Perceptual Speaker Identification Using Monosyllabic Stimuli

- Effects of the Nucleus Vowels and Speaker Characteristics Contained in Nasals -

Kanae Amino and Takayuki Arai

Department of Electrical and Electronic Engineering, Sophia University, Tokyo, Japan

amino-k@sophia.ac.jp, arai@sophia.ac.jp

Abstract

The goal of our research is to find out the acoustical correlates of human perception of speaker identity. In this study we investigated the effects of the stimulus contents on perceptual speaker identification. Forty-eight monosyllables were used as the stimuli for identifying four male speakers. The results showed that the syllables containing a coronal nasal yielded higher identification accuracies than the syllables without it, and the syllables with a back vowel gained significantly better scores than those with a front vowel. We also found speaker-dependent characteristics in the velar movements in articulation of nasal consonants.

Index Terms: perceptual speaker identification, speaker's individuality, nasals, vowels, energy onset

1. Introduction

In human communication, we perceive and recognise various kinds of information conveyed by speech signals. It is not disputed that information carried by human speech includes not only the phonological information but also the information about the speakers.

In some studies, phonological and speaker information are thought to be independent from each other [1-2], though, on the other hand, there are interactions between them [3-4]. An example of this interaction is the differential effects of the stimuli on perception of the speaker identity. Specifically, speaker identification performances are influenced by the phonemic variations of the stimuli, and it is known that vowels and voiced consonants are relatively more effective for accurate speaker identification [5-6].

In our previous experiments, we conducted perceptual speaker identification tests using monosyllabic stimuli, and we found that the stimuli containing a nasal sound obtained consistently higher identification rates than oral stimuli, in both familiar [7-10] and unknown [11] speaker identifications, and in both syllable onset and coda positions [10]. However, in those previous research designs, only one vowel /a/ was examined in order to make the experiments simple. Even the monosyllables are subject to coarticulation in spontaneous speech, and effects of the following vowels on the acoustical properties of the syllable onset consonants are expected.

In this study, we carried out another set of speaker identification experiments using all the five vowels of Japanese. We also conducted acoustical analyses of the stimuli, in order to find speaker-dependent characteristics.

2. Experiment

2.1. Speech materials

Speech materials of four male speakers were selected from JEIDA Japanese Common Speech Data Corpus [12] and used in the experiment. Information on the speakers is shown in Table 1. In this table, information on speakers' ages and heights were cited from the tutorial of the JEIDA corpus, and the mean fundamental frequencies (f0) and their standard deviations were analysed by the authors. The analyses were performed manually using Praat [13]. The mean f0 is the average value as for the vowel portions of all the stimulus monosyllables. These four speakers were selected, because they are all native speakers of Tokyo Japanese, and recordings of their speech were held in a quiet room compared to other speakers in the corpus.

Out of 110 entries of monosyllables in JEIDA corpus, we selected 48 syllables. First, we selected the following fourteen consonants taking into account the consistency with our previous experiments [7-11]; /t/, /d/, /ɾ/, /s/, /z/, /t͡ʃ/, /t͡s/, /d͡ʒ/, /d͡z/, /m/, /n/, /ɲ/, /j/, and /w/. Then all the phonotactically possible syllables containing those consonants were employed. The list of the syllables is presented as Table 2. All the sound materials were digitised at the sampling frequency of 48 kHz with 16 bit resolution.

Japanese has five vowels, that is, three back vowels, /a/, /o/, and /ɯ/, and two front vowels, /i/ and /e/. When the close vowels follow the fricatives /s/ and /z/, and the stops /t/ and /d/, the consonants are realised as their allophones, and become either postalveolar consonant, /ʃ/, or the affricates, /t͡ʃ/, /t͡s/, /d͡ʒ/, and /d͡z/.

Three tokens for each syllable uttered by each of the speakers were used in the experiment. The total number of the stimulus syllables was 576, that is corresponding to 48 monosyllables, three tokens, and four speakers.

2.2. Procedures

Fifteen volunteers participated as the listeners in the perception tests. None of them had heard the four speakers' speech before. They were all native speakers of Japanese, and their mean age was 23.4 years old. No one had known hearing problems.

All the speech materials were played on a computer through headphones (SONY MDR-Z700). First, the participants listened to the sample words of each speaker. The sample words were: /hoɾʲɯː/ (保留, suspension), /kaigʲoː/ (改行, creating a new line), and /heɴkaɴ/ (変換, conversion). These words were again selected from the JEIDA corpus on the basis that they do not contain a syllable that is used as the stimuli in the experiment.

Since the speakers were unknown to the participants, the participants had to get familiarised with the four speakers first. They listened to the sample words introduced above as many times as they wanted. After the participants showed some confidence, they practised the experimental task using the same sample words. We repeated familiarisation and practice

until the participants could tell the speakers with more than 90% accuracy.

Test session followed after the practice. The stimuli were presented pseudo-randomly by using Praat Multiple Forced Choice programme [13]. The participants listened to the stimuli and answered the speakers by IDs. No replays were allowed, and sample words were no longer accessible once the test session began.

2.3. Results and Discussion

Speaker identification results according to the consonants and the vowels are shown in Figures 1(a) and 1(b), respectively. We find that for all the consonants and vowels identification performances were better than the chance level (25% correct).

In Figure 1(a), the following tendencies can be seen:

- Performances with onsetless syllables were the worst.
- Coronal nasals, /n/ and /ɲ/, obtained the highest scores, though bilabial nasal, /m/, did not.
- Voiced consonants, /d/ and /z/, were more effective than their voiceless counterparts, /t/ and /s/.
- Palatalised sounds, /ʃ/ and /ɲ/, were better than their alveolar counterparts, /s/ and /n/, though the difference is slight in the latter case.

All of these tendencies above were also found in our previous experiments [7-11]. Differences among the consonants showed a significant tendency in one-way ANOVA (p = 0.058), and the difference between nasal and non-nasal consonants was significant in Mann-Whitney U-test (p = 0.045).

From the results for the vowels, illustrated in Figure 1(b), we gained the following outcomes:

- Back vowels gained higher identification accuracies than front vowels.
- The feature [±back] is more important for the perception of the speaker identity than [±high].

Front-back difference among the vowels was significant in Mann-Whitney U-test (p = 0.003), although high-low difference was not significant in the same test.

Effects of the following vowels on the results of the preceding consonants are illustrated in Figure 2 as for the nasal consonants. Again, the tendency that the back vowels are more effective than the front vowels was proved, regardless of the places of articulation.

The syllables containing nasal consonants were effective because they may reflect speaker's anatomical characteristics more than the syllables with only oral sounds. Nasal consonants are similar to approximants in that they have an uninterrupted airflow that does not pass through a constriction [14]. This makes the nasals have both source and resonance characteristics. The production of nasals involves resonances in nasal cavity, velopharyngeal cavity and paranasal sinuses. Morphological individualities of these cavities are reported in previous research [15]. Also, in our previous studies [9, 11], there were greater inter-speaker variations in the spectral properties of the nasals compared to those of non-nasals, and the inter-speaker cepstral distances correlated with perceptual confusions among the speakers.

On the other hand, articulation of nasal consonants is similar to oral stops. The only difference is that nasals have another pathway, the nasal tract, and the articulation of nasal sounds involves the raising and lowering of the velum. The movements of the velum are difficult to control in a brief interval, though the movement itself is voluntary [16]. Hence, the timing of the velar movements related to nasal articulation may differ among speakers. In order to examine the individualities in the timing of velic actions, we conducted acoustical analyses.

Table 1. *Speaker ensemble*

Speaker ID	Sex	Age	Height [cm]	Mean f0 [Hz]	*S.D.* [Hz]
#1	Male	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. *List of the stimulus monosyllables*

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

3. Acoustical Analyses

3.1. Methods

In order to find speaker-specific characteristics in nasals, stimuli used in the experiment, containing nasal consonants /m/ and /n/, were analysed. The analysis targets were the following ten monosyllables: /mi/, /me/, /ma/, /mo/, /mɯ/, /ni/, /ne/, /na/, /no/ and /nɯ/. Three tokens for each were used in the analysis.

The analysis parameter was the transition of the energy across the time. This parameter was selected because it captures abrupt spectral change well [17], thus we thought it reflects velar movements in nasal-vowel transitions.

First, the stimulus syllables were down-sampled from 48 kHz to 16 kHz. Then we calculated the energy for each stimulus syllable by frames of 30 ms length with a shift of 10 ms. The energy vector for each syllable was normalised by the total energy, and was plotted across time as in Figure 3.

3.2. Results and discussion

Figure 3 shows the energy transition contours for each speaker's utterances containing either /m/ (above) or /n/ (below). For all speakers, the contours seem to be quite reproducible. In speakers #1, #2 and #4, we can see that intra-speaker variations are smaller in /n/ than in /m/. This may explain the difference between the labial and coronal nasals in the effectiveness in perceptual speaker identification.

In Figure 3, movements of the velum are reflected in the left side of the curve, or in the energy onset. We calculated the linear approximations of the energy onsets, and compared their slopes among the speakers. Summary of the slope analysis is shown in Table 3.

One-way ANOVA showed a significant difference among the four speakers' slope values. This implies that the controls of velar movements in syllable onset nasals are one of the speaker-specific characteristics. The analysis targets here are monosyllables, and we can see in Figure 3 that the velic action occurs in relatively short durations, in the range of 50-100 ms. It is known that the controls of intentional movements may take 70-100 ms [18]. This seems to indicate that the movement of the velum is out of the control by the brain for some speakers. Therefore, nasal articulation may reflect speakers' characteristics, especially the individual habits in the control of the velum. Another possibility is that the velar movements reflect speakers' physiological properties. The velar movements may be determined by the organic properties of the velum itself, that is, its mass and elasticity.

Small intra-speaker variations, greater inter-speaker variations, and factors that cannot easily be controlled by the speakers are all important factors in speaker identification [19]. In this sense, velar movements in nasal articulation satisfy all these three criteria.

4. Summary and Conclusions

In this study, we conducted perceptual speaker identification experiment in order to see the differential effects of the stimulus contents on the accuracy of the identification performances. The results showed that coronal nasals and back vowels were effective for identifying speakers. Coronal nasals have been always the most effective sounds in the series of our previous experiments despite different sets of speakers and listeners [7-11]. Spectral differences in nasal sounds were also greater than those in oral sounds in our previous research [9, 11].

This time we performed analyses on the transition of energy focusing on the syllable onset nasals. The slopes of the energy contours varied significantly among speakers, and it was implied that the velar raising movements show speaker individualities.

Our future tasks are to find more elegant and suitable ways to capture inter-speaker differences in nasal articulation, and to investigate the energy transitions in intervocalic and postvocalic nasals. These will lead to find a way to discriminate among speakers by focusing on nasals. The availability of the back vowels has not been explained yet. Also, it is predictable that human beings exploit not only a single cue but several cues in order to identify speakers. For a better understanding of the human perception of speaker identity, we need to investigate the strategies of human perception and recognition.

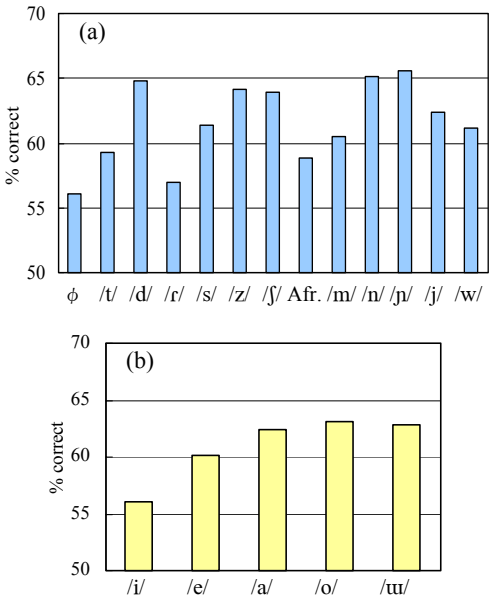

Figure 1: *Speaker identification accuracies (percent correct); (a) according to the syllable onset consonant; (b) according to the nucleus vowels.*

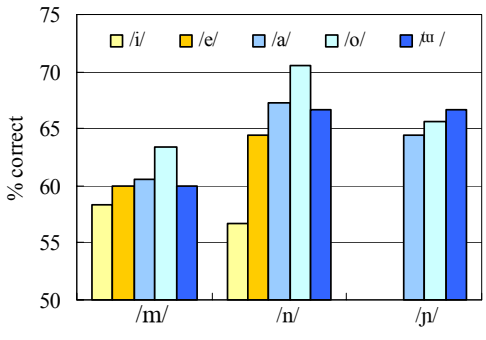

Figure 2: *Speaker identification accuracies (percent correct), according to nasals and the following vowels.*

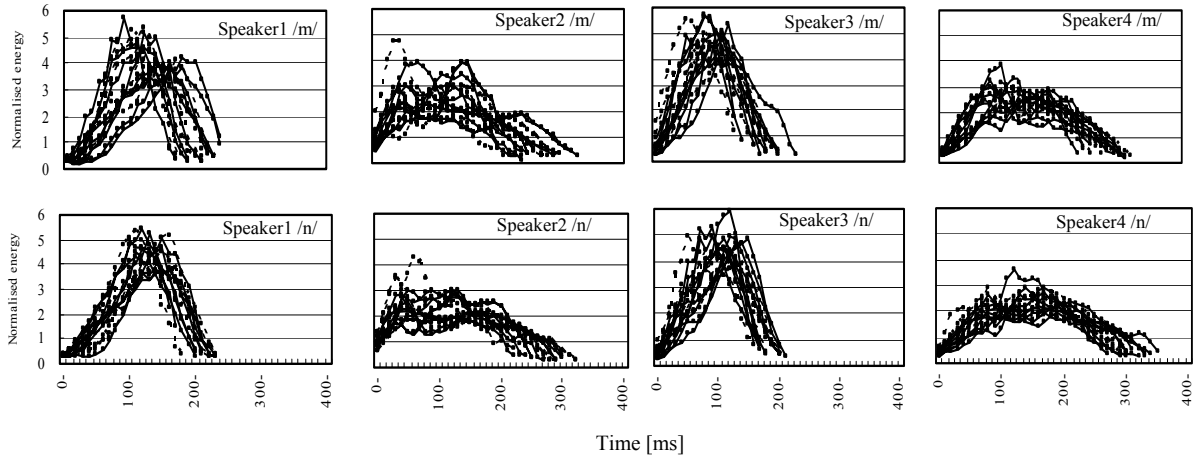

Figure 3: *Energy transition contours of the four speakers; contours of the syllables containing /m/ (above) and /n/ (below).*

Table 3. *Mean slope values of energy onset curves.*

Speaker ID	Mean slope	*S.D.*
#1	0.148	0.064
#2	0.011	0.125
#3	0.248	0.128
#4	0.010	0.077

5. Acknowledgements

This study was supported by Sophia University Open Research Centre.

6. References

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

[2] Bricker, P. and Pruzansky, S., "Speaker recognition," in Lass, N. [Ed.], Experimental Phonetics, 295-326, Academic Press, London, 1976.

[3] Nygaard, L., "Perceptual integration of linguistic and nonlinguistic properties of speech," in Pisoni, D., and Remez, R. [Eds], The Handbook of Speech Perception, 390-413, Blackwell Publishing, Oxford, 2005.

[4] Pollack, I., Pickett, J.M., and Sumby, W.H., "On the Identification of Speakers by Voice," J. Acoust. Soc. Am., 126: 403-406, 1954.

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

[6] Ramishvili, G., "Automatic Voice Recognition," Engineering Cybernetics, 5: 84-90, 1966.

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

[8] Amino, K., "Properties of the Japanese Phonemes in Aural Speaker Identification," IEICE Tech. Rep., 104: 49-54, 2004.

[9] Amino, K., Sugawara, T., and Arai, T., "Correspondences between the Perception of the Speaker Individualities Contained in Speech Sounds and Their Acoustic Properties," Proc. Interspeech, 2025-2028, 2005.

[10] Amino, K., Sugawara, T., and Arai, T., "Effects of the Syllable Structure on Perceptual Speaker Identification," IEICE Tech. Rep., 105: 109-114, 2006.

[11] Amino, K., Arai, T., and Sugawara, T., "Phoneme-dependency of Accuracy Rates in Familiar and Unknown Speaker Identification," J. Acoust. Soc. Am., 120: 3291, 2006.

[12] JEIDA Japanese common speech data corpus, http://www.sunrisemusic.co.jp/dataBase/fl/voicebase01_fl.html

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

[14] Ladefoged, P., and Maddieson, I., The Sounds of the World's Languages, Blackwell Publishers Ltd., Oxford, 1996.

[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., 100(5): 3374-3383, 1996.

[16] Tortora, G.K., and Grabowski, S.R., Introduction to the Human Body –The Essentials of Anatomy and Physiology, 5th ed., Japanese translation, Maruzen Co. Ltd., Tokyo, 2002.

[17] Pruthi, T, and Epsy-Wilson, C.Y., "Acoustic Parameters for Automatic Detection of Nasal Manner," Sp. Comm., 43(3): 225-239, 2004.

[18] Hollerbach, J., "Computers, Brains and the Control of Movement," Trends in Neuroscience, 5: 189–192, 1982.

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