Robust Speech Recognition based on Selective Use of Missing Frequency Band HMMs

Takayoshi Kawamura, Kazuya Takeda and Fumitada Itakura

Center for Integrated Acoustic Information Research, Nagoya University, Japan
takeda@muse.nagoya-u.ac.jp

Abstract

In this paper, we propose a multi-stream approach that selectively uses Missing Frequency Band HMMs (MFB-HMM) that is trained on the band-eliminated speech. This makes the model insensitive to the noise in the missing frequency band. With multiple MFB-HMMs of different missing frequency bands, the proposed recognition system is robust in various type of noise conditions.

Recognition experiments show that the selective use of the MFB-HMMs is very effective in narrow band noise condition even if the noise is unstationary, however, the improvements of the performance to general noisy conditions, e.g. in-car noise and music sound, are not as high as in the narrow band noise case. The results of the experiments also show that the optimal selection of the MFB-HMM significantly improves the performance regardless of the type of the noise; therefore, the model selection measure is the key issue in this method.

1. Introduction

In order to implement the missing feature theory[1] in a speech recognition system, we need to 1) find which segment of the speech is corrupted with noise, and 2) disregard the noisy part removed from decoding. A sub-band based approach is one of the most common approaches of the localizing and elimination of the unknown noise in speech[2]. In sub-band based recognition systems, the simple division of the input signal into frequency sub-bands is the basic strategy for feature diversity.

The drawback of such sub-band based feature representations is that the global shape of the speech spectrum cannot be utilized in the decoding stage because each feature stream is modeled as a statistically independent signal. In order to deal with this problem some research efforts include the sub-band cepstrum representation [3] and the non-uniform overlapping sub-band design [4].

We propose the Missing Frequency Band HMM (MFB-HMM), which is the reverse process of the ordinary sub-band HMM that takes a spectral shape of speech without a particular sub-band. Since the speech recognition performance is not affected by this band-elimination as far as the eliminated sub-band remains relatively narrow with respect to its center frequency, MFB-HMM is not sensitive to the noise in the eliminated sub-band. Therefore, if we can select an optimal MFB-HMM from a set of MFB-HMMs of various missing frequency bands, we can realize a noise robust speech recognition system to various narrow band noises.

Furthermore, in general noise conditions, it is reasonable to assume that the noise signal is colored, and that the power of the signal is concentrated in a frequency band or a few frequency bands at each analysis frame. Thus, if optimal MFB-HMM can be selected on a frame-by-frame basis, the proposed method is expected to deal with a wide variety of noisy conditions.

The rest of this paper discusses the proposed method as follows. In Section 2, the basic algorithm is described. In Section 3, the experimental evaluations of the proposed method and their results will be detailed. In Section 4, the likelihood normalization method for the better selection of MFB-HMM is discussed.

2. Algorithm

The scheme of the proposed method is as shown in Figure 1. For each missing frequency band, the feature vector for the MFB-HMM is calculated by

\[ c_i^{(m)} = \sum_{j=1, j \neq M(m)}^j c_j^{(m)} p_j, \]  

where \( M(m) \) is the spectral channel number of the \( m \)-th missing frequency band, and \( p_j \) is the log amplitude spectrum of the speech at the \( j \)-th spectral channel.

Because diagonal covariance Gaussian distributions are used as the pdf of MFB-HMMs, a linear transformation by \( C = \{c_{ij} \} \) is introduced for orthogonalization the feature vectors. In this paper, we test KLT and DCT for the feature extraction. In the DCT case, the feature extraction procedure can be regarded as the band-eliminated version of MFCC analysis. However, it should be noted that discontinuities may occur around the eliminated band. Hence, the cosine series expansion does not directly extract the envelope of the speech spectrum.

Each MFB-HMM takes the corresponding feature vector as input and calculates the state likelihood \( g_i^{(m)}(a_i^{(m)}) \). The optimal model selection is performed based on the likelihood. Utterance-based and frame-by-frame-based model selections can be implemented. In the utterance-based strategy, the cumulative likelihood along the most probable state sequence is calculated for each HFB-HMM, and then, the word sequence obtained by the most probable model is selected as the recognition result.

The frame-by-frame-based strategy decodes the speech based on the output and the transition probabilities of the HMM.
3. Experimental Evaluations

3.1. Setup

The analysis conditions for the experiments are listed in Table 1. 24-channel mel-filter-bank is used for the basic spectral analysis. Nine missing frequency bands listed in Table 2 are used. Each missing frequency band covers consecutive two filter channels of the mel-filter-bank. The frequency region higher than 4052 Hz (the 19\textsuperscript{th} filter channel) is used for calculating feature vectors although no missing frequency band is in the frequency region.

For training the KLT matrix in feature extraction process, 5,000 sentences uttered by 50 male and 50 female speakers are used. Note that the zeroth order parameter is not included in the DCT feature vector whereas the weight for the first principle component is included for the KLT representation.

43 monophone HMMs are trained by 40,000 sentences uttered by 100 male and 100 female speakers (JNAS standard training set[5]). Each monophone HMM has three states with 32 mixtures. The recognition task is the discrimination of 50 phonetically balanced sentences with the vocabulary size of 311, without any grammar. A male speaker who is not included in the training speakers utters the test utterances. The recognition performances are evaluated by word correct rate (%Correct).

The noisy version of the test utterances are generated by adding speech and noise signals at 0 dB SNR. As for the noisy signals, band-limited (867 - 1421 Hz) white noise, a chirp tone, in-car noise and a piece of solo piano composition are used.

3.2. Results

3.2.1. Basic Performance of MFB-HMM

In Figure 2, the basic performances of the MFB-HMMs of recognizing clean speech are listed. As shown in this figure, as far as the recognition condition matches the training condition, eliminating the narrow band information does not affect the recognition performance regardless of the eliminating frequency band. Therefore, it is confirmed that if the noise band is limited within an eliminated frequency band, we can avoid the degradation of recognition performance by using the MFB-HMM.

3.2.2. Performance on Narrow Band Noise Conditions

In Figure 2, the basic performances of the MFB-HMMs of recognizing clean speech are listed. As shown in this figure, as far as the recognition condition matches the training condition, eliminating the narrow band information does not affect the recognition performance regardless of the eliminating frequency band. Therefore, it is confirmed that if the noise band is limited within an eliminated frequency band, we can avoid the degradation of recognition performance by using the MFB-HMM.

In Figure 3 (a), the recognition performance of the various missing frequency band HMMs and the results of the model selection based on the utterance-based maximum likelihood measure are shown for the band limited white noise case. Because the noise band is not used in the fifth MFB-HMM, word error rate can be reduced down to 1/100 of that of full-band HMM with the fifth MFB-HMM is used. From the results, it is clarified that model selection based on the utterance based maximum likelihood measure is shown for the band limited white noise case. Because the noise band is not used in the fifth MFB-HMM, word error rate can be reduced down to 1/100 of that of full-band HMM when the fifth MFB-HMM is used. From the results, it is clarified that model selection based on the utterance based maximum likelihood can find the optimal model, i.e., the fifth MFB-HMM, and reduce the word error rate significantly.

In Figure 3 (b), the results on the speech corrupted with a chirp tone signal are shown. Since a chirp tone is a time varying signal, we have tested the frame-by-frame based model selection method. As seen in the figure, the selective use of MFB-HMM reduces the word error rate to 1/4 of that of the full band results even no MFB-HMM outperforms the full-band model.

From these results, it is confirmed that the selective use of HFB-HMMs is very effective as far as the noise spectrum is limited in the narrow band. Furthermore, it is also confirmed that the maximum likelihood criterion can select the optimal model only in simple conditions.

3.2.3. Performance in Real Environment

Unlike the noisy conditions tested in the previous section, noise signals in the real environment are not always

| Table 2: Missing frequency bands for the experiments. (Lower and higher bounds in Hz) |
|-----------------|-----|-----|-----|-----|-----|
| MFB-HMM id     | 1 | 2 | 3 | 4 | 5 |
| lower          | 0 | 156 | 347 | 581 | 867 |
| higher         | 1245 | 420 | 718 | 1034 | 1421 |
| MFB-HMM id     | 6 | 5 | 8 | 9 | |
| lower          | 1245 | 1646 | 2170 | 2811 |
| higher         | 895 | 2475 | 3884 | 4052 |
ited in a narrow frequency band. Thus, we have tested the performance of the proposed method using the speech corrupted with real noise signals, i.e., in-car noise and a piano composition.

The results of the in-car noise are shown in the Figure 3 (c). Since the power of the in-car noise is concentrated in the lower frequency, it is reasonable that the first MFB-HMM, which does not use the information lower than 250 Hz, outperforms the full-band model. The superiority of the model selection is not clear in this condition. The model selection result is better than the full-band result for the 16-order KLT representation; however, no significant difference is found for MFCC the 20-order KLT. Furthermore, the results of the lowest MFB-HMM (MFB-HMM of the lowest missing frequency band) are better than that of the model selection.

The result marked as 'Max' shows the manually calculated performance when the 'optimal' model is given, i.e., the upper bound of the model selection performance. As the upper bound performance is much higher than that of the full-band and the lowest MFB-HMM, it can be concluded that the model selection procedure rather than the feature extraction method has considerable room for improvement.

The same experiment is performed using the solo piano piece as additive noise. Since this noise is not stationary, frame-based selection strategy is tested for the selection method. The results are shown in Figure 3 (d). Although no MFB-HMM outperforms the full-band result, the results of the MFB-HMM selection are higher than the full-band result. This is a natural consequence of the nonstationarity of the piano signal. However, it is noted that the optimal selection results show much better performance than other methods. Therefore, a more sophisticated model selection procedure is expected to improve this performance.

4. Improving Model Selection Method

In the previous section, it was clarified that the proposed method can improve upon the conventional results, and it is suggested that the sophistication of the model selection method brings further improvements to the performance. In this section, we discuss the normalization of the likelihood across the different MFB-HMMs in order to build a better model selection method and compare some normalization methods.

4.1. Likelihood Normalization

In the proposed recognition method, the model selection is based on the likelihood calculated by each MFB-HMM. However, since this likelihood function is defined independently for each feature space, they cannot be compared directly. Thus, the following three normalization methods were tested.

4.1.1. A-posteriori Probabilities (AP)

Normalize the likelihood of the \(m^{th}\) MFB-HMM by dividing the sum of the likelihoods over HMM states. In this case, the normalized likelihood for observation of \(o^{(n)}\) at
the measure for the feature selection. This measure can be calculated by
\[ P^{(m)}(o_t^{(m)}) = - \sum_j p_j^{(m)}(o_t^{(m)}) \log P_j^{(m)}(o_t^{(m)}), \] (4)
where \( p_j^{(m)}(o_t^{(m)}) \) is the a posteriori probabilities calculated by 2.

4.2. Evaluation

The above three normalizing methods are evaluated by the same experiments as the previous section. The results are shown in Figure 4 (a) and (b) for the in-car and the piano noise, respectively.

As for the in-car noise case, the relative likelihood to the training condition (marked as RT) attains the best performance when applied to the utterance-based selection strategy. After this normalization, the lowest MFB-HMM is selected for all test sentences. For the piano noise case, the relative likelihood also performs best when applied to the frame-based selection strategy. In this case, 4-7% improvement is obtained from the full-band results.

These results clarify that the relative likelihood to the training data is the best likelihood normalization method, and that the performance improvement can be obtained by the selective use of the MFB-HMMs regardless of the type of noise after the likelihood normalization.

5. Summary

A multi-stream approach that selectively uses Missing Frequency Band HMM (MFB-HMM) was proposed. Recognition experiments show that this method is very effective in narrow band noisy conditions, even if the noise is unstationary. Furthermore, the effectiveness of the likelihood normalization through taking relative probabilities to the training data is confirmed in the recognition experiments where the recognition performance under the in-car and the piano noise are improved by at most 10%.

6. References