Sub-Band Weighted Projection Measure for Robust Sub-Band Speech Recognition

Babak Nasersharif, Ahmad Akbari

Computer Engineering Department
Iran University of Science and Technology
{nasser_s, akbari}@iust.ac.ir

Abstract

In recent years, sub-band speech recognition has been found useful in robust speech recognition, especially for speech signals contaminated by band-limited noise. In sub-band speech recognition, full band speech is divided into several frequency sub-bands and then sub-band feature vectors or their generated likelihoods by corresponding sub-band recognizers are combined to give the result of recognition task. In this paper, we concatenate sub-band feature vectors, where we extract phase autocorrelation (PAC) MFCC, as noise robust features, from each sub-band. Furthermore, we extend a model adaptation method, named sub-band weighted projection measure (SWPM), to adapt HMM Gaussian mean vectors to concatenated sub-band feature vectors in noisy conditions. The experimental results indicate that the proposed method significantly improves the sub-band speech recognition system performance in presence of additive noise.

1. Introduction

The problem of robustness in ASR systems against contamination with noise is considered as a mismatch between the training and testing conditions. Common approaches used to reduce the mismatch can be divided into three main categories: data-driven methods, model-based techniques and sub-band approach. Data-driven methods try to compensate noise effects on speech or speech features, where model-based approaches modify acoustic models instead of speech signal or its features. Sub-band technique, viewed as a new architecture for ASR systems, can be usually applied to noises which cause partial corruption of signal frequency spectrum. Data-driven methods usually are divided into two main categories: speech signal enhancement approaches and feature compensation techniques. The enhancement methods process the noisy speech signal directly and try to estimate clean speech signal from noisy speech signal and reduce the mismatch in this way. Spectral subtraction [5] and wavelet thresholding [10] are two instances of speech enhancement schemes. Feature compensation techniques usually decrease the mismatch in two ways. In the first methods, a transformation is applied to features to remove noise effects such as, cepstral mean normalization (CMN) [5] and RASTA PLP [13]. In the second method, new features are extracted to become more robust to noise effects such as, phase autocorrelation features (PAC) [2].

Model-based methods modify environment statistical model so that it adapts to new properties of environment, for example, noisy conditions. This adaptation has the advantage that no decisions or hypotheses about speech are necessary. Some examples of such approaches are: parallel model combination (PMC) [9] and maximum likelihood linear regression (MLLR) [11]. In the sub-band approach, the speech signal is first divided into several frequency bands. Then in each sub-band, a feature vector is extracted. After further processing, the sub-band feature vectors can be treated in two ways: they are concatenated and used to replace the original feature (feature combination) [7], or each of them is processed by a separate sub-band recognizer which is trained on respective sub-bands. In this case, each sub-band recognizer generates a probability estimate. After this, a statistical formalism is used to recombine the respective probability estimates. This approach is named probability combination or model combination [3] [4] [8].

In this paper, we propose a combination of sub-band technique, a feature compensation approach and an extended model-based method called sub-band weighted projection measure (SWPM). In this way, we use PAC-MFCC [2] as a robust feature in each sub-band and construct a feature vector by concatenating sub-band feature vectors. In next step, we define SWPM and use it to adapt HMM Gaussian mean vectors to concatenated sub-band feature vectors. The remainder of this paper is organized as follows. Section 2 discusses the phase autocorrelation based features and feature compensation techniques. The remainder of this paper is organized as follows. Section 2 discusses the phase autocorrelation based features and feature compensation techniques. The remainder of this paper is organized as follows. Section 2 discusses the phase autocorrelation based features and feature compensation techniques. The remainder of this paper is organized as follows. Section 2 discusses the phase autocorrelation based features and feature compensation techniques. The remainder of this paper is organized as follows. Section 2 discusses the phase autocorrelation based features and feature compensation techniques. The remainder of this paper is organized as follows. Section 2 discusses the phase autocorrelation based features and feature compensation techniques. The remainder of this paper is organized as follows. Section 2 discusses the phase autocorrelation based features and feature compensation techniques.

2. Phase autocorrelation based features

A short review of the Phase AutoCorrelation (PAC), firstly presented in [2], is as follows: if $s$ represents a speech frame given by:

$$s = \{s[0], s[1], \ldots, s[N-1]\}$$

(1)

where $N$ is frame length, and

$$x_0 = \{s[0], s[1], \ldots, s[N-1]\}$$

$$x_k = \{s[k], s[k+1], \ldots, s[N-1], s[0], \ldots, s[k-1]\}$$

(2)

then the autocorrelation coefficients are computed using dot product given by:

$$R[k] = x^T \cdot x_k$$

(3)

On the other hand,

$$R[k] = |x|^2 \cos(\theta_k)$$

(4)

Where $|x|^2$ denotes the energy of the frame and $\theta_k$ represents the angle between vectors $x_0$ and $x_k$ in $N$ dimensional space. PAC coefficients are derived from autocorrelation coefficients using following equation:
\[ P[k] = \theta_k = \text{Arc cos} \left( \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| \cdot |\vec{b}|} \right) \] (5)

As angle gets less affected in noise than dot product, PAC coefficients are more robust to noise than the regular autocorrelation coefficients [14]. The Fourier equivalent of PAC coefficients in frequency domain is called PAC spectrum. Similar to the features extracted from the regular spectrum, a class of features named PAC based features can be extracted from the PAC spectrum. Mel frequency cepstral coefficients extracted from PAC spectrum is called PAC-MFCC. Experimental results in [1] show that PAC-MFCC is very robust to noise but deteriorates in clean conditions that this is its drawback.

In current work, we apply PAC-MFCC to sub-band speech recognition where we use feature combination approach. We extract PAC-MFCC (as a robust feature to noise) from each sub-band and construct the sub-band feature vector called sub-vector. Then, we concatenate these sub-vectors to obtain a new feature vector. By using this method, we can increase performance of sub-band approaches in presence of wide-band noises which contaminate all frequency sub-bands.

3. Sub-band weighted projection measure

The theory behind of weighted projection measure (WPM) is based on the observation of Mansour and Juang [14] that the norms of cepstral vectors are reduced by additive white noise. From this, a computationally efficient measure based on the projection operation was formulated which significantly improved DTW speech recognition performance in presence of noise. Carlson and Clements extended the projection measure to be used in a CDHMM-based recognition system [12]. They incorporated a scale factor into the CDHMM state distribution or equivalently into the Gaussian likelihood score to compensate for the reduction in vector norm. They used MFCC features instead of cepstral coefficients. The measure was found to significantly improve speaker dependent isolated word recognition rate in presence of several noise types, including white, jittering white and broadband colored noise [12]. Their compensated expression for Gaussian distribution can be more accurately represented as follows:

\[ b_{i,j}(c_t) = \mathcal{N}(c_t|\mu_{i,j}, \Sigma_{i,j}) = \exp\left(-\frac{1}{2}(c_t - \mu_{i,j})^T \Sigma_{i,j}^{-1} (c_t - \mu_{i,j}) \right) \] (6)

where:
- \( c_t \) : observation vector for frame \( t \)
- \( \mu_{i,j} \) : mean vector of \( j \)-th Gaussian mixture in state \( i \)
- \( \Sigma_{i,j} \) : covariance matrix of \( j \)-th Gaussian mixture in state \( i \)
- \( \lambda_{i,j} \) : scale factor for frame \( t \) in \( j \)-th Gaussian mixture in state \( i \)
- \( b_{i,j}(c_t) \) : generated probability by \( j \)-th Gaussian mixture for observing vector \( c_t \) in state \( i \)

In the Viterbi algorithm, an appropriate matching measure between the observation and Gaussian mixture distribution function can be found from the log likelihood of above Gaussian function. This results in the following:

\[ \log b_{i,j}(c_t) = (c_t - \lambda_{i,j}\mu_{i,j})^T \Sigma_{i,j}^{-1} (c_t - \lambda_{i,j}\mu_{i,j}) + \log |\Sigma_{i,j}| + N \log(2\pi) \] (7)

From the orthogonality principle, the optimal \( \lambda_{i,j} \) value is the projection of \( c_t \) onto \( \mu_{i,j} \) weighted in the space spanned by \( \Sigma_{i,j}^{-1} : \)

\[ \lambda_{i,j,t} = c_t^T \Sigma_{i,j}^{-1} \mu_{i,j} \] (8)

With this value of \( \lambda_{i,j,t} \), the above log likelihood becomes what is called as WPM. In relations (7) and (8), Gaussian mixture means are scaled and transformed by the equalization factor \( \lambda_{i,j,t} \). We can compare this to basic MLLR. In basic MLLR, the Gaussian means parameters are transformed according to:

\[ \mu = A \mu + b \] (9)

where \( A \) is a \( n \times n \) matrix and \( b \) is an \( n \) dimensional vector (and \( n \) is dimensionality of observations) [11]. The transformation matrix \( A \) and vector \( b \) are estimated by expectation maximization (EM) algorithm such that the likelihood of adaptation data is maximized. We can notice that WPM is a simple form of MLLR with \( A = \lambda \ I, b = 0 \), where \( \lambda \) is an \( n \times n \) identity matrix and \( \lambda \) is a scalar. Furthermore, WPM does not need adaptation data and off-line training for estimation of coefficient which can be easily computed for every observation vector in relation (8). So, we generalize WPM to use it with more complex observation vectors.

We initially represent our generalization in term of relation (9) and then relate it to WPM. We replace \( A \) by a diagonal matrix in relation (9) and assign a zero vector to \( b \). In this way, different values can be assigned to \( \lambda \) for different elements of observation vector, in spite of WPM which same value is assigned to all elements of observation vector. We define \( A \) as follows to insert it instead of \( \lambda_{i,j,t} \) in relation (7):

\[ A_{i,j,t} = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix} \] (10)

Assuming that Gaussian covariance matrix is diagonal, if we replace \( \lambda_{i,j,t} \) with \( A_{i,j,t} \) in (7) and rewrite the relation in summation form, we have:

\[ \log b_{i,j}(c) = \log |\Sigma_{i,j}| + N \log(2\pi) + \sum_{k=1}^{n} \frac{(c_{i,k} - \lambda_{i,k})^2}{\sigma_k^2} \] (11)

where \( c_{i,k} \) and \( \sigma_k^2 \) are \( k \)-th elements of \( c_i \) and \( \sigma_k^2 \) respectively.

We can write the \( \sum \) term in right-hand side of relation (11) as follows:
\[
\sum_{k=1}^{R} \left( c_\lambda - \mu_{j,k} \right)^2 \sum_{k=1}^{R} \frac{\left( \mu_{j,k} - \lambda \mu_{j,k} \right)^2}{\delta^2_{j,k}} + \sum_{i=1}^{\infty} \frac{\left( e_i - \lambda \mu_{j,i} \right)^2}{\delta^2_{i}} + \sum_{i=1}^{\infty} \frac{\left( e_i - \lambda \mu_{j,i} \right)^2}{\delta^2_{i}}
\]

According to (12), we divide the observation vector (feature vector) to R smaller independent parts (sub-vectors) based on the effects of noise on them. In the same way, Gaussian mean vector and covariance matrix are also divided into R mean sub-vectors and R covariance sub-matrix corresponding to feature vector. We can illustrate it as:

\[
c_i = \begin{bmatrix} x_1 \\ x_2 \\ \ldots \ldots \\ x_n \end{bmatrix}, \quad X_{r,j} = \begin{bmatrix} x_{i,j},_{1} \\ x_{i,j},_{2} \\ \ldots \ldots \\ x_{i,j},_{n} \end{bmatrix}, \quad \mu_{j,i} = \begin{bmatrix} m_1 \\ m_2 \\ \ldots \ldots \\ m_n \end{bmatrix}, \quad M_r = \begin{bmatrix} m_{r,1} \\ m_{r,2} \\ \ldots \ldots \\ m_{r,n} \end{bmatrix}
\]

where \( X_{r,j}, M_r \) are r-th feature sub-vector and r-th mean sub-vector, respectively and \( V_r \) is r-th covariance sub-matrix.

Corresponding to each feature sub-vector, a different \( \lambda \) must be computed. Then we rewrite equation (12) as:

\[
\sum_{k=1}^{R} \left( c_\lambda - \mu_{j,k} \right)^2 \sum_{k=1}^{R} \frac{\left( \mu_{j,k} - \lambda \mu_{j,k} \right)^2}{\delta^2_{j,k}} + \sum_{i=1}^{\infty} \frac{\left( e_i - \lambda \mu_{j,i} \right)^2}{\delta^2_{i}} + \sum_{i=1}^{\infty} \frac{\left( e_i - \lambda \mu_{j,i} \right)^2}{\delta^2_{i}}
\]

By inserting (14) in (11) and deriving from (11) with respect to \( \lambda \) (r = 1, 2, ..., R) and rewriting it in a matrix form \( \lambda_r \) can be computed as follows:

\[
\lambda_r = \frac{X^T \Sigma \gamma V, \mu^T_{r,j}}{M_r^T \Sigma \gamma V, \mu^T_{r,j}} \quad r = 1, 2, ..., R
\]

As indicated in (15), we do not need any adaptation data and off-line training to compute each \( \lambda_r \). In addition, any a-priori information about type of noise and SNR is not required to compute \( \lambda_r \). With inserting these \( \lambda_r \) in relation (14) and then (11), the log likelihood in (11) becomes what we call sub-band weighted projection measure (SWPM).

4. Experiments and results

We report our results on TIMIT database for isolated word recognition. Two sentences from speakers in two dialect regions were selected and were segmented into words. In this way, we have 21 words spoken by 151 speakers including 49 females and 102 males. These speakers were divided into train and test speakers according to 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 which is trained on clean speech. Two types of additive noises were used: white and factory noises selected from NOISEX92 database. We added two noises to both training and testing sets. We chose 4 sub-bands as in [4] [6] and used the discrete wavelet transform for speech decomposition into 4 sub-bands with dyadic bandwidths: 0-1 kHz, 1-2 kHz, 2-4 kHz, and 4-8 kHz. This selection is based on our previous work results [1]. We used 5-th order Daubechies wavelet as wavelet decomposition filter because of its smoothness and compact support [1]. In feature extraction phase, we divided the 24 Mel filter between 4 sub-bands. In this manner, we applied 6 Mel filters for each sub-band and then extracted 3 PAC-MFCC (or 3 MFCC) and delta-PAC-MFCC (or delta-MFCC) features from each sub-band. Hence, the length of each feature vector is 6. In the full-band system, feature vector contains 12 PAC-MFCC (or 12 MFCC) and 12 delta-PAC-MFCC (or delta-MFCC) features and so its length is 24.

Fig. 1 shows results of feature combination (FC) and the effect of applying WPM and SWPM to it in presence of white noise. The results are reported on SNR values of 10 and 0 dB, for 3216 utterances of testing and training noisy database, in terms of word error rates (WER). In Fig. 1, the word "full" shows the full-band system WER. As Fig. 1 shows, in both SNR values of 10 and 0 dB, PAC-MFCC performs better than MFCC in both full-band and FC systems. Moreover, FC system has better performance than full-band system. According to Fig. 1, if we apply WPM and SWPM to FC system, its word error rates decrease. The difference between FC and full-band system performance increases. It can be seen that SWPM improves FC system performance more than WPM. In case of PAC-MFCC, when we use SWPM for FC system, the improvements in performance are almost about 4% and 12% for SNR values of 10 and 0 dB, respectively. But, by applying WPM to FC, these improvements are about 1.3% and 3% for SNR values of 10 and 0 dB, respectively.

Fig. 2 illustrates the evaluation results of feature combination system and applying WPM and SWPM to it in presence of factory noise. As Fig. 2 displays, in both SNR values of 10 and 0 dB, PAC-MFCC shows better results than MFCC in both full-band and FC systems. Further, FC system performs better than full-band system. If we compare FC results with full-band results for using PAC-MFCC, we can see that its results have improved about 1% and 10% for SNR values of 10 and 0 dB, respectively. On the other hand, when we utilize MFCC, FC results are about 6% better than full-band results in SNR value of 0 dB. The difference between FC and full-band word error rates for MFCC is about 3.5% in SNR value of 10 dB. It can be seen from Fig. 2 that when we apply WPM and SWPM to FC system, its word error rates decrease. The amounts of decreases for SWPM are more than WPM, when we use PAC-MFCC. They are almost about 3% and 16% for using SWPM in SNR values of 10 and 0 dB, respectively. On the contrary, the improvements for using WPM are almost about 1.2% and 11% for SNR values of 10 and 0 dB, respectively. In case of MFCC, WPM has a little better effect than SWPM on FC system performance.
5. Conclusion

In this paper, we used a robust feature, named PAC-MFCC, instead of MFCC for sub-band feature combination. Our results showed that combination of PAC-MFCC sub-vectors were more effective than combination of MFCC sub-vectors. We also defined and extended a model adaptation method called SWPM to apply it to concatenated feature vectors which consist of PAC-MFCC (or MFCC) sub-vectors extracted from sub-bands. In this way, we compensated remained noise effects on each sub-vector separately. Our results showed that SWPM improved performance of feature combination system more than WPM, when we used PAC-MFCC sub-vectors.

As future work, we plan to use a priori information about noise in each sub-band in order to optimize SWPM for different types of noise and use PAC-MFCC only for noisy sub-bands.

6. References