Estimation of Perceptual Spaces for Speaker Identities Based on the Cross-Lingual Discrimination Task

Minoru Tsuzaki [1], Keiichi Tokuda [2], Hisashi Kawai, Jinfu Ni [3]

[1] Faculty of Music, Kyoto City University of Arts, Japan
[2] Department of Computer Science, Nagoya Institute of Technology, Japan
[3] Knowledge Creating Communication Research Center, NICT, Japan

minoru.tsuzaki@kcua.ac.jp, tokuda@nitech.ac.jp, {hisashi.kawai,jinfu.ni}@nict.go.jp

Abstract

This paper reconfirms that talker identity can be transmitted across languages. Talker discrimination was examined in the ABX paradigm, where the stimuli A and B were utterances by different talkers in the same language and the stimulus X was an utterance by either of A or B in the different language. The average hit rate of this discrimination task was as high as 0.89. The mutual distance matrices were generated using the discrimination index, d' . By applying the multidimensional scaling, three-dimensional perceptual spaces were estimated. The features related with loudness and spectral centroid had high contribution to the perceptual dimensions.

Index Terms: talker discrimination, bilingual corpus, MDS, auditory model

1. Introduction

One of the most challenging tasks for the current text-to-speech (TTS) synthesis technology is how to realize a reasonable variety in synthesized voices. This variety includes the within-speaker variations reflecting the internal state, i.e., the emotional change, or attitude of a specific speaker. It also includes the between-speaker variation reflecting the variations among different personalities. The importance of the latter variation will increase under multi talker circumstances, such as meetings, conferences, and cocktail parties, because speech (spoken language) uses the acoustic channel where the mixture of the informations from different sources is unavoidable [1]. The voice personality can provide a signature of the particular information source. It helps listeners to dissect the surrounding world, that is, to understand "who is talking what".

The hidden Markov model (HMM) based speech synthesis technique has been applied to achieve a TTS system with a wide range of talker variation [2, 3]. Compared to the unit concatenation speech synthesis system that requires a full set of speech corpus for each talker, the HMM system can potentially reduce the cost to add a new talker by applying the speaker adaptation technique. One can apply the extension of this speaker adaptation technique to corpora of different languages [4]. The outline is as follows: (a) training of the speaker-independent model for each of the input and output languages; (b) mapping the corresponding states between the two speaker-independent models with the criterion of minimum Kullback-Leibler Divergence; (c) applying the conversion matrix of the target speaker to the mapped output average voice.

In parallel to the development of this cross-lingual speaker adaptation technique, we need to establish the appropriate method to evaluate each new system. It is insufficient to check simply the standard mean opinion scores on speech quality and/or naturalness. At the first stage, it is necessary to evaluate how similar the output voices sound as the target speaker by perceptual evaluations. It is, however, an open question whether human listeners can identify a certain person's voice even if he/she speaks a different language without having any experience of hearing him/her speaking that language.

Kuwabara and Ohgushi argued that the perceptual cues for speaker identification mainly existed in the first and second formant frequencies [5]. If this is the case, one might predict that the cross linguistic speaker identification would be difficult because the first two formants could largely changes depending on languages. On the other hand, Kitamura and Saitou argued that the cues would be the frequencies of higher formants that stayed reasonably stable against different configurations of articulatory systems [6]. Recently, it was reported that human listeners could successfully make talker discrimination across languages [7], while it has also been suggested that there exist language dependent cues for speaker identification [8].

The first purpose of the current paper was to reexamine whether features characterizing speakers could be transferred across languages by using bilingual (English and Japanese) speech corpora. The second purpose was to estimate perceptual spaces for talker identity and to explore the underlying auditory features.

2. Perceptual Experiments

The basic paradigm of perceptual experiments was a talker discrimination task with the ABX method. Listeners heard a triplet of speech tokens in each trial. The triplet wad composed of three successive intervals. The intervals of A and B contained of utterances spoken by a pair of different talkers, and the interval X contained of an utterance by either A or B. The tokens AB and the token X were uttered in different languages. For example, when the tokens AB was uttered in Japanese, the token X was uttered in English.

If the talkers' characteristics were transmitted across the different languages, listeners could make correct judgments above the chance level (0.5). The rate of correct responses for each combination of talkers was then converted to the distance measure d' in the frame work of the theory of signal detection [9]. Thus, matrices of mutual distances between each pair of the talkers was obtained, which could be a base data for a further multi dimensional scaling and a cluster analysis.

2.1. Method

a. Stimuli

All the stimuli were selected from bilingual corpora which contains utterances of 42 bilinguals, (27 Japanese-English bilinguals; 15 Japanese-Chinese bilinguals). Utterances of 27

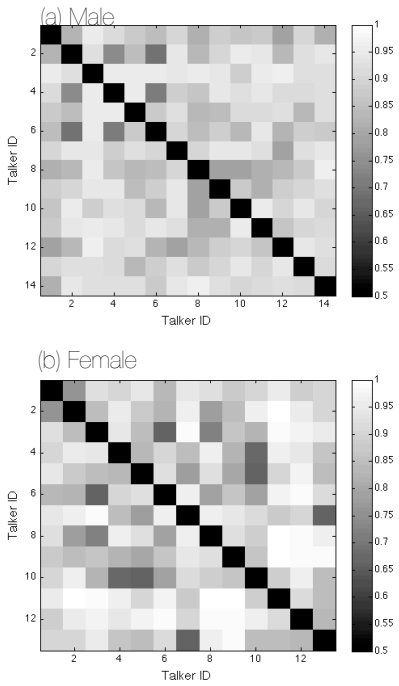

Figure 1. Correct choice rate in the ABX talker discrimination task. (a) The results for the male voice set. (b) The results for the female voice set.

Japanese-English bilinguals were used. Since the experimental task was the discrimination between talkers, it is not so informative to ask listeners to distinguish between male voices and female voices. Therefore, the experimental sets were divided into the male-voice and female-voice tests. For each talker, five phrases were selected from the phoneme balanced sentences, and five phrases were selected from the public address by President Obama.

b. Task and Procedure

Listeners was required to distinguish a pair of talkers by the ABX paradigm. In each trial, three intervals, namely, A, B, and X were presented. The intervals, A and B, contained of utterances spoken by a pair of different talkers, and the interval X contained of an utterance by either A or B. The tokens A and B and the token X were uttered in different languages. Listeners' task was to selected which of A and B interval was perceived to be spoken by the same talker as X.

Listeners were divided into four groups depending on the voice set, male or female, and the language combination, Japanese to English, or English to Japanese. In the Japanese-to-English language direction, the intervals A and B were in Japanese, and the interval X was in English. In the English-to-Japanese language direction, vice versa.

For each listener, all the combination of the talkers were presented four times, that is the combination of the order of A and B, and whether X was the same talker as A or B. Thus, the total number of trials for each listener was 364 ($_2C_{14}\times$ 4) for the male set, and 312 ($_2C_{13}\times$ 4) for the female voice set, respectively. The phrases were randomly selected from the 10 phrases to each interval, A, B, and X, with the limitation that they should be different from each other.

Each experimental session lasted about three hours including breaks. The experimental session was controlled by a laptop computer, Apple MacBook, and the stimuli were presented through a headphone, SONY MDR-Z900.

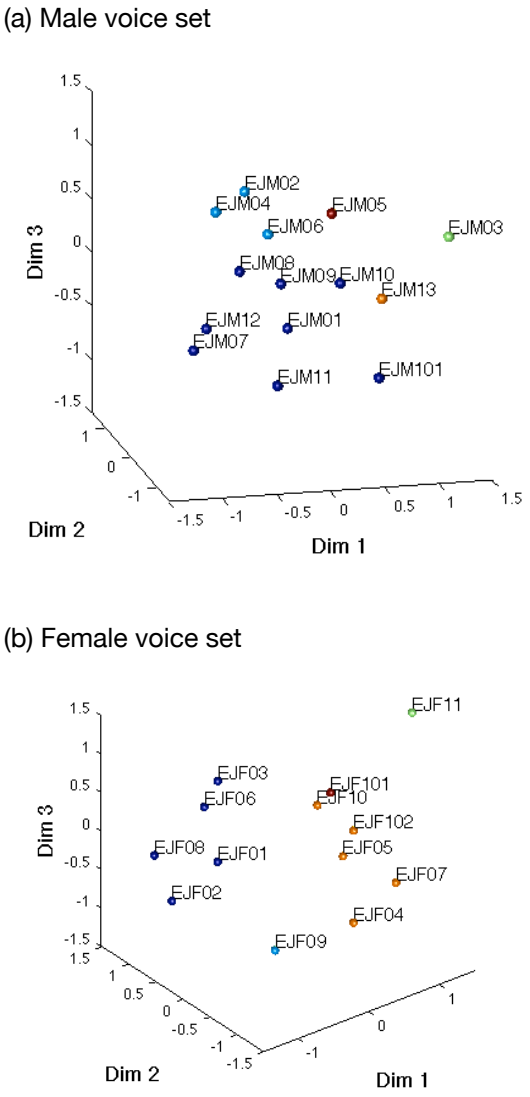

Figure 2. Three dimensional talker spaces estimated by applying MDS to the mutual distance matrices of the ABX discrimination results. (a) The configuration for the male voice set. (b) The configuration for the female voice set.

c. Listeners

Forty Japanese listeners without any significant hearing problem participated in the experiments. They were divided into four groups: (a) Male voice, English-to-Japanese; (b) Male voice, Japanese-to-English; (c) Female voice, English-to-Japanese; (d) Female voice, Japanese-to-English. The number of listeners for each group was ten. They were paid 6,000 yen per hour for their participation.

2.2. Results

Since no prominent difference was observed in the correct rate of the choice depending on the language direction, this factor and the individual listeners are pooled and the correct choice rates for each combination of the talkers are depicted in Fig. 1 (Fig. 1a for the male voice set; Fig. 1b for the female voice set). The average correct rate was 0.89 for the male voice set, and 0.89 for the female voice set. The number of talker pair for which the correct choice rate was below 0.75 was only three for the male voice set, and five for the female voice set.

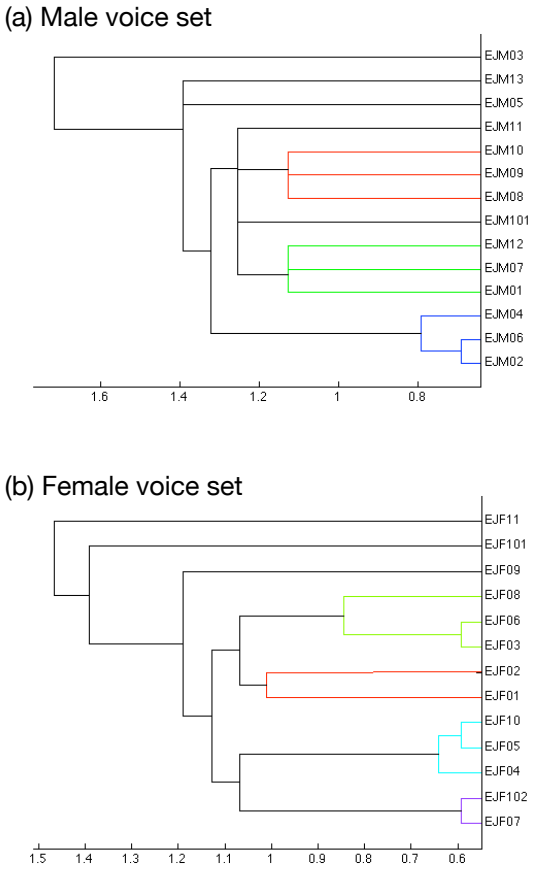

Figure 3. Dendrograms obtained by the cluster analysis for the mutual distance matrices of the ABX discrimination results. (a) For the male voice set. (b) For the female voice set.

Each correct choice rate can be converted into a discrimination index, d' [9]. These scores are considered as perceptual distances between each pair of talkers. Thus, a distance matrix was constructed for each voice set. Non-metric multidimensional scaling (MDS) was performed based on these scores for each voice set using MATLAB, Statistical Toolbox (Mathworks). The distance matrices were also submitted to the cluster analysis.

Three dimensional solution was adopted for the MDS. Figure 2 depict the perceptual talker spaces: (a) for the male voice set; (b) for the female voice set. The color indicates the clusters based on the cluster analysis, of which the dendrograms were shown in Fig. 3.

3. Auditory Feature Estimation

Several auditory features were extracted for each speech samples to investigate what perceptual cues were used for across language talker discrimination. Auditory Image Model (AIM) [10, 11] introduced by Patterson and his colleagues were used as a fundamental tools to simulate a plausible auditory processing.

The outline of this simulated auditory processing was: (1) bandpass filtering reflecting the frequency response of the outer and middle ear; (2) frequency analysis by the dynamic compressive gammachirp filter bank [12] mimicking the mechanical filtering by the basilar membrane; (3) half-wave rectification simulating the phase locked response of the inner hair cell; (4) obtaining time interval histograms by the strobed temporal integration.

Table 1. Coefficients of multiple regression analyses for each dimension by the auditory features.

	Male			Female		
Feature	**Dim 1**	**Dim 2**	**Dim 3**	**Dim 1**	**Dim 2**	**Dim 3**
loudness_mean	**-3.51**	**2.45**	**2.96**	3.99	**2.40**	**3.57**
loudness_sd	**6.09**	-0.76	**-0.93**	**-4.39**	**-2.91**	-0.48
F0_mean	**-0.46**	-0.11	**-0.81**	0.86	-0.22	-0.12
F0_sd	**-0.77**	**0.63**	0.23	-0.07	0.25	0.23
pitch_salience_mean	**-1.73**	0.29	0.40	0.73	0.31	-0.70
pitch_salience_sd	**0.41**	-0.28	-0.22	**-1.15**	-0.12	**1.18**
centroid_mean	**4.38**	**-2.16**	**-2.37**	**-5.65**	**-2.15**	**-2.79**
centroid_sd	**-7.01**	**1.00**	**1.40**	**5.84**	**2.28**	-0.22
speaking_rate	**-0.71**	-0.12	-0.28	-0.16	0.15	-0.71
speaking_rate_sd	**-1.23**	-0.04	0.20	0.29	-0.13	-0.29

These processing produced an image of multichannel time interval histograms of the auditory neural activity for a certain acoustic signal with a 5 ms frame rate, where the channels corresponded to an array of the center frequencies of the filter bank aligned at an equal distance on the ERB rate. This two dimensional activity pattern can be summarized in two directions. First, an "auditory" spectral profile can be obtained by pooling over the time interval. Second, a periodicity profile can be obtained by pooling over the center frequency. An estimate corresponding to loudness can be obtained by integrating the auditory spectral profile. An example of this auditory image is displayed in Fig. 4.

Thus, one can obtain following four estimates for every frame, (a) loudness, (b) F0, (c) pitch salience, and (d) spectral centroid. The F0 estimate can be calculated by taking the inverse of the time interval where the periodicity profile reaches a peak. The pitch salience can be calculated by a

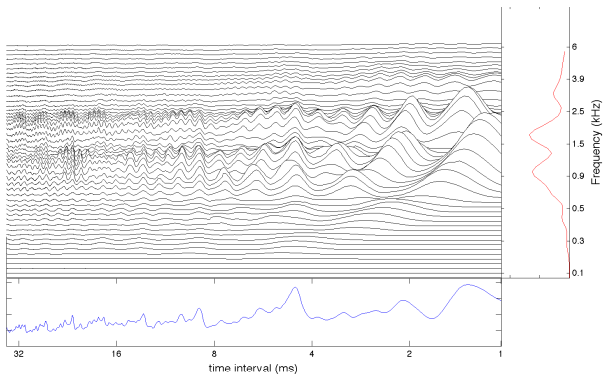

Figure 4. An example of the auditory image. Each line in the center panel depicts the time interval histogram for a specific center frequency. The profile at the bottom is the periodicity profile. The profile at the right is the auditory spectrum.

normalized peak height of the periodicity profile, where the normalization factor is the loudness.

The means and standard deviations (SDs) of these estimates were calculated only for the sonorant segments of speech samples. The decision of sonorousness was performed by an arbitrary thresholding based on the loudness and pitch salience. Accordingly, each speech sample could be divided into two types of segment, sonorant, or non-sonorant. Only the statistics for the sonorants were collected for each talker. Two additional statistics were calculated: one was the speaking rate, which was defined by the average inter-onset interval between the starting points of consecutive sonorants; the other was its standard deviation.

Multiple regression analysis was performed using these auditory feature statistics as explanatory parameters for each coordinate on the first to third dimension of the male and female talker spaces. Tables. 1 shows the results of the regression analysis. The tendencies were almost similar between the male and female voice. The loudness SD, centroid mean, and centroid SD had large contribution to the first dimension. The loudness mean and centroid mean had large contribution both to the second and to the third dimensions.

4. Discussion

Although the experiments using two different, i.e., male and female, voice sets were done independently, and the voice quality can be assumed to be quite different depending on the gender, the auditory features used to distinguish talkers appeared to be similar. The reason that the second and third dimensions had almost similar regression coefficients might be that there would be other hidden auditory cues available to distinguish talkers. Although the cue used to differentiate talkers were similar, it would not necessarily mean that the subjective impression to describe each of dimension was the same both for the male and female voice. For example, as the results of informal listening, the first dimension seemed to correspond to the degree of matureness for the female voice, but it seemed to correspond to the degree of intonation for the male voice.

The authors would not insist that the current experiment has revealed all the perceptual cues for the identification of talkers. Because of the experimental conditions used in this study, the talker identification performed was that for unfamiliar talkers. For familiar talkers, different cues might be used. However, the cues revealed in the current study would be adequate for the cross lingual talker adaptation. The usefulness of such a system will be demonstrated when we need voices of a specific person speaking in a language which that person cannot actually use. Imagine that Queen Elizabeth must make an address in Japanese. If she were a fluent speaker of Japanese, why don't we ask her to speak in Japanese? Accordingly, the situation to identify a person for the talker adapted synthesis is limited to the case of identifying unfamiliar voices, even if the target person is of a certain familiarity.

5. Conclusions

The ABX discrimination task of talkers in the cross lingual situation revealed that the talker identity could be transmitted across languages. The comparison of talker spaces estimated for the male and female voice sets in terms of their relation to the auditory features indicated that the information on loudness and spectral centroid would be the factors which contributed mainly for the cross lingual speaker identification.

Acknowledgements

This research was partly funded by the Strategic Information and Communications R&D Promotion Program (SCOPE), Ministry of Internal Affairs and Communication, Japan; and by KAKENHI 21330170, Japan Society for the Promotion of Science.

References

[1] A. S. Bregman, Auditory Scene Analysis: The Perceptual Organization of Sound. Massachusetts: MIT Press, 1990.

[2] [H. Zen, et al., "Statistical parametric speech synthesis," Speech Communication, vol. 51, pp. 1039-1064, 2009.

[3] J. Yamagishi, et al., "Roles of the average voice in speaker-adaptive HMM-based speech synthesis," in Proceedings of INTERSPEECH 2010, Makuhari, Japan, 2010.

[4] Y. Wu, et al., "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," in Proceedings of INTERSPEECH 2009, Brighton, U. K., 2009.

[5] H. Kuwabara, and K. Ohgushi, "The role of formant frequencies and bandwidths in the perception of speaker," [In Japanese] IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. J69-A, pp. 509-517, 1986.

[6] T. Kitamura, and T. Saitou, "Effects of acoustic modifications on perception of speaker characteristics," [In Japanese with English Abstract] Technical Report of IEICE, vol. SP-2006-167, pp. 1-6, 2007.

[7] M. Wester, "Cross-lingual talker discrimination," in Proceedings of INTERSPEECH 2010, Makuhari, Japan, 2010, pp. 1253-1256.

[8] S. J. Winters, et al., "Identification and discrimination of bilingual talkers across languages," Journal of Acoustical Society of America, vol. 123, pp. 4524-4538, 2008.

[9] D. M. Green and J. A. Swets, Signal Detection Theory and Psychophysics. New York: Wiley, 1966.

[10] R. D. Patterson, et al., "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," Journal of the Acoustical Society of America, vol. 98, pp. 1890-1894, 1995.

[11] S. Bleeck and R. D. Patterson. aim-mat. Available: http://www.pdn.cam.ac.uk/cnbh/aimmanual/index.html

[12] T. Irino and R. D. Patterson, "A time-domain, level-dependent auditory filter: The gammachirp," Journal of the Acoustical Society of America, vol. 101, pp. 412-419, 1997.