Identification of Child Users via Web-Based Voice Interface using Auditory Filterbank

Ryuichi Nisimura, Shoko Miyamori, Erika Okamoto, Toshio Irino, Hideki Kawahara
Auditory Media Laboratory, Wakayama University, Japan, nisimura@sys.wakayama-u.ac.jp

1. Introduction

In our past studies [1][2], a method to identify child speakers was developed on the basis of an automatic speech recognition (ASR) algorithm, which uses an acoustic hidden Markov model (HMM) and a support vector machine (SVM). To enhance the proposed system for use as a web application, we applied our original voice-enabled web framework to the front-end interface of the proposed system.

However because the voices of the majority of teenagers have a large variation in acoustic features, it is difficult even for a human being to identify their age groups exactly. To deal with the variations, we have introduced a new acoustic features derived from a gammachirp auditory filterbank (GCFB) [3] into the HMM classifier. We demonstrated that the GCFB-based feature outperformed the mel-frequency based features in the vocal tract length (VTL) estimation [4]. It would increase the reliability to identify children because the VTL is roughly proportional to the height of speaker which is also a function of the age of child.

In this presentation, we demonstrate our voice-enabled web application, and evaluate the HMM classifier improved by GCFB in comparison with a traditional mel-frequency based feature (MFCC).

2. Overview of our system

Figure 1 shows screen shots of our system —a web application running on a typical web browser. A web user can easily record his or her voice via the PC's microphone. The captured voice signals are transmitted to our web server where programs identify whether the speaker is an adult or a child. Finally, our system displays the result of the identification (child or adult) automatically like other cloud computing applications. The voice-enabled web system is composed by a simple pure Java applet and server-side programs. Our system can run on major operating systems and web browsers without installing special programs.

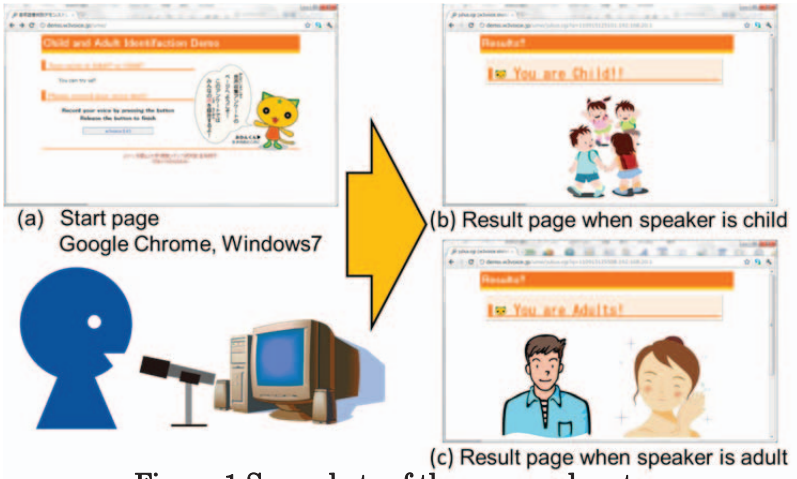

Figure 1 Snap shots of the proposed system

3. Evaluations

We made 3-states HMMs (128 Gaussians), which were constituted from 3 classes (adult-male, adult-female and child). GCFBs (# of channels: 25, 50, 100) and delta GCFBs were evaluated. The traditional features composed of 12-dim. MFCCs, delta MFCCs, and delta power were also tested. We used 2,360 utterances as the test data from our Japanese voice collection. To collect the voices in real home environments, we have completed the public testing the voice-enabled web site by the Internet users [1]. We performed a 10-fold cross-validation in which the speakers used for the evaluation were excluded from the training data.

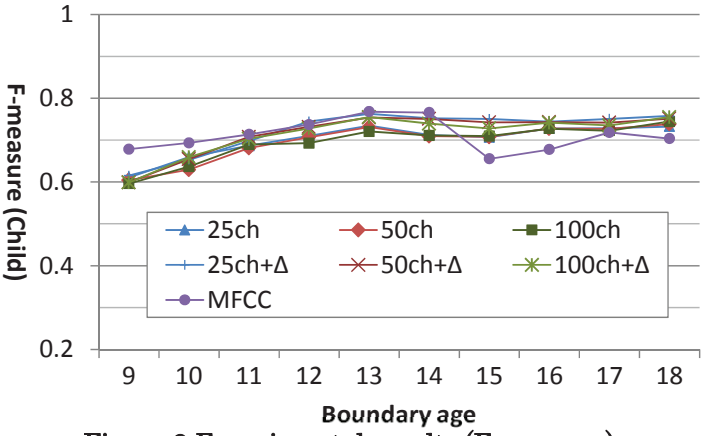

Figure 2 Experimental results (F-measure)

Figure 2 indicates experimental results showing the F-measures calculated from recall and precision in distinguishing child voices from the whole utterances. GCFBs could keep the high accuracies even when 15 years or more in the mid-teens was considered as a boundary. MFCCs, on the other hand, showed the tendency of remarkable performance decrement. Moreover, we could obtain good performance even when the number of GCFBs channels was small so that calculation resources could be reduced. This study proved that GCFBs have high potentials to identify child speakers. To improve accuracies, further optimizations of identification algorithm for GCFBs are being planned.

References

[1] Ryuichi Nisimura, et al., "Development of Web-Based Voice Interface to Identify Child Users Based on Automatic Speech Recognition System," Lecture Notes in Computer Science, vol. 6764 , pp.607-616, 2011.
[2] Shoko Miyamori, et al., "Real world utterance collection using voice-enabled web system for child speaker identification," Proc. 13th Oriental COCOSDA Workshop, 2010.
[3] Toshio Irino and Roy D. Patterson, "A dynamic compressive gammachirp auditory filterbank," IEEE Trans. Audio, Speech, and Language Process., 14(6), pp.2222-2232, Nov. 2006.
[4] Erika Okamoto, et al., "Auditory Filterbank Improves Voice Morphing", Proc. Interspeech2011, pp.2517-2520, 2011.