Automatic Speech Recognition In Noisy Environments Using Wavelet Transform

WEAAM ALKHALDI, WALEED FAKHR and NADDER HAMDY
Electronics and Communications Department, Faculty of Engineering
Arab Academy for Science and Technology
P.O.Box: 1029, Miami, Alexandria
EGYPT

Abstract: - The performance of speech recognition systems is mainly determined by the used acoustic feature extraction technique. Two techniques are known, namely the full-band approach and the multi-band approach using filter banks. Systems using either approach usually suffer from performance degradation in the presence of noise. In this paper, the multi-band approach using Wavelet transform is suggested for speaker-independent isolated word recognition in noisy environments. Moreover, it has been found that combining the acoustic features obtained using both the full-band approach and the Wavelet transform- based multi-band approach has led to an improvement in the achievable recognition rates especially under mismatched conditions at low signal-to-noise ratio situations.

Key-Words: - Speech Processing, Wavelet Transform and Automatic Speech Recognition.

1 Introduction

Recognition rates of speech recognizers decrease with increasing background noise levels of the input speech to be recognized. Robustness to noise conditions is an important issue in the field of Automatic Speech Recognition (ASR), in order to gurantee high recognition rates in practical applications, e.g. for voice dialing in a car (car noise) or in the case of interference from other speakers (babble noise) [1]. Furthermore, speech recognizers can not be trained to cover all potential background noise conditions during training [2].

In this paper, two ASR systems, namely the full-band ASR system and the Discrete Wavelet Transform (DWT)- based multi-band ASR system, are implemented. They are tested under matched/mismatched acoustic conditions between training and testing. The recognition rates of both systems were identical, 99.2%, under matched conditions (clean speech for both training and testing). On the other hand, a significant degradation in performance has been observed upon testing both systems under mismatched conditions (clean speech for training and noisy speech for testing). It has been observed that one system performs better than the other for some word classes and vice versa. This means that the two systems complement each others under mismatched conditions. A third system, an ASR proposed system, which combines the acoustic features of the above two systems is implemented. It has given the same recognition rate, 99.2%, under matched conditions while it has improved the recognition rate under mismatched conditions, especially at low signal-to-noise ratio (SNR) situations.

This paper is organized as follows; section 2 discusses the full-band ASR approach, section 3 discusses the multi-band ASR approach and section 4 describes the implementation of the systems. In section 5, the obtained results are presented. Finally, the conclusion is given in section 6.

2 Full-Band ASR

Traditionally, ASR is performed by recognizing an extracted set of acoustic feature vectors, which are calculated from the whole frequency band of input speech. The mel-frequency cepstral coefficients (MFCCs) are the acoustic features of choice for many speech recognition applications [3]. They are calculated using the Discrete Cosine Transform as given in Equation (1).

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

where:
c_i : the i^{th} MFCC.
N : the number of filterbank channels covering the whole frequency band of input speech.
m_j : the log filterbank amplitude of the j^{th} channel.

A drawback of the full-band approach is that: even if only a part of the speech frequency band is corrupted by an additive noise, all the MFCCs are affected.

3 Multi-Band ASR

Recently, several researches have been published on multi-band ASR in which acoustic features are extracted independently from a number of frequency sub-bands. In the feature recombination (FC) technique [1], the acoustic features of the frequency sub-bands are combined before computing the global acoustic likelihood for classification. Some motivations of the multi-band approach are [1,4]:

- Human speech perception is multi-band by nature.
- Only some frequency sub-bands are affected in the case of speech corrupted by an additive colored noise.
- Multi-band ASR is well suited for parallel architectures.

In this respect, the DWT is introduced as an efficient tool for decomposing signals into frequency sub-bands.

3.1 The Discrete Wavelet Transform (DWT)

In Wavelet transform, the fixed bandwidth of the Fourier transform (FT) is replaced by several bandwidths that are proportional with the frequency. This allows better time resolution at high frequencies than the FT [5]. Any function *f(t)* can be represented as a superposition of time functions called wavelets, that are derived from dialations and translations of a single mother wavelet $\psi(t)$. A mother wavelet is a time function with finite energy and fast decay. The different versions of a single wavelet are orthogonal to each other. Discretizing the scale parameter (a) and the translation parameter (b) in a dyadic fashion, leads to octave-band filter banks [6], giving :

$$W(j,k) = \sum_{j}\sum_{k} f(k) e^{-j/2} \psi(2^{-j} n - k) \qquad (2)$$

which is the DWT of f(t).

This scheme can be implemented efficiently using quadrature mirror filters (QMFs) at each scale to give a low-pass version of the signal (approximation) and a high-pass version (detail). The DWT presents a good model for the human auditory system since it has decreasing frequency resolution for increasing frequencies. Using the fast pyramidal filtering algorithm of Mallat [7], speech signals can be decomposed with approximations being decomposed successively.

3.2 Feature Recombination (FC) of the DWT- Based Acoustic Features

After analyzing speech into its frequency sub-bands using DWT with three levels of decomposition, linear-frequency cepstral coefficients (LFCCs) are computed, as in Equation (3), for each individual frequency sub-band [8].

$$c_i = \frac{1}{N_S} \sum_{k=0}^{N_S - 1} \log \left| S(k) \right| e^{j 2\pi k i / N_S} \qquad (3)$$

where:

c_i : the i^{th} LFCC.

N_S : the number of points in the Fourier transform.

The feature recombination (FC) technique is then utilized to combine the LFCCs, computed for each frequency sub-band, into a single acoustic features set [8].

4 Implementation

In this paper, three isolated word recognition systems are implemented. They are:

1) The full-band system (FBS).
2) The DWT- based multi-band system (DWTS).
3) The proposed system (PS), a system combining the acoustic features of the above two systems.

The recognition rates of the three systems are compared under matched/mismatched conditions.

A 10 word database is used, the Arabic digits (0-9). The training data set is consisted of 500 utterances (50 utterances per word) by 50 speakers, 29 of them are males and 21 are females. The testing data set is consisted of 130 utterances (13 utterances per word) by 13 male speakers, different from the ones used for training.

The training and the testing data sets are recorded with a close-talk microphone in laboratory environments (the data sets are considered clean) using a 8 KHz sampling rate (telephone quality). Each digitized waveform is analyzed with a 25 ms Hamming window, that is shifted by 10 ms intervals.

Hidden Markov models (HMM)- based recognizers are used for all systems utilizing the HTK package [3]. Each system contained 10 HMMs, one per word. Each HMM contained 5 states, 7 Gaussian mixtures per state.

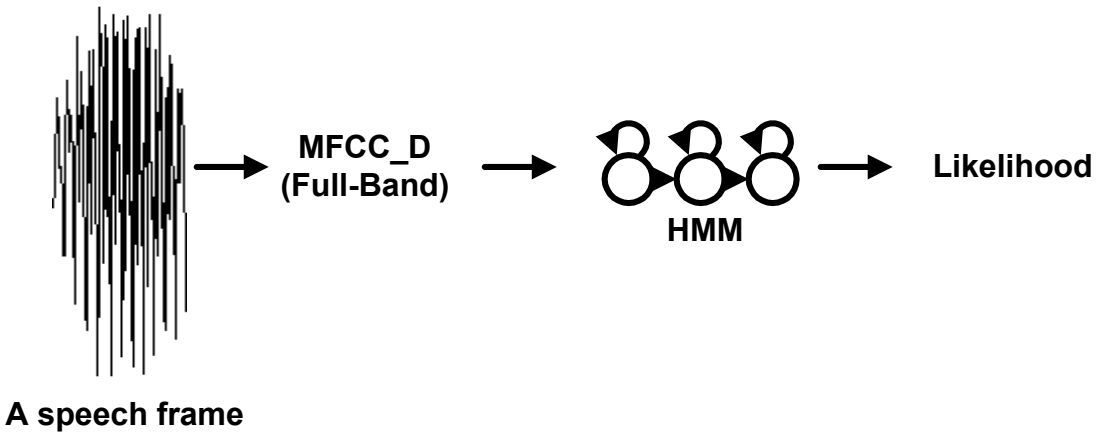

Fig. (1): A schematic diagram for the full-band ASR system.

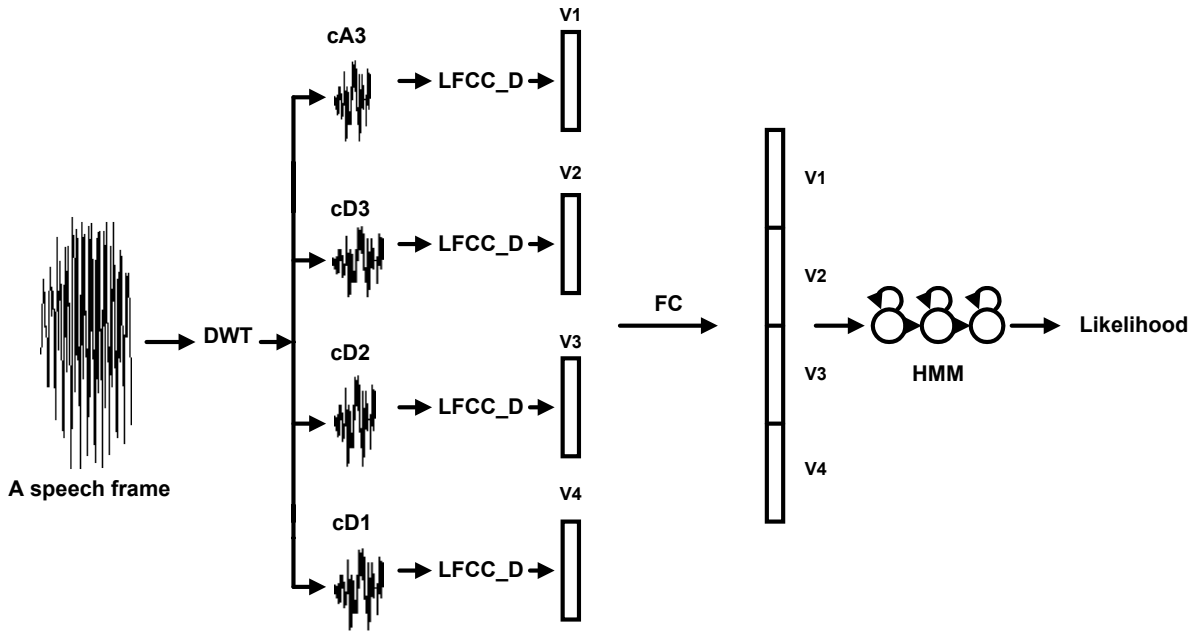

Fig. (2): A schematic diagram for the DWT- based multi-band system.

In all experiments here, four different types of noise are added only to the testing data set to test the recognizers under mismatched conditions. All HMMs are trained under no noise conditions.

4.1 The Full-Band System

The front end here is based on the mel-cepstral analysis of the input speech using 26 channels filter bank. The feature vector of each frame is made of 24 components (MFCC_D), consisting of 12 MFCCs with their 1st derivatives, at a 100 Hz frame rate. Figure (1) shows a schematic diagram system.

4.2 The DWT- Based Multi-Band System

There are several issues that should be considered in the design of any multi-band ASR system, e.g. the number of frequency sub-bands, the type and number of features to be used in each sub-band [9]. As a front end for the DWT- based multi-band system here, the mother wavelet 'db6' is used to decompose speech signals into four frequency sub-bands (one level of approximation coefficients 'cA3' and three levels of detail coefficients 'cD3', 'cD2' and 'cD1'). A schematic diagram for the system is given in Figure (2).

For each frame, a number of LFCCs is calculated from each frequency sub-band individually before combining them into one feature

vector with their 1st derivatives, at a 100 Hz frame rate.

The LFCCs are considered the features of choise for this system and the number of LFCCs computed for each frequency sub-band was a point of research in this paper. In a previous research [8], 12 LFCCs were computed for each frequency sub-band and the recognition rate was about 86.15% under matched conditions.

In this paper, feature vectors of lengths equal to and less than 12 LFCCs, with their 1st derivatives (LFCC_D), for each frequency sub-band are tried to measure the recognition rate versus the length of the feature vector. The maximum recognition rate is obtained for feature vector length of 32 components (4 LFCCs per band with their 1st derivatives). A silence model is included in this experiment. Figure (3) illustrates the recognition rates versus these different lengths of the feature vector, under matched conditions.

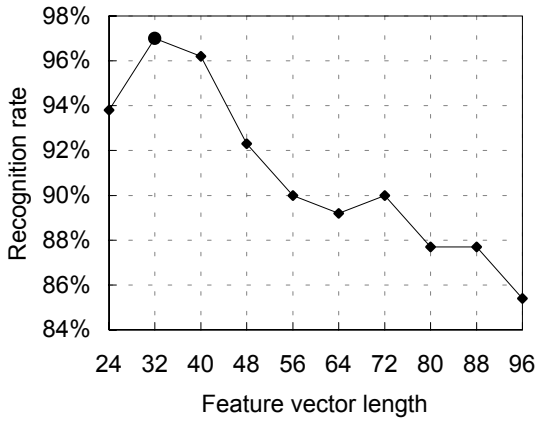

Fig. (3): Recognition rate vs feature vector length for the DWT- based multi-band system.

4.3 The Proposed System

Upon testing the above two systems under mismatched conditions, it has been observed that one system outperforms the other for some word classes and vice versa. This suggests that both systems are complementary in noisy environments and hence combining their acoustic features can lead to an improvement in the recognition rates. Each frame feature vector is made of 56 components (24 MFCC_D components from the full-band system and 32 LFCC_D components from the DWT- based multi-band system), at 100 Hz frame rate. During training, the HMM is enforced to consider each feature vector consisting of two independent data streams.

5 Results

To compare the performance of the above three systems, testing is performed under matched conditions (clean speech) and under mismatched conditions (speech corrupted by an additive noise).

5.1 Testing Under Matched Conditions

Table (1) summarizes the obtained recognition rates of the three systems. It is to be noted that no silence model is assigned to the third (the proposed) system in contrast to the other two systems.

Table (1): The total recognition rates under matched coditions for all systems.

System	With the silence model	Without the silence model
FBS	**97.7%**	**99.2%**
DWTS	**97%**	**99.2%**
PS	**-----**	**99.2%**

From the above Table, it is clear that disregarding the silence models for both the FBS and the DWTS has improved their recognition rates. Also, the performance of the PS is exactly equivalent to the performance of the other two systems and this indicates that this combining strategy does not lead to any loss of information.

5.2 Testing Under Mismatched Conditions

Four different types of noise are used to corrupt speech: 1) "lp-white" noise (a type of noise corrupts only the frequency range 0-500 Hz of speech), 2) babble noise, 3) destroyer-engine noise and 4) volvo noise. The last three types are taken from the NOISEX-92 database and they are added to the testing data set at SNRs ranging from 5 db to 25 db. Figures 4, 5, 6 and 7 below depict the recognition rates of the three systems for each noise type.

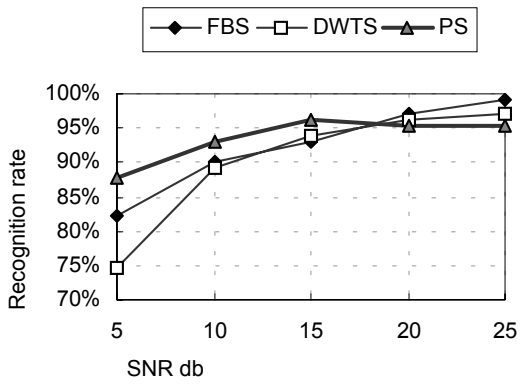

Fig. (4): Recognition rates for "lp-white" noise.

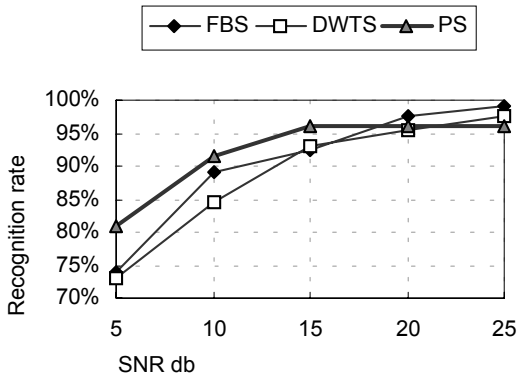

Fig. (5): Recognition rates for babble noise.

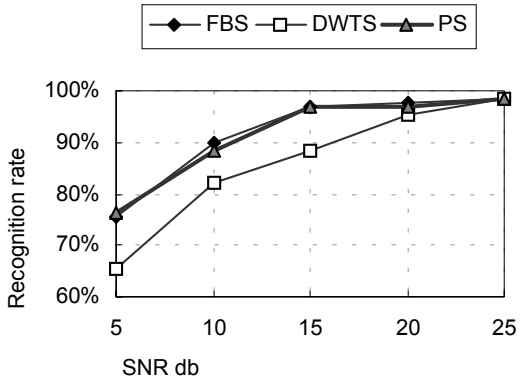

Fig. (6): Recognition rates for destroyer-engine noise.

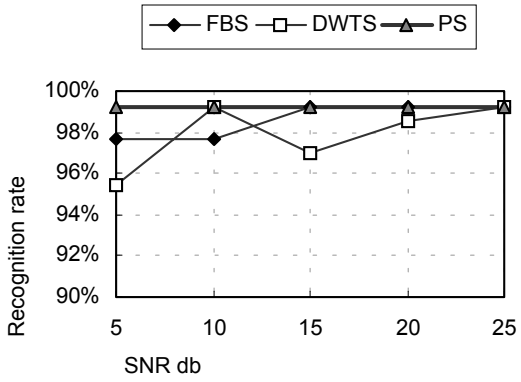

Fig. (7): Recognition rates for volvo noise.

From the Figures, it is evident that the proposed system has provided better recognition rates than the other two systems for "lp-white" noise, babble noise and volvo noise especially at 5 db and 10 db SNRs.

6 Conclusion

The mismatch between the acoustic features during training and testing of ASR systems significantly degrades their recognition rates, especially at low SNRs.

In this paper, a full-band ASR system and a DWT- based multi-band ASR system are implemented and their recognition rates are measured under matched/mismatched conditions. It has observed that the two systems have the same recognition rates under matched conditions while they have complemented each others under mismatched conditions. Accordingly, their acoustic features are combined into an ASR proposed system which has given the same recognition rate of them under matched conditions while it has improved the recognition rates under mismatched conditions for three types of noise, especially al low SNRs.

It is planned to apply the DWT- based multi-band approach in speaker recognition.

References:

[1]S. Okawa, E. Bocchieri and A. Potomiaos."Multi-Band Speech Recognition in Noisy Enviornments". *In Proc. of The ICASSP, Vol. 2, pp. 641-644, Seatle, WA*, May 1998.

[2] H.G. Hirsch, P. Meyer, H.W. Rühl. "Improved speech recognition using high-pass filtering of sub-band envelopes", *Eurospeech, pp. 413-416*, 1991.

[3] Steve Young et al. *The HTK Book: HTK Tools and Reference Manuals,* Version 2.2, Entropic, 1999.

[4] N. Mirghofori and N. Morgan. "Transmissions and Transitions: A Study of Two Common Assumptions in Multiband ASR". *In Proc. of The ICASSP, Vol. 2, pp. 713-716, Seatle, WA*, May, 1998.

[5] O. Rioul and M. Vetterli. "Wavelets and Signal Processing", *IEEE Signal Processing Mag., Vol.8, pp. 14-38*, Oct. 1991.

[6] Murali Krishnan et al. "Wavelet Transform Speech Recognition using Vector Quantization, Dynamic Time Warping and Artificial Neural Networks", Preprint, 1994.

[7] Randy K. Young, *Wavelet and Its Applications,* Kluwer Academic Publishers, 1993.

[8] Weaam Alkhaldi, Waleed Fakhr and Nadder Hamdy. “Multi-Band Based Speech Recognition of Spoken Arabic Numerals using Wavelet Transform”, *In Proceedings of The 19th National Radio Science Conference, pp. 224-229, Alexandria, Egypt*, March 2002.

[9] S. Tibrewala and H. Hermansky. “Sub-Band Based Recognition of Noisy Speech”. *In Proceedings of the ICASSP, pp. 1255-1258, Munich*, April 1997.