Sub-band Information Fusion Based on Wavelet Thresholding for Robust Speech Recognition

Babak Nasersharif [a,b] *, Ahamd Akbari [b]

[a] School of Computer Engineering, Faculty of Engineering, University of Guilan, Rasht, Iran
[b] Audio and Speech Processing Lab, Department of Computer Engineering, Iran university of Science and technology, Tehran, Iran

Received 24 January 2010; revised 24 May 2010; accepted 5 June 2010

Abstract

In recent years, sub-band speech recognition has been found useful in addressing the need for robustness in speech recognition, especially for the speech contaminated by band-limited noise. In sub-band speech recognition, the full band speech is divided into several frequency sub-bands, with the result of the recognition task given by the combination of the sub-band feature vectors or their likelihoods as generated by the corresponding sub-band recognizers. In this paper, we draw on the notion of discrete wavelet transform to divide the speech signal into sub-bands. We also make use of the robust features in sub-bands in order to obtain a higher sub-band speech recognition rate. In addition, we propose a likelihood weighting and fusion method based on the wavelet thresholding technique. The experimental results indicate that the proposed weighting methods for likelihood combination and classifiers fusion improve the sub-band speech recognition rate in noisy conditions.

Keywords: Recognition, Wavelet, Sub-band, Likelihood Combination.

1. Introduction

The issue of robustness against contamination with noise is considered as a mismatch between the training and testing conditions in automatic speech recognition (ASR) systems. The most common approaches for alleviating this mismatch can be divided into three main categories: data-driven methods, model-based techniques and the sub-banding approach. While data-driven methods try to compensate for the noise effects on speech or its features, the model-based approaches modify, instead, the acoustic models of the environment. The sub-band technique, on the other hand, is deemed as a new architecture for ASR systems, and can usually be applied to noises underlying the partial corruption of the frequency spectrum of the signal.

Data-driven methods are, in turn, usually divided into two main categories: speech signal enhancement approaches and feature compensation techniques. The enhancement methods reduce the mismatch by processing the noisy speech signal directly and trying to estimate clean speech from a noisy signal. Spectral subtraction [2] and wavelet thresholding [4, 20] are two instances of the speech enhancement schemes. Feature compensation techniques, on the other hand, usually decrease the mismatch in two ways; the first by applying a transformation to features for removing the noise effects such as: cepstral mean and variance normalization (CMVN) [8] and RASTA PLP [7], the second by extracting the new features to become more robust against the noise effects (e.g., the phase autocorrelation features (PAC) [9]).

The model-based methods modify the environment's statistical model so that it adapts to the changing conditions, for example, to noisy situations. This adaptation has the advantage that no decision or hypothesis about the speech is necessary to be made. Two known examples of such approaches are: parallel model combination (PMC) [5] and maximum likelihood linear regression (MLLR)[11].

The sub-band approach is theoretically based on the Fletcher's work [1, 22]. Fletcher et al. [1] suggested that in human auditory perception, the linguistic message gets decoded independently in different frequency sub-bands and the final decoding decision is derived from merging the decisions associated with the sub-bands. With this understanding, in the sub-band approach, the speech signal

* Corresponding Author. Email: nasersharif@guilan.ac.ir

is initially split into several frequency bands. Next, a feature vector is extracted from each sub-band. The sub-band feature vectors can be used in two ways. In feature combination schemes [6, 15, 17], the resultant features are concatenated to be used instead of the original full-band features. The second approach, known as model combination or sometimes likelihood combination, is effectively a variant of classifier fusion [6, 17]. In particular, each sub-band feature vector is processed by a sub-band recognizer to obtain a probability estimate corresponding to this sub-band. Then, a statistical formalism is used to fuse the classifiers and to recombine the probability estimates of all the sub-bands in order to get the final recognition results. Figure1 shows the overall process of classifier fusion and the model combination system.

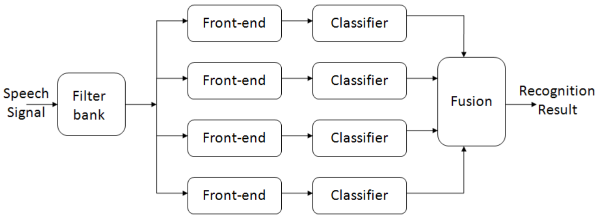


Fig 1. General schematic diagram for model combination and classifier fusion.

In this paper, we draw on the notion of discrete wavelet transform as a filter bank for decomposing the speech into sub-bands. Next, in the front-end section, we extract the phase autocorrelation-based as well as the modified-group-delay-based features from each sub-band so as to achieve more robustness against noise. We also propose a new weighting method based on wavelet thresholding for linear likelihood combination (as well as for classifier fusion) in sub-band speech recognition.

The remainder of the paper is organized as follows. Section 2 elaborates on the robust feature extraction methods used in sub-bands. The classifier fusion method is explained in Section 3. In Section 4, we introduce our proposed weighting method for combining likelihoods. Section 5 reports on the conducted experiments and the evaluation results. The paper concludes in Section 6.

2. Front-end and Feature Extraction

Traditional speech features are typically extracted from the power or the amplitude spectrum of the speech signal. Therefore, the changes of the speech spectrum due to additive noise can have a negative impact on the spectrum-based features, i.e. it deteriorates the performance of the speech recognition system in proportion to the noise power.

Conventional sub-band approaches overcome the band-limited noise effects through dividing the speech signal into sub-bands. However, the specifics of the extraction of features from the sub-bands do not get changed. In this paper, we propose to extract from the sub-bands those features which are robust to noise. This way, we can also use the sub-band approach as a robust method against full-band noises.

As pointed out earlier, several techniques have been proposed to reduce the sensitivity of the features to external noise. There are schemes which work at the spectral level, and try to reduce the effect of the additive noise on the speech spectrum for subsequent feature extraction. Spectral subtraction [2] and different spectral filtering techniques are amongst the well-known examples. In spectral subtraction, an estimation of the noise spectrum is subtracted from the speech power spectrum to remove the noise effects. Phase autocorrelation (PAC) is a comparable technique that is recently introduced [9]. It tries to make the autocorrelation coefficient less sensitive to additive noise [9, 15]. The group delay function (GDF), the negative derivative of the speech phase spectrum, is another technique used for speech spectrum estimation [23] and robust feature extraction [24]. With GDF, features are derived from the modified speech phase spectrum and not from the speech power or amplitude spectrum [13, 24]. In this paper, we make use of PAC and GDF-based features in the sub-bands, as discussed in the following two sub-sections.

2.1. Phase Autocorrelation-Based Features

The traditional autocorrelation function is computed as the dot product of the time-delayed speech vectors. Recently, an alternative measure of autocorrelation, namely phase autocorrelation (PAC), has been introduced, which is based on the angle between the vectors in the signal vector space [9]. The rationale behind the use of the angle is that, compared to the dot product, it typically gets less affected by the noise [12].

Here, we give a brief overview of the specifics of Phase AutoCorrelation (PAC), first presented in [9]. Consider a speech frame *s* as:

$$s=\left\{s[0],s[1],...s[N-1]\}\right\} \tag{1}$$

where *N* is the frame length. Suppose two vectors x_0 and x_k as:

$$\begin{aligned} x_0 &= \left\{s[0],s[1],...s[N-1]\}\right\} \\ x_k &= \{s[k],...,s[N-1],s[0],...,s[k-1]\} \end{aligned} \tag{2}$$

Using dot product, the autocorrelation coefficients of the speech frame are computed by:

$$R[k]=x_0^T x_k \tag{3}$$

R[k] can also be shown by:

$$R[k]=|\,x|^2\cos(\theta_k) \tag{4}$$

where $|x|^2$ denotes the energy of the frame and Θ_k represents the angle between vectors x_0 and x_k in the N dimensional space. The PAC coefficients are derived from the autocorrelation coefficients using the equation below:

$$P[k] = \theta_k = Arc\ \cos\ (\frac{R[k]}{|x|^2}) \tag{5}$$

Given that compared to the dot product, the angle is less affected by noise, PAC coefficients are more robust than the regular autocorrelation coefficients [12]. The Fourier equivalent of the PAC coefficients in frequency domain is referred to as the PAC spectrum. The computation of the PAC coefficients from the autocorrelation coefficients using (5) involves two operations: energy normalization and inverse cosine. As has been explained in [9], the inverse cosine transformation has the effect of enhancing the spectral peaks out of spectral valleys. PAC enhances the spectral peaks, on the one hand, and gives less weight to some high frequency information of the spectrum, on the other.

Similar to the features extracted from the regular spectrum, a class of features can also be extracted from the PAC spectrum. The Mel frequency cepstral coefficients (MFCC), extracted from the PAC spectrum, are called PAC-MFCC. Experimental results in [9] and [15] show that PAC-MFCC, though highly robust to noise, does not work well in clean speech conditions. In this paper, PAC-MFCC is extracted from each sub-band as the robust features to noise.

2.2. Group Delay Function-based Features

It is widely perceived that the magnitude spectrum represents the speech spectral information much better visually than the phase spectrum. Interestingly enough, its negative derivative, i.e. GDF [18, 23, 24], unlike the case with the phase spectrum, can be effectively used to extract the various parameters of a minimum phase speech signal. This is due to the fact that the magnitude spectrum of a minimum phase signal and its GDF are similar to each other. GDF is defined as:

$$\tau_p(\omega) = -\frac{d(\theta(\omega))}{d\omega} \tag{6}$$

where $\theta\ (\omega)$ is the unwrapped phase function. GDF can also be calculated from the speech signal by:

$$\tau_p(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2} \tag{7}$$

where the subscripts R and I indicate the real and the imaginary parts, respectively and $X(\omega)$ and $Y(\omega)$ represent the Fourier transforms of $x(n)$ and $nx(n)$, respectively. GDF requires that the signal be of minimum phase or that the poles of the transfer function be within the unit circle. GDF becomes spiky in nature due to the pitch peaks, noise and window effects. This has been illustrated in [13] and [18]. It is also noticeable that the denominator in equation (7) vanishes at zeros that are located close to the unit circle, making it necessary to somehow suppress the zeros. The spiky nature of the group delay spectrum can be overcome by replacing the denominator of GDF with its cepstrally smoothed version $S(\omega)$. This gives the modified GDF (MGDF) as follows [13, 18]:

$$\tilde{\tau}_p(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{(S(\omega))^2} \tag{8}$$

In [24], Zhu and Paliwal have defined the *product spectrum* as the product of the power spectrum and GDF as follows:

$$Q(\omega) = |X(\omega)|^2\ \tau_p(\omega) = X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega) \tag{9}$$

The product spectrum, referred to in this work as *PG*, is affected both by the magnitude spectrum and the phase spectrum. It enhances the region at the formants over the MGDF and has an envelope comparable to that of the power spectrum. The product spectrum represents well the details of clean speech power spectrum, but it is not capable of enhancing spectral peaks as well as PAC.

It is shown in [13] and [24] that by using the MFCCs extracted from the product spectrum, (hereinafter referred to as *PG-MFCC*), a higher recognition rate can be obtained compared to when using MFCCs extracted from MGDF. Hence, in this work, we make use of PG-MFCC as the robust features in the sub-bands.

3. Classifier Fusion: Likelihood Combination

As pointed out earlier, in the model-combination approach, each sub-band region is treated as a distinct source of information. A given sub-band recognizer generates probability estimates which must be combined at some level of the time segmentation such as the phoneme, the syllable or the word level. The specifics of combining the probability estimates from the different sub-band recognizers essentially influence the performance of the combined system. Depending on the nature of the sub-band recognizer, i.e. whether it is likelihood-based such as HMM or posterior-based like HMM/ANN hybrid classifier, the statistical formalism changes [3, 6]. This statistical formalism can take on a linear or nonlinear form. The Dempster–Shafer (DS) combination rule [21, 22] has been recently proposed for neural network classifiers.

In case of HMM recognizers, the likelihoods, as returned by HMMs, can linearly be recombined using sub-band weighting based on the following equation [3, 11, 14]:

$$S(x, M) = \sum_{b=1}^{B} \alpha_{b,M} P(x\,|\,M, b) \tag{10}$$

where S represents the score of the utterance x with model M, $P(x\ |\ M,\ b)$ is the likelihood returned by the HMM corresponding to the model M in sub-band b and finally, B is the number of sub-bands.

The difficulty within the sub-band weighting approach is the estimation of weighting factors $\alpha_{b,M}$ such that the recognizers associated with the cleaner and more reliable sub-bands be given a higher weight. The most common weighting factors are: SNR estimation in each sub-band as reported in [3, 19] and inverse HMM entropy in each sub-band as discussed in [11, 14]. In this paper, we propose a new weighting factor based on wavelet thresholding.

4. Sub-band Weighting Based on Wavelet Thresholding

The removal of noisy components by thresholding the wavelet coefficients is based on the observation that in many signals, speech no different, energy is mostly concentrated within a small number of coefficients. These coefficients are relatively large compared to the other coefficients or to any other signal (especially noise) that has its energy spread over a large number of coefficients. Hence, by setting the smaller coefficients to zero, one can eliminate noise while preserving the important information of the original signal [4]. Accordingly, the wavelet coefficients are compared to a threshold and their values are changed based on a threshold function such as hard, soft or semi-soft thresholding functions [20]. The threshold value can be estimated in many ways. For instance, Donoho [4, 20] has suggested a well-known estimation method with the following relation:

$$T = \hat{\sigma}_n \sqrt{2\log(N)} \tag{11}$$

where T is the threshold value, N denotes the length of the noisy signal and $\hat{\sigma}_n$ is a robust estimate of the noise level based on the median absolute deviation of the wavelet coefficients at the finest resolution level with index n. The reason for considering only the finest level is that its corresponding wavelet coefficients constitute most of the noise [20].

Alternatively, we have also used the *minimax* principle as the other threshold estimation technique. Since the de-noised signal can be assimilated into the estimator of the unknown regression function, the *minimax* estimator is defined to be the one that produces the minimum value for the maximum mean square error corresponding to the worst function in a given set [18].

Based on the wavelet thresholding idea, we can safely assume that the coefficients below the threshold contain noise information and those above the threshold encompass the information of the speech signal. Therefore, we can determine the amount of noise contamination in wavelet sub-bands via comparing the energy or the number of the aforementioned coefficients with each other. In particular, we propose the following measure for determining the sub-band contamination and reliability:

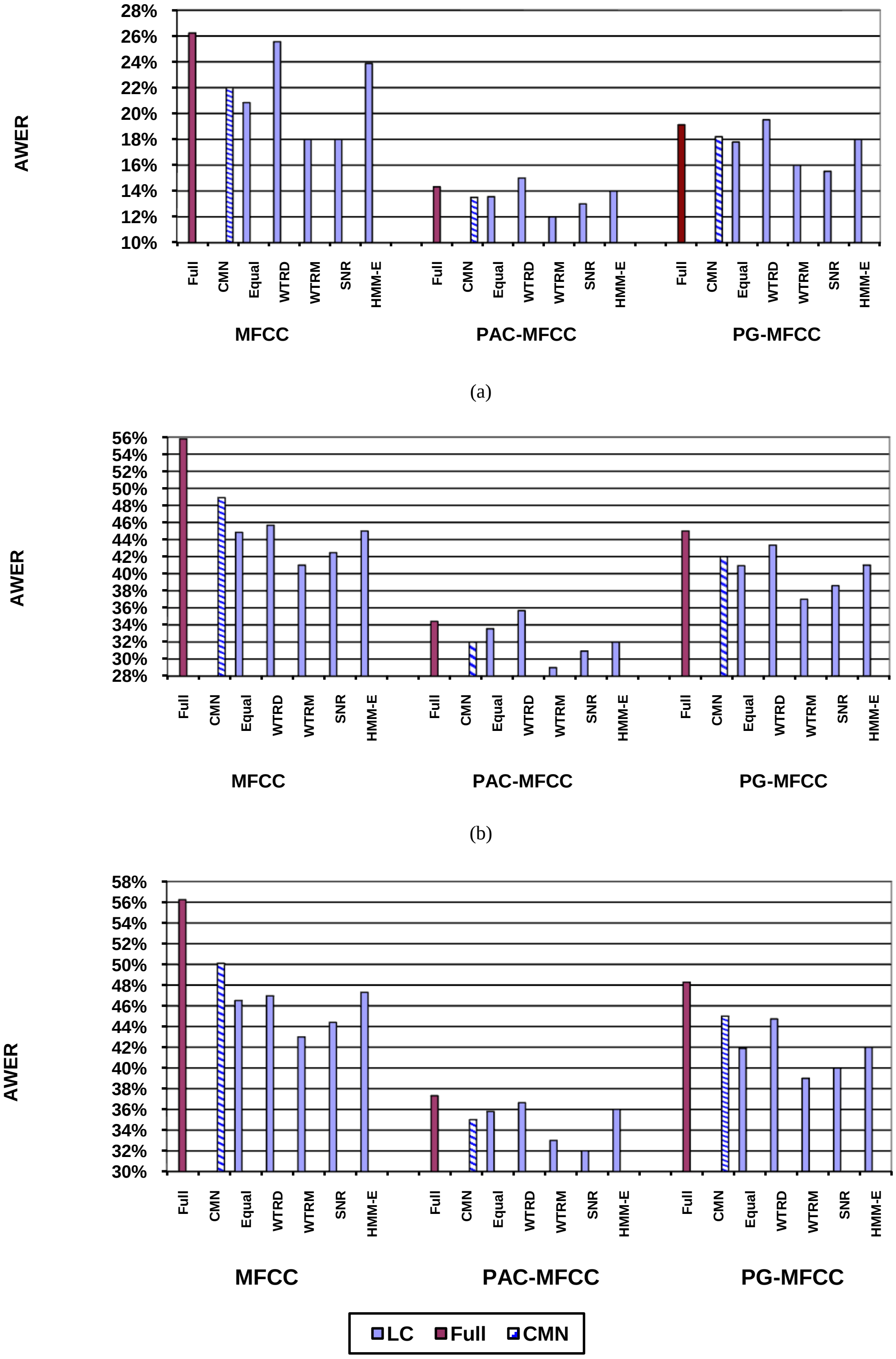
$$WTR(T_j) = \frac{\sum_{i=1}^{N} G(w_i, T_j)}{N}$$
$$G(w_i, T_j) = \begin{cases} 1 & for\ all\ \ |w_i| > T_j \\ 0 & otherwise \end{cases} \tag{12}$$

where N is the number of wavelet coefficients in the j-th wavelet sub-band, w_i is the i-th wavelet coefficient in the j-th wavelet sub-band and T_j is the threshold value in the j-th sub-band. We substitute $\alpha_{b,M}$ in equation (10) with this measure for the corresponding sub-band. The WTR value in equation (12) relies on accurate noise and threshold estimations. It is of note that by this thresholding, the wavelet coefficients are not changed in the sub-bands; instead, we basically compute WTR and plug it as $\alpha_{b,M}$ into equation (12).

5. Experiments and Results

We evaluate the proposed method on TIMIT database used as a benchmark dataset for isolated word recognition. Two sentences from the speakers in two dialect regions have been selected and segmented into words; in particular, we have 21 words spoken by 151 speakers including 49 females and 102 males. These speakers have been divided into train and test speakers according to the TIMIT speakers division. Our training set contains 2349 utterances spoken by 114 speakers. The testing set includes 777 utterances spoken by 37 speakers. Our recognizer is CDHMM with 6 states and 8 Gaussian mixtures per state and trained on clean speech. Three types of additive noises have been used: pink, white and factory noises, selected from NOISEX92 database. We have added these three noises to both training and testing sets. We have chosen four sub-bands and used the discrete wavelet transform for decomposing the speech into four sub-bands with dyadic bandwidths: 0-1 kHz, 1-2 kHz, 2-4 kHz, and 4-8 kHz. This selection has been justified by our observations from previous work [16]. We have used the 20-th order Daubechies wavelet as the wavelet decomposition filter. Over the course of the feature extraction phase, we have divided the 24 Mel filter into four sub-bands. We extracted four features (PG-MFCC or PAC-MFCC or MFCC) and four delta features (delta-PG-MFCC or delta-PAC-MFCC or delta-MFCC) from each of the first three sub-bands. In the full-band system, the feature vector contains 12 features (PG-MFCC or PAC-MFCC or MFCC) and 12 delta features (delta-PG-MFCC or delta-PAC-MFCC or delta-MFCC), making it of a total length of 24.

In Figure 2, the results associated with the full-band speech recognition system have been marked as “Full”, while “LC” identifies the results of the sub-band speech recognition with likelihood combination. “CMN”, on the other hand, denotes the cepstral mean subtraction method as applied to the MFCCs extracted from the full-band speech signal. The four abbreviations "Equal", "SNR", "HMM-E" and "WTR" represent four different likelihood weighting methods based on: equal weights, sub-band signal to noise ratio, HMM entropy and our proposed technique based on wavelet thresholding, respectively.



(a)

(b)

(c)

Fig. 2. Average word error rate (AWER) for three noise types (white, factory and pink); (a) SNR= 10 dB, (b) SNR= 5 dB, (c) SNR= 0 dB.

Our weighting initiative based on wavelet thresholding is marked with “WTR_M” when threshold estimation is done with reference to the *minimax* method, and with “WTR_D” to denote our alternative use of the Donoho method [4] instead. As can be seen in the figure 2, for all feature types, the LC system as a feature compensation method has better recognition results compared to the full-band system and the full-band CMN method. In addition, the LC system with PG_MFCC and PAC-MFCC in sub-bands outperforms the same system with the use of MFCC instead, indicating that using robust features in sub-bands pays off with a superior performance for the LC system. It is also worth noting that PAC-MFCC happened to be more effective in this case for increasing the performance of the LC system, as compared to PG-MFCC.

As for the weighting methods, it can be seen that in most cases, the *minimax* estimated threshold outperforms its counterparts. This is while weighting based on the notion of Donoho’s estimated threshold leads to a lower recognition rate in comparison with both the SNR weighting method as well as weighting based on *minimax* threshold. This can be attributed to the fact that the Donoho’s threshold has been primarily proposed for the white noise [4]. We may, thus, conclude that our proposed weighting method relies on the appropriate choice of noise and threshold estimation techniques. Still, our results suggest that the *minimax* method is promising to give a better threshold compared to the Donoho’s technique for all noise types.

6. Conclusion

In this paper, we have shown that using robust features in sub-bands, such as PAC-MFCC or PG-MFCC, results in a higher recognition rate for the likelihood combination system in the presence of noise. We have also demonstrated that the LC system, by using robust features in sub-bands, outperforms CMN as a conventional full-band feature compensation method. In addition, we have proposed a new likelihood and sub-band weighting method based on the notion of wavelet thresholding. Experimental results reveal that should a suitable threshold estimation technique be used, a higher sub-band noisy speech recognition rate can be obtained using the proposed weighting method in comparison to the other weighting approaches. Also suggested by the outcome of the experiments is that *minimax* thresholding estimates a better threshold in comparison to the Donoho’s threshold estimation approach.

References

[1] J.B.Allen, How do human process and recognize speech, IEEE Trans. on acoustics, speech and signal processing 2 (4), pp. 567-577, 1994.

[2] S. F. Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. on acoustics, speech and signal processing 27 (2), pp. 113-120, 1979.

[3] C. Cerisara, D.Fohr, Multi-band automatic speech recognition, Computer Speech and Language 15 (2), pp. 151-174, 2001.

[4] D. L. Donoho, Denoising by soft thresholding, IEEE Trans. on Information Theory 41 (3), pp. 613-627, 1995.

[5] M.J.F.Gales, S.J.Young, Robust continuous speech recognition using parallel model combination, IEEE Trans. on acoustics, speech and signal processing 4 (5), pp. 352-359, 1996.

[6] A. Hagen, A. Morris, Recent advances in the multi-stream HMM/ANN hybrid approach to noise robust ASR, Computer Speech and Language 19 (1), pp. 3-30, 2005.

[7] H. Hermansky, N. Morgan, RASTA processing of speech, IEEE Trans. on acoustics, speech and signal processing..2 (4), pp. 578-589, 1994.

[8] X. Huang, A.Acero, H. Hon, Spoken language processing, Prentice Hall, 2001.

[9] S. Ikbal, H. Misra, H. Bourlard,, Phase autocorrelation derived robust speech features, In: Processing of IEEE Int. Conf. on Acoustics, Speech, and Signal processing, 2003.

[10] Y. Kessentini, T. Paquet, A. B. Hamadou, Off-line handwritten word recognition using multi-stream hidden Markov models, Pattern Recognition Letters, 31 (1), pp. 60-70, 2010.

[11] C.J. Leggetter, P.C. Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, Computer speech and language 9 (2), pp. 171-185, 1995.

[12] D. Mansour, B.Juang, A family of distortion measure based upon projection operation for robust speech recognition, IEEE Trans. on acoustics, speech and signal processing 37 (11), pp. 1659-1671, 1989.

[13] H.A Murthy, V.Gadde, The modified group delay function and its application to phoneme recognition, In: Proc. of IEEE Int. Conf. on Acoustic, Speech, and Signal processing, 2003.

[14] B.Nasersharif, A.Akbari, Improved HMM entropy for robust sub-band speech recognition, In: Proc. of 13th European Signal Processing Conferences (EUSIPCO), (2005).

[15] B.Nasersharif, A.Akbari, Sub-band weighted projection measure for sub-band speech recognition in noise, IEE Electronics letter. 42, (14), pp. 829-831, 2006.

[16] B.Nasersharif, A.Akbari, Application of wavelet transform and wavelet thresholding in robust sub-band speech recognition, In: Proc. of European Signal Processing Conference, 2004.

[17] S. Okawa, E. Boochieri, A. Potamianos, Multi-band speech recognition in noisy environment, in: Proceeding of IEEE Int. Conf. on Acoustics, Speech, and Signal processing, 1998.

[18] K.Paliwal, L.Alsteris, Usefulness of phase spectrum in human speech perception, In: Proc. of EUROSPEECH, 2003.

[19] X. Shao, J. Barker , Stream weight estimation for multistream audio–visual speech recognition in a multispeaker environment ,Speech Communication, 50, (4), pp. 337-353, 2008.

[20] K.P.Soman, K.I.Ramachandran, Insight into wavelets: From Theory to Practice, Second Edition, Prentice-Hall of India, 2005.

[21] F. Valente, H. Hermansky, Combination of acoustic classifiers based on Dempster–Shafer theory of evidence, In. Proc. ICASSP 2007.

[22] F. Valente, Multi-stream speech recognition based on Dempster–Shafer combination rule, Speech Communication, 52, (3), pp. 213-222, 2010.

[23] B.Yegnanarayana, H.A Murthy, Significance of group delay functions in spectrum estimation, IEEE Trans. on Acoustic, Speech and signal processing 40 (9), pp. 2281-2289, 1992.

[24] D.Zhu, K Paliwal, Product of power spectrum and group delay function for speech recognition, In: Proceeding IEEE Int. Conf. on Acoustics, Speech, and Signal processing, 1, pp. 125-128, 2004.