Detecting child speaker based on auditory feature vectors for VTL estimation

Ryuichi Nisimura*, Shoko Miyamori*, Erika Okamoto*, Hideki Kawahara*, and Toshio Irino*
* Wakayama University, Wakayama, Japan
E-mail: nisimura@sys.wakayama-u.ac.jp Tel/Fax: +81-73-457-8527

***Abstract*—We introduce novel auditory features in the hidden Markov model (HMM) system for detecting child speakers. The features derived by the gammachirp auditory filterbank (GCFB) have been demonstrated to be suitable for vocal tract length (VTL) estimation, both theoretically and experimentally. We performed numerical experiments to distinguish between child and adult speakers using HMMs trained on 2,360 speech samples collected through a web-based query interface, and we compared the performance of the common mel-frequency cepstral coefficients (MFCC) and the GCFB-based feature vectors. We also introduced the modulation features as the substitution of delta parameters. It has been clearly demonstrated that the error rate distinguishing a child from an adult is reduced by GCFB. To enhance our method for use as a web application, we applied our original voice-enabled web framework to the front-end interface of the proposed system.**

I. Introduction

To protect children from harmful content on the Internet, including violence and sexual matter, a reliable technique for confirming the age group to which a user belongs is required. Recently, parental control has become necessary to ensure a child's safety in the world of web networking. However, such parental control method are not reliable because they can be bypassed using several methods. While age group confirmation using information related to human behavior, such as facial images, is being researched, automatic speech recognition (ASR) has the potential to realize a friendly human-machine interface for children using information related to the speaker's voice, which is gathered while he or she is speaking naturally[1]. In addition, it is useful to know the attributes of the user of a system that incorporates spoken dialogue.

In our previous studies [2][3], a method of detecting child speakers was developed on the basis of ASR, which uses an acoustic hidden Markov model (HMM) and a support vector machine (SVM). However, when identifying speakers in the second half of their teenage years, we were not able to achieve sufficient accuracy. It is necessary to improve the accuracy of the system, especially when dealing with teenagers whose voices change frequently. Figure 1 indicates our preliminary examination results of human's hearing sense of distinguishing child and adult speakers. The line shows correct rates in which the target was correctly distinguished as the voice spoken from a child speaker. In this test, we conducted the subjective evaluation by 5 human subjects (2 males and 3 females). 260 utterances (male voice: 146; female voice: 114) were evaluated. The subjects listened to a recorded voice from a loudspeaker directly. It is difficult even for a human being to identify the age group of a teenage speaker because the voices of most teenagers vary widely in terms of their acoustic features. Our preliminary results also showed that the automatic approach is often erroneous, identifying a child's samples as belonging to an adult female. For details, you can refer to the Reference [3].

Fig. 1. Correct rate [%] in distinguishing child and adult speakers by humans evaluation[3]. "Boundary age (*x*-axis)" indicates the decision of the speaker age threshold that acts as a boundary between adults and children.

Several approaches to speech-based age estimation using ASR have been investigated. In traditional methods, acoustical feature vectors composed of the mel-frequency cepstral coefficients (MFCCs) were often used[6]. Wada et al. suggested methods using maximum a posteriori (MAP) adapted Gaussian mixture model (GMM) supervector features and a maximum likelihood linear regression (MLLR) transform vector[7]. Although we could achieve a certain performance using these features, which are suitable for ASR, it is also necessary to investigate the acoustical features based on the theoretical background of auditory studies.

To deal with the variations mentioned above, we have introduced a novel acoustic feature derived from a gammachirp auditory filterbank (GCFB)[8]. We demonstrate that the GCFB-based feature outperforms the mel-frequency-based features in vocal tract length (VTL) estimation[9]. This GCFB-based feature would increase the reliability of the HMM system for finding children because the VTL is roughly proportional to the height of a speaker, which is a function of the speaker's age.

II. Background of the gammachirp

The wavelet transform has been used to simulate the auditory filterbank as the first-order approximation above 500 Hz.

The optimal kernel function was derived as the "gammachirp" function, which satisfies the minimum uncertainty relationship between the joint representation of time and scale[12]. The scale representation is derived using the Mellin transform, which normalizes the scale variability. We proposed the stabilized wavelet-Mellin transform to extract the information about the vocal tract shape information and to segregate it from the information about the VTL[13], as a model for the early auditory systems. As an extension of the linear system, without losing the essential optimality, the gammachirp filterbank (GCFB) was developed to simulate nonlinearity known to exist in the auditory periphery, such as level-dependent auditory filter shape, fast-acting compression, and two-tone suppression[8].

A. *Successful VTL estimation by GCFB*

In the Reference [9], we performed numerical experiments to evaluate the stability of the VTL estimation by calculating the VTL ratios for all combinations of 28 speakers. Figure 2 shows the errors that occurred in VTL estimation when using various filterbanks. $\mathrm{GCFB_{dyn}}$ (leftmost) is GCFB with nonlinear dynamics. $\mathrm{GCFB_{lin}}$ is a linear version of GCFB in which the nonlinear circuit is cut off. $\mathrm{GTFB_{100}}$, $\mathrm{GTFB_{50}}$, and $\mathrm{GTFB_{25}}$ represent popular linear gammatone auditory filterbanks with the bandwidth of 100%, 50%, and 25% relative to the standard bandwidth, $\mathrm{ERB_N}$[14]. $\mathrm{MFFB_{STR24}}$, $\mathrm{MFFB_{STR40}}$, and $\mathrm{MFFB_{STR120}}$ represent mel-frequency filterbanks derived from F0-independent STRAIGHT spectrograms, where the number of filterbanks are represented in the suffix. $\mathrm{MFFB_{STFT24}}$, $\mathrm{MFFB_{STFT40}}$, and $\mathrm{MFFB_{STFT120}}$ (rightmost) represent standard, STFT-based mel-frequency filterbanks, which are commonly used as the preprocessor to calculate the MFCCs. The results clearly demonstrate that $\mathrm{GCFB_{dyn}}$ is the best filterbank for the VTL estimation. Since it performs better than $\mathrm{GCFB_{lin}}$ (which is the second best), it can be concluded that the nonlinearity in the auditory filterbank improves VTL estimation performance. The results imply that the use of GCFB is also effective for detecting child speakers because the VTL is highly correlated with the body height[15], which is smaller for a child than for an adult (see Fig. 3).

III. Experiments of distinguishing between child and adult speakers based on HMM

We compared the acoustic features derived from the GCFB and MFCC. The experimental conditions are shown in Table I. We have introduced three-type GCFB-based feature vectors: $\mathrm{GCFB_{25}}$, GCCC, and GCMC. $\mathrm{GCFB_{25}}$ represents a simple frequency domain feature, which consists of 25 channel spectral values of GCFB. The first order regression coefficients (Δ) are appended to $\mathrm{GCFB_{25}}$ as dynamic information. GCCC is a cepstral domain feature generated by the same generation process as MFCC. The logarithmic gammachirp spectral values are transformed to the cepstral domain by means of discrete cosine transform (DCT). 12-dimensional cepstral coefficients $\mathrm{C_1}$ to $\mathrm{C_{12}}$ are produced, and Δ are appended.

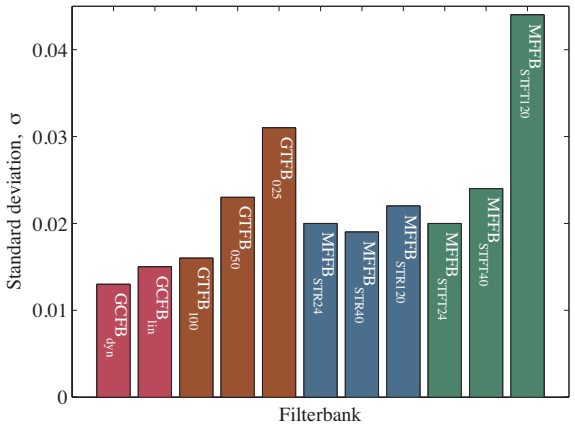

Fig. 2. Error (represented as standard deviation σ) in the VTL estimation by various filterbanks[9].

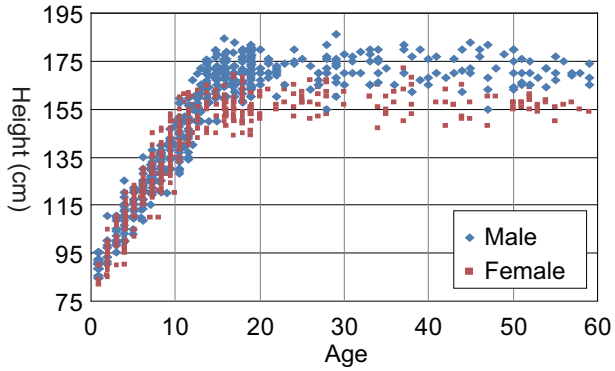

Fig. 3. Speakers age and height in our collection.

As shown in Figure 4, GCMC introduces the modulation features of each coefficient[16] as the substitution of Δ. The modulation spectrum is calculated from GCCC ($\mathrm{C_0}$ to $\mathrm{C_{12}}$) (100 Hz sampling, 160 ms window with a 10-ms frame shift). Cumulated energies for the frequencies between 2 to 16 Hz are computed as $\mathrm{C_{13}}$ to $\mathrm{C_{26}}$ after applying the bandpass filter. Introducing the modulation feature is motivated by findings that the linguistic information of speech is distributed in a limited modulation frequency region[17], [18], [19]. We could improve robustness due to the modulation features given by low frequencies between 2 to 16 Hz, which is relevant to important linguistic information of speech signal[16] Thus, environmental changes according to recording conditions would be reduced by GCMC.

When GCFB was used in our original features, input speech sound was analyzed as a two-dimensional cochlear spectrogram by the dynamic compressive GCFB[8]. The center frequencies of the gammachirp filters were equally spaced on the $\mathrm{ERB_N}$-number[14] axis between 100 and 6,000 Hz. A Hamming window of 30 ms with a 10-ms frame shift was applied to the power of the GCFB output to derive the smoothed spectrogram.

The acoustic features for reference were 12-dimensional

TABLE I
EXPERIMENTAL CONDITIONS.

Acoustic features	
$GCFB_{25}$	GCFB (25 dims.) + Δ
GCCC	cepstral domain GCFB (12 dims.) + Δ
GCMC	cepstral domain GCFB (12 dims.) + modulation
MFCC	MFCC (12 dims.) + Δ + Δ power
Configurations	
test data	2,360 utterances from 2 to 59 years (10-fold cross validation)
HMM type	three class HMMs three state 128 Gaussians, left-to-right
Tool	HMM Builder: HTK3.4.1[10] Classifier: Julius4.1.4[11]

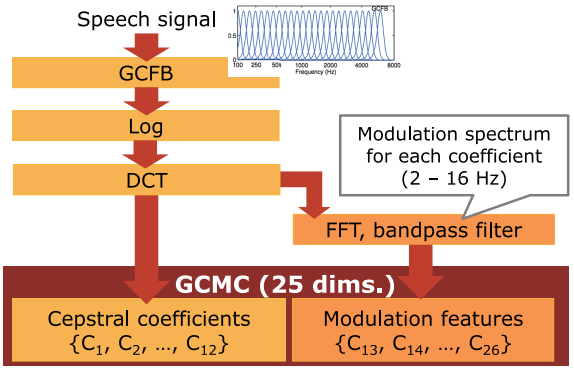

Fig. 4. Processing for GCMC feature, which consists of the gammachirp cepstral coefficients and modulation coefficients.

MFCC, Δ and Δ power, which are commonly used in ASR.

In order to compare the features by a time-tested technique, we adopted a simple HMM-based method that was developed for speaker recognition[20]. It had been developed on the basis of a isolated word ASR approach using the speech recognizer Julius[11]. Because this approach is widely used as the speaker recognition method, we think that it is suitable for measuring the performance of the features impartially. The three class HMMs with three state GMMs (Gaussian Mixture Model) were built to classify three categories: *Child*, *Adult (Female)*, and *Adult (Male)*, which were trained for correspondence between the acoustic features and class label for each utterance. In order to consider diversities of acoustical characters after the age at which teenager's voice changes, female and male of adults were set as another HMM class. The number of Gaussian we used was 128.

We set the boundary age to categorize a child or adult based on the original personal age data: it is the decision that acts as a boundary between adults and children. We set eight boundary ages, i.e., one for every year between 13 and 20 years, to find the best performance. For example, when the boundary age was set as 15 years, the speakers in the age group of 0 to 14 years were categorized as children while speakers above 15 years of age were considered as adults.

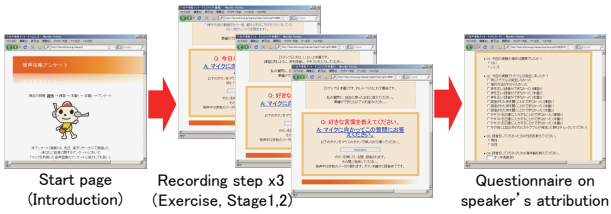

Fig. 5. Constitution of the voice-enabled website used for collecting utterance.

Fig. 6. Screen capture of the comic-like interface to introduce our system.

A. *Test data collected via the Internet*

In the experiments, a web-based system enabled us to evaluate the utterances collected from real home environments via the Internet. We assumed our approach will be adopted at home in everyday life. Therefore, we had to organize a collection of voices recorded in home environments in order to measure the performances[3].

Figure 5 illustrates the constitution of the website used for collecting the utterances. In order to provide an easy and friendly interface for recording the voices of children, we developed our website by introducing a comic-like interface and Flash animations in the introduction section of the experiment as shown in Figure 6.

This website has recorded the trial user's voice three times as shown as "Exercise", "Stage 1" and "Stage 2". In the exercise stage, the user was familiarized with the recording interface. The trial user would utter the answer after a simple question was displayed in Japanese. The questions that were displayed in Stage 1 and Stage 2 were as follows.

- Stage 1: "Could you tell me your favorite food?"
- Stage 2: "Please tell me your favorite words."

All captured voices would be automatically uploaded to our web server. After completing three recording steps, trial users and their parents were requested to fill surveys to report their genders, ages, and hometowns. When it is difficult for child users to operate PCs and fill the reports, we requested to their parents vicarious operations excluding speaking acts.

The trial users were recruited via the Internet monitor

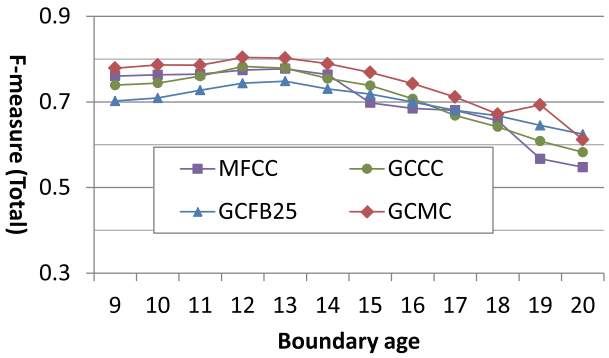

Fig. 7. Experimental results (F-measure).

invitation service of the Rakuten research company[1], which is one of the web-based crowdsourcing service provider in Japan. For speech studies, the crowdsourcing is a useful approach that involves outsourcing tasks to a distributed group of people[4], [5]. In order to cover a wide variety of attributions such as age groups and genders in the collection, the trial users of our studies were adjusted by a prior screening on the basis of preliminary surveys by the Rakuten's service.

We succeeded in collecting utterances from 1,152 trial web users through the public test of our website. The author confirmed the captured voices manually since these voices could include invalid recording data. We collected 3,053 utterances of 1,050 speakers, excluding the invalid data. Figure 3 shows the distribution of the speakers' age and body height. We were able to achieve an adequate balance between the number of samples of child voices and those of the adult voices; 59.7% of the total 3,053 samples were uttered by children. To evaluate the proposed method, we used 2,360 utterances from our collection as the test data, and we performed a 10-fold cross-validation. The speakers used for the evaluation were excluded from the training data.

B. Results

Figure 7 indicates experimental results showing the F-measure, where the horizontal (x-)axis represent the boundary age. The F-measure values indicating the total accuracy in distinguishing between children and adults were calculated as follows:

$$F - measure = \frac{2 \times Precision \times Recall}{(Precision + Recall)}. \quad (1)$$

Precision is a measure of exactness or fidelity given by Eq. (2), and recall is a measure of completeness given by Eq. (3).

$$Precision = \frac{Correct\ result}{Correct\ result + Unexpected\ result} \quad (2)$$

$$Recall = \frac{Correct\ result}{Correct\ result + Missing\ result} \quad (3)$$

In the figure, we achieved the best performance when using the feature vectors of GCMC (red line) in comparison with other features. GCMC (red) has marked improved accuracies on average 0.06 from MFCC (purple line), and 0.03 from GCCC (green line). Although the performance of the MFCC tended to decrease when the boundary age was set in the late-teens (15 years or more), the GCFB-based features, such as GCMC (red), GCCC (green), and GCFB_{25} (blue), show comparatively stable performances. This implies that the GCFB-based feature can outperform the MFCC in distinguishing the majority of teenagers, whose acoustic features vary widely. GCMC (red) in comparison with GCCC (green) proved that the modulation features yielded further improved accuracies by enhancing the speech intelligibility in the signal.

The further details about classified results when the boundary age was set as 18 years are shown in Figure 8 (MFCC) and Figure 9 (GCMC). The figures illustrate classification rates (%) of all samples, where the horizontal (x-)axises indicate the speaker's age on the basis of the original personal age data. Because the boundary age was 18 years, the samples in the age group of 0 to 17 years should be classified into the green area (child). That is, it shows good performance that the green area from 0 to 17 years (x-axises) is large.

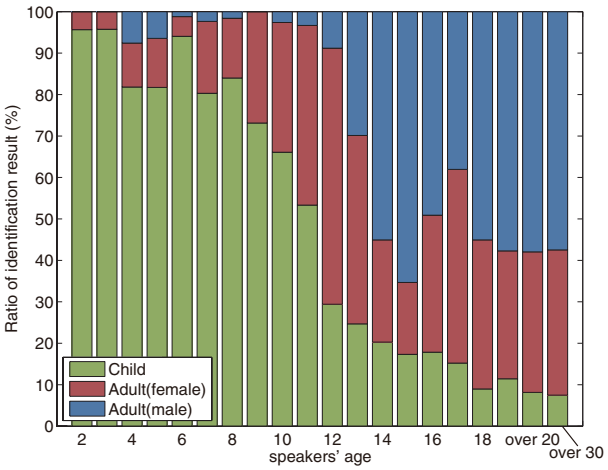

Fig. 8. Distribution of classified results. (MFCC, Boundary age: 18 years old)

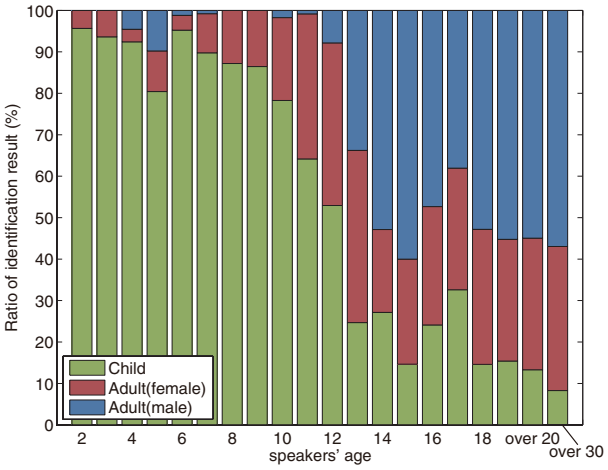

Fig. 9. Distribution of classified results. (GCMC, Boundary age: 18 years old)

[1]http://research.rakuten.co.jp/

Especially, we proved that the GCMC is able to reduce the error rates by 12 % on the average in the age group of 8 to 12 years, and an improvement of 5 % in the age group of 13 to 17 years is obtained. In distinguishing between adult and child speakers, there was a serious problem in that the majority of teenagers voices are often confused because of the acoustical diversity of their voice characteristics. It was clearly demonstrated that the GCFB-based feature can conduce distinguishing the acoustical features of the teenagers voices.

IV. Web-based prototype system

We have developed a prototype system to demonstrate and evaluate the proposed method through public testing by Internet users. Figure 10 shows screen shots of our system running on a typical web browser as a web application. A web user can easily record his or her voice using 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 consist of a simple pure Java applet and server-side programs. As described in [21], our voice-enabled web system can run on all major operating systems and web browsers without the installation of special programs. We have a plan to release the prototype system as a free software for the beta-test by Internet users.

V. Conclusion

Auditory feature vectors derived from the gammachirp auditory filterbank (GCFB) have been tested for detecting child users. As the GCFB is suitable for VTL estimation, both theoretically and experimentally, a comparison of the HMM-based method using the common MFCC and the GCFB-based features showed that the GCFB significantly improves the accuracy. As future works, computing of the GCFB needs to be accelerated to develop fast and smooth interaction interface. To improve the accuracy, instead of a text-independent system, a text-dependent system could be considered. We intend to introduce linguistic information[22] representing characteristic differences between child and adult speakers.

Acknowledgment

This study was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (KAKENHI), Japan, and 2011 Original Research Support Project of Wakayama University.

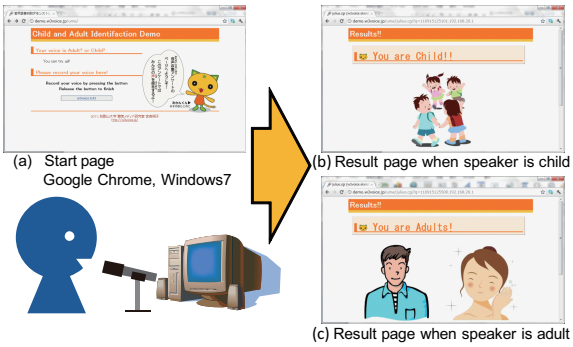

Fig. 10. Snap shots of the prototype system running on a web browser.

References

[1] Gerosa, M., et al., "A review of ASR technologies for children's speech," Proc. 2nd Workshop on Child, Computer and Interaction, pp. 1–8, 2009.
[2] Nisimura, R., 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.
[3] Miyamori, S., et al., "Real world utterance collection using voice-enabled web system for child speaker identification," Proc. 13th Oriental COCOSDA Workshop, 2010.
[4] Callison-Burch, C. and Dredze, M., "Creating speech and language data with Amazon's Mechanical Turk," Proc. the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 1–12, 2010.
[5] Parent, G. and Eskenazi, M., "Speaking to the Crowd: looking at past achievements in using crowdsourcing for speech and predicting future challenges," Proc. Interspeech, pp. 3037–040, 2011.
[6] Müller, C. and Burkhardt, F., "Combining Short-term Cepstral and Long-term Pitch Features for Automatic Recognition of Speaker Age," Proc. Interspeech, pp. 2277–2280, 2007.
[7] Wada, T., et al., "Investigations of Features and Estimators for Speech-based Age Estimation," Proc. APSIPA, pp. 470–473, 2010.
[8] Irino, T. and Patterson, R.D., "A dynamic compressive gammachirp auditory filterbank," IEEE Trans. Audio, Speech, and Language Process., vol. 14, no. 6, pp. 2222–2232, 2006.
[9] Okamoto, E., et al., "Auditory Filterbank Improves Voice Morphing," Proc. Interspeech, pp. 2517–2520, 2011.
[10] Young, S.J., et al., "The HTK book version 3.4," Cambridge University Engineering Department, Cambridge, UK, 2006.
[11] Lee, A., et al., "Julius – An Open Source Real-Time Large Vocabulary Recognition Engine," Proc. Eurospeech, pp. 1691–1694, 2001.
[12] Irino, T. and Patterson, R.D., "A time-domain, level-dependent auditory filter: the gammachirp," J. Acoust. Soc. Am., vol. 101, no. 1, pp. 412–419, 1997.
[13] Irino, T. and Patterson, R.D., "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The Stabilised Wavelet Mellin Transform," Speech Communication, vol. 36, issues 3–4, pp. 181–203, 2002.
[14] B.R. Glasberg and B.C. J.Moore, "Derivation of auditory filter shapes from notched-noise data," Hear.Res., vol. 47, no. 4, pp. 103–138, 1990.
[15] W.T. Fitch, J. Giedd, "Morphology and development of the human vocal tract: a study using magnetic resonance imaging," J. Acoust. Soc. Am., vol. 106, pp. 1511–1522, 1999.
[16] HariKrishna, M. and Marco, M., "A Level-dependent Auditory Filterbank for Speech Recognition in Reverberant Environments," Proc. Interspeech, pp. 685–688, 2011.
[17] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am., vol.95, no.2, pp. 1053–1064, 1994.
[18] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," J. Acoust. Soc. Am., vol.95, no.5, pp. 2670–2680, 1994.
[19] N. Kanedera, T. Arai, and T. Funada, "Robust Automatic Speech Recognition Emphasizing Important Modulation Spectrum," The Institute of Electronics, Information and Communication Engineers, vol.J84-D-2, no.7, pp.1261–1269, 2011.
[20] D.A.Reynolds, R.C.Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. on Speech and Audio Processing, vol. 3, no. 1, pp. 72–83, 1995.
[21] Nisimura, R., et al., "Development of Speech Input Method for Interactive VoiceWeb Systems," Lecture Notes in Computer Science, vol. 5611, pp. 710–719, Springer, 2009.
[22] Nisimura, R., et al., "Public Speech-Oriented Guidance System with Adult and Child Discrimination Capability," Proc. ICASSP, vol.1, pp.433–436, 2004.