euronoise

Differential effects of the phonemes on identification of previously unknown speakers

K. Amino and T. Arai

Dept. of Electrical and Electronics Engineering, Sophia University, 7-1 Kiyoi-cho, Chiyoda-ku, 102-8554 Tokyo, Japan
amino-k@sophia.ac.jp

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 human speaker identification, and that coronal sounds are more effective than labial ones. This tendency was observed in the identifications of both familiar and previously unknown speakers. In this present study, we investigated the effects of the speech contents again, by focusing on the vowels in CV monosyllabic stimuli. Through the experiment we obtained several findings: 1) stimuli containing a nasal gained significantly higher accuracy compared to stimuli with only oral sounds; 2) coronal sounds were more effective than labial sounds; 3) palatalisation caused an improvement in performance; and 4) back vowels were more effective than front vowels significantly. These results can be explained by speaker-specific morphologies of the nasal cavity and the paranasal sinuses, and also of the pharyngeal cavity. We will also show analyses and discussions on the acoustical properties of the stimuli and the performance differences among the subjects.

1 Introduction

Human speech sounds convey both linguistic and non-linguistic information, and these two are known to interact with each other [1]. One of the examples is the interaction between the phonological contents and the speaker identity of an utterance. It is reported that the intelligibility of an utterance increased with the familiarity to the speakers [2].

Also, it is known that the accuracy of perceptual speaker identification depends on the phonological contents of the stimuli presented to the listeners [3, 4]. In previous studies, vowels and voiced consonants have been reported to yield high speaker identification rates [5, 6, 7].

Some studies report the availability of the liquids in speaker recognition [8, 9]. In our previous experiments, we conducted several speaker identification tests with various monosyllabic stimuli, and we found that the stimuli containing a nasal sound were effective for the judgment of the speaker identity [10-14]. Summary of our experiments are shown in Table 1. The tendency that the nasals are effective was observed despite the different sets of speakers and listeners, the syllabic structures, and the familiarity to the speakers.

Availability of the nasals is shown in automatic speaker recognition, too [15, 16]. Spectral individualities of the nasals were also observed in our study [12]. Inter- speaker cepstral distances were greatest in nasals.

However, we notice in Table 1 that the effects of the following vowels on the availability of the nasals were not yet examined. In this present study, we conducted another speaker identification experiments where the CV monosyllabic stimuli of various consonants and vowels were investigated.

2 Experiment

2.1 Participants and Speech Materials

Fifteen listeners identified the four male speakers. All of the speakers and listeners were native speakers of Japanese, and had no known hearing problems. Mean age of the listeners was 23.4 years old.

The speakers were unknown to the listeners. Information on the speakers is shown in Table 2. The speakers were selected from JEIDA (Japanese Electronic Industry Development Association) speech corpus [17]. Out of one hundred speaker entries, these four were selected because they all spoke the Japanese of Tokyo dialects, and their recordings were made in relatively quiet environments.

Forty-eight monosyllables shown in Table 3 were selected from the speech corpus and used in the experiment. Thirteen coronal consonants were selected in accordance with our previous experiments [10-14], and used in combination with phonotactically possible vowels.

Reference	No. of Speakers and Listeners	Speaker Sex	Familiarity	Stimulus syllables	Effective sounds
[10]	3, 14	Female	Familiar	CV monosyllables (isolated)	Nasals
[11]	3, 18	Both	Familiar	CV monosyllables (excerpted)	Nasals and voiced coronal consonants
[12]	10, 5	Male	Familiar	CV monosyllables (excerpted)	Nasals
[13]	8, 8	Male	Familiar	CV, CVV, CVN, V, VV, VN monosyllables (isolated)	Nasals in onset / coda positions
[14]	10, 16	Male	Unknown	CV monosyllables (excerpted)	Nasals

Table 1 Summary of our previous experiments

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

Table 2 Four male speakers of this experiment and their mean F0; mean F0 represents the average values of all the utterances used in this experiment.

Three tokens for each were used as the stimuli. The affricates /t͡ʃ/, /t͡s/, /d͡ʒ/ and /d͡z/ appears as allophones for /t/ and /d/, when they are combined with high vowels, /i/ and /ɯ/. Also, word-initial /z/ may be realised as an affricate /d͡z/. The voiceless fricative /s/ are realised as palatal fricative /ʃ/ in combination with /i/. Thus, we had different numbers of syllables for each of the consonants.

2.2 Procedure

Since the listeners had not known the speakers, they got familiarised with the speakers before starting the experiment. In the familiarisation session, the listeners heard each speaker uttering three sample words; /horʲɯː/ (保留, suspension), /kaigʲoː/ (改行, creating a new line), and /heɴkaɴ/ (変換, conversion). These words were selected from the same corpus, on the basis that they do not contain any of the stimulus syllables. Participants listened to these three words of the four speakers as many times as they wanted.

After they showed some confidence, they practised the experimental task using these sample words. All the sound files were presented to the listeners on a computer through headphones (SONY MDR-Z 700). Feedback was given after each trial. We repeated the familiarisation and practice sessions until the participants could tell the speakers at more than 90% accuracy.

Then test session followed, and this time the listeners identified the speakers by monosyllabic stimuli. The experimental task was conducted with Praat MFC (Multiple Forced Choice) programme [18]. No feedback was given, and no replays of the stimuli or access to the sample words were allowed during the test. The listeners answered a speaker who they thought the stimulus belonged to, and evaluated the degree of confidence for each trial. Confidence was rated by four degrees, where scale 1 indicated no confidence and scale 4 showed confidence.

The total number of trials was 576, that is corresponding to forty-eight syllables, three tokens for each and four speakers. The participants took breaks after every 192 trials. The total experiment took them about an hour.

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/ /d͡ʒ/ /d͡z/	-	-	/t͡ʃi/ /d͡ʒi/	-	/t͡sɯ/ /d͡zɯ/
Nasals	/m/	/ma/	/me/	/mi/	/mo/	/mɯ/
	/n/	/na/	/ne/	/ni/	/no/	/nɯ/
	/ɲ/	/ɲa/	-	-	/ɲo/	/ɲɯ/
Approximants	/j/	/ja/, /wa/	-	-	/jo/	/jɯ/

Table 3 Stimulus Monosyllabes

3 Results and Discussion

3.1 Effects of the stimulus contents

The results of the experiment were evaluated by percent correct speaker identification. The identification rates according to the consonants and the vowels of the stimuli are shown in Figures 1(a) and 1(b), respectively. As can be seen, all of the consonants and vowels obtained higher scores than the chance level, which is 25% correct.

In figure 1(a), we can see that coronal nasals /n/ and /ɲ/ yielded the best identification performances. The voiced alveolar stop /d/, and then the coronal fricatives /s/ and /ʃ/ followed them. Palatalised sounds, /ʃ/ and /ɲ/, were slightly better than their non-palatal counterparts, /s/ and /n/.

Results of the one-way ANOVA as for the consonant showed a significant tendency (F (11, 575), p = 0.058). The difference between nasal and non-nasal consonants was significant in Mann-Whitney U-test (p = 0.045). The difference between sononrants and obstruents in the same test was not significant (p = 0.85).

The syllables without a consonant obtained the lowest scores, and this was also seen in our previous experiment [12]. Effectiveness of the nasals and coronal-labial asymmetry in nasals have also been consistently observed in our previous works [10-14]. It is reported that phonemic variations in the stimuli are more important for speaker individuality than the duration of the stimuli [7, 19, 20], and this explains the worse performances with the onsetless syllables. Nasals are effective for identifying speakers, because the resonance cavities involved in nasal articulation have morphological variations among speakers [21], thus their resonance feature reflect individual differences [13].

Among the five vowels, the back vowels /a/, /o/ and /ɯ/ were more effective for identifying the speakers than the front vowels /i/ and /e/. The difference between back and front vowels was significant in Mann-Whitney U-test (p = 0.003). The difference between open and close vowels was not significant in the same test (p = 0.36).

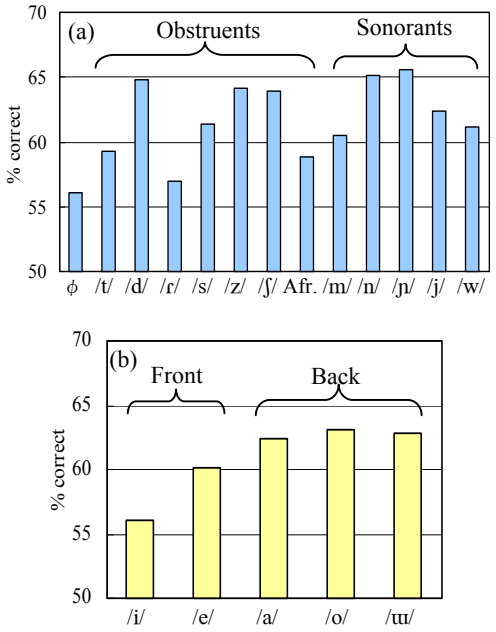

Fig.1 Results of the speaker identification experiment; (a) according to the consonants; (b) according to the vowels.

The difference between the back and front vowels is that the former has its second formant in lower frequency region. When we look at the four speakers first and second formant frequencies for the five vowels shown in Table 4, the boundary in the second formant frequency is around 1500 Hz.

The effects of the formant frequencies in the higher bands are reported in some studies [22, 23, 24]. However, low second frequencies appeared to be effective for perceptual speaker identification in this study, although the reason for this is not clear.

3.2 Participant factors

Confusion matrix among speakers is shown in Table 5. We can see that speaker #3 gained the highest identification score, and the confusion between the speakers #2 and #4 occurred most frequently. Both of these can be explained by the fundamental frequencies.

Formant frequencies [Hz]	Speaker #1		Speaker #2		Speaker #3		Speaker #4	
	F1	F2	F1	F2	F1	F2	F1	F2
/i/	278.2	2297.6	246.2	2356.2	309.1	2245.9	251.5	2443.7
/e/	432.7	2073.8	411.4	2079.1	481.8	2005.0	457.8	2226.4
/a/	854.9	1244.8	784.4	1274.6	956.2	1148.35	829.3	1267.4
/o/	432.7	864.3	411.4	800.34	482.5	809.5	442.9	844.2
/ɯ/	283.5	1375.8	256.9	1045.4	324.0	1240.5	299.3	893.7

Table 4 First and second formant frequencies of the four speakers; averaged among three tokens; analysed manually from the FFT spectra and the spectrograms using the computer software Praat [18]

Speaker	#1	#2	#3	#4
#1	63.4	13.4	11.9	10.7
#2	25.3	46.5	6.9	32.3
#3	8.2	0.8	79.4	0.9
#4	3.1	39.3	1.8	56.1

Table 5 Confusion matrix among the four speakers; percent response of perceived speakers (shown in the column) for the actual speakers (shown in the row)

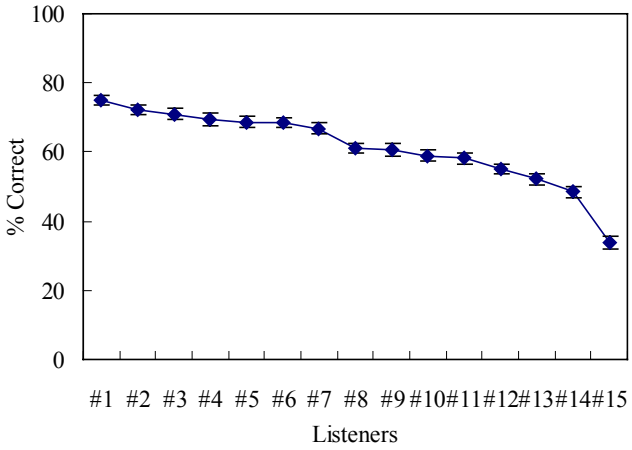

Fig.2 Each listener's speaker identification performances; average percent correct identification and the standard deviation; the whiskers indicate +/- 1SD.

The importance of the average F0 in speaker identification is pointed out both in automatic speaker identification [22, 25] and in human speaker identification [26, 27]. In this study, too, the effect of the average fundamental frequency was observed.

As for the listener factors, Figure 2 shows the performance differences among the listeners in the speaker identification experiment. The difference among the listeners was significant (F (719, 14), $p < .001$). We can see that the average percent correct identification ranged from 75% and 34% correct. The ability difference among individuals is pointed out in previous study [3], although the listener sets of more than twelve people can be considered to yield a typical result [28, reviewed in 3].

3.3 Confidence evaluations

Correlation between evaluated confidence values and the accuracy of the performances was examined. The ratings of scales 1 and 2 were considered as "not confident," and 3 and 4 as "confident." Correlation analysis was performed between the identification accuracy and the confidence, and the results showed a significant correlation.

Unlike face recognition, there are the reports on the correlation between performance and confidence in speaker recognition [29], and the results of this experiment did not contradict them.

4 Conclusions

In this study, perceptual speaker identification experiment was conducted in order to investigate the effects of the stimulus contents on the identification performances. Especially, we focused on the effects of the nucleus vowels in monosyllabic stimuli.

Four male speakers were identified by fifteen listeners who had not known the speakers before. The results showed the effectiveness of the coronal nasals, which has been observed consistently in our previous experiments [10-14]. Among the five Japanese vowels, the back vowels were significantly better than the front vowels. Nasals are effective, because they accompany resonances in nasal cavity and paranasal sinuses that are idiosyncratic in their morphology [21]. The effectiveness of the back vowels remains unexplained, but the lower second formant frequencies may be a clue for it.

As pointed out in previous research [22, 25, 26, 27], there was a great influence by the average fundamental frequency in the confusion among speakers. The difference in listener's ability to identify speakers was significant, and correlation was observed between the speaker identification performance and the listener's confidence for the judgments. These tendencies were also reported in previous experiments [28, 29].

Acknowledgments

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

References

[1] 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.), 390-413, Blackwell Publishing, Oxford (2005)

[2] J. Goggin, C. Thompson, G. Strube, and L. Simental, "The role of language familiarity in voice identification", *Memory and Cognition* 19, 448-458 (1991)

[3] P. Bricker and S. Pruzansky, "Speaker recognition", Chap. 9 in *Experimental Phonetics*, N. Lass (Ed.), 295-326, Academic Press, London (1976)

[4] I. Pollack, J. M. Pickett, and W. H. Sumby, "On the identification of speakers by voice", *J. Acoust. Soc. Am.* 26, 403-406 (1954)

[5] 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.* 44, 1596-1607 (1968)

[6] G. Ramishvili, "Automatic voice recognition", *Engineering Cybernetics* 5, 84-90 (1966)

[7] P. Bricker and S. Pruzansky, "Effects of stimulus content and duration on talker identification", *J. Acoust. Soc. Am.* 40, 1441-1450 (1966)

[8] F. Nolan, *The Phonetic Basis of Speaker Recognition*, Cambridge Studies in Speech Sci. and Comm., Cambridge (1983)

[9] C. Zhang, J. van de Weijer and J. Cui, "Intra- and inter-speaker variations of formant pattern for lateral syllables in standard Chinese", *Forensic Sci. Int.* 158, 117-124 (2006)

[10] K. Amino, "The characteristics of the Japanese phonemes in speaker identification", *Proc. Sophia Univ. Linguistic Soc.* 18, 32-43 (2003)

[11] K. Amino, "Properties of the Japanese phonemes in aural speaker identification", *IEICE Tech. Rep.* 104, 49-54 (2004)

[12] K. Amino, T. Sugawara, and T. Arai, "Effects of the syllable structure on perceptual speaker identification", *IEICE Tech. Rep.* 105, 109-114 (2006)

[13] K. Amino, T. Sugawara and T. Arai, "Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties", *Acoust. Sci. Tech.* 27 233-235 (2006)

[14] K. Amino and T. Arai, "Effects of stimulus contents and speaker familiarity on perceptual speaker identification", *Acoust. Sci. Tech.* 28, 128-130 (2007)

[15] L.S. Su, K.P. Li and K.S. Fu, "Identification of speakers by use of nasal co-articulation", *J. Acoust. Soc. Am.* 56, 1876-1882 (1972)

[16] S. Nakagawa and T. Sakai, "Feature analyses of Japanese phonetic spectra and considerations on speech recognition and speaker identification", *J. Acoust. Soc. Jpn.* 35, 111-117 (1979)

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

[18] P. Boersma and D. Weenink, "Praat: doing phonetics by computer, version 4.5.14", Retrieved from "http://www.praat.org/", Computer programme (2005)

[19] H. Hollien, *Forensic Voice Identification*, Academic Press, San Diego (2002)

[20] R. Roebuck and J. Wilding, "Effects of vowel variety and sample length on identification of a speaker in a line-up", *Appl. Cogn. Psychol.* 7, 475-481 (1993)

[21] J. Dang and K. Honda, "Acoustic characteristics of the human paranasal sinuses derived from transmission characteristics measurement and morphological observation", *J. Acoust. Soc. Am.* 100, 3374-3383 (1996)

[22] M. Sambur, "Selection of acoustic features for speaker identification", *IEEE Trans. ASSP.* 23, 176-182 (1975)

[23] P. Ladefoged, *A Course in Phonetics*, 4th Ed., Heinle and Heinle, Boston (2001)

[24] S. Hayakawa and F. Itakura, "The influence of noise on the speaker recognition performance, using the higher frequency band", *Proc. ICASSP* 1, 321-324 (1995)

[25] H. Kuwabara and Y. Sagisaka, "Acoustic characteristics of speaker individuality: control and conversion", *Sp. Comm.* 16, 165-173 (1995)

[26] M. Hashimoto, S. Kitagawa and N. Higuchi, "Quantitative analysis of acoustic features affecting speaker identification", *J. Acoust. Soc. Jpn.* 54, 169-178 (1998)

[27] T. Kitamura and T. Saitou, "Effects of acoustic modification on perception of speaker characteristics for sustained vowels", *Acoust. Sci. Tech.* 28, 434-437 (2007)

[28] C.E. Williams, "The effects of selected factors on the aural identification of speakers", in Sect.3 of *Report ESD-TDR-65-153 Electronics Systems Division*, Air Force Systems Command, Hanscom Field (1964)

[29] H. Kasahara and K. Ochi, "In the confusion matrix among speakers", *Tech. Rep. of Jpn. Cog. Sci. Soc.* 60 (2007)