the identification depends on what sounds are presented to the listeners. In previous studies, sonorants, such as vowels and voiced consonants, were reported to be effective for perceptual speaker identification [13-19]. Especially, vowels and nasals are found to be effective [13, 14]. The same results were obtained in automatic speaker recognition tests [20-22].

Investigating the differential effects of the sounds on speaker identification enables us to know about the effects of the physical and physiological speaker variations in speech production, and at the same time, leads to a better understanding of human cognition. In this present study, we carried out two experiments in order to see the differences among the sounds in perceptual speaker identification tests. In the first experiment [23], the differences among the onset consonants in monosyllables which were excerpted from sentences were inspected, and in the second experiment [24], the stimuli with various syllabic structures were compared. The results showed that nasals and coronals in the onset position were effective for the identification.

2 Experiment 1

2.1 Methodology

Speakers and subjects. In human speaker identification tests, the selection of speakers and subjects is one of the most important and difficult tasks. The size of the speaker ensemble is concerned with the difficulty of the test task, and a homogeneous

subject group is also necessary for reliable data. Also the speakers' ages, genders and accents must be consistent [12].

Taking these things into account, we selected ten male speakers and five male subjects. All of them were undergraduate students at Sophia University, and they had lived in the same dormitory for more than four years. They were all native speakers of Japanese and none of them had hearing impairments.

Speech materials. The speech materials used in this study were CV Japanese monosyllables excerpted from the carrier sentences. The onset consonants were six oral consonants articulated at coronal area, and three nasal consonants /m/, /n/, and /ɲ/. The vowel was controlled to be /a/, because this vowel was the most effective vowel for speaker identification in previous studies [12-14, 16, 18], and also for making the experiment simple.

The recording sessions were held in a soundproof room. The speakers uttered the sentences shown in Table 1. All the materials were recorded onto a digital audiotape at the sampling frequency of 48 kHz with 16 bit resolution, using the electret-condenser microphone (SONY, ECM-MS957) and DAT recorder (SONY, TCD-D8).

The uttered sentences were /'aCaCaCa' to: o ʃiʒi ʃimasɯ̥/, as shown in Table 1. The carrier sentence means "I support 'aCaCaCa' (political) party," and the first four syllables, /'aCaCaCa'/, are the names of the fictional political parties. The symbol "a" is the Japanese /a/, and "C" stands for one of the following consonants: /t/, /d/, /s/, /z/, /r/, /j/, /m/, /n/ and /ɲ/. The reason for using the names of parties is that the suffix "/-to:/ (party)" forms compound words that do not have accentual nucleus (pitch fall) [25], thus /'aCaCaCa'/ is uttered with relatively stable accent pattern in the last two syllables.

The fourth syllables of the uttered sentences were manually excerpted for making the stimuli presented to the listeners. The excerption was conducted based on the waveform, using the computer software Cool Edit Ver.96 (Syntrillium Software Corporation).

Five tokens for each consonant were selected to be used in the test, and the total number of the stimuli was 450, i.e. corresponding to five tokens, nine consonants and ten speakers. The speech samples were randomly presented to the subjects, and a 500 ms portion of white noise was inserted before each stimulus in order to degrade the auditory memory of the preceding stimulus [26].

Procedures. The experiment was also conducted in the soundproof room. The subjects listened to the stimuli through binaural headphones (SONY, MDR-Z400) at a comfortable loudness level.

The subjects were first informed of the names of the ten speakers, listened to several sample files as to each speaker, and practised the task by use of these samples. These files were different from the samples used in the actual test, and the subjects listened to them and practised only once. During the test, 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. List of the recorded sentences in Experiment 1. 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/	

2.2 Results

The results of the identification test are shown in Table 2. Just as with the results in the previous experiments [13, 14], the nasals are the most effective sounds for the identification of the speakers, followed by the fricatives and the oral stops. Moreover, in the voiceless-voiced pairs of the same places and manners of articulation, /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 [13, 14, 18, 21].

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 ($p = 0.36$).

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

Stimulus	Percent Correct (%)
/na/	86.0 (215/250)
/ɲa/	85.6 (214/250)
/ma/	80.8 (202/250)
/za/	
/sa/	78.8 (197/250)
/ja/	78.4 (196/250)
/da/	78.0 (195/250)
/ɾa/	74.4 (186/250)
/ta/	73.6 (184/250)

3 Experiment 2

3.1 Methodology

In Experiment 1, the effectiveness of the nasals in monosyllabic stimuli was found out. However, only the nasals in the onset position were examined, and the nasals in the coda position were not dealt with. Moreover, the stimuli had an onset consonant followed by a nucleus vowel and therefore the effects of the vowel part or the transition to the following vowel were not inspected.

In Experiment 2, we carried out another speaker identification test in order to investigate the effects of the syllable structures and the contributions of the onset consonant and the transition to the vowel to the speaker identification.

Speakers and speech materials. Eight male students in the age range 22-25 years old (average 23.1 years old) served as the speakers in this experiment. All of them spoke Tokyo Japanese as their native language and had normal hearing.

The recording procedure was exactly the same as in Experiment 1. As shown in Table 3, the speech materials used in this experiment were Japanese non-sense monosyllables of various structures. 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 them 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 –VN, which were cut out from recorded CN and CVN. These two types were edited manually on the computer, using the software Praat [27]. 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 –VN contained the transition parts to the nucleus vowel. We will indicate it by the notation '–.'

Subjects. Eight students, two males and six females, who belonged to the same research group as the speakers participated in the experiment. 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.

Procedures. The procedure of the second perception test was almost the same as in the test in Experiment 1 except that the test sessions were performed on a computer. The subjects listened to the 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 stimuli and 5 different tokens for each stimulus. The total test time was about an hour, and the subjects took breaks after every 230 trials.

Table 3. List of the stimuli used in Experiment 2. Consonants in the parentheses were cut off manually from corresponding samples in CV and CVN.

Syllable structure	Stimuli
V	/a/
VV	/aa/
VN	/aN/
CV	/ba/ /da/ /ma/ /na/
CVV	/baa/ /daa/ /maa/ /naa/
CVN	/baN/ /daN/ /maN/ /naN/
–V	/(b)-a/ /(d)-a/ /(m)-a/ /(n)-a/*
–VN	/(b)-aN/ /(d)-aN/ /(m)-aN/ /(n)-aN/*

3.2 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.

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 scores of these onsetless syllables were the worst of all structures, though it still gained more than 90 % correct identification.

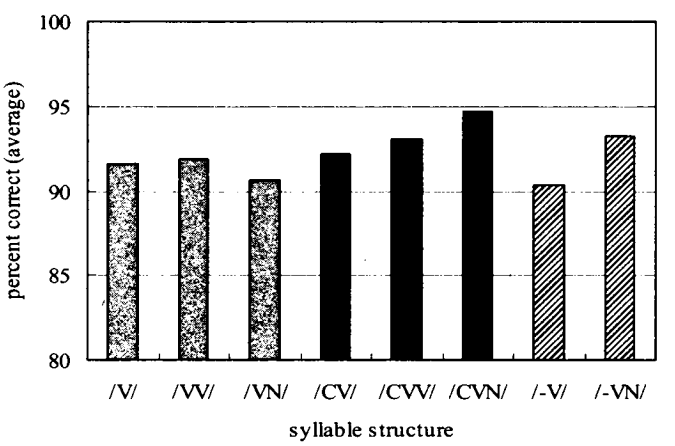

Fig. 1. Percentages of correct speaker identification (as to the syllable structure)

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.

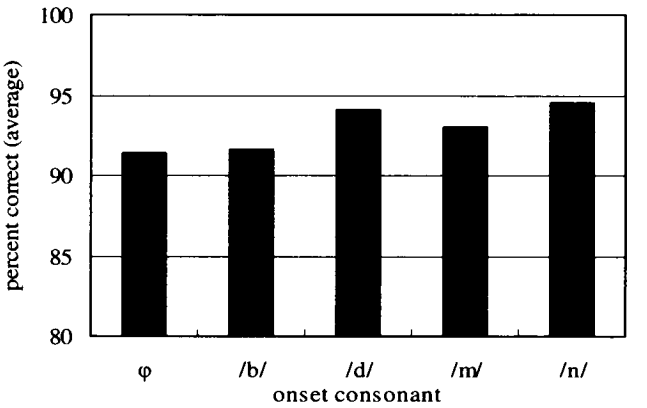

Fig. 2. Percentages of correct speaker identification (as to the onset consonant)

4 General Discussion

In this study, two perception experiments were conducted in order to investigate the differential effects among the stimulus contents on familiar speaker identification by listening.

In Experiment 1, ten male speakers were identified by five listeners who knew the speakers very well, and the identification rate was the highest when the stimuli containing the nasal sounds were presented. The difference between the nasal stimuli and the oral stimuli was significant. In Experiment 2, eight male speakers were identified by eight familiar subjects, and the effects of the syllable structures were examined. The results showed that the structures with onset consonant obtained higher identification scores than onsetless structures, and coda nasals were found to be effective for speaker identification. Furthermore, the identification was more accurate when the onset consonant was one of the nasals, and also the coronal consonants were better than the labial consonants.

In summary, the results in this study yield the following conclusions:

- Onset consonants are important for perceptual speaker identification.
- Nasals are effective for speaker identification both in onset and coda positions.
- Coronal consonants convey more individuality than labial consonants.

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 of the stimuli.

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 [28]. In addition, the shapes of these resonators cannot be changed voluntarily. This means that the resonance 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 necessarily 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.

Japanese coda nasal /N/ in the word-final position has been said to be articulated at the uvula, but recent work [29] reports that the place of articulation of /N/ differs among speakers, and this sound is not always uvular. This variation among the speakers explains the results in this study, too.

Coronal consonants. This tendency is what was seen in our previous experiment, too [14]. 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 among the stop sounds. This may lead to inter-speaker differences in articulation of alveolar sounds.

Phonological unmarkedness and speaker individuality. It is interesting that the effective structures or the sounds shown above are all linguistically unmarked ones. The structures with onset consonants are universally unmarked [30]; vowels and nasals are the sounds that children learn in the early stage of language acquisition. There is also a typologically universal tendency that coronal is the dominant place of articulation compared to labial and guttural [31].

The relationship between the linguistically unmarked structures/sounds and speaker individuality has not been clarified yet, but the obtained results here imply that unmarked structures/sounds are more effective for perceptual speaker identification than other structures/sounds. One reason for this may be that unmarked sounds occur more frequently in natural language.

The final goal of our study is to delimit the speaker individuality conveyed by speech sounds 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 explaining the effectiveness of the phonologically unmarked sounds. We must also test on different kinds of vowels, in order to examine the effects of co-articulation. Speaker identification experiments with reversed speech may also be useful for revealing the properties of human perception. As to the nasal sounds, the acoustic properties 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.

Acknowledgments. This work was supported by a Grant-in-Aid for Scientific Research (A) 16203041, and by a Grant-in-Aid for JSPS Fellows (17·6901).

References

1. Abercrombie, D.: Elements of General Phonetics. Edinburgh Univ. Press, Edinburgh (1967)
2. Laver, J.: The Phonetic Description of Voice Quality. Cambridge Univ. Press, Cambridge (1980)
3. Ball, M., Rahilly, J.: Phonetics. Arnold, London (1999)
4. Hollien, H.: The Acoustics of Crime. Plenum, New York (1990)
5. Furui, S.: Acoustic and Speech Engineering. Kindai Kagaku-sha Publishing Company, Tokyo (1992)
6. Hashimoto, M., Kitagawa, S., Higuchi, N.: Quantitative Analysis of Acoustic Features Affecting Speaker Identification. J. Acoust. Soc. Jpn. 54(3), 169–178 (1998)
7. Kuwabara, H., Sagisaka, Y.: Acoustic Characteristics of Speaker Individuality: Control and Conversion. Speech Com. 16, 165–173 (1995)
8. O'Shaughnessy, D.: Speech Communications –Human and Machine–. 2nd edn. Addison-Wesley Publishing Company, New York (2000)
9. Nygaard, L.: Perceptual Integration of Linguistic and Nonlinguistic Properties of Speech. In: Pisoni, D., Remez, R. (eds.) The Handbook of Speech Perception, pp. 390–413. Blackwell Publishing, Oxford (2005)
10. Kreiman, J., van Lacker, D., Gerratt, B.: Perception of Voice Quality. In: Pisoni, D., Remez, R. (eds.) The Handbook of Speech Perception, pp. 338–362. Blackwell Publishing, Oxford (2005)
11. Goggin, J., Thompson, C., Strube, G., Simental, L.: The Role of Language Familiarity in Voice Identification. Memory and Cognition 19, 448–458 (1991)
12. Bricker, P., Pruzansky, S.: Speaker Recognition. In: Lass, N. (ed.) Experimental Phonetics, pp. 295–326. Academic Press, London (1976)
13. Amino, K.: The Characteristics of the Japanese Phonemes in Speaker Identification. Proc. Sophia Univ. Ling. Soc. 18, 32–43 (2003)
14. Amino, K.: Properties of the Japanese Phonemes in Aural Speaker Identification. Tech. Rep. IEICE 104(149), 49–54 (2004)
15. Bricker, P., Pruzansky, S.: Effects of Stimulus Content and Duration on Talker Identification. J. Acoust. Soc. Am. 40(6), 1441–1449 (1966)
16. Kitamura, T., Akagi, M.: Speaker Individualities in Speech Spectral Envelopes. J. Acoust. Soc. Jpn (E) 16(5), 283–289 (1995)
17. Matsui, T., Pollack, I., Furui, S.: Perception of Voice Individuality Using Syllables in Continuous Speech. In: Proc. Autumn Meet. Acoust. Soc. Jpn, pp. 379–380 (1993)
18. Nishio, T.: Can We Recognise People by Their Voices? Gengo-Seikatsu 158, 36–42 (1964)
19. Stevens, K., Williams, C., Carbonell, J., Woods, B.: Speaker Authentication and Identification: A Comparison of Spectrographic and Auditory Presentations of Speech Material. J. Acoust. Soc. Am. 44(6), 1596–1607 (1968)
20. Nakagawa, S., Sakai, T.: Feature Analyses of Japanese Phonetic Spectra and Consideration on Speech Recognition and Speaker Identification. J. Acoust. Soc. Jpn. 35(3), 111–117 (1979)
21. Ramishvili, G.: Automatic Voice Recognition. Engineering Cybernetics 5, 84–90 (1966)
22. Sambur, M.: Selection of Acoustic Features for Speaker Identification. Proc. IEEE Trans. ASSP 23(2), 176–182 (1975)

23. Amino, K., Sugawara, T., Arai, T.: Correspondences between the Perception of the Speaker Individualities Contained in Speech Sounds and Their Acoustic Properties. In: Proc. Interspeech, pp. 2025–2028 (2005)
24. Amino, K., Sugawara, T., Arai, T.: Effects of the Syllable Structure on Perceptual Speaker Identification. IEICE Tech. Rep. 105(685), 109–114 (2006)
25. Kindaichi, H., Akinaga, K. (eds.): Meikai Japanese Accent Dictionary, 2nd edn. Sanseido, Tokyo (1981)
26. Repp, B., Healy, A., Crowder, R.: Categories and Context in the Perception of Isolated Steady-State Vowels. J. of Exp. Psychol.: Human Perc. Perf. 5(1), 129–145 (1979)
27. Boersma, P., Weenik, D.: Praat Doing Phonetics by Computer. Ver.4.3.14 (Computer Program) (2005), retrieved from http://www.praat.org/
28. Dang, J., Honda, K.: Acoustic Characteristics of the Human Paranasal Sinuses Derived from Transmission Characteristic Measurement and Morphological Observation. J. Acoust. Soc. Am. 100(5), 3374–3383 (1996)
29. Hashi, M., Sugawara, A., Miura, T., Daimon, S., Takakura, Y., Hayashi, R.: Articulatory Variability of Japanese Moraic-Nasal. In: Proc. Autumn Meet. Acoust. Soc. Jpn, pp. 411–412 (2005)
30. Spencer, A.: Phonology. Blackwell, Oxford (1996)
31. Whaley, L.: Introduction to Typology. Sage Publications, London (1997)