
2006 RISP International Workshop
on Nonlinear Circuits and Signal Processing (NCSP'06)
Waikiki Beach Marriott, Honolulu,
Hawaii, USA, March 3-5, 2006

# Effects of vowel types on perception of speaker characteristics of unknown speakers

Tatsuya Kitamura and Parham Mokhtari

ATR Human Information Science Laboratories
2-2-2 Hikaridai, "Keihanna Science City," Kyoto 619-0288, Japan
E-mail: {kitamura, parham}@atr.jp

**Abstract**

The aim of this study is to explore possible speaker characteristics common to speech sounds, through two psychoacoustic experiments. In the first experiment, sustained Japanese vowels produced by four adult male speakers were used, and ABX tests were carried out to confirm whether speaker individualities common to sustained vowels exist by testing whether participants can verify unknown speakers using speaker characteristics obtained from other vowels. The experimental results show that there are some speaker characteristics common to the vowels and that pitch frequency is one of the primary cues for identification of unknown speakers of the sustained vowels. In the second experiment, ABX tests were conducted using three Japanese sentences produced by four adult male speakers. The results indicate that the participants can identify the speakers even though the mean of the pitch frequency is speaker-normalized, implying that dynamic properties of speech are important for auditory-perceptual speaker identification.

## 1. Introduction

Humans can identify speakers of speech sounds even though the acoustic properties of speech sounds vary drastically within a speaker. Since speech sounds convey speaker's biological information, there are probably such cues not affected by intra-speaker variations. If perceptual cues for speaker identification adapting to intra-speaker variation are clarified, these parameters would facilitate the development of advanced speech signal processing techniques; for example, the utilization of such cues would make speaker recognition systems more robust to intra-speaker variation. The perceptual cues for speaker identification not depending on intra-speaker variation, however, have not been revealed in previous studies.

Furui and Akagi[1] showed that speaker individualities exist mainly in the frequency range from 2.5 kHz to 3.5 kHz of time-averaged spectra. They, however, did not investigate correlations between the frequency range and perception of speaker individuality. Zhu and Kasuya[2] reported through psychoacoustic experiments that averaged features of vocal tract characteristics are more important for speaker identification than dynamic features. However, since vocal tract shape varies by pitch frequency shift[3], the average vocal tract feature may also be expected to vary according to the speaker's pitch frequency. It is therefore likely that when perceiving a speaker's identity, humans rely on acoustic features that are robust against various kinds of intra-speaker variation.

In the present study, we carry out two preliminary psychoacoustic experiments which are founded on the assumption that there are common perceptual cues not depending on intra-speaker variation. Experiment 1 shows the existence of speaker individualities common to sustained vowels and Experiment 2 shows effects of dynamic features of speech on identifying the speaker of different sentences.

## 2. Experiment 1

### 2.1. Method

#### *2.1.1. Stimuli*

Four male native Japanese speakers recorded five sustained vowels at a sampling rate of 16 kHz with 16-bit resolution using a microphone (SONY ECM-77B) and an IC recorder (Marantz PMD-670) in a soundproof room. In order to avoid the influence of the duration of stimuli on the speaker identification test, the speakers were forced to keep the duration of their voices to the same length as that of a 0.6-sec white noise presented through headphones. They were not asked to tune their pitch frequency. Each vowel was uttered five times, of which two tokens were used in the experiment. The mean and standard deviation of the pitch frequency of the speech waves were 131.5 Hz and 11.5 Hz, respectively. In Experiment 1, the following two types of stimuli were used:

**V1** speech waves with normalized amplitude,

**V2** speech waves with normalized amplitude and pitch frequency.

The pitch frequencies of stimuli V2 were tuned to a mean value of 131.5 Hz by using the STRAIGHT analysis-

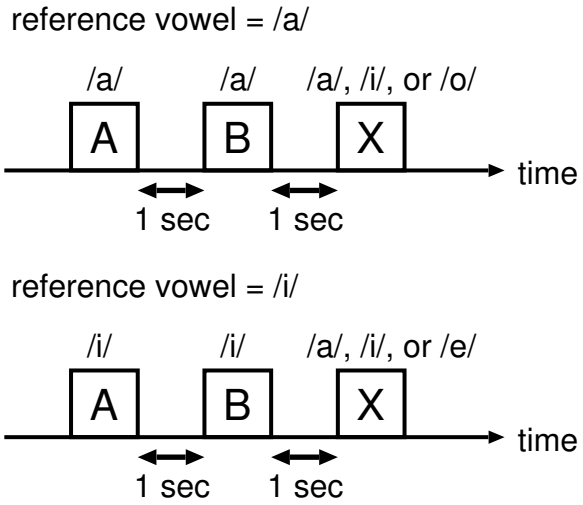



Figure 1: ABX sequences for Experiment 1. A and B are stimuli produced by difference speakers and X is a stimulus produced by one of the speakers.

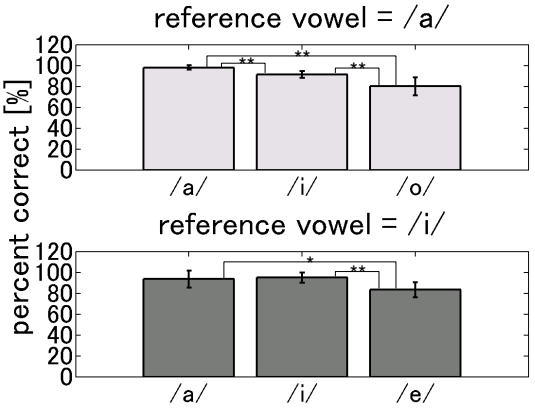
synthesis system[4], while retaining the original shapes of the pitch contour of each stimulus.

### *2.1.2. Participants*

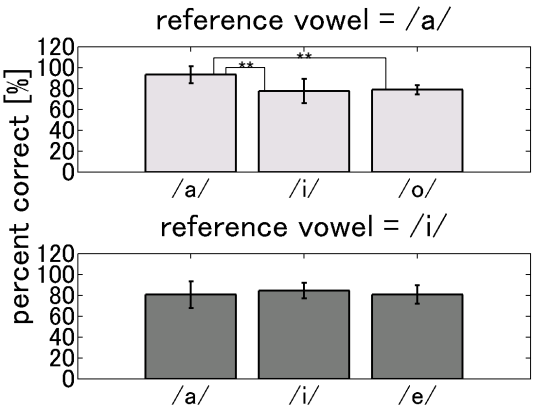
Nine listeners (two males and seven females) participating in Experiments 1 and 2 had never met with the speakers and had never listened to the speakers' voices. All were native speakers of Japanese and had no known hearing impairments.

### *2.1.3. Procedure*

Experiment 1 was conducted by an ABX test procedure. In the experiments, the subjects listened to triplets of the stimuli (A, B, and X) randomly at intervals of 1 sec as shown in Fig. 1. The first (A) and second (B) stimuli of a triplet were vowels (/a/ or /i/) produced by different speakers and the third one (X) was one of the vowels (/a/, /i/, /e/, or /o/) produced by one of the speakers of the first two stimuli. Different tokens were used as those three stimuli. The vowel of the first two stimuli is referred to as the "reference vowel." Participants were asked to select which of the first two speakers produced the third stimulus. The stimuli were presented in ABX and BAX orders to counterbalance any effects due to the order of presentation. The participants were allowed to listen to each triplet up to three times. The stimuli were presented through binaural earphones (Sennheizer HDA200) at a comfortable loudness level. Speaker identification rates for the stimuli were averaged across subjects. This experimental procedure is also used in Experiment 2.




(a) stimulus V1




(b) stimulus V2

Figure 2: Speaker identification rates for stimuli V1 and V2. An asterisk (*) represents a significant difference at the level ($p < .05$) and two asterisks (**) represent significant difference at the level ($p < .01$) between the speaker identification rates.

## 2.2. Results and discussion

The speaker identification rates for stimuli V1 and V2 are shown in Fig. 2. The speaker identification rates were tested by ANOVA ($F(1, 38; 0.01) = 7.44$ and $F(1, 70; 0.01) = 7.01$). The results lead to the following conclusions.

1. The participants can somewhat identify the speaker of a vowel using speaker characteristics extracted from another vowel (the speaker identification rate ranges from 80.1 % to 93.5 % for the non-reference vowels of V1). This implies that there are some speaker characteristics common to the vowels.

2. The mean of the pitch frequency is one of the primary cues for identification of unknown speakers of the sus-

tained vowels ($F(1, 38) = 11.58$ between V1 and V2 for the reference vowels and $F(1, 70) = 12.80$ between V1 and V2 for the non-reference vowels).

3. There is no notable difference between the front (/i/) and back (/a/) vowels as the reference vowel.

## 3. Experiment 2

### 3.1. Method

#### *3.1.1. Stimuli*

Four male native Japanese speakers, who are different from the speakers in Experiment 1, recorded the following three Japanese sentences at a sampling rate of 16 kHz with 16-bit resolution in a soundproof room.

**Sentence 1** /aɾajɯɾɯ geNʒitˢɯ o sɯbeteʒibuN no hoːe neʒimagetanoda/

**Sentence 2** /iʃʃɯːkaN bakarei njɯːyoːkɯ o ʃɯzai ʃita/

**Sentence 3** /teɾbi geːmɯ ja pasokoN de geːmɯ o ʃite asobɯ/

Each sentence was uttered five times and two of them were used in the experiment. The mean of the pitch frequency of the speech waves across the speakers was 117.1 Hz. In the experiment, the following two types of stimuli were used:

**S1** speech waves with normalized amplitude,

**S2** speech waves with normalized amplitude and pitch frequency.

Methods for making stimuli S1 and S2 were the same as for stimuli V1 and V2, respectively.

#### *3.1.2. Procedure*

Experiment 2 was also conducted by the ABX test procedure. The first (A) and second (B) stimuli of a triplet are sentence 1 produced by different speakers and the third one (X) is one of the three sentences produced by one of the speakers of the first two stimuli. Different tokens were used as those three stimuli.

### 3.2. Results and discussion

The speaker identification rates for Experiment 2 are shown in Fig. 3. The speaker identification rates were tested by ANOVA ($F(2, 24; 0.01) = 5.61$ and $F(1, 52; 0.01) = 7.15$). There are no significant differences among the sentences ($F(2, 24) = 1.68$ for S1 and $F(2, 24) = 1.37$ for S2) and between S1 and S2 ($F(1, 52) = 2.43$). The results clearly

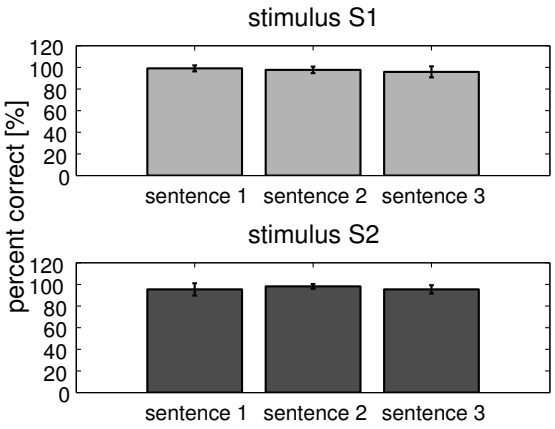



Figure 3: Speaker identification rates for Experiment 2.

indicate that, in contrast to the sustained vowels, the participants can identify the unknown speakers using speaker characteristics obtained from another sentence even though the mean of the pitch frequency is normalized across the speakers.

The difference of the results of Experiments 1 and 2 could be explained by three possible hypotheses: (1) the participants could obtain dynamic features of each speaker from the different sentences, which were absent in the sustained vowel stimuli; (2) they needed speech sounds with dynamic variations in order to obtain invariant static features as cues to speaker identification; (3) they identified the speakers using speaker characteristics obtained from phonemes common to the sentences. Akagi and Ienaga[5] reported that pitch frequency contours convey speaker individualities. Many researchers also have shown the effectiveness of dynamic features of vocal tract characteristics (e.g. delta-cepstrum) for automatic speaker recognition[6][7]. While it therefore seems likely that dynamic features contributed to identifying the speakers in Experiment 2, questions of what kind of features were used and how the participants could obtain such features from the different utterances remain to be investigated. On the other hand, Zhu and Kasuya[2] showed that vocal tract characteristics are more significant to perceive speaker individualities than the voice source and static features of the vocal tract are more important than dynamic features, implying that the second hypothesis may be more promising.

## 4. Conclusions

In order to investigate mechanisms of the human ability to identify speakers from their voices while adapting to intra-speaker variations, we carried out two psychoacoustic experiments assuming that there are common perceptual cues not

depending on intra-speaker variation. In this study, we focused on vowel and sentence variations and studied whether humans can identify unknown speakers using speaker characteristics obtained from different sustained vowels or sentences. The experimental results show that there are some speaker characteristics common to the vowels and sentences, and that dynamic features of speech contribute to speaker identification. Our future work is aimed at specifying common perceptual cues and investigating whether such cues are also the primary ones when humans identify speakers while adapting to other types of intra-speaker variation such as pitch frequency, vocal effort, and speaking styles.

**Acknowledgments**

This research was supported by the Ministry of Internal Affairs and Communications on their Strategic Information and Communications R&D Programme.

**References**

[1] S. Furui and M. Akagi, "Perception of voice individuality and physical correlates," *Tech. Rep. Hear. Acoust. Soc. Jpn.*, H85-18, pp. 1–8, 1985.

[2] W. Zhu and H. Kasuya, "Perceptual contributions of static and dynamic features of vocal tract characteristics to talker individuality," *IEICE Trans. Fundamentals*, Vol. E81-A, No. 2, pp. 268–274, 1998.

[3] T. Kitamura, P. Mokhtari, and H. Takemoto, "Changes of vocal tract shape and area function by F0 shift," *Proc. MAVEBA2005*, 85-88, 2005.

[4] H. Kawahara, I. Matsuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, Vol. 27, No. 3–4, pp. 187–207, 1999.

[5] M. Akagi and T. Ienaga, "Speaker individualities in fundamental frequency contours and its control," *J. Acoust. Soc. Jpn(E)*, Vol. 18, No. 2, pp. 73–80, 1997.

[6] S. Sagayama, and F. Itakura, "On individuality in a dynamic measure of speech," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 589–590, 1979.

[7] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans.* Vol. ASSP-29, No. 2, pp. 254–272, 1981.