SPEECH FEATURE EXTRACTION METHOD REPRESENTING PERIODICITY AND APERIODICITY IN SUB BANDS FOR ROBUST SPEECH RECOGNITION

Kentaro Ishizuka and Noboru Miyazaki

NTT Communication Science Laboratories, NTT Corporation

ABSTRACT

This paper proposes a feature extraction method that represents both the periodicity and aperiodicity of speech for robust speech recognition. The development of this feature extraction method was motivated by findings in speech perception research. With this method, the speech sound is filtered by Gammatone filter banks, and then the output of each filter is comb filtered. Individual comb filters designed for each output signal of the Gammatone filter are used to divide the output of each filter into its periodic and aperiodic features in the sub band. The power suppressed by comb filtering is considered to be a periodic feature, whereas the power of the residue after comb filtering is considered to be an aperiodic feature. This method uses both features as the feature parameters for automatic speech recognition. A preliminary experiment using a five vowel recognition task designed to compare the proposed approach with the conventional MFCC-based feature extraction method shows that the proposed method improves vowel recognition rates by as much as 14.7 % in the presence of pink noise or a harmonic complex tone interferer. An evaluation experiment undertaken using the Aurora-2J database (Japanese noisy digit recognition database) to compare the proposed approach with the MFCC-based conventional (baseline) feature extraction method shows that the proposed method reduces the word error rate by as much as 59.62 %, with an average value of 18.21 %.

1. INTRODUCTION

After Davis and Mermelstein reported that Mel-frequency cepstral coefficients (MFCCs) provided better performance than other features in 1980 [1], the MFCC has been widely used as the feature parameter for automatic speech recognition (ASR). However, the MFCC is not robust enough in noisy environments, which suggests that the MFCC still has insufficient sound representation capability. This has created a need for feature extraction methods designed to represent more robust features.

Most of these methods are based on findings related to the psychology or physiology of the human auditory system e.g. GSD [2], PLP [3], EIH [4], ZCPA [5], and ALSD based on GSD [6]. These methods aim to simulate the speech processing of the human auditory system, and most focus on the neural firing cycle or some similar periodicity. From the engineering viewpoint, SBCOR [7] also focuses on periodicity, especially on the center frequency of the sub band filter, in a way similar to GSD or ALSD. Since most of these methods have the advantage of representing the periodicity, they can improve ASR performance in noisy environments. Recently, psychological research has also revealed that the human auditory system is very sensitive to the harmonicity that is related to the periodicity of sound [8].

However, these feature extraction methods have no advantage when it comes to representing aperiodic sound, besides they lose aperiodic information about quasi- or non-periodic sound.

By contrast, speech perception research has revealed that the human auditory system is also sensitive to aperiodicity. In concurrent vowel recognition research, de Cheveigné et al. [9] showed that the target vowel was perceived more easily when the interferer vowel was harmonic rather than inharmonic sound. This result suggests the existence of a mechanism similar to the comb filter for canceling the harmonicity of sound in the human auditory system, and that the human auditory system may perceive the target vowel after harmonic canceling. Therefore, it is conceivable that the human auditory system may represent both the harmonic i.e. periodic feature and the residue after canceling the harmonicity i.e. aperiodic feature, which deviates from the dominant periodicity. Recently, Ishizuka and Aikawa [10] showed that very small fundamental frequency (F0) fluctuations of vowels improve human vowel identification. Their results also support the importance of aperiodicity. However, in terms of engineering, methods of employing both periodicity and aperiodicity have not been well studied except by Jackson et al. [11].

This paper proposes a feature extraction method that represents both periodic and aperiodic features for each sub band using Gammatone filter banks and comb filters. Unlike previous studies [11], our method can improve the ASR performance in noisy environments without precise F0 estimations from clean speech or voicing detections. The proposed method is described in detail in section 2. In section 3, a preliminary experiment confirms that the proposed method improves the vowel recognition rates in the presence of interferers. In addition, an experiment with the Aurora-2J database (Japanese noisy digit recognition database) shows that the proposed method can reduce the word error rate in real noise environments.

2. METHOD

Figure 1 shows a block diagram of our proposed method. In the first step, the input speech is divided into sub band signals by Gammatone filter banks [12]. The center frequencies and bandwidths for each filter in the filter banks are decided in terms of the equivalent rectangular bandwidth (ERB) scale. In our example, we use 24 filters whose frequency characteristics are shown in Fig. 2. In the second step, the output signal for each filter is divided into frames with a certain temporal length, e.g. 30 ms, and are shifted a certain temporal length, e.g. 10 ms. In the third step, the dominant periodicity is detected in each frame. The periodicity is calculated using the same method as that used in the autocorrelation method for F0 estimation. The method

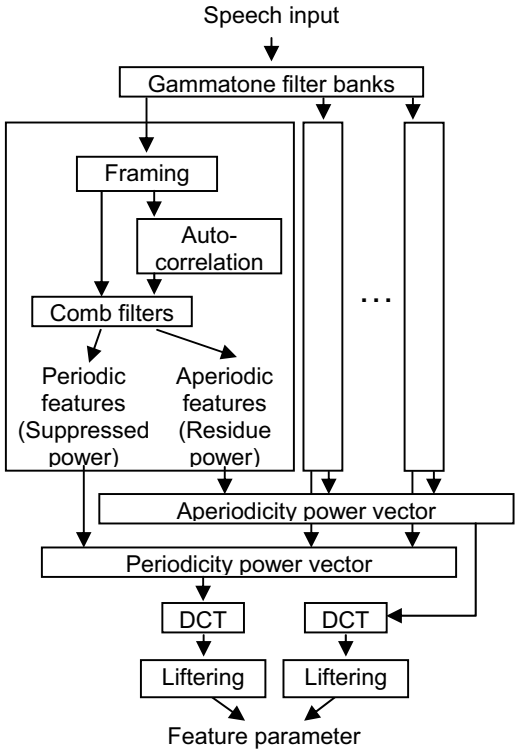

***Figure 1**: Block diagram of the proposed method.*

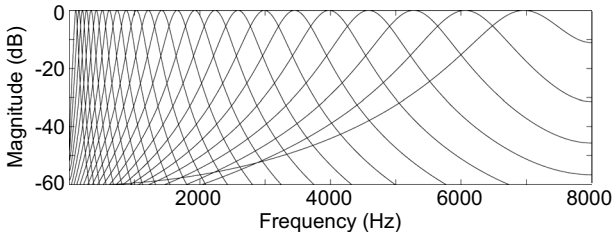

***Figure 2**: The frequency characteristics of 24-channel Gammatone filter banks.*

calculates the autocorrelation function of the signal in the frame and searches for the maximum peak of the function within the search range e.g. from 80 to 200 Hz. In the fourth step, the signal in the frame is comb filtered using the periodicity detected in the third step. The characteristic of the comb filter is given by $H(z)$, where n indicates the period with the maximum value detected in the third step.

$$H(z) = 1 - z^{-n}$$

In the fifth step, the power suppressed by the comb filtering and the power of the residue signal in the frame after the comb filtering are calculated as the sum of the square of the signals. The power suppressed by the comb filtering is calculated as the difference between the signal powers before and after the comb filtering. After the fifth step, the power suppressed by the comb filtering is considered to be the periodic feature, and the power of the residue signal is considered to be the aperiodic feature. Then, the powers across the sub bands at the same frame shift point are combined for each feature and considered to be vectors. Figure 3 shows the output power patterns from Gammatone filter

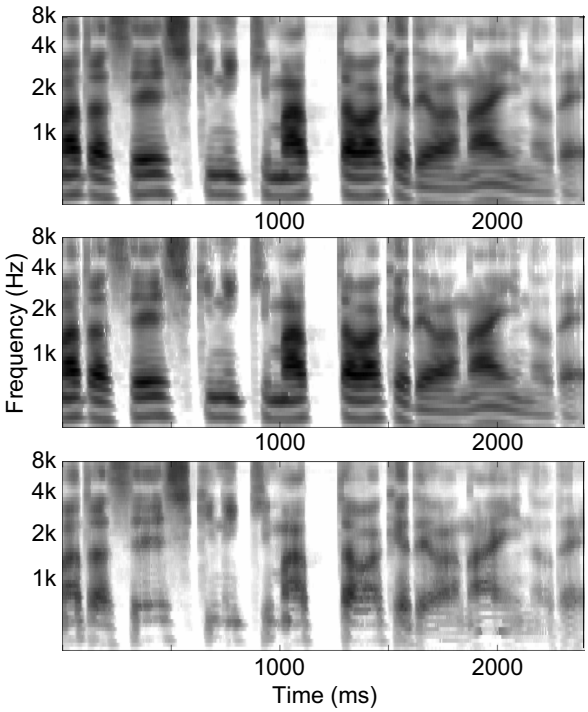

***Figure 3**: The excitation patterns i.e. the output powers from Gammatone filter banks (top), and the periodic (middle) and aperiodic features (bottom) of the Japanese 'mada seisiki ni kimatta wake de wa nai no de' read by a male speaker. The dark color indicates the power intensity in the region. The periodic feature well represents the power of the stable periodicity, whereas the aperiodic feature clearly represents the power changing part e.g. sound onset and formant transition.*

banks (i.e. similar to excitation patterns) and the vectors of its periodic and aperiodic features obtained after analyzing a speech sentence with the 48-channel Gammatone filter banks and the 30-ms frame shifted by 10 ms. As seen in Fig. 3, the periodic features have a power pattern that well represents stable periodicity, whereas the aperiodic features show the fluctuations of sound that deviate from the periodicity, such as sound onset, rapid frequency or amplitude changes in a frame, more clearly than the excitation pattern. In the final step, each power vector calculated in the fifth step is discrete cosine transformed into cepstral coefficients. We use the transformation shown below, where N is the number of Gammatone filters, m_j is the power vector for filter number j, and c_i is the i-th coefficient.

$$c_i = \sqrt{\frac{N}{2}} \sum_{j=1}^{N} \log(m_j) \cos\left(\frac{\pi i}{N}(j - 0.5)\right)$$

This transformation is the same as that used with the MFCC method. These coefficients are calculated for each feature, and only certain low order coefficients (e.g. the first to 12th coefficients) are used as the feature parameters for ASR. Then, both features are combined as the feature parameter, that is, if the coefficients from the first to 12th order are used then the total number of feature parameters is 24.

The key to this method is the representation of both periodic and aperiodic features. In addition, the periodicity in a channel is calculated in a more adaptive manner compared with

GSD [2] or SBCOR [7] which only focus on the center frequency of the band pass filter. Our method is similar to the method proposed by Jackson et al. [11] as regards the key points. However, the first step of their method depends strongly on the accuracy of the pitch-scaled harmonic filter (PHIF), and so the effect of any failure to decompose the harmonicity in the first step may become very large. Therefore, in their experiment, they used F0 values estimated from clean speech data to decompose the periodicity and aperiodicity of noisy speech data. By contrast, because our method employs a band pass filter before dividing the input speech into periodic and aperiodic sound, it is expected to have such advantages as being able to recover a failed harmonicity estimation and to cope with an interferer whose energy is not distributed evenly in the frequency region. Such failures in some channels do not affect the other channels, so our method can correctly extract speech features in channels with high local signal to noise ratios (SNRs). Therefore, our method does not need clean F0 estimation.

Our method also can be regarded as an enhanced MFCC method in terms of sound representation capability. When Mel-scale filter banks are used instead of Gammatone filter banks, the sum of the periodic and aperiodic features in each channel provides the same representation as the MFCC. However, the division into two features is expected to reduce the influence of the power pattern distortion in noisy environments. In addition, the proposed approach is expected to improve robustness since the representation it provides reflects properties that the MFCC cannot deal with because of the sound onset and rapid frequency changes in a frame with an aperiodic feature.

3. EXPERIMENTS

3.1. Preliminary vowel recognition experiment

To evaluate the proposed method, we first conducted a preliminary experiment. The vowel recognition rates in the presence of pink noise or a harmonic complex tone interferer were measured at various SNRs. We measured the robustness by comparing the rates obtained with the proposed method and a conventional MFCC method. In this experiment, we used the hidden Markov model (HMM) as a pattern classifier. The proposed method uses 24-channel Gammatone filter banks, a 30 ms frame shifted by 10 ms, and 12-order coefficients for each feature i.e. 24-dimension feature parameters. By comparison with the proposed method, the MFCC method uses 24-channel Mel-scale filter banks, the same frame length and shift as the proposed method, and 12-order coefficients. In addition, the MFCC method uses dynamic features (ΔMFCC), and the dimension of the feature parameters was 24.

The training speech data consisted of 105 sentences from JNAS Japanese speech corpora, which are sentences extracted from a newspaper database and read by a male speaker in a clean environment. 64-Gaussian mixture monophone HMMs were trained using HTK [13]. The test speech data were five Japanese vowels /a/, /i/, /u/, /e/, and /o/ extracted from the training data, which were correctly recognized by the trained HMM in a clean environment. The test set of vowels, whose shortest length was 50 ms, were randomly extracted from the training data, and 1,191 vowels were used for the test. These vowels were used as isolated vowels, that is, each test sound included only one vowel. Test sounds were generated by adding pink noise or a harmonic complex tone interferer to the test speech data at SNRs of 20, 10, and 5 dB. The harmonic complex tone consisted of a 100 Hz pure tone and its 50 harmonics whose powers decreased at -3 dB/oct., that is, the same as the pink noise. The F0 of the harmonic complex tone interferer was constant at 100 Hz. The vowel recognition rates were measured through a single vowel recognition task. We used the Japanese speech recognizer Julian [14] with a network grammar that allows sentences containing only one of the vowels. The speech condition was closed, and the interferer conditions were open.

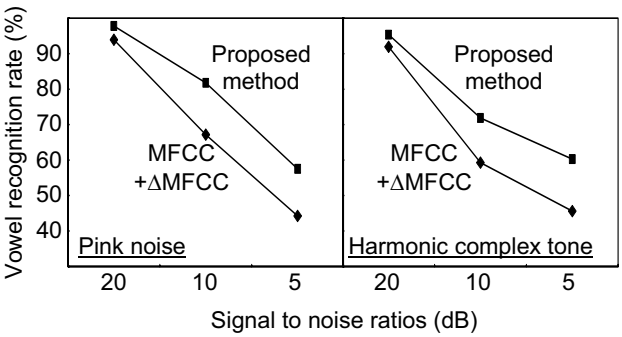

***Figure 4**: Vowel recognition experimental results. The vowel recognition rates in the presence of pink noise (left) and a harmonic complex tone (right). The abscissa indicates SNRs, and the ordinate indicates the vowel recognition rate.*

Figure 4 shows the results for each interferer. In both interferers at all SNRs, the vowel recognition rates obtained with the proposed method were always higher than the rates obtained with MFCC+ΔMFCC. The maximum improvement in the recognition rates was 14.7 %.

As shown in Fig. 4, although the feature parameters provided by the proposed method were static parameters (without deltas), the experimental results showed that the proposed method is more robust in terms of vowel recognition in the presence of pink noise or a harmonic complex tone interferer than MFCC+ΔMFCC, which includes dynamic feature parameters. This result confirms that the proposed method can improve ASR performance in the presence of interferers.

3.2. Noisy digit recognition experiment with Aurora-2J

We also conducted an evaluation experiment with the Aurora-2J Japanese noisy digit recognition database. The evaluation category was 5 because only the feature extraction process was changed. We measured the robustness by comparing the word accuracies obtained with the proposed method and baseline feature parameters i.e. 12-order MFCCs and a log power, and their deltas and accelerations (39 dimensions). The baseline scripts were used unchanged for training and testing. That is, we used 16-state 20-Gausiaan mixture HMMs as a pattern classifier. The proposed method used the same Gammatone filter banks, frames, and 12-order coefficients as in section 3.1 i.e. 24-dimension feature parameters. In addition, the dynamic features (delta parameters of the coefficients calculated in the same way as the ΔMFCC) were used, and the total dimension of the feature parameters was 48. In this paper, only clean training HMMs were used throughout the recognition experiments.

Figure 5 and Table 1 show the averaged word accuracies with the baseline MFCC-based features and with the proposed features, and the reduction in the word error rate (WER) from the baseline features realized by the proposed method. The

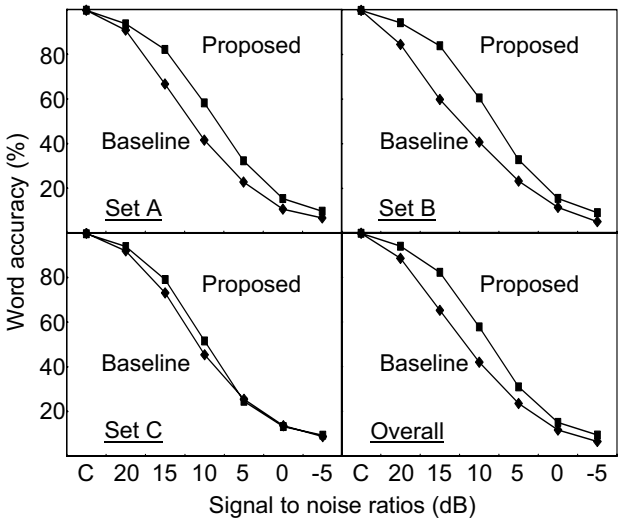

***Figure 5**: Experimental results with Aurora-2J. The word accuracies for set A (top left), B (top right), C (bottom left) and their averages (bottom right). The abscissa indicates SNRs, and the ordinate indicates the word accuracy. Note that SNR "C" on the abscissa means "clean" conditions.*

maximum reduction in WER was 77.9 % (airport noise at 20 dB SNR, not shown in Figure 5 or Table 1), and the average value was 18.21 % at SNRs of 0 to 20 dB. At an SNR of 15 dB, the proposed method shows the best WER reduction performance.

Without any noise reduction techniques, the proposed method not only maintained (or improved) the accuracy in clean environments, but also improved its accuracy especially at SNRs of 20 to 10 dB. In this SNRs region, the averaged WER reduction was 36.49 %. It is also expected that the performance of this method will improve further in conjunction with certain noise reduction methods. Also in this experiment, although the feature parameters provided by the proposed method were only static parameters and its deltas, the experimental results show that the proposed method is more robust in terms of digit recognition in noisy environments than the MFCC-based features, which use its acceleration and log power parameters added to static and its delta. The experimental results confirm that the feature representation provided by our proposed method is also useful for robust ASR in real noise environments.

4. CONCLUSION

This paper proposed a speech feature extraction method representing periodic and aperiodic features for robust ASR. The method uses Gammatone filter banks and comb filters to divide speech signals into two features. An evaluation experiment with the Aurora-2J database showed that the proposed feature extraction method provides better performance in the presence of noise than the conventional MFCC-based feature extraction method. The results indicate that such an enhancement in sound representation can improve the robustness of an ASR system.

Acknowledgements: The authors thank Dr. Tomohiro Nakatani and Dr. Yasuhiro Minami (NTT Communication Science Labs.) for very helpful advice and discussions in relation to this work. This research employed the noisy speech recognition evaluation environment (Aurora-2J) produced by the IPSJ SIG-SLP noisy speech recognition working group.

***Table 1**: Experimental results with the Aurora-2J. The word accuracies with the baseline features and the proposed features (top) and the WER reduction rate (bottom) at an SNR of 15 dB and the average value at SNRs of 0-20 dB. At an SNR of 15 dB, our method showed the best performance in WER reduction.*

Word Accuracy (%)					
Method	**SNR**	**Set A**	**Set B**	**Set C**	**Overall**
Baseline	**Average**	**46.51**	**43.98**	**49.90**	**46.17**
	15 dB	**66.67**	**59.83**	**73.10**	**65.22**
Proposed	**Average**	**56.33**	**57.38**	**52.46**	**55.98**
	15 dB	**82.15**	**83.84**	**79.08**	**82.21**
WER reduction rate (%): Relative performance					
Proposed	**Average**	**18.37**	**23.92**	**5.10**	**18.21**
	15 dB	**45.65**	**59.62**	**20.77**	**46.26**

REFERENCES

[1] Davis, S. B. and Mermelstein, P. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. on Acoustic, Speech and Signal Processing, **ASSP-28**, No. 4, 1980.

[2] Seneff, S. "A joint synchrony/mean-rate model of auditory speech processing," J. of Phonetics, **16**, 55-76, 1988.

[3] Hermansky, H. "Perceptual linear predictive (PLP) analysis of speech," J. of Acoust. Soc. Am., **87**, 1738-1752, 1990.

[4] Ghitza, O. "Auditory models and human performance in tasks related to speech coding and speech recognition," IEEE Trans. on Speech and Audio Processing, **2**, No.1, 115-132, 1994.

[5] Kim, D. S., Lee, S. Y. and Kil, R. M. "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," IEEE Trans. on Speech and Audio Processing, **7**, No.1, 55-69, 1999.

[6] Ali, A. M., Spiegel, J. V. and Mueller, P. "Robust auditory-based speech processing using the average localized synchrony detection," IEEE Trans. on Speech and Audio Processing, **10**, No.5, 279-292, 2002.

[7] Kajita, S. and Itakura, F. "Robust feature extraction using SBCOR analysis," *Proc. of ICASSP*, 421-424, 1995.

[8] Darwin, C. J. and Carlyon. R. P. "Auditory grouping," in *Hearing*, Academic Press, San Diego, 387-424, 1995.

[9] de Cheveigné, A., McAdams, S., and Marin C. M. H. "Concurrent vowel identification. II. Effects of phase, harmonicity, and task," J. Acoust. Soc. Am. **101**, 2848-2856, 1997.

[10] Ishizuka, K. and Aikawa, K. "Effect of F0 fluctuation and amplitude modulation of natural vowels on vowel identification in noisy environments," *Proc. of ICSLP*, 1633-1636, 2002.

[11] Jackson, P. J. B., Moreno, D. M., Russell, M. J. and Hernando, J. "Covariation and weighting of harmonically decomposed streams for ASR," *Proc. of Eurospeech*, 2321-2324, 2003.

[12] Patterson, R. D. and Moore, B. C. J. "Auditory filters and excitation patterns as representations of frequency resolution," in *Frequency Selectivity in Hearing*, Academic Press, London, 123-177, 1986.

[13] http://htk.eng.cam.ac.uk/

[14] http://julius.sourceforge.jp/