Study of perceptual factors for speaker identification focusing on perceptual similarity of speaker characteristics

Tsuyoshi Izumida^{1,*} and Tatsuya Kitamura^{2,†}

¹*School of Information Science, Japan Advanced Institute of Science and Technology, 1-1, Asahidai, Nomi, 923-1292 Japan*

²*Faculty of Intelligence and Informatics, Konan University, 8-9-1 Okamoto, Higashinada, Kobe, 658-8501 Japan*

(Received 14 January 2011, Accepted for publication 25 February 2011)

Keywords: Speaker individuality, Phoneme duration, Pitch frequency contour, Speech spectra
PACS number: 43.71.Bp [doi:10.1250/ast.32.216]

1. Introduction

In this study, we aim to find perceptual clues for the identification of a speaker uttering sentences, focusing on the perceptual similarity of speaker characteristics. If a pair of speakers have similar voices, some of the acoustic features that contribute to perceptual speaker identification must be similar to one another. We can therefore identify acoustic features that are important for speaker identification by comparing similar voices. Such acoustic features can be used for voice quality control and speaker recognition.

Perceptual clues for the identification of a speaker uttering sentences have been investigated in several papers. Itoh and Saito [1] carried out speaker identification experiments using sentences uttered by speakers familiar to the subjects. They reported that the spectral envelope is more dominant for speaker identification than the pitch frequency and the tempo. Hashimoto *et al.* [2] compared the contributions of the spectrum, the pitch frequency, and the phoneme duration to speaker identification. They showed that the first two acoustic features contribute significantly and that their contribution rates depend on the distance between the acoustic features of speakers. Ohno and Akagi [3] demonstrated that the pitch frequency contours have high speaker individualities. Yamashita and Matsumoto [4] reported the relationship between adjective ratings and acoustic features in detail.

The previous studies concerned attempts to investigate perceptual clues for speaker identification, focusing on differences in speaker characteristics; however, in the present study, we focus on the similarity of speaker characteristic. We first compare sentences produced by pairs of speakers who have similar voices, and extract acoustic features that are similar for both speakers. We then examine the perceptual contribution of the acoustic features to speaker identification by an XAB test.

2. Preliminary experiment

An experiment was carried out to select speakers who have similar speaker characteristics to obtain perceptual clues for speaker identification from participants' comments.

2.1. Stimuli

Five male native Japanese speakers (HF, HH, HT, MT, and TK) recorded five Japanese sentences from the ATR speech database [5] in a soundproof room. Two of the speakers, MT and TK, are often confused with each other. The five speakers uttered each sentence five times, and the speech sounds were recorded at a sampling rate of 48 kHz with 16-bit resolution using a DAT recorder (Sony, TCD-D10 Pro II). The data were downsampled to 16 kHz and the maximum amplitude was normalized between the speakers. We used one sentence, "*arayuru geNjitsuwo subete jibuN no houe nejimagetanoda* (I twisted all the facts to suit myself)," of each speaker as the stimuli in this experiment.

2.2. Methods

Five male and five female listeners participated in the experiment. They had never met the speakers or listened to the speakers' voices. All were native speakers of Japanese and had no known hearing impairments. The participants were presented with a pair of sentences uttered by two of the speakers. The number of pairs was 10 ($= {}_5C_2$) and all pairs were presented once in a random order. The stimuli were presented through binaural headphones (Sennheiser HDA200) at a comfortable loudness level using a D/A converter (Roland UA-25) in a soundproof room. The participants were asked to evaluate the similarity between the speakers of each pair of sentences on a four-level scale of "similar," "rather similar," "not very similar," and "dissimilar." After evaluating all the pairs, they were asked to report any subjective perceptual clues they used to rate the similarity in the experiment.

2.3. Results

Table 1 lists the perceptual similarity ratings for each pair of speakers in the four-level rating experiment. Seven participants rated the speakers TK and MT as being "similar" and the other three rated them as "rather similar," showing that there is a marked similarity between the speakers. In addition, two participants rated the speakers HF and HT as being "similar" and six participants rated them as "rather similar." Most participants reported that the pitch and speech rate are perceptual clues for rating similarity and some of the participants reported that the voice quality is a perceptual clue.

*e-mail: tizumida@jaist.ac.jp

†e-mail: t-kitamu@konan-u.ac.jp

Table 1 Results for the four-level rating of perceptual similarity of pairs with the scale “similar”(A), “rather similar”(B), “not very similar”(C), “dissimilar”(D).

pairs of speakers	similarity			
	A	B	C	D
HF and HH	0	1	5	4
HF and HT	2	6	2	0
HF and MT	0	2	5	3
HF and TK	0	1	2	7
HH and HT	0	5	2	3
HH and MT	0	1	0	9
HH and TK	0	1	4	5
HT and MT	0	1	2	7
HT and TK	0	0	3	7
MT and TK	7	3	0	0

3. Analysis of F_0 contour

On the basis of the participants' comments, we compared the pitch frequency contour of the speech data between the speakers. The pitch frequency was extracted using the Pitch Contour function of WaveSurfer [6] with its default parameters; the method used was ESPS [7], the frame length was 7.5 ms, and the frame period was 10 ms. The obtained pitch frequency was corrected manually. Figure 1 shows comparisons of the pitch frequency contours of speakers TK and MT, HF and HT, and HF and HH. The shapes of the pitch frequency contours are similar for the first two pairs, and thus possibly contribute to the perception of similarity.

4. Experiment

An experiment was carried out to clarify the contributions of the phoneme duration and the pitch frequency contour to speaker identification.

4.1. Stimuli

The stimuli were synthesized speech consisting the sentences of speakers HF, HH, and TK, which were used in the preliminary experiment. The results of the preliminary experiment showed that there is less similarity between these three speakers. The following three types of stimuli were synthesized from the speech data of two of the speakers, referred to as the original and target speakers, using the STRAIGHT analysis-synthesis system [8].

Stimulus 1: resynthesized speech waves of the original speaker.

Stimulus 2: stimulus 1 with the phoneme duration replaced with that of the target speaker.

Stimulus 3: stimulus 1 with the phoneme duration and pitch frequency contour replaced with those of the target speaker.

Time alignment between the speech data of the original and target speakers was carried out by DP matching with the path limitation and weighting shown in Fig. 2. Twentieth-

Fig. 1 Pitch frequency contours of speakers (a) MT and TK, (b) HF and HT, and (c) HF and HH (The frequency scale is logarithmic). **Fig. 2** DP path limitation and weighting for time alignment between sentences of two speakers.

order cepstral coefficients calculated from STRAIGHT spectra were used as feature parameters. The number of pairs of original and target speakers is 6 ($= {}_3P_2$).

Fig. 3 Combinations of triplets of stimuli in the XAB test. (a) Stimulus X is stimulus 1. (b) Stimulus X is stimulus 2 or 3.

4.2. Participants

Ten male and ten female native Japanese speakers participated in the experiment. They were different from the participants of the preliminary experiment and had never met the speakers or listened to the speakers' voices. None of the participants had hearing impairments.

4.3. Procedure

This experiment was conducted following an XAB test procedure. In the experiment, the participants listened to triplets of the stimuli (X, A, and B) at intervals of 0.5 s. The combinations of the triplets of stimuli used in the experiment are illustrated in Fig. 3. In Fig. 3(a), stimuli X, A, and B are all stimulus 1, and in Fig. 3(b), the first stimulus (X) is stimulus 2 or 3 and the second and third stimuli are stimulus 1. The speakers of stimuli A and B are different in both cases. The triplets in the two cases were mixed and presented randomly. The participants were asked to select which of the latter two speakers was similar to the speaker of the first stimulus. The stimuli were presented in both XAB and XBA orders to counterbalance any effects due to the order of presentation. The number of triplets was 108. The participants were allowed to listen to each triplet once. The stimuli were presented through binaural headphones (Sennheiser HDA200) at a comfortable loudness level using a D/A converter (Roland UA-25) in a soundproof room.

4.4. Results and Discussion

Figure 4 shows the percentages of answers where the speaker of stimulus X was similar to that of stimulus B. The results were averaged across the participants. The difference between the results for stimuli 1 and 2 indicates the perceptual contribution of the phoneme duration, and the difference between the results for stimuli 2 and 3 indicates that of the pitch frequency contour. The percentages were tested by ANOVA ($F(1,238; 0.05) = 3.881$). There is a significant difference between the results for stimuli 1 and 2 ($F(1,238) = 9.293$) and between the results for stimuli 2 and 3 ($F(1,238) = 8.653$), supporting the findings of a previous study that showed that the pitch frequency contours contribute to the perception of speaker individuality [3].

The percentages of answers where the speakers of stimuli 2 and 3 are similar to the target speaker are 10.1% and 19.6%,

Fig. 4 Percentages of answers where the speaker of stimulus X is similar to that of stimulus B.

Fig. 5 Percentages of answers when the target speaker is speaker HH. The original speaker is speaker TK (a) or HF (b).

respectively. The results showed that acoustic features other than the phoneme duration and pitch frequency contour mainly contribute to the perception of speaker individuality. A possible acoustic feature that contributes to speaker individuality is the spectrum; Itoh and Saito [1] reported that the spectral envelope is more dominant for speaker identification than the pitch frequency and tempo.

Figure 5, which shows subsets of the results in Fig. 4, illustrates the results for stimuli when the target speaker is HH. Figures 5(a) and 5(b) show the results for stimuli when the original speaker is TK or HF, respectively. There are significant differences between the results for stimulus 2 in Figs. 5(a) and 5(b) ($F(1,38; 0.05) = 4.098$, $F(1,38) = 19.477$) and between the results for stimulus 3 in Figs. 5(a) and 5(b) ($F(1,38) = 7.961$). The results suggest that the dominant perceptual factors for speaker characteristics depend on the speaker pairs. In the case of speakers TK and HH, the percentage for stimulus 3 in Fig. 5(a) is 13.3%, showing that the dominant perceptual factor is probably the spectrum rather than the phoneme duration or the pitch frequency contour. In the case of speakers HF and HH, however, the percentage for

Fig. 6 Relationship among speakers HF, HH, and TK in perceptual space spanned by axes of perceptual distances between the speech spectra and prosody.

stimulus 3 in Fig. 5(a) is 38.3%, which is closer to the chance level (50%), indicating that the spectra of the speakers are similar. Despite this, there was less similarity between the two speakers in the preliminary experiment. Therefore, the dominant perceptual factors for discrimination between speakers HF and HH are the phoneme duration and pitch frequency contour.

On the basis of the above consideration, we schematically depict the relationship among speakers HF, HH, and TK in perceptual space with axes corresponding to the perceptual distances between the speech spectra and the prosody (the phoneme duration and pitch frequency) in Fig. 6. The perceptual distance between the prosody was large for speakers HF and HH, whereas that between the speech spectra was small; the participants thus confused the speaker of stimulus 3 when the original speaker was HF and the target speaker was HH. On the other hand, the participants did not confuse the speaker of stimulus 3 when the original speaker was TK and the target speaker was HH, because the perceptual distance between the speech spectra was large. The result that the relative contribution of each acoustic factor depends on its perceptual distance between speakers is consistent with the conclusions of Hashimoto *et al.* [2]; they reported that the contribution of the mean pitch frequency to speaker identification depends on its difference between speakers.

5. Conclusions

In this study, we explored perceptual factors contributing to speaker identification, focusing on perceptual similarity for

spoken sentences. The shapes of the pitch frequency contours of the sentences were similar for speakers with similar voices. However, the perceptual contribution of the pitch frequency contour was small, as was that of the phoneme duration, implying that other acoustic factors, probably those related to the speech spectrum, are dominant for the perceptual similarity of speaker characteristics. Also, the relative contribution of each factor depended on the pair of speakers, suggesting that the acoustic features that differ considerably between two speakers in the perceptual space contribute to speaker identification.

Our future works are to investigate which acoustic features in the speech spectrum are important for perceptual speaker identification. It is also necessary to clarify the relationship between the perceptual distance and the distance between acoustic features.

Acknowledgments

The authors thank Professor Hideki Kawahara of Wakayama University for providing the STRAIGHT analysis-synthesis system. This research was partly supported by SCOPE (No. 071705001) of the Ministry of Internal Affairs and Communications, Japan, and KAKENHI (Nos. 16700199, 21300071, and 21330170).

References

- [1] K. Itoh and S. Saito, "Effects of acoustical feature parameters of speech on perceptual identification of speakers," *Trans. IEICE*, **J65-A**, 101–108 (1982) (in Japanese).
- [2] M. Hashimoto, S. Kitagawa and N. Higuchi, "Quantitative analysis of acoustic features affecting speaker identification," *J. Acoust. Soc. Jpn. (J)*, **54**, 169–178 (1998) (in Japanese).
- [3] H. Ohno and M. Akagi, "Speaker individuality in fundamental frequency contours of sentences," *Tech. Rep. IEICE*, **97** (586), 89–96 (1998) (in Japanese).
- [4] Y. Yamashita and H. Matsumoto, "Acoustical correlates to adjective ratings of speaker characteristics of adults in reading style," *J. Acoust. Soc. Jpn. (J)*, **62**, 856–864 (2006) (in Japanese).
- [5] M. Abe, Y. Sagisaka, T. Umeda and H. Kuwabara, "Speech database user's manual," *Tech. Rep. ATR*, TR-I-0166 (1990) (in Japanese).
- [6] WaveSurfer, <http://www.speech.kth.se/wavesurfer/>
- [7] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," In *Speech Coding and Synthesis*, W. B. Kleijin and K. K. Paliwal, Eds. (Elsevier, New York, 1995), pp. 495–518.
- [8] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based on F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, **27**, 187–207 (1999).