SPEAKER CHANGE DETECTION CONSIDERING MOBILE DEVICE ENVIRONMENT

Masashi TATENO and Eiji KAMIOKA

Graduate School of Engineering and Science, Shibaura Institute of Technology

ma14071@shibaura-it.ac.jp, kamioka@shibaura-it.ac.jp

ABSTRACT

Generally, to achieve the speaker identification correctly in a plural speaker environment, a high performance computation to process tons of training data via wireless network is required. However, it is unrealistic to prepare the training data in advance that might be used in near future. In this paper, a speaker change detection method as an introduction of accurate speaker verification in mobile device environment is proposed. To achieve accurate speaker change detection, three threshold values are introduced. Firstly, similarity value which is utilized to detect speaker change in terms of characteristics of individual voice. Here, a combined feature which is composed of vocal tract parameter and fundamental frequency is investigated to calculate the similarity value. Secondly, fundamental frequency value which is utilized to detect the change of speaker's gender. Generally, the distribution of fundamental frequency for adult male is largely different from the one for female. Hence, the use of fundamental frequency value can improve the accuracy of speaker change detection. Thirdly, speaker change restriction time which is utilized to prohibit detecting the speaker change within a short time. The use of the similarity value and the fundamental frequency value has some possibilities to detect a speaker change within such a short time even if the speaker change does not occur. Thus, the appropriate restriction time of speaker change detection is investigated. As for the threshold of similarity, the value of 180 determined in the previous work was used in this evaluation. As for the rest 2 thresholds, 160Hz for the fundamental frequency threshold and 0.6 seconds for the restriction time threshold were determined from the evaluation analyses. As a result, the proposed speaker change detection algorithm achieved the accuracy of more than 80% when the speaker changed to the different gender.

1. INTRODUCTION

GMM[1][2][3] is a promising technology to achieve an accurate speaker change detection to mixed speakers' voice data. However, it needs each speaker's training voice data. It means that GMM cannot work if any speaker is not identified. Suppose that a user is recording a meeting voice data using a smartphone and he/she tries to identify each speaker's voice. However, if there are unknown participants in the meeting, GMM cannot perform the speaker change detection since their training voice data does not exist. Apart from the training data, there are many issues to be solved in mobile device environments. For instance, the accuracy of speaker change detection is influenced by the reverberation sound as noise. Also the characteristics of individual voice becomes weak when the speaker is far from the microphone used for recording.

This study proposes an accurate speaker change detection method which does not use each speaker's training data but uses characteristics of individual voice. More concretely, vocal tract parameter (MelCeps), fundamental frequency (F0), Mel-Frequency Cepstrum Coefficients (MFCC) and each characteristic's Δ and $\Delta\,\Delta$ parameters [5] are considered. The vocal tract parameter can be extracted from the first 10 dimensional characteristics of the MelCeps. The MelCeps is obtained from the cepstrum weighted by Mel Scale. Here, the Mel Scale is a scale which shows the voice frequency that human perceives, which is different from the real frequency. The following equation (1) indicates how to calculate the MelCeps from the Cepstrum (Ceps).

$$MelCeps(f) = log_2(1 + f/1000) * Ceps(f) \quad (1)$$

The 9 characteristics mentioned above include strong individual features, but they are tend to transform with the change of vocalized sound. Therefore, the use of several

characteristics of individual voice has been considered [4]. The combined characteristics were obtained as a waveform and the waveform pattern was investigated if it is similar to a speaker's one or the others using DP matching which is one of the pattern matching techniques. Then, the accuracy of speaker change detection was calculated. As a result, the use of multiple characteristics achieved a more accurate speaker change detection than a single use of characteristics. The best combination of the characteristics was MelCepsF0, which is the combination between MelCeps and F0. In the analysis, a threshold value of DP distance, which is 180, was determined to decide if the speaker changed or not. However, the distribution of DP distance varies so much for this purpose in general.

In this study, 3 types of threshold values, will be discussed in order to detect the speaker change more accurately. They are the threshold values of DP distance, fundamental frequency and restriction time. The first one is the DP distance as stated above. The second one, fundamental frequency, is effective to detect the change of speaker's gender. The third one, restriction time, is a duration time within which the speaker will not change because it is too short. Figure 1 shows the flow of the speaker change detection proposed in this study.

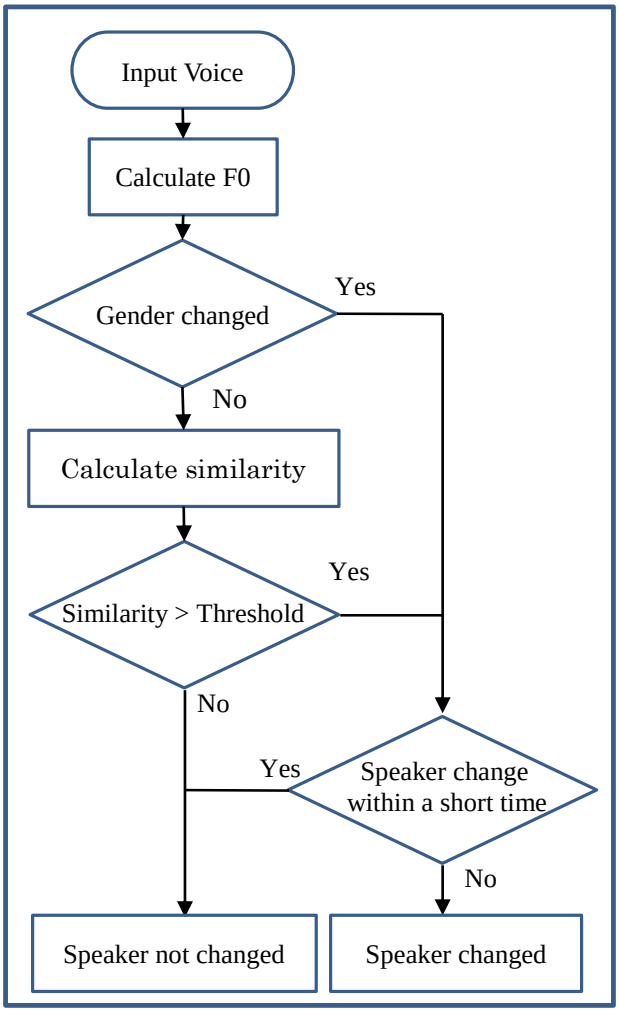

Figure 1 Speaker change detection flow using 3 thresholds

2 FUNDAMENTAL FREQUENCY THRESHOLD TO DETECT GENDER CHANGE DETECTION

Voice pitch, loudness and spectrum of waveform are affected by the vocal organ and the way to articulate. Formants are parameters which define these feature. Fundamental frequency is one of the formants. Generally, the fundamental frequency is different between males and females. Specifically, the average value of fundamental frequency and its standard deviation for adult males are about 125Hz and 20.5Hz, respectively. For adult females, those values are about 250Hz and 50Hz. In addition, the fundamental frequency tends to be stable for the same person. Hence, the threshold value of fundamental frequency is introduced to detect the speaker change in gender.

3 RESTRICTION TIME SHRESHOLD TO PROHIBIT DETECTING SPEAKER CHANGE

In General, one talk does not finish within 1 or 2 seconds. However, the proposed approach has some possibilities that it detects a speaker change within such a short time even if the speaker change does not occur. Therefore, the threshold value of restriction time is introduced to reduce this wrong decision. If the threshold value is too short, the above wrong decision would occur, and thus the effectiveness will not be large. On the other hand, if the threshold value is too long, it would decrease the accuracy of speaker change detection. Hence, the appropriate threshold value of restriction time should be determined.

4 EVALUATION ANALYSIS

Some analyses using voice data were performed to validate the effectiveness of the multiple characteristics of MelCepsF0. Firstly, 50 males and 50 females voice data were prepared from Corpus of Spontaneous Japanese. Recording time of each voice data was 10 seconds and the sampling frequency was 16,000Hz. Subsequently, 3 types of mixed voice data, in which 2 persons were talking, were prepared. The details are as follows: (1) 2 males mixed voice data (Male and Male), (2) 1 male and 1 female voice data (Male and Female) and (3) 2 females voice data (Female and Female). Therefore, the combination number of speakers is 1,225 for (1), 625 for (2), and 1,225 for (3). Then, the waveform of MelCepsF0 for each mixed data was generated. Each frame length was 0.25 seconds and frame shift was 0.05 seconds.

In this analyses, 2 threshold values, which are the fundamental frequency threshold and the restriction time threshold, were investigated. For the threshold value of DP distance, the value of 180 determined in the previous work was used in this evaluation.

4.1 FUNDAMENTAL FREQUENCY THRESHOLD

The fundamental frequency of each speaker was extracted from the voice data of 50 males and 50 females. Figure 2 shows the frequency distributions of the fundamental frequency for males and for female.

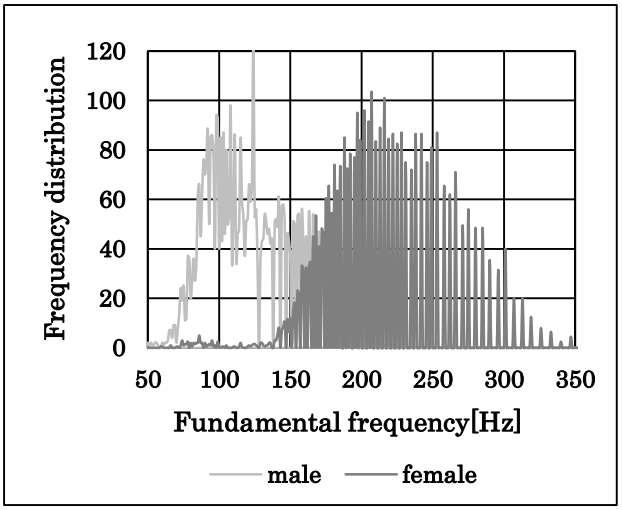

Figure 2 Fundamental frequency distributions of male and female

The threshold value of fundamental frequency must be set at the boundary between the male frequency distribution and the female frequency distribution. If the threshold value is too small, the accuracy of male-to-male speaker change detection would decrease. In contrast, if the threshold value is too large, the accuracy of female-to-female speaker change would decrease. In this analysis, the threshold value was varied from 150 to 200Hz to investigate how much the speaker change detection accuracy would improve. Figure 3 shows the improvement of speaker change accuracy to the case without introducing the threshold.

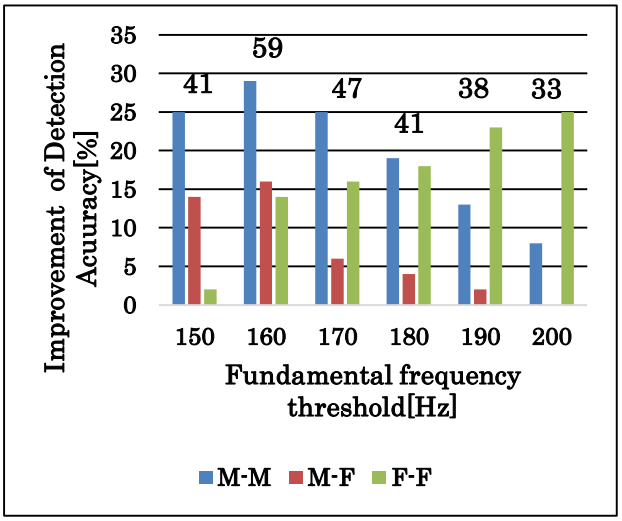

Figure 3 Improvement of accuracy vs. fundamental frequency threshold

The values shown above each histogram are the summation values of M-M, M-F and F-F. In this analysis, these values are evaluated since the effectiveness must be discussed for all the combination of speakers regardless of the gender. Figure 3 shows that the summation value for the fundamental frequency of 160 Hz is the largest, and thus it is concluded that the most appropriate threshold value of fundamental frequency is 160 Hz.

4.2 RESTRICTION TIME THRESHOLD

The threshold value of restriction time defined in Section 3 was evaluated varying the value from 0.1 to 10 to determine the appropriate the threshold value of restriction time.

In this analysis, two evaluation metrics were introduced, which are (1) the number of wrong detections when speaker did not change and (2) the number of wrong detections when speaker changed. For (1), it shows that the proposed method detected the speaker change even though the speaker did not change. For (2), it shows that the proposed method did not detect the speaker change even though the speaker changed.

Figure 4 shows the number of wrong detections when speaker did not change. It is obvious that as the restriction time increases, the number of wrong detection when speaker did not change decreases. This number tends to decrease drastically from 0.1 to 0.6. When the restriction time is 0.6, the number is almost 1/3 compared to the case when the restriction time is 0. In addition, this number decreases slightly from 0.6 to 1.0.

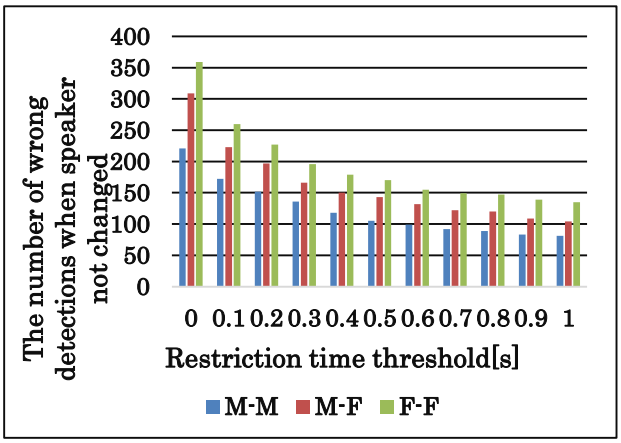

Figure 4 The number of wrong detection when speaker not changed

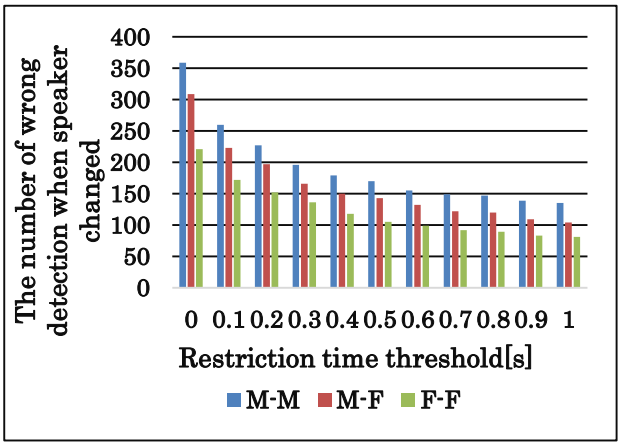

Figure 5 The number of wrong detection when speaker changed

Figure 5 shows the number of wrong detections when speaker changed. The tendency of the change is the same as the one in Fig.4. Therefore, it is clear that the threshold

values from 0.7 to 1.0 are not really effective, and thus it is concluded that the most appropriate threshold value of restriction time is 0.6.

Figure 6 shows the accuracy of speaker change detection when the proposed 3 threshold values were introduced. The proposed method achieved the speaker change detection accuracy of 83% when the speaker changed to the different gender. However, when the speaker changed to the same gender, namely, male-to-male and female-to-female, the accuracies are 53% and 41%, respectively. The vocal tube feature and the fundamental frequency are greatly different between males and females. However, the difference between the same genders is not much. In addition, the influence of the fundamental frequency threshold is not large for the speaker change detection between the same genders.

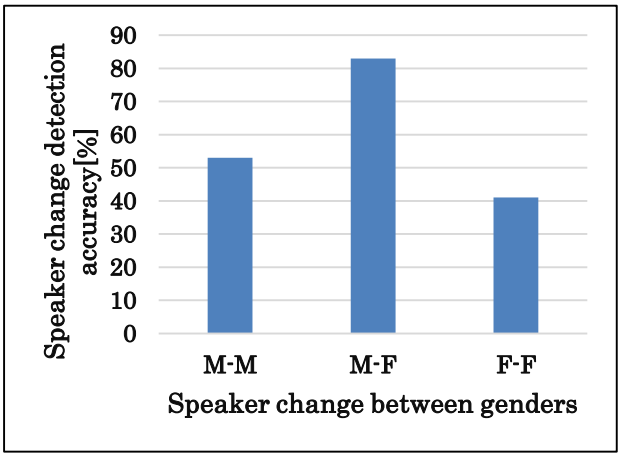

Figure 6 Speaker Change Detection Accuracy

5 CONCLUSION

In this paper, an accurate speaker change detection algorithm as the initial step of speaker verification was proposed. The proposed method improved the accuracy of speaker change detection using 3 thresholds which are newly introduced. Those thresholds are the threshold values of DP distance, fundamental frequency and restriction time. The threshold of DP distance was used to compare the waveforms of MelCepsF0 and to detect the speaker change. The threshold of fundamental frequency was used to distinguish the change of speaker's gender. The threshold of restriction time was used to prohibit detecting the speaker change within a short time.

As for the threshold of DP distance, the value of 180 determined in the previous work was used in this evaluation. As for the rest 2 thresholds, 160Hz for the fundamental frequency threshold and 0.6 seconds for the restriction time threshold were determined from the analyses. As a result, the proposed speaker change detection algorithm achieved the accuracy of more than 80% when the speaker changed to the different gender. However, when the speaker changed to the same gender, the accuracy deteriorated to around 50%. The solution of this will be considered as a future work.

REFERENCES

[1] L. Wang, M. Kazue, Y. Kazumasa, N. Seiichi, "Evaluation of speaker identification / verification method using phase information," Proc. SLP, vol.108, no.338, pp.173-178, 2008.

[2] W. H. Tsai, C. Che and W. W. Chang, "Text-Independent Speaker Identification Using Gaussian Mixture Bigram Models," Proc. ICSLP, vol. 2, pp. 314-317, 1999.

[3] C. Miyajima, Y. Hattori and K. Tokuda, "Text Independent Speaker Identification Using Gaussian Mixture Models Based on Multi-Space Probablity Distribution," Proc. IEICE, vol. E84-D, no.7, pp. 847-855, 1998.

[4] M. Tateno, E. Kamioka, "Detection of Speaker Change with Mel-Cepstrum Considering Speaker-Microphone Distance," Proc. WIT, vol.114, no.447, pp.21-26, 2015.

[5] M. Tateno, E. Kamioka, "Realtime Detection of Speaker Change Considering Mobile Device Environment," Proc. WIT, vol.115, no.354, pp.7-12, 2015.

[6] NINJAL, "Corpus of Spontaneous Japanese, " http://pj.ninjal.ac.jp/corpus_center/csj/

[7] M. Akagi, T. Ienaga, "Speaker Indivisuality in fundamental frequency contours and its control," Proc. IEICE, vol.18, no.2. pp.73-80, 1997.

BIBLIOGRAPHY

Masashi Tateno is currently a master course student at Graduate School of Engineering and Science, Shibaura Institute of Technology. He received his bachelor degree from Department of Communications Engineering, Shibaura Institute of Technology.

Eiji Kamioka is a Professor at Shibaura Institute of Technology. He received his B.S., M.S. and D.S degrees in Physics from Aoyama Gakuin University, Japan. Before joining SIT, he was working for SHARP Communications Laboratory, Institute of Space and Astronautical Science (ISAS) as a JSPS Research Fellow and National Institute of Informatics (NII) as an Assistant Professor. His current research interests encompass mobile multimedia communications and ubiquitous computing.