Combining Multi-Band and Frequency-Filtering Techniques for Speech Recognition in Noisy Environments

Peter Jančovič, Ji Ming, Philip Hanna, Darryl Stewart, Jack Smith

School of Computer Science, The Queen's University of Belfast
Belfast BT7 1NN, Northern Ireland, UK
Tel: +44+(0)1232 +274072, Fax: +44+(0)1232 +666520
Email: p.jancovic@qub.ac.uk

Abstract. While current speech recognizers give acceptable performance in carefully controlled environments, their performance degrades rapidly when they are applied in more realistic situations. Generally, the environmental noise may be classified into two classes: the wide-band noise and narrow band noise. While the multi-band model has been shown to be capable of dealing with speech corrupted by narrow-band noise, it is ineffective for wide-band noise. In this paper, we suggest a combination of the frequency-filtering technique with the probabilistic union model in the multi-band approach. The new system has been tested on the TIDIGITS database, corrupted by white noise, noise collected from a railway station, and narrow-band noise, respectively. The results have shown that this approach is capable of dealing with noise of narrow-band or wide-band characteristics, assuming no knowledge about the noisy environment.

1. Introduction

For speech recognition, the most common feature representation of a frame is the logarithm filter-bank energy (logFBE). However, this parameterization is highly correlated, what does not fit the assumptions in Hidden Markov Models (HMM) using diagonal covariance matrices. Thus, a linear transformation is needed to decrease the correlation. In the full-band approach, usually, the Discrete Cosine Transform (DCT) is applied to the frequency dimension over the entire logFBE feature vector, leading to the cepstral representation. However, in the case of narrow-band noise, which affects only some local frequency regions, this transformation will cause the noise to be spread across all the feature components. This is not desirable. In order to avoid this spreading, the entire logFBE feature vector can be split into several sub-vectors, each representing the feature of a sub-band and subsequently, by applying individual DCT within each sub-vector the correlation can be decreased. This is the typical parameterization used in the multi-band approach.

Recently, an alternative way to de-correlate the logFBE features has been studied in [4,5]. In this method filtering is applied to the frequency dimension of each logFBE feature vector. When this method is used in the full-band approach, this yields better results than using the DCT technique for wide-band noisy speech recognition. Another advantage of the frequency filtering technique is that it usually employs FIR filters of a length of 2 or 3 samples. This means that after performing frequency filtering (FF) a narrow-band noise will not be spread over entire feature vector as was usually encountered in the DCT approach.

In this paper, we investigate the employment of the feature vectors obtained after FF in a multi-band speech recognition system. For the combination of sub-bands the probabilistic union model was used [3]. The paper is organized as follows: in the next section, the frequency filtering technique and its employment in the multi-band approach is described. In section 3, the probabilistic union model is explained and the results of the experiments are shown in section 4.

2. Frequency Filtering Parameterization Employed in Multi-Band Approach

The frequency filtering technique, studied in [4,5], was defined as a filtering operation performed on the frequency dimension of a logFBE feature vector. It can be seen as a liftering operation of cepstral parameters, but performed in the spectral domain. The frequency filters are designed to equalize the variance of the cepstral coefficients. In this work, we use the FIR frequency filter proposed in [4], which has a transfer function given bellow

$$H(z) = z - z^{-1} \quad (1)$$

The frequency characteristic of this filter is depicted in Figure1, which shows that this filter corresponds to an emphasizing of the cepstral coefficients with the middle indices.

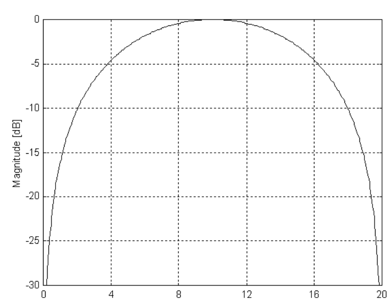


Fig. 1. Frequency characteristic of filter (1)

As the impulse response of (1) is *h(k)={1,0,-1}*, the filtering operation consists of a subtraction of the two logFBEs of the bands adjacent to the current one. This means that if the speech signal is corrupted by narrow-band noise that affects *i* components of the logFBE vector, after performing the frequency filtering only *i+2* components of the vector will be affected. Because of such a small spread effect of the noise this filtering approach is particularly suitable for being used in the multi-band system. Each feature vector is split into several sub-vectors – each corresponding to a specific sub-band.

3. Probabilistic Union Model

In the multi-band approach an important problem is to find an optimal strategy for combining the sub-bands. One possible way to solve this problem was recently introduced in [3], i.e. the probabilistic union model, where the sub-band combination is base on the union of random events.

If we assume a system with N sub-band feature streams $o=(o_1,o_2,\ldots,o_N)$, where o_n represents the feature stream from the n'th sub-band, in traditional HMM the overall likelihood is computed using the "and" operator (i.e. as product) of the individual likelihoods:

$$p(o) = p(o_1 \wedge o_2 \wedge \cdots \wedge o_N) = p(o_1) \cdot p(o_2) \cdots p(o_N) \tag{2}$$

When some sub-bands are corrupted by noise, applying "and" operator may destroy the ability of the model to discriminate between correct and incorrect word classes. We can alternatively assume that, in a given feature streams $o=(o_1,o_2,\ldots,o_N)$, the useful features that characterize the speech utterance may be any of the o_n's, $n=1,\ldots,N$. This can be expressed, based on the disjunction (i.e. inclusive "or") operator $\vee$, as

$$o_{\vee} = o_1 \vee o_2 \vee \cdots \vee o_N \tag{3}$$

where o_V is a combined observation based on $\vee$. In [3] a disadvantage of (3) was observed, i.e. it effectively averages the ability to discriminate between correct and incorrect words. Thus, a combination of the "and", "or" operators between the sub-bands has been suggested.

Specifically, if the noise occupies M-bands (M<N), then there exist (N-M) bands of clean speech. These can be combined with the "and" operator to accumulate their discriminative information. The "or" operator is taken over all possible (N-M) out of N combinations and we obtain a model

$$o_{\vee} = \bigvee_{n_1 n_2 \cdots n_{N-M}} (o_{n_1} \wedge o_{n_2} \wedge \cdots \wedge o_{n_{N-M}}) \tag{4}$$

which is called the Union Model with order M. Implementing the union model in a conventional HMM means only a modification in the calculation of the observation probability during the decoding stage. This can be approximated as [3]

$$p(o|\lambda) = p(o_{\vee}|\lambda) \approx \sum_{\forall (n_1 n_2 \cdots n_{N-M})} p(o_{n_1}) \cdot p(o_{n_2}) \cdots p(o_{n_{N-M}}) \tag{5}$$

A union model with *order=0* and *order=N-1* corresponds to the product and sum combination of the sub-bands, respectively. Moreover, it is important to note that in the particular case of frequency filtered parameterization *order=0* means de facto the full-band approach.

4. Experiments

The above described frequency filtering technique has been combined with the probabilistic union model in the multi-band approach. The experiments have been carried out using TIDIGITS speaker-independent database consisting of 11 words (‘one’, ‘two’, … ‘nine’, ‘oh’, ‘zero’). The models were trained on clean speech and applied to both clean and noisy speech recognition.

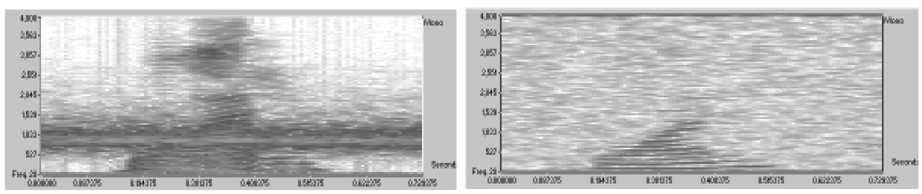
In the front-end preemphasis was used and the input signal was divided into 30ms long, Hamming windowed frames obtained every 10ms. For comparison, the experiments were carried out with the frequency filtered (FF) parameterization as well as the typical multi-band parameterization (defined below) and the full-band (FB) parameterization. In the case of full-band and the FF, a filter-bank analysis with 20 Mel-scaled filters was used to achieve logFBE feature vector for each frame. In the full-band parameterization, DCT was applied on this vector to achieve mfcc-parameters (the c_0 was excluded). In the case of FF, the logFBE feature vector was frequency filtered by using the filter defined in (1) and split into 5 sub-vectors corresponding to 5 sub-bands, each of a length of 4 components. This constitutes 5 streams. For the typical multi-band approach a filter bank analysis with 35 Mel-scaled filters was used; the logFBE vector was divided into 5 sub-bands (each consisting of 7 components); within each sub-band, the DCT was performed and the first 4 DCT coefficients composed the sub-band feature vector. We again have 5 streams. In order to include dynamic spectral information, delta parameters were computed in both cases and formed the next 5 streams. The probabilities of these 10 streams were merged at the frame level using the probabilistic union model. In the full-band approach the delta parameters was simply added to the static parameters. A 12-state HMM is estimated for each word, with the first and the last states being tied among all the vocabulary words to account for the silence part of the utterances.

Firstly, tests were performed for clean speech recognition. Based on (4), for each model with N sub-bands, recognition can be performed with different model orders (i.e. M) from 0 to N-1. Table1 presents the recognition results obtained by using the typical multi-band and the frequency filtered parameterization for clean speech. If the order in union model is low the typical approach achieves slightly better results. However, a very interesting point is the robustness of these two approaches to the order of the union model. For the typical approach as the order increases the recognition score decreases, and for *order=6* and higher the performance degradation is rapid. On contrary, the FF approach shows a very robust performance towards the order variation and keeps the rate around 97%.

Table 1. Recognition results of the multi-band system for clean TIDIGITs achieved by the FF and the typical multi-band approach

Multi-band method	Order of Union Model									
	0	1	2	3	4	5	6	7	8	9
FF	98.4	97.6	97.4	97.2	97.2	97.0	96.9	97.0	97.1	97.1
Typical	99.2	99.4	99.4	99.3	99.0	98.6	96.7	92.7	82.7	65.8
Full-band (FB): static+delta / delta	98.0 / 95.4									

Next, experiments in the presence of stationary narrow-band noise and wide-band noise were conducted. The narrow-band noise, added to the speech, is generated by passing the Gaussian white noise through a band-pass filter. We fix the bandwidth of the noise at 100Hz and vary the central frequency of the noise. For wide-band noise corruption, white, railway station and pub noises were added to the clean speech utterances, respectively. Figure2 shows an example of the noisy speech.



(a) (b)

Fig. 2. Spectrograms of an utterance, “one”, corrupted by stationary narrow-band noise with central frequency at 900Hz (a), and white noise (b)

The recognition results for wide-band noise and narrow-band noise are presented in Table2 and Table3, respectively. From Table2 it can be seen that in the case of wide-band noise, the FF approach achieved in all cases the best results for *order=0* and highly outperformed the typical multi-band parameterization. Results in Table3 show that both approaches achieved similar results for narrow-band noise corruption and the high order robustness with the FF approach is observed.

Table 2. Recognition results of the multi-band system in wide-band noise (SNR=10dB) for the FF approach / the typical multi-band approach

Order of Union Model	White	Railway station	Pub
0	76.2 / 12.4	93.3 / 45.7	71.1 / 38.5
1	55.7 / 15.3	88.0 / 44.0	57.3 / 40.4
2	50.7 / 20.7	85.4 / 32.8	58.6 / 41.8
3	47.1 / 18.8	82.6 / 22.5	55.1 / 37.0
4	44.7 / 18.4	81.2 / 19.5	54.5 / 30.6
5	43.5 / 37.9	79.3 / 74.3	53.6 / 67.1
6	42.7 / 23.6	78.2 / 59.4	53.1 / 57.9
FB: static+delta / delta	33.2 / 64.8	91.8 / 83.9	59.4 / 72.9

Table 3. Recognition results of the multi-band system in stationary narrow-band noise (SNR=0dB) for the FF approach / the typical multi-band approach

Order of Union Model	Noise central frequency (Hz)			
	900	1200	1800	2400
0	9.1 / 22.2	9.5 / 21.6	16.1 / 35.5	9.7 / 23.2
1	9.2 / 53.6	11.2 / 63.1	36.7 / 45.6	26.8 / 22.2
2	52.5 / 73.1	45.8 / 64.4	79.4 / 42.5	88.7 / 54.2
3	85.3 / 78.9	88.9 / 69.7	83.5 / 44.5	89.7 / 50.9
4	84.4 / 84.8	86.9 / 78.2	85.7 / 56.7	89.9 / 57.2
5	84.0 / 84.9	85.2 / 86.4	85.2 / 85.3	90.7 / 85.0
6	84.0 / 74.7	84.4 / 75.5	85.0 / 72.3	90.7 / 73.2
7	83.7 / 61.4	83.7 / 60.7	84.9 / 57.6	90.0 / 59.8
8	83.5 / 41.3	83.6 / 40.3	84.6 / 36.7	90.0 / 39.5
9	82.9 / 19.0	83.5 / 20.0	84.5 / 19.1	90.0 / 21.0
FB: static+delta / delta	16.0 / 79.7	15.3 / 81.9	16.5 / 78.0	18.3 / 81.9

5. Summary

An employment of frequency filtering parameterization and the probabilistic union model in a multi-band approach has been described in this paper. As a consequence of the properties of the frequency filtering (FF) and the union model, this technique has a considerable potential to cope with additive noise with no knowledge about the noise characteristics (e.g. bandwidth). The experiments with FF parameterization have shown similar results as typical multi-band parameterization in narrow-band noise corruption, and the results for wide-band noisy speech have shown a significant improvement. Moreover, high robustness on the order of the union model was observed for the FF approach. However, for various noise environments the best order of the union model can differ. This will be explored in the future.

This work was supported by UK EPSRC grant GR/M93734.

References

[1] Tibrewala, S., Hermansky, H., "*Sub-band Based Recognition of Noisy Speech*", Proc. ICASSP'97, Munich, Germany, 1997, pp.1255-1258

[2] Bourlard, H., "*Non-Stationary Multi-Channel (Multi-Stream) Processing Towards Robust and Adaptive ASR*", Proceedings of Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland, 1999, pp.1-10

[3] Ming, J., Smith, J., "*Union: A New Approach for Combining Sub-band Observations for Noisy Speech Recognition*", Proceedings of Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland, 1999, pp.175-178

[4] Nadeu, C., Hernando, J., Gorricho, M., "*On the Decorrelation of the Filter-Bank Energies in Speech Recognition*", Proc. Eurospeech'95, pp.1381-1384

[5] Macho, D., Nadeu, C.*:* "*On the Interaction between Time and Frequency Filtering of Speech Parameters for Robust Speech Recognition*", Proc. ICSLP'98, pp.1487-1490