



JCA2007

4-6 June

The Japan-China Joint Conference of Acoustics 2007

CONTRIBUTION OF CONSONANTS AND VOWELS TO THE PERCEPTION OF SPEAKER IDENTITY

Kanae Amino and Takayuki Arai
Department of Electrical and Electronics Engineering,
Sophia University, Tokyo, Japan

ABSTRACT

In many studies, it is reported that stimulus contents affect the perceptual speaker identification. In our previous research, we showed that nasal sounds are effective for accurate speaker identification.

In this present study, we investigate the contributions of the syllable onset consonants (C) and the syllable nucleus (V) to the perception of the speaker identity by using the hybrid CV monosyllables where C and V are uttered by two different speakers.

The results showed that perceived speaker of the hybrid CV syllables was inclined to be the speaker of V, and this tendency was prominent especially with the stimuli containing nasal consonants. This suggests that vowels mainly convey speaker individuality, and nasalised vowels contain more speaker information than oral vowels.

INTRODUCTION

It is apparent that the primary information conveyed by speech sounds is the phonological information, which is the message of the utterance [1-3], but this is not the only information conveyed. Speech sounds also contain other kinds of information such as expressive quality including speaker's emotions and attitudes, speaker's individuality including physical and stylistic variations, and so on [4, 5].

Research on the perception of speaker identity has been important for military and forensic purposes since early 20th century [1, 3, 6]. Nowadays, acoustic and perceptual properties of speaker individuality are also exploited in the field of speech technologies, mainly for applications such as speaker conversion, automatic speaker recognition, and speaker adaptation [7-9].

In perceptual speaker identification, it is reported that there is an interaction between the perception of speaker identity and that of phonological information [10], i.e. the accuracy of speaker identification depends on the stimulus contents presented to the listeners [1, 9, 11-15]. A series of our study has shown that the stimuli containing nasal sounds are effective for speaker identification by listening, both for identifying familiar speakers and previously unknown speakers [16-19]. We also found that there are correspondences between the spectral properties of the stimuli and the perception of the speaker identity, and inter-speaker spectral distances are greater in nasal sounds than in oral sounds [16, 17, 20]. Further analyses showed that the listeners exploit information about speakers contained in the onset consonant part (C) of a monosyllable when it is a nasal; otherwise only the nucleus vowel part (V) had relatively high correlations with

the perception [20]. This present paper discusses the roles that C and V each play when we perceptually identify previously unknown speakers.

BACKGROUNDS

In our previous experiment [21], the effects of the syllable structures of the stimuli on perceptual speaker identification were examined. We found that not only the syllable duration but also phonemic variation was needed for more accurate speaker identification, as was pointed out in other studies [6, 12]. The identification results for each of the syllable structures and those for each of the CV stimuli and C-V stimuli are shown in Figures 1 and 2. The CV stimuli are the originally recorded speech, while C-V stimuli are modified speech where the onset consonants were cut off and only the remainders were presented to the listeners.

Specifically, this work showed that:

1. Onset consonants are important for speaker identification.
2. Alveolar consonants convey more individuality than bilabial consonants.
3. Nasals are effective for speaker identification both in onset and coda positions.

The first conclusion was elicited by comparing the results for CV stimuli and C-V stimuli. The second conclusion was brought out from the fact that the syllables containing /d/ and /n/ gained higher identification scores than those containing /b/ or /m/, respectively. The same tendency was also observed in other experiments [17, 19, 20]. Finally, the third conclusion indicates that the structures containing nasal sounds yielded better performance than those without a nasal, either in the onset position ($C_{\text{nasal}}V > C_{\text{oral}}V$) or in the coda position ($CVN > CVV > CV$).

Moreover, our analysis data showed that spectral distances among the speakers were relatively large in nasal consonants, but not in oral consonants [20]. During the vowel parts, inter-speaker distances were large for both nasal and oral environments [20, 22]. In this paper, we further examine the contributions of C and V to the perception of speaker identity in order to see the effectiveness of the nasals.

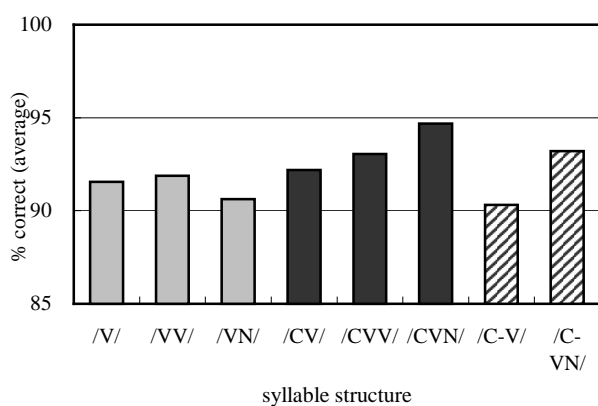


Fig. 1 Identification rates for each syllable structure

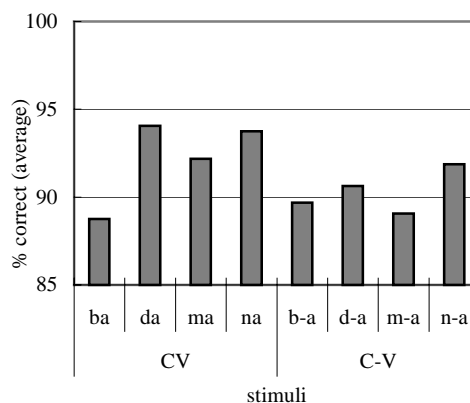


Fig. 2 Results for CV and C-V stimuli

METHODOLOGY

Speakers and speech materials Speech materials of two male speakers were used. They were both native speakers of Tokyo Japanese and both twenty years old at the time of the recordings. Speakers with relatively similar voice quality and of similar fundamental frequency were selected, in order to prevent the situation where one of the speakers can easily be identified due to an idiosyncratic voice quality or due to the pitch property of the utterance.

The recordings were conducted in a soundproof room, and all the utterances were recorded onto a digital audio tape at the sampling frequency of 48 kHz with 16-bit resolution. The speakers uttered eight monosyllables carried in a common phrase, where the accent pattern was controlled. The eight syllables are: /da/ /ma/ /na/ /nja/ /ra/ /sa/ /ta/ and /za/. Two tokens for each syllable and for each speaker were used in the experiment.

In order to create the test stimuli, the following steps were taken: first, the boundary between C and V was determined using the D-parameter introduced in [23]. Then C and V were exchanged between and within speakers. To simplify the experiment and to be sensitive to the discontinuities that arise in formant transitions, we exchanged C only with the same consonants.

Combination patterns of C and V are shown in Table 1. We had two tokens for two speakers, thus sixteen combinations were possible for each syllable. There were three types of the stimuli:


1. The original speech: There was no splicing and no exchange for these tokens.
2. Speech with intra-speaker exchange: We spliced C and V between two different utterances of each syllable by the same speaker.
3. Speech with inter-speaker exchange: C and V were cross-spliced between the two different speakers.


The numbers of stimuli for each type are shown below the description of each combination type appearing under Table 1. For original speech, we used other three tokens for each speaker apart from the two tokens that underwent the exchange operations.


Table 1. Combination patterns of C and V; C is one of the following consonants: /d, m, n, nj, r, s, t, z/; V is /a/. Sp1 is speaker 1 and Sp2 is speaker 2. C1 and V1 are C and V of token 1, and C2 and V2 are those of token 2, respectively.

↓ C· V→	Sp1-V1	Sp1-V2	Sp2-V1	Sp2-V2
Sp1-C1	Sp1-C1 + Sp1-V1	Sp1-C1 + Sp1-V2	Sp1-C1 + Sp2-V1	Sp1-C1 + Sp2-V2
Sp1-C2	Sp1-C2 + Sp1-V1	Sp1-C2 + Sp1-V2	Sp1-C2 + Sp2-V1	Sp1-C2 + Sp2-V2
Sp2-C1	Sp2-C1 + Sp1-V1	Sp2-C1 + Sp1-V2	Sp2-C1 + Sp2-V1	Sp2-C1 + Sp2-V2
Sp2-C2	Sp2-C2 + Sp1-V1	Sp2-C2 + Sp1-V2	Sp2-C2 + Sp2-V1	Sp2-C2 + Sp2-V2

Combination types and the numbers of the stimuli

 Original speech (no exchange operated); Ten for each consonant; four shown above and six other tokens

 Speech with intra-speaker exchange (exchange between different tokens within the same speaker); Four patterns shown in the table

 Speech with inter-speaker exchange (exchange between different speakers); Eight patterns shown in the table

Procedures Perception experiment was conducted in the same soundproof room as the recordings. Thirteen listeners who had never known the speakers participated in the experiment. They were all native speakers of Japanese and had normal hearing.

Since the task was speaker “identification,” not “discrimination,” the listeners had to get familiarised with the two speakers in the first part of the experiment. They listened to the speakers’ sample utterances as many times as they wanted. After they felt confident enough, they had a practice using the same sample files. Feedback was given after each trial during the practice. Familiarisation and practice sessions were repeated until the listener could identify the speakers with 100% accuracy. It took about five minutes on the average for the listeners to be able to distinguish the two speakers’ voices.

The test session had 352 trials and the stimuli were presented in a pseudo-random order. The session took about twenty minutes and the listeners took a break once in the halfway. They were not allowed to listen to the speakers’ sample utterances, once they started the experiment.

RESULTS AND DISCUSSION

The results are shown in Figures 3 to 5. Thirteen listeners participated in the experiment, thus the numbers of evaluations for each type of stimuli were 130 for the original speech, 52 for speech with intra-speaker exchange, and 208 for the speech with inter-speaker exchange.

In Figure 3, we can see that speaker identification rate for original speech was higher than the chance level (50%) in all syllables. The score was the highest in /nja/, and /da/ and /sa/ followed it. In the results for speech with intra-speaker exchange, shown in Figure 4, there were prominent declines in the identification scores of /da/ and /ra/ compared to their original counterparts.

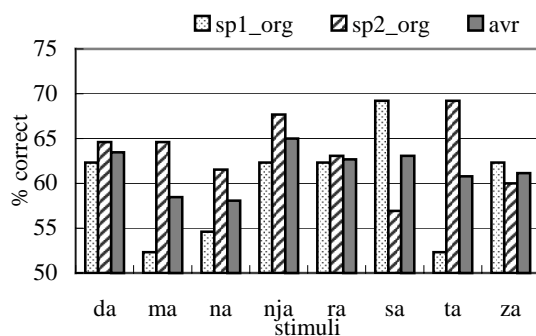


Fig. 3 Results for original speech

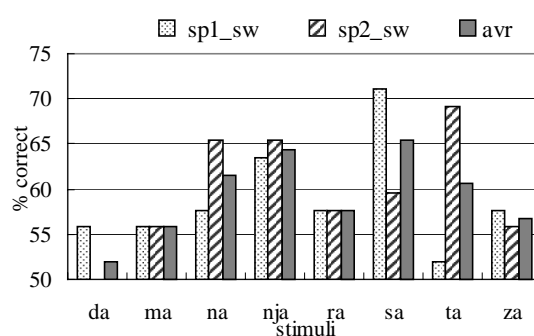


Fig. 4 Results for speech with intra-speaker exchange

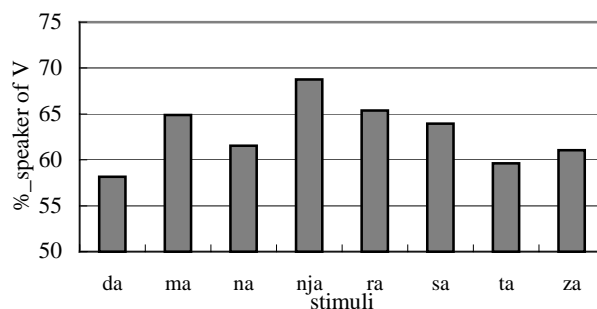


Fig. 5 Results for speech with inter-speaker exchange

The rankings of the consonants were different from those observed in our previous experiments [16, 17, 20-22], and not all nasals gained high scores. However, we can not simply compare the results, since the test tasks were also different [24]. In addition, the syllables /sa/ and /ta/ have large inter-speaker variations, in both Figures 3 and 4.

There are no correct answers for the hybrid stimuli with different speakers for C and V; therefore the percentages of the response for the speaker of V are shown in Figure 5. Here we can see that all the syllables obtained more than 50%, and this means that the perception of the speaker identity is more influenced by the vowels than by the consonants. The effectiveness of the vowels can be explained by greater energy and sonority, and also longer duration.

As for the consonant rankings in Figure 5, the syllable /nja/ was most influenced by its vowel part, and then /ra/ and /ma/ followed it. When we focus on the manners of articulation, they can be arranged in to the following order: nasals, fricatives and stops. This ranking is the same as the rankings of speaker identification accuracies in our previous experiments, although there are slight differences in the consonant orders.

The effect of the consonants was statistically meaningful, and the difference between /nja/ and /da/ was significant ($p < 0.01$). The syllable /nja/ obtained high identification score and also the effect of the vowel part was the greatest among other syllables. This suggests that nasalised vowel contain a lot of speaker individuality. Articulation of nasal sounds involves several resonators that reflect speakers' physiological information as pointed out in some studies [25]. Moreover, nasalised vowels contain not only resonance property but also the timing property of the velic action.

In this study, the importance of the feature [nasal] in perceptual speaker identification was again confirmed. Our future tasks will be to further prove it on stimuli with inter-speaker exchange among different consonants, and to test on other vowels.

ACKNOWLEDGMENT

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

REFERENCES

- [1] Y. Niimi, *Speech recognition*. (Kyoritsu Shuppan Publishing Company, Tokyo, 1979)
- [2] P. Ladefoged and D. Broadbent, "Information conveyed by vowel", *J. Acoust. Soc. Am.*, **vol. 29**, pp.98-104 (1957)
- [3] S. Furui, *Acoustic and speech engineering*. (Kindai Kagaku-sha, Tokyo, 1992)
- [4] H. Traunmueller, "Modulation and demodulation in production, perception, and imitation of speech and bodily gestures", *Proc. FONETIK 98*, pp.40-43 (1998)
- [5] H. Joh, K. Sato, Y. Saito, T. Fukumori, and H. Matsuzaki, *Computer onseigaku (Computer phonetics) Pedagogy of Japanese series 3*. (Ohfu Shuppan, Tokyo, 2001)
- [6] H. Hollien, *The acoustics of crime*. (Plenum, New York, 1990)
- [7] H. Kuwabara and T. Takagi, "Acoustic parameters of voice individuality and voice-quality control by analysis-synthesis method", *Speech Communication*, **vol.10**, pp.491-495 (1991)
- [8] H. Kuwabara and Y. Sagisaka, "Acoustic characteristics of speaker individuality: control and conversion", *Speech Communication*, **vol.16**, pp.165-173 (1995)
- [9] D. O'Shaughnessy, *Speech communications –human and machine–*, second ed. (Addison-Wesley Publishing Company, New York, 2000)

- [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 (eds.), pp.390-413 (Blackwell Publishing, Oxford, 2005)
- [11] T. Nishio, "Can we recognise people by their voices?", *Gengo-Seikatsu*, vol.158, pp.36-42 (1964)
- [12] P. Bricker and S. Pruzansky, "Effects of stimulus content and duration on talker identification", *J. Acoust. Soc. Am.*, vol.40, pp.1441-1449 (1966)
- [13] 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)
- [14] 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)
- [15] T. Kitamura and M. Akagi, "Speaker individualities in speech spectral envelopes", *J. Acoust. Soc. Jpn. (E)*, vol.16, pp.283-289 (1995)
- [16] 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)
- [17] K. Amino, T. Sugawara, and T. Arai, "Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties," *Acoust. Sci. Tech.*, vol.27, pp.233-235 (2006)
- [18] K. Amino, T. Arai, and T. Sugawara, "Phoneme-dependency of accuracy rates in familiar and unknown speaker identification", *J. Acoust. Soc. Am.*, vol. 120, p.3291 (2006)
- [19] K. Amino and T. Arai, "Effects of stimulus contents and speaker familiarity on perceptual speaker identification", *Acoust. Sci. Tech.*, vol.28, pp.128-130 (2007)
- [20] K. Amino, T. Sugawara, and T. Arai, "Speaker similarities in human perception and their spectral properties", *Proc. of WESPAC 2006* (2006)
- [21] 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).
- [22] K. Amino and T. Arai, "Speech similarity in perceptual speaker identification", *Proc. of 2006 autumn meet. Acoust. Soc. Jpn.*, pp.273-274 (2006)
- [23] S. Furui, "On the role of spectral transition for speech perception", *J. Acoust. Soc. Am.*, vol.80, pp.1016-1025 (1986)
- [24] P. Bricker and S. Pruzansky, "Speaker recognition," in *Experimental Phonetics*, N. Lass ed., pp. 295-326 (Academic Press, London, 1976)
- [25] J. Dang and K. Honda, "Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation", *J. Acoust. Soc. Am.*, vol.100, pp.3374-3383 (1996)