2012 International Workshop on Nonlinear Circuits,
Communications and Signal Processing
NCSP'12, Honolulu, Hawaii, USA. March 4-6, 2012

Study on hearing impression of speaker identification focusing on dynamic features

Tsuyoshi Izumida and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292 Japan
Phone/FAX:+81-761-51-1391/+81-761-51-1149
Email: {t_izumida, akagi}@jaist.ac.jp

Abstract

In this study, relationships between speaker identification and amount of dynamic features were investigated focusing on hearing impression. A three-layered model was adopted to model the hearing impression. First, relationships between speaker identification (first layer) and hearing impression (second layer), and those between hearing impression and acoustic features (third layer) were constructed with top down strategy. The results show that, "brisk" is a major factor in hearing impression of speaker identification, and slope of fundamental frequency (F_0) and dynamic range of spectral slope were correlated with the degrees of "brisk." Slope of F_0 and dynamic range of spectral slope were amount of dynamic features. Since slope of F_0 and dynamic range of spectral slope were correlated with the degrees of "brisk," "brisk" is hearing impression of speaker identification, correlated with dynamic features. Next, influences on speaker identification in the first layer from varied acoustic features in the third layer were investigated from bottom to top. The results show that, varied acoustic features for "brisk" affected speaker identification. Thus, it revealed that amount of dynamic features affects speaker identification.

1. Introduction

Human can determine who speaks in speech communication. In order to understand this ability, it is necessary what acoustical features in speech become cues of speaker individuality. Previous studies on speaker identification [1, 2, 3] reported that a variety of acoustical features contribute to perception of speaker individuality. Features in these studies are categorized into two groups, that is, averaged amount (static features) and varied amount (dynamic features). However, it is difficult to say in current research that relationships between speaker identification and dynamic features have been investigated enough. The dynamic features are derived from movements of speech organs. Acoustic features related to the movements also vary each other. Thus, it is necessary to consider combinations of several acoustic features to investigate the relationships between speaker identification and dynamic features. Focusing on hearing impressions of speech such as voice quality and speaking style, is beneficial to integrate several acoustic features. For example, relationships between perception and acoustic features on non-linguistic areas such as emotional speech and singing voice were modeled using three-layer models [4, 5, 6].

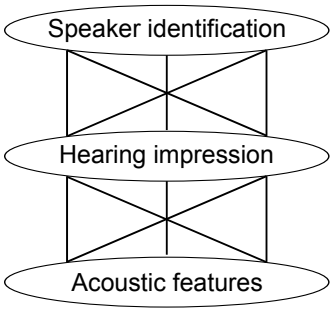

Figure 1: A three-layer model

This paper reports results discussed about relationships between speaker identification and dynamic features using a three-layered model, in which relationships between speaker identification (first layer) and hearing impression (second layer), and the second layer and acoustic features (third layer) are constructed from top to bottom. Figure 1 shows a three-layer model used in this study. Relationships between the first and the second layers are obtained by taking the following two steps. First, a perceptual space for speakers is estimated from similarity measurements of speakers' characteristics using the multi dimensional scaling (MDS). Next, degrees of speaker impressions are estimated by the Semantic Differential test (SD test). The relationships between the second and the third layers are found out by the correlation analysis between the acoustic features and the degrees of hearing impressions. Furthermore, influences on speaker identification in the first layer from varied acoustic features in the third layer are investigated from bottom to top.

2. Analysis of relations between first and second layer

To determine hearing impressions (second layer) related to speaker identification (first layer), a perceptual space for speakers was estimated, and relationships between the space and hearing impressions were investigated. Hearing impressions were described using adjectives. The space was estimated from similarity measurements of speakers' characteristics using the MDS. Collection of similarities of speakers for the MDS was carried out to estimate similarity measurements among speakers. To investigate relationships between the space and hearing impressions, degrees of speaker impressions were estimated by the SD test.

Table 1: Results of multiple correlation analysis. The r is coefficient of correlation.

Adjectives	r	Adjectives	r
Fluent	0.89	High	0.86
Brisk	0.85	Calm	0.82
Detached tone	0.78	Spirited	0.78
Clear	0.78	Powerful	0.78
Staccato	0.77	Polite	0.77
Forceful tone	0.76	Nasal	0.72
Deep	0.66	Tongue may slip	0.63
Old	0.60		

2.1. Speech data

Speech data uttered by fourteen male native Japanese speakers were used from the ATR speech database (C set) [7]. The used sentence was "reiboudeha hiesugiga moNdaininaru (The problem is too cold on air conditioning)." Maximum amplitude was normalized.

2.2. Collection of similarity of speakers

Collection of similarity of speakers was carried out to estimate similarity measurements among speakers. Ten male listeners participated in the experiment. The participants were presented with a pair of sentence uttered by two speakers. They were asked to evaluate similarity between the speakers of each pair of sentences on a five-level scale of "0: dissimilar," "1: not very similar," "2: rather similar," "3: similar," "4: same speaker." The stimuli were also presented in reverse sequence to counterbalance any effect due to the order of presentation and pairs of stimuli were presented randomly. Each pair of speakers were evaluated twice. The number of pairs was 392 (=14 $\times$ 14 $\times$ 2). Answers were provided to click the answer button on the monitor in a soundproof room. The stimuli were presented at a comfortable loudness level through binaural headphones (STAX SR-404) using a D/A converter(YAMAHA DP-U50). The participants were allowed to listen to each pair once.

2.3. SD test

Degrees of speaker impressions were estimated by the SD test using fifteen adjectives ("fluent," "high," "brisk," "calm," "detached tone," "spirited," "clear," "powerful," "staccato," "polite," "forceful tone," "nasal," "deep," "tongue may slip," "old"). The adjectives were collected in the previous study [8] and questionnaire. Ten male listeners participated in the experiment. The participants were presented one speaker's voice. The each speaker was evaluated the degrees of hearing impression on a seven-level scale of "-3, 3: very," "-2, 2: considerably," "-1, 1: rather or somewhat," "0: neither." Impressions of each speaker were evaluated thrice for each adjective. The number of trials was 630 (=14 $\times$ 15 $\times$ 3). The SD test was performed separately adjective by adjective, order of adjectives was randomized to among participants. Other experimental conditions were the same as experiment 2.2.

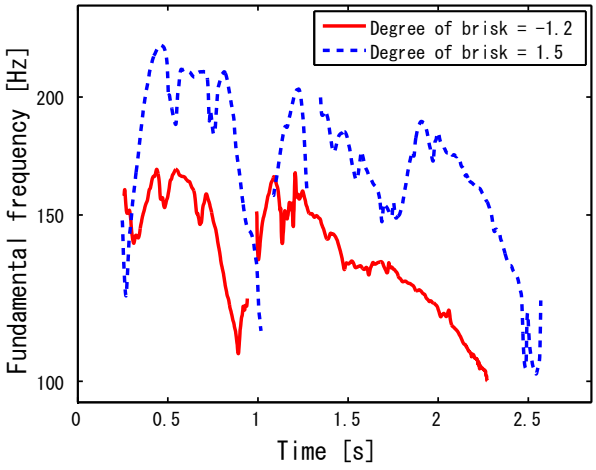

Figure 2: F_0 contours of briskly voice(briskly score = 1.5) and non-briskly voice(-1.2) (The frequency scale is logarithmic).

2.4. Results and discussions of analyses

Results of similarity of speaker individuality were analyzed using the MDS. Results show that, case of six-dimension, the stress becomes lower than 10%. Thus, the speakers were superimposed on the six-dimensional perceptual space. Relationships between hearing impressions and the perceptual space were analyzed using multi correlation analysis. Table 1 shows the results of multi correlation analysis. Results show that, "fluent (r = 0.89)," "high (r = 0.86)," "brisk (r = 0.85)" and "calm (r = 0.82)" were hearing impression that highly correlated with the perceptual space (r is coefficient of correlation). These are able to be thought of as major factors in hearing impressions for speaker identification. In addition, "brisk" may be hearing impression corresponding to dynamic feature. Thus, we focused "brisk," hereafter.

3. Analysis of relations between second and third layer

Acoustical features (third layer) were found corresponding to "brisk (second layer)." Relationships between extracted acoustical features and results of the SD test were estimated using correlation analysis.

3.1. Slope of F_0

F_0 contours ware extracted using STRAIGHT [9]. The obtained F_0s farther were corrected manually. Figure 2 shows typical F_0 contours of briskly voice (degrees of brisk is 1.5) and non-briskly voice (degrees of brisk is -1.2). What parameters of F_0 contours contribute to "brisk" was investigated. Average, maximum, minimum, dynamic range and slope [4, 5] of F_0 were extracted. Results of correlation analysis show that, all parameters except minimum were correlated with "brisk." Figure 3(a) shows relationship between degrees of brisk and slope of F_0.

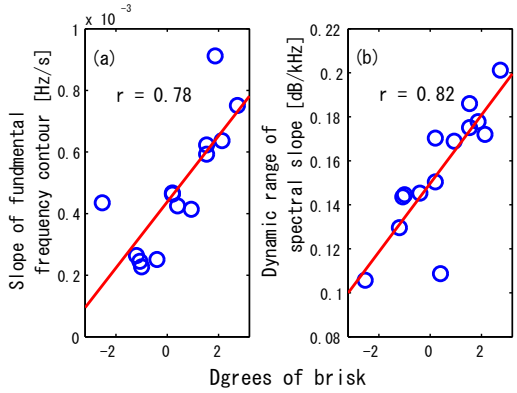

Figure 3: The relationship between degrees of brisk and (a) slope of F_0, (b) dynamic range of spectral slope.

3.2. Dynamic range of spectral slope

Since spectral distances during phonemes were correlated with "brisk" from preliminary study, spectral slopes were extracted. The spectral slopes are correlated with vibrational properties of the glottal source and radiational properties based on Source-Filter Theory [10, 11]. It is expected that "brisk" is correlated with vibrational properties of the glottal source and radiational properties. Thus, "brisk" correlated with F_0 contour. Average, maximum, minimum and dynamic range of spectral slopes were extracted and analyzed correlation with "brisk." Results of correlation analysis show that, maximum and dynamic range of spectral slopes were correlated with "brisk." Figure 3(b) shows relationship between degrees of brisk and dynamic range of spectral slope.

4. Evaluation of the model

From analysis of the previous sections, "brisk" is a major factor in hearing impression of speaker and correlated with dynamic features. Hence, influences on speaker identification in the first layer from varied acoustic features in the third layer were investigated from bottom to top. Four types of stimuli were synthesized controlling slope of F_0 and dynamic range of spectral slope to evaluate the model. First, influences on hearing impression in the second layer from varied acoustic features in the third layer were investigated in Evaluation experiment I. Range of degrees of "brisk" was investigated by controlled slope of F_0 and dynamic range of spectral slope. Next, influences on speaker identification in the first layer from varied acoustic features in the third layer were investigated in Evaluation experiment II. Then, influence on speaker identification was investigated by controlled slope of F_0 and dynamic range of spectral slope.

4.1. Stimuli

Stimuli were synthesized from speech data of three speakers (Speaker 108 , 419 and 702) in the fourteen speakers using the STRAIGHT analysis-synthesis system [9]. Each voice of three speakers was evaluated, as "brisk (Speaker 702, 1.5)," "non-brisk (Speaker 108, -1.2)" and "neither (Speaker 419, 0.2)," by the SD test on Chapter 2 (Numbers in parentheses is degrees of brisk). The following four types of stimuli were synthesized.

Stimulus 1: resynthesized speech waves of the original,
Stimulus 2_X: contorting slope of F_0 so that the degrees of "brisk" is X,
Stimulus 3_X: contorting dynamic range of spectral slope so that the degrees of "brisk" is X,
Stimulus 4_X: contorting slope of F_0 and dynamic range of spectral slope so that the degrees of "brisk" is X,

Stimuli 2–4 were synthesized, in which degrees of "brisk" is controlled to -2, 0 or 2. All stimuli were given with pink noise (S/N ratio 25 dB) to avoid risk of an adverse effect on the degradation of sound quality to experiments. Maximum amplitude was normalized.

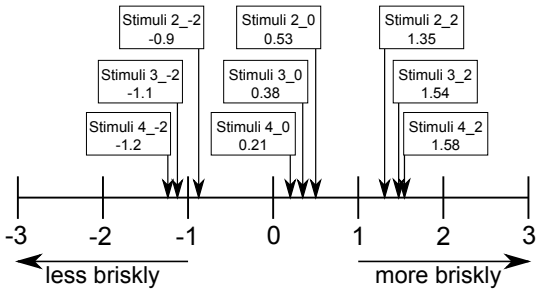

Figure 4: Results of evaluation experiment I

4.2. Evaluation experiment I

Evaluation experiment I was carried out using the SD test. Eight male listeners participated in the experiment. They were evaluated degrees of "brisk" for Stimuli 1-4. Each speaker impression was evaluated thrice. The number of trials was 90 (=3 $\times$ 10 $\times$ 3). Other experimental conditions were the same as the experiment 2.3.

4.3. Results and Discussions I

Figure 4 shows results of Evaluation experiment I. The results were averaged across the participants and speakers. Figure 4 showed degrees of "brisk" are varied by controlling slope of F_0 and dynamic range of spectral slope. Additionally, the results of Stimuli 4 were approached to each X rather than Stimuli 2 and 3.

4.4. Evaluation experiment II

Next, influence on speaker identification by varying degrees of "brisk," was investigated. Evaluation experiment II was carried out by X-A test. Eight male listeners participated in the experiment. The participants were presented a pair of the stimuli X and A. They were asked to evaluate whether paired stimuli are from the same speaker or not. Stimulus A is Stimulus 1, Stimulus X is Stimulus 1 or 4, and Stimuli X and A were different sentences. Degrees of "brisk" are -2, 0 or 2 on Stimuli 4. The stimuli were presented both X-A and A-X orders to counterbalance any effect due to the order of presentation. Pairs of stimuli were presented randomly. Each pairs of speaker were evaluated twice. The number of pairs

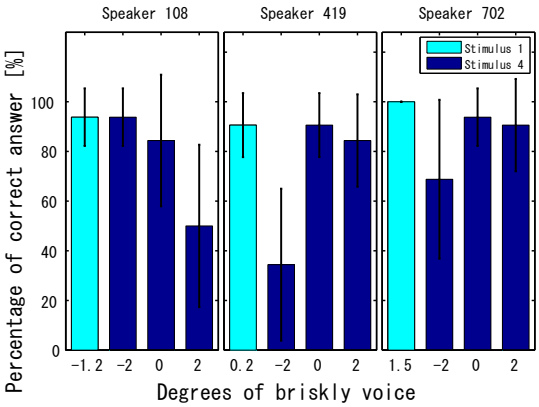

Figure 5: Results of evaluation experiment II

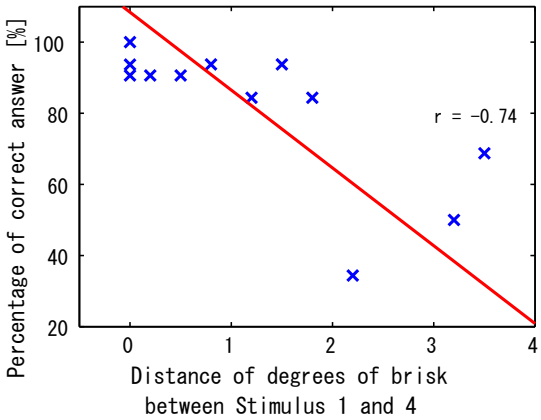

Figure 6: The relationship between distances of degrees of brisk between Stimulus 1 and 4, and percentage of correct answer

was 144 (=3 × 4 × 3 × 2 × 2). Other experimental conditions were the same as other experiments.

4.5. Results and Discussions II

Figure 5 shows results of Evaluation experiment II. The results were averaged across the participants. Figure 5 shows high percentage of correct answers case of Stimulus X is Stimulus 1 at all speakers. Also, Figure 5 shows low percentage of correct answers case of Stimulus X is Stimulus4_2 of Speaker 108, Stimulus4_-2 of Speaker 419 and 702. Degrees of "brisk" between these stimuli and original (Stimuli 1) are far away. Figure 6 shows relationship between distances of degrees of "brisk" during stimulus 1 and 4, and percentage of correct answer. Figure 6 shows percentage of correct answer were related with the distance (r = -0.74). The participants perceived as different speaker due to greatly varying degrees of "brisk" from original speaker.

Stimulus 4 was synthesized contorting slope of F_0 and dynamic range of spectral slope. By controlling the slope of F_0, which also changes average of F_0. Previous studies [2, 3] have shown that average of F_0 affects the speaker identification. Results of this study support these studies. Additionally, Akagi and Ienaga [1] has shown that F_0 contour affects the speaker identification . This study suggested that slope of F_0 is important. On the other hand, Kitamura and Saitou [3] has shown that spectral slope affect the speaker identification. This study suggested that dynamic range of spectral slope is also important. Thus, dynamic features affect speaker identification.

5. Conclusions

In this study, relationships between speaker identification and amount of dynamic features were investigated, focusing on hearing impressions with the three-layer model. The results show that, "brisk" was a major factor in hearing impression of speaker identification, and slope of F_0 and dynamic range of spectral slope were correlated with the degrees of "brisk." Additionally, varying acoustic features for "brisk" affected speaker identification. Thus, it revealed that amount of dynamic features affect speaker identification. Additionally, it was suggested that degrees of hearing impressions affect speaker identification.

References

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

[2] M. Hashimoto, S. Kitagawa and N. Higuchi, "Quantitative analysis of acoustic features affecting speaker identification, " J. Acoust. Soc. Jpn. (J), Vol. 54, No. 3, pp. 169–178, 1998.

[3] T. Kitamura and T. Saitou, "Effects of acoustic modification on perception of speaker characteristics for sustained vowels," Acoust. Sci. & Tech., Vol. 28, No. 6, pp. 434–437, 2007.

[4] C-F. Huang and M. Akagi, "A Multi-Layer fuzzy logical model for emotional speech perception," *Proc. EuroSpeech 2005*, pp. 417–420, Lisbon, Portugal, 2005.

[5] C-F. Huang and M. Akagi, "A three-layered model for expressive speech perception," *Speech Commum.*, Vol. 50, No. 10, pp. 810–828, 2008.

[6] T. Saitou, N. Tsuji, M. Unoki and M. Akagi, "Analysis of proper acoustic features to singing voice based on a perceptual model of "singing-ness"," J. Acoust. Soc. Jpn. (J), Vol. 64, No. 5, pp. 267–277, 2008.

[7] M. Abe, Y. Sagisaka, T. Umeda and H. Kuwabara, "Speech database usr's manual," Tech. Rep. ATR, TR-I-0166, 1990.

[8] Y. Yamashita and H. Matsumoto, "Acoustical correlates to adjective ratings of speaker characteristics of adults in reading style," J. Acoust. Soc. Jpn. (J), Vol. 62, No. 12, pp. 856–864, 2006.

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

[10] G. Fant: Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations, Mouton, 1970.

[11] K. N. Stevens: Acoustic Phonetics, The MIT Press, 2000.