International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol. 2, No. 2, June 2009

A Robust Front-End Processor combining Mel Frequency Cepstral Coefficient and Sub-band Spectral Centroid Histogram methods for Automatic Speech Recognition

R. Thangarajan[1] and A.M. Natarajan[2]
[1]*Assistant Professor, Department of Information Technology*
Kongu Engineering College
Perundurai – 638 052, Erode, Tamilnadu State, INDIA
[2]*Chief Executive and Professor*
Bannari Amman Institute of technology
Sathyamangalam – 638 401, Erode, Tamilnadu State, INDIA
thangs_68@yahoo.com[1], *amnatarajan2006@yahoo.co.in*[2]

Abstract

Environmental robustness is an important area of research in speech recognition. Mismatch between trained speech models and actual speech to be recognized is due to factors like background noise. It can cause severe degradation in the accuracy of recognizers which are based on commonly used features like mel-frequency cepstral co-efficient (MFCC) and linear predictive coding (LPC). It is well understood that all previous auditory based feature extraction methods perform extremely well in terms of robustness due to the dominant-frequency information present in them. But these methods suffer from high computational cost. Another method called sub-band spectral centroid histograms (SSCH) integrates dominant-frequency information with sub-band power information. This method is based on sub-band spectral centroids (SSC) which are closely related to spectral peaks for both clean and noisy speech. Since SSC can be computed efficiently from short-term speech power spectrum estimate, SSCH method is quite robust to background additive noise at a lower computational cost. It has been noted that MFCC method outperforms SSCH method in the case of clean speech. However in the case of speech with additive noise, MFCC method degrades substantially. In this paper, both MFCC and SSCH feature extraction have been implemented in Carnegie Melon University (CMU) Sphinx 4.0 and trained and tested on AN4 database for clean and noisy speech. Finally, a robust speech recognizer which automatically employs either MFCC or SSCH feature extraction methods based on the variance of short-term power of the input utterance is suggested.

Keywords: *Auditory models, dominant-frequencies, feature extraction, MFCC, Noise robustness, SSCH.*

1. Introduction

Robustness is an essential feature for practical automatic speech recognition (ASR) systems in order to avoid severe degradation of performance when mismatch between training and deployment conditions occur. Many variations like ambient background noise, channel and microphone variations, as well as speaker variations viz. dialect, age and gender cause degradation in performance. These environmental conditions are highly variable and unpredictable, and therefore cannot be accounted during training. Much research has been done and efficient methods have been developed [1] [2] [3].

Speech feature vectors used in ASR should contain relevant information for discriminating different speech sounds. Commonly used speech features are MFCC and LPC. It has been

shown that both MFCC and LPC are significantly affected by environmental variations [4]. Hence speech features which are less sensitive to the changes in environmental factors and retain good discriminative properties helps to increase robustness.

Human speech perception has inspired ASR research community to develop feature extraction methods which are based on detail modeling of human auditory processing system. These methods have shown significant improvement in robustness against noise when compared to other standard methods. The success of methods which model human auditory system is due to the use of information about the dominant frequencies or spectral peak positions in speech signals. These spectral peaks are generally not affected by environmental noise provided noise spectrum does not have strong spectral peaks.

1.1. Use of dominant frequency information using time-domain analysis

The concept of synchrony spectrum which is a speech feature based on detailed modeling of processes in the human auditory system is proposed in [5]. The outputs from a set of generalized synchrony detectors, one for each sub-band, which measures the extent of dominance of the periodicities at sub-band center frequencies, are included in the feature. Therefore, the sub-bands that have their center frequencies close to the spectral peaks obtain the highest scores. In this way, the information on dominant frequencies in the speech signal is included into the feature vectors. Ensemble Interval Histogram (EIH) is another method which is based on computing a set of level-crossing rates in each sub-band [6]. The sub-bands are obtained by filtering the speech signal by a cochlear filter. It is obvious that the level crossing rates are related to the dominant sub-band frequency. Moreover, the number of levels crossed is related to the sub-band signal power. Thus EIH combines dominant sub-band frequency information with sub-band power information. Zero crossings with peak amplitudes (ZCPA) method is simpler compared to EIH method [7]. In ZCPA feature extraction, the cochlear filter is replaced with a simple band-pass filter and the set of level crossing detectors for each sub-band are replaced with a zero crossing detector. Sub-band power is measured as amplitudes between subsequent zero crossings. Thus ZCPA histograms combine sub-band power and sub-band dominant frequency in the features. Sub-band autocorrelation analysis has been successfully applied to feature extraction [8]. This technique is based on a simple band-pass filtering followed by the computation of autocorrelation coefficients for the sub-band signals at a time.

$$\tau = 1/F_c$$

Where, F_c is the sub-band center frequency.

This method also uses dominant frequency in a sub-band because a spectral peak at frequency F_c gives rise to peaks in autocorrelation function at integer multiples of $1/F_c$. Therefore the value of sub-band autocorrelation coefficient at time $1/F_c$ indicates the extent of dominance of the sub-band center frequency in the sub-band signal.

All these approaches are robust to noise because they use dominant frequency in the sub-band. However, the robustness comes at a cost. All these methods are computationally expensive due to time domain analysis and not preferred for practical ASR systems.

1.2. Use of dominant frequency information using short-term power spectrum

Another feature extraction technique based on sub-band spectral centroids (SSC) is proposed [9]. SSC are closely related to spectral peak positions in clean and noisy speech.

They can be computed efficiently from short-term speech power spectrum i.e. frequency domain. Subsequently this method was extended to a new approach called spectral sub-band centroid histograms (SSCH) [10] [11]. SSCH is a better method of integrating dominant frequency information provided by the SSC with the sub-band power. This is achieved by constructing histogram bins similar to ZCPA method. However, the major limitation of this method is that it is effective only for background additive noise without spectral peaks.

It is further found that SSCH and MFCC have the same order of time complexity. SSCH performs better than MFCC when input speech is noisy. For clean speech, MFCC performs even better than SSCH. Hence in this paper, we design a front-end processor of ASR system which automatically employs MFCC or SSCH features depending upon environment noise.

2. MFCC and SSCH features

Both MFCC and SSCH feature extraction methods involve several steps that are common viz. spectral estimation, sub-band filtering, and sub-band power computation. SSCH method incorporates two more steps, namely centroid computation and histogram construction. In this section we review both the techniques.

2.1. MFCC Feature extraction

The processes in MFCC feature extraction are shown in figure 1. First short-term spectral estimation is computed for the input speech frame followed by filtering by overlapping triangular band pass filters. For each sub-band, only the log-energy is computed which is later subjected to discrete cosine transform (DCT).

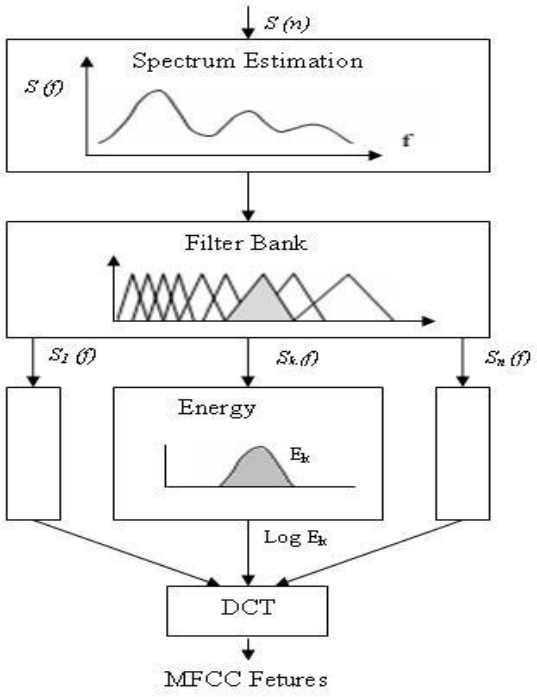

Figure 1. MFCC feature Extraction

2.2. SSCH Feature Extraction

The uniqueness of SSCH feature lies in the information about sub-band spectral peak positions with conventional sub-band power in order to increase ASR robustness against additive background noise. Thus it is more robust than MFCC and greater improvements occur for noise types characterized by relatively flat spectral density. The SSCH feature extraction procedure is shown in figure 2. Initially FFT-based power spectrum estimate S (f) for the given speech frame is computed and passed through a set of K overlapping band pass filters with amplitude responses H (k), for k = 1… N

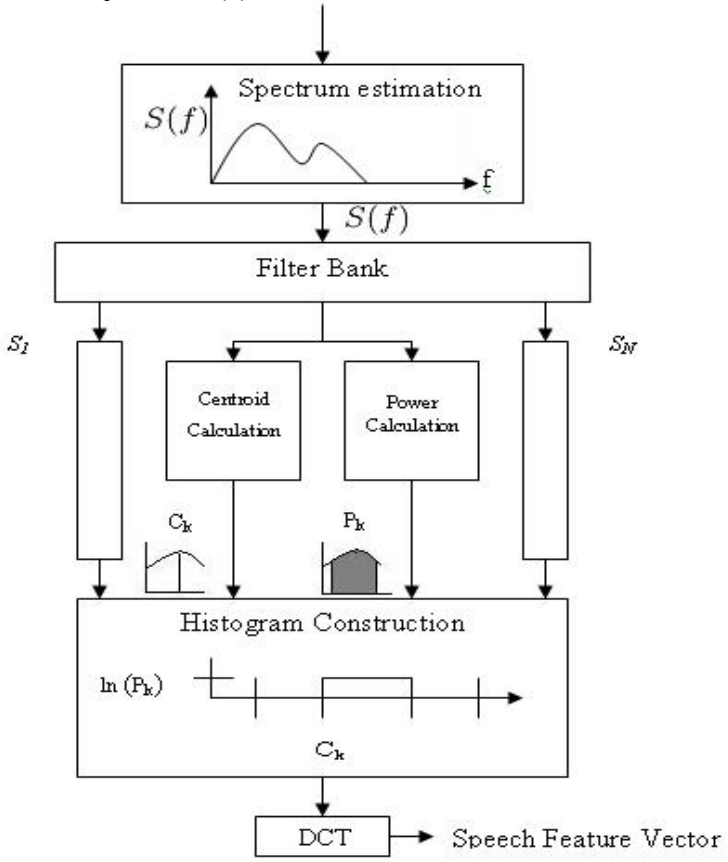

Figure 2. SSCH feature extraction

SSC is then computed for each sub-band using the expression shown in equation (1)

$$C_k = \frac{\sum fH_k(f)S(f)}{\sum H_k(f)S(f)}, k = 1,\ldots,n \quad (1)$$

Where, the summation is performed for all frequency samples n in the FFT. Simultaneously power estimates of each sub-band are also computed using the expression shown in equation (2)

$$p_k = \sum H_k(f)S(f), k = 1,\ldots,K$$

(2)

where the summation is performed for the frequency range of 1 Bark centered on sub-band centroids. This ensures more robust sub-band power estimates, since the effect of additive noise in log-spectral domain is smallest around spectral peaks. Next, histogram bins of the SSC are constructed by dividing the speech frequency range into bins Rj for j = 1... J. The number of bins is computed by the expression shown in equation (3)

$$\text{count(j)} = \sum_{k} \psi_j \{C_k\} \text{j} = 1,\ldots,\text{J} \tag{3}$$

Where

$$\psi_j \{C_k\} = \begin{cases} \ln(P_k / N_k), C_k \in R_j \\ 0, otherwise \end{cases}.$$

Where N_k is the number of frequency samples in the histogram bin *k*.

3.0 Combination of MFCC and SSCH feature extraction methods

The Figure 3 shows the recognition system that is evolved as a result of combination of both MFCC and SSCH feature extraction methods. In this method the spectrum estimation is carried out for the speech signal and it is identified whether the signal is clean or it is mixed with noise.

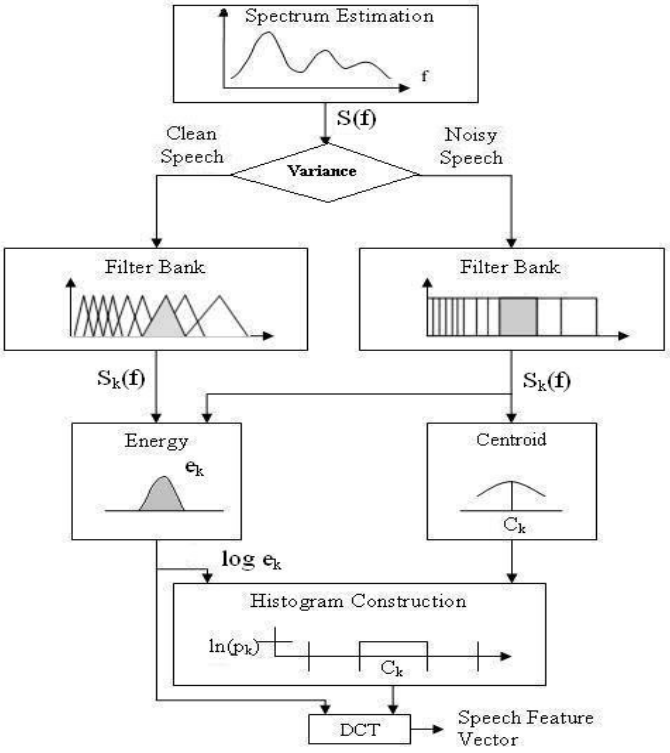

Figure 3. MFCC and SSCH combined Method

If the speech signal is clean then the MFCC feature extraction method is used for speech recognition. If the speech signal is noisy then the SSCH feature extraction method is used for speech recognition and hence the word accuracy for both clean as well as noisy speech can be made maximum at the same time. The computational cost as well as the computational complexity remains the same for both methods. The only difference is that in the proposed method the variance of every 3000 samples is calculated before deciding the method to be used for the feature extraction.

3.1 Identification of noisy speech from clean speech

The input speech signal of different speakers is sampled and the variance of first 3000 samples is calculated and tabulated in table 1. It was found that variance was lower than 10-7 for speech without noise. For speech with 25 db SNR the value lies in the range 10-6 to 10-5. For 15 db SNR the value lies in the range 10-5 to 10-4 and for 5 db SNR the value was found be greater than 10-4. The variance of samples is taken as the threshold value for deciding whether the input signal is noisy or not. If the variance is below 10-7 then MFCC method is carried out for those samples and if it is above 10-007 then SSCH method is carried out for speech recognition.

Table 1. Variance values of different speech files with different noise levels

Speakers	Clean speech	SNR		
		25 db	15 db	5 db
1	5.4772×10^{-7}	1.0933×10^{-6}	6.0627×10^{-5}	5.4920×10^{-4}
2	6.9289×10^{-7}	3.7544×10^{-6}	3.2131×10^{-5}	2.9981×10^{-4}
3	1.8093×10^{-7}	9.0061×10^{-6}	7.5938×10^{-5}	7.3575×10^{-4}
4	2.0540×10^{-7}	3.3079×10^{-6}	3.1099×10^{-5}	3.0458×10^{-4}
5	1.7653×10^{-7}	3.0423×10^{-6}	2.8294×10^{-5}	2.9304×10^{-4}
6	4.6125×10^{-7}	2.6975×10^{-6}	2.2520×10^{-5}	2.1930×10^{-4}

4. Recognition tasks

In order to evaluate the proposed method and the existing methods, an alphanumeric database called AN4 is used. AN4 database consists of utterances of personal information of users such as their name, address, telephone numbers and date of birth, etc. More details of the database can be taken from [12]. Model training was performed using the CMU *SphinxTrain* [13] using both MFCC and SSCH features.

4.1. Preparation of Database for recognition tasks

Utterances from 74 speakers (21 female and 53 male) were used for model training, while utterances from an additional 20 speakers (10 female and 10 male) were used for evaluation.

Each triphone was modeled by a three state left-to-right hidden Markov model (HMM) with eight Gaussian components per state and no skip transitions.

4.2. Evaluation and Results

For the purpose of evaluating robustness against environmental noise, three different types of noises namely, white Gaussian noise, factory noise, and background speech were added to the test data at several SNR. White Gaussian noise was generated using a random noise generator, while factory noise and background speech were taken from the NOISEX database, where they are referred to as factory1 and babble noise, respectively. Noisy speech utterances were generated as follows. For each speech utterance in the test database, a noise segment of length equal to the length of the speech utterance was randomly extracted, multiplied by a gain factor and added to the speech utterances. The gain factor was computed in accordance with the required SNR which is shown in equation (4)

$$SNR[dB] = 10\log_{10}\left(\frac{p_s^{\max}}{g^2 p_n}\right) \tag{4}$$

Where Psmax is the maximal frame power of the given speech utterance and Pn is the noise power estimated over the noise segment. The SNR is thus measured as the ratio between maximal speech power and average noise power. This computation method makes SNR independent of both the phonetic content of speech utterance and the length of silent intervals surrounding the speech utterance.

After the models were trained, the evaluation was carried out on CMU Sphinx 4.0 decoder. A front-end processor for MFCC feature extraction was already available in CMU Sphinx 4.0 but front-end processor for SSCH feature extraction was implemented and added to CMU Sphinx 4.0. The ASR performance was measured in terms of word accuracy as given in equation (5).

$$WAC = \left(\frac{N - S - D - I}{N}\right) \times 100 \tag{5}$$

where N is the total number of words in the test set, S is the number of substitution errors, D is the number of deletion errors, and I is the number of insertion errors. The results of the tests are shown in table 2.

Table 2. ASR Results

Feature Type	Word Accuracy [%]			
	Clean Speech	Speech with Noise SNR [dB]		
		25	15	5
MFCC	98.32%	96.90%	91.19%	47.37%
SSCH	96.54%	98.54%	95.57%	75.77%
Combined model	98.32%	98.54%	95.57%	75.77%

5. Conclusion and Future work

Robustness to environmental noise is a major issue in ASR. MFCC is the preferred method of feature extraction in all practical ASR. However, MFCC features are not robust to noisy speech. The accuracies of the ASR systems using MFCC features degrade considerably in the presence of noise. On the other hand, SSCH features are resilient to noise but are inferior to MFCC features when clean speech is used. It is also shown that MFCC and SSCH have many steps in common and the time complexity is almost same. Therefore, in this paper, a novel front-end processor which combines both MFCC and SSCH methods is proposed, designed and implemented. The resulting front-end is tested on CMU Sphinx 4.0 decoder on a publicly available AN4 database. The test set included both clean speech and noisy speech with three different types of noises. White Gaussian noise was added to the speech with SNR of 25 db, 15 db and 5 db. Initially recognition accuracy was calculated using MFCC as the front-end for speech without noise and SSCH method was used for speech with different noise levels and the combination of both MFCC and SSCH feature extraction methods for clean and noisy speech. It has been found that the proposed front-end dynamically adapts to both clean and noisy speech.

However the system also has some drawbacks. It will degrade if the noise has spectral peaks. It is tested only on a small vocabulary task. It may be tested on a large vocabulary and continuous speech task for more reliable results.

References

[1] X. Huang, A. Acero, and H.W. Hon, “Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.” Englewood Cliffs, NJ: Prentice-Hall, 2001.
[2] J.C. Junqua and J.P. Haton, “Robustness in Automatic Speech Recognition – Fundamentals and Applications”, Norwell, MA: Kluwer, 1996.
[3] Y.Gong, “Speech recognition in noisy environments: A survey”, Speech Communications, vol. 16, no. 3, 1995, pp. 261–291.
[4] J. P. Openshaw and J. S. Mason, “On the limitations of cepstral features in noise”, Proc. ICASSP, vol. 2, 1994, pp. 49–52.
[5] S. Seneff, “A joint synchrony/mean-rate model of auditory speech processing”, Journal of Phonetics, vol. 16, no. 1, pp. 55–76, Jan. 1988.
[6] O. Ghitza, “Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment”, Journal of. Phonetics, vol. 16, no. 1, pp. 55–76, Jan. 1988.
[7] D.S. Kim, S.Y. Lee, and R. M. Kil, “Auditory processing of speech signals for robust speech recognition in real-world noisy environments”, IEEE Trans. Speech Audio Processing, vol. 7, no. 1, pp. 55–69, Jan. 1999.
[8] S. Kajita and F. Itakura, “Sub-band-autocorrelation analysis and its application for speech recognition”, in Proc. ICASSP, vol. 2, 1994, pp. 193–196.
[9] K. K. Paliwal, “Spectral sub-band centroid features for speech recognition”, Proc. ICASSP, vol. 2, May 1998, pp. 617–620.
[10] B. Gajic and K. K. Paliwal, “Robust feature extraction using sub-band spectral centroid histograms”, Proc. ICASSP, vol. 1, May 2001, pp. 85–88.
[11] B. Gajic and K. K. Paliwal, “Robust speech recognition in noise environments based on Sub-band spectral centroid histograms”, IEEE Transactions on Audio, Speech and Language Processing Volume 14, No. 2, March 2006
[12] AN4 Speech database website: http://www.speech.cs.cmu.edu/databases/an4/
[13] Lamere P., Philip Kwok, William Walker, Evandro Gouvea, Rita Singh, Bhiksha Raj and Peter Wolf, “Design of the CMU Sphinx-4 Decoder” in EUROSPEECH 2003.