IMPROVEMENT IN ROBUSTNESS OF SPEECH FEATURE EXTRATION METHOD USING SUB-BAND BASED PERIODICITY AND APERIODICITY DECOMPOSITION

Kentaro Ishizuka*, Noboru Miyazaki*, Tomohiro Nakatani† and Yasuhiro Minami†

1NTT Communication Science Laboratories and 2NTT Cyber Space Laboratories, NTT Corporation
*2-4 Hikari-dai, Keihanna Science City, Kyoto 619-0237 Japan / †1-1 Hikarino-oka, Yokosuka City, Kanagawa 239-0847 Japan

ishizuka@cslab.kecl.ntt.co.jp, miyazaki.noboru@lab.ntt.co.jp, {nak, minami}@cslab.kecl.ntt.co.jp

ABSTRACT

This paper shows improvements in robustness of a speech feature extraction method using Sub-band based Periodicity and Aperiodicity Decomposition or SPADE. With SPADE, the speech signal is divided into sub-band signals through bandpass filter banks, after which the sub-band signal is decomposed into its periodic and aperiodic features by the comb filter. The comb filters are designed individually based on estimated periodicities of each sub-band signal. Both the periodic and aperiodic features are used as speech feature parameters. The evaluation experiment conducted with AURORA-2J (Japanese AURORA-2) shows that SPADE certainly reduces the averaged word error rate (WER) under clean-speech training. SPADE also improves the performance under multicondition-speech training when the noise condition of the test data is open. However, SPADE degrades the performance when the channel condition of the test data is open. To cope with this problem, in this paper we apply the cepstral mean normalization (CMN) to SPADE. The result shows that CMN greatly improves the performance not only for test data under the open-channel condition but also for data under the closed-channel condition. SPADE with CMN achieves an averaged word accuracy of 89.96 %, and an averaged WER reduction of 28.61 %. This word accuracy is better than that achieved by using MFCC with CMN.

1. INTRODUCTION

The mel-frequency cepstral coefficient (MFCC) has succeeded as a speech feature parameter for automatic speech recognition (ASR) for last two decades. However, MFCC is still not robust enough in noisy environments because its speech feature representation is easily deteriorated by noise. Hence, speech feature extraction methods have been proposed that are designed to represent more robust speech features.

Periodicity has often been employed by robust speech feature extraction methods to improve their noise robustness because the periodicity is inherently robust to interferer sounds. For example, such methods based on auditory nerve responses to periodic signals as GSD [1], ALSD [2], EIH [3], and ZCPA [4] significantly improve the noise robustness of ASR. GSD [1] and ALSD [2] focus on the synchrony around the characteristic frequencies (CFs) of the auditory filter. SBCOR [5] also focuses on the same point from an engineering perspective. These methods presume the auditory filter is a frequency analyzer, and they improve noise robustness by enhancing the spectral peak around CFs. From the different view of the auditory filter, EIH [3] and ZCPA [4] are based on the characteristics of auditory nerve firings phase-locked to lower frequencies different from their CFs in the time domain. The latter methods, based on the periodicity in the time domain, improve the ASR performance in noise more than the former methods based on CFs [4].

However, speech signals not only consist of completely periodic signals such as steady vowels, but of also aperiodic signals such as fluctuations of vowels, stops, fricatives, and affricates. The methods above lose the information about such aperiodicities; therefore, they often degrade the performance of ASR in clean or low-noise environments. From the perspective of psychoacoustics, the finding by de Cheveigné [6] suggests 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. Ishizuka and Aikawa [7] also showed the importance of aperiodicity of the speech signal even in noise. In addition, Jackson et al. [8] showed that the ASR accuracy in noisy environments is improved by using both periodic and aperiodic features of speech signals in terms of engineering.

Recently, we proposed a speech feature extraction method using Sub-band based Periodicity and Aperiodicity Decomposition (SPADE) [9]. SPADE represents both periodic and aperiodic features for each sub-band signal using bandpass filter banks and comb filters. The decomposition method based on comb filters is motivated by the auditory comb filter hypothesis [6], and its periodicity representation is based on the auditory nerve characteristics in the time domain. In addition, unlike previous studies [8], SPADE can improve the ASR performance in noisy environments without depending heavily on precise fundamental frequency (F0) estimations from clean speech or voicing detections.

This paper introduces a brief explanation about SPADE and shows the result of an evaluation experiment with the AURORA-2J database (Japanese AURORA-2; noisy Japanese digit recognition database) in Section 2. The experimental result shows that SPADE can reduce the word error rate (WER) under clean-speech training in comparison with MFCC-based method. However, SPADE cannot reduce the averaged WER under multicondition-speech training; in particular, it degrades the performance when the channel conditions are different between training and test data. To account for such a problem, in Section 3 the cepstral mean normalization (CMN) [10] is applied to SPADE. The result shows that CMN is effective at improving robustness of SPADE.
The characteristic of the comb filter is given by the output signal of each bandpass filter into its periodic and aperiodic features. The comb filters are designed individually for each sub-band. The comb filter decomposes the input speech sound into bandpass filter banks such as Gammatone filter banks [11], after which the output signal for each filter is divided into frames of fixed temporal length that shift with a certain temporal length. The dominant periodicity estimated for the comb filter is estimated by employing a periodicity estimation method such as the autocorrelation method for F0 estimation. The comb filters are used for designing the comb filter is estimated by employing a shift with a certain temporal length. The dominant periodicity is divided into frames of fixed temporal length that are used for training, whereas for test sets the periodicities are estimated practically. Under the clean-speech training condition, the noise condition is always open. Half of the noise types in test set C are closed only under multicondition-speech training, and the other is open.

\[ c_i = \frac{1}{N} \sum_{j=1}^{N} \log(m_j) \cos \left( \frac{\pi}{N} (j - 0.5) \right) \]

This transformation is the same as that used with MFCC. These coefficients are calculated individually for each feature, and only certain low-order coefficients are used as the feature parameters for ASR. Finally, both features are combined, and that combination is used as the speech feature parameter.

### 2.2. Noise robustness evaluation with AURORA-2J

We conducted an evaluation experiment with the AURORA-2J. The evaluation category was 0 (no changes in its backend scripts); that is, only the feature extraction process was changed. AURORA-2J, which is the noisy Japanese digit speech recognition database, includes three types of test data sets. These test data sets are spoken under different channel and noise conditions related to the training data sets, and the differences are shown in Table 1. AURORA-2J also includes two types of training data sets. One is the clean-speech training data set, which includes speech data spoken in a clean (noiseless) environment. The other is the multicondition-speech training data set, which includes speech data spoken in noises the same as those in test set A. Henceforth, the term “clean-speech training” means that the recognizer is trained using a clean-speech training data set. The term “multicondition-speech training” means that the recognizer is trained using a multicondition-speech training data set.

We measured the robustness by comparing the word accuracies obtained with the proposed method and baseline feature parameters, namely 12-order MFCCs and a log power, and their deltas and accelerations (39 dimensions in total). SPADE used the 24 Gammatone filter banks as the bandpass filter banks, 25-ms length temporal frames with 10-ms shifting, and 12-order coefficients for each feature, giving 24 feature parameters, namely 12-order MFCCs and a log power, and their deltas and accelerations (39 dimensions in total). Therefore the total number of dimensions of the feature parameters was 75. Although only the results under clean-speech training were reported in [9], this paper shows these under the multicondition-speech training because such a type of training is used in practice for ASR. Under multicondition-speech training, the periodicities estimated for clean speeches are used for training, whereas for test sets the periodicities are estimated practically.

#### 2.2.1. Result under clean-speech training

The top graph in Fig. 2 shows the result under clean-speech training. SPADE certainly improves the performance in...
2.2.2. Result under multicondition-speech training

In comparison with the baseline MFCC feature, SPADE achieves an averaged word accuracy of 53.30 %, and an averaged WER reduction of 13.24 %. This result indicates the effectiveness of SPADE under clean-speech training. However, an averaged WER reduction of -4.30 %. However, note that SPADE was able to improve for test set B (open-noise condition), indicating that SPADE is robust to differences between noises in the training and test data. However, SPADE’s performance deteriorated for test set C (open-channel condition), which suggests that SPADE is not robust enough to the differences between channel characteristics.

The reason why SPADE’s performance improved for test set B is explained below. Under multicondition-speech training, the aperiodic feature of the training data may include mostly information about voice. On the contrary, the periodic features of the training data may include mostly information about voice speech in the training data. When the noise in the test data set differs from that in the training data set, the difference between the aperiodic features of the training and test data indeed becomes large, whereas the difference between the periodic features remain small. In the case of MFCC, the difference between the noises in the training and test data directly distorted all parameters then degraded the word accuracy, because sound representation of MFCC does not consider the signal characteristics but the estimated spectrum of the whole sound signal. Therefore, SPADE is more robust than MFCC under open noise conditions.

The reason why SPADE is not robust enough to channel distortions is discussed below. In the case of MFCC, when the feature parameter is distorted, it is difficult to identify what type of distortion (i.e. noise or channel) occurs. This means that the training data distorted by noise is often able to account for the test data distorted by a channel, and vice versa. On the other hand, because SPADE decomposes the periodic and aperiodic features, these features have different properties related to noise and channel distortions. Although the periodic features are robust to noise, they are easily transformed by channel characteristics. In addition, the periodic features in the multicondition-speech training data are not so heavily distorted because of their robustness to noise. Therefore, the differences between the periodic features in training and test sets may be larger than MFCC features. To cope with this problem, we investigate whether CMN can be applied to SPADE.

3. EFFECT OF APPLYING CEPSTRAL MEAN NORMALIZATION TO “SPADE”

CMN [10] is a widely used technique to reduce the effects of differences in channel characteristics. This section presents a method for applying CMN to SPADE and its effect through an evaluation experiment.

3.1. Cepstral mean normalization for SPADE

The method for applying CMN to SPADE is described as follows. Because the speech feature representation of SPADE is similar to MFCC except for periodicity-aperiodicity decomposition, CMN is applicable to SPADE by the same method as that applied to MFCC. The coefficient of speech features calculated by SPADE is presented by $C(n,t)$, where $n$ is the index number of dimensions and $t$ is the time stamp of the temporal frame. The mean cepstral coefficient $M(n)$ is calculated as below, where $T$ is the number of frames in one speech segment.

$$M(n) = \frac{1}{T} \sum_{t=1}^{T} C(n,t)$$

The normalized coefficients $N(n,t)$ are calculated as below.

$$N(n,t) = C(n,t) - M(n)$$

$N(n,t)$ is calculated for all $n$ and $t$. Both the training data and the test data are normalized.

3.2. Evaluation Experiment

We experimentally evaluate the effect of applying CMN to SPADE with AURORA-2J under multicondition-speech training. Figure 3 shows the result. SPADE with CMN achieves an averaged word accuracy of 89.96 % and an averaged WER reduction of 28.61 % in comparison with the baseline MFCC. These results indicate that CMN can improve the performance effectively under multicondition-speech training, especially for test set C (an improvement of 10.10 %).
Applying CMN to SPADE improves the performance greatly not only under open-channel condition, but also under closed-channel condition, and the performance is better than for the MFCC with CMN[12].

It can be considered that applying CMN to SPADE can effectively absorb the influence of channel distortion. Note also that applying CMN to SPADE can improve the performance not only under open-channel conditions (test set C), but also under closed-channel conditions (test sets A and B). This indicates that applying CMN to SPADE is also effective at improving noise robustness under closed-channel conditions.

To compare the result with MFCC using certainly enhancement methods for improved noise robustness, we refer to the results by Fujimoto and Ariki [12]. They not only applied CMN to MFCC, but also other noise reduction methods such as the Gaussian mixture model of noise, ARMA filtering for noise suppression, and voice activity detections for frame drop to reduce false alarms. These results are also shown in Fig. 3 for comparison, indicating that SPADE with CMN achieves better word accuracies for all test sets. The improvement by CMN is almost the same for test set B. For test set C, the improvement achieved by applying CMN to SPADE is significantly greater than that by applying CMN to MFCC.

4. DISCUSSION

Applying CMN to SPADE improves the performance greatly not only under open-channel condition, but also under closed-channel condition, and the performance is better than for the MFCC with CMN. The reason is considered as follows.

SPADE initially improves the performance under the open-noise condition, possibly because the periodic features that hold most of the information about voiced speech signals can absorb the effect caused by the difference of noises in training and test data. However, the periodic features are easily affected by channel distortion. In addition, the periodic features are robust to noise, hence the training data distorted by noise cannot account for the test data with channel distortion. As a result, SPADE is easily affected by channel distortion.

Applying CMN to SPADE can effectively reduce the influence of differences in channel characteristics in the periodic feature. Consequently, SPADE with CMN improves the performance greatly for test set C. On the other hand, the aperiodic features include information about not only aperiodic speech signals, but also most of noise. Note that the noises included in AURORA-2J are almost stable; the influence of such stable noise is also expected to be suppressed by CMN. CMN may also effectively reduce the influence of noise in the aperiodic feature because this feature essentially includes most of the noise energy. Namely, the influence of channel and noise can be suppressed effectively because SPADE decomposes the signal into periodic and aperiodic features. As a result, the effect of applying CMN to SPADE becomes more effective than that of MFCC.

5. CONCLUSION

The speech feature extraction method SPADE decomposes the signal into periodic and aperiodic features in sub-bands, and it can certainly improve the performance of ASR under clean-speech training in the AURORA-2J task. Under multicondition-speech training, SPADE also improves the performance of the test data under the open-noise condition. However, SPADE degrades the performance of the test data under the open-channel condition. To cope with this problem, CMN is applied to SPADE. Applying CMN to SPADE greatly improves the performance under multicondition-speech training, with the effect being greater than when applying CMN to MFCC.

Acknowledgements: The authors wish to thank the members of the Spoken Dialogue System Group at NTT Cyber Space Laboratories for their useful suggestions on the evaluation experiments. This research employed the noisy speech recognition evaluation environment (AURORA-2J) produced by the IPSJ SIG-SLP noisy speech recognition working group.

REFERENCES