IMPROVED HMM ENTROPY 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 use continuous density hidden Markov model (CDHMM) as recognizer and propose a weighting method based on HMM entropy for likelihood combination. We also use an HMM adaptation method, named weighted projection measure, to improve HMM entropy and its performance in noisy environments. The experimental results indicate that the improved HMM entropy outperforms conventional weighting methods for likelihood combination. In addition, results show that in SNR value of 0 dB, proposed method decreases word error rate of full-band system about 20%.

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), 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].

In this paper, we propose a combination of sub-band technique and a model-based method called weighted projection measure (WPM) [12]. In this way, we first use a new weighting method for probability combination based on HMM entropy. In next step, we use WPM to improve performance of this proposed weighting method.

The remainder of this paper is organized as follows. Section 2 discusses the likelihood combination and defines the HMM entropy for probability combination. In Section 3, the weighted projection measure is introduced. Section 4 includes our experiments and results. Finally, our conclusions are given in Section 5.

2-LIKELIHOOD COMBINATION IN SUB-BAND SPEECH RECOGNITION

As mentioned above, in model-combination approach, each sub-band region is treated as a distinct source of information. Each sub-band recognizer generates probability estimates which must be combined at some level of time segmentation such as phone, syllable or word level. The right choice in how to combine the probability estimates from the different sub-band recognizers essentially influences the performance of the combined system. Depending on the nature of sub-band recognizer, whether they are likelihood-based such as HMM, or posterior based like HMM/ANN hybrid classifier, the statistical formalism changes [3]. This statistical formalism can be in a linear or nonlinear form.

In the case of HMM recognizers, likelihoods, returned by HMMs can be linearly recombined using sub-band weighting based on the following equation [4]:

$$S(x,M) = \sum_{b=1}^{B} \alpha_{b,M} P(x \mid M, b) \quad (1)$$

where S represents the score of 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.

One problem in sub-band weighting approach is the estimation of weighting factors $\alpha_{b,M}$ in order to give more weight to recognizers corresponding to more reliable sub-bands. The most common weighting factors are SNR estimation in each sub-band [8] and inverse conditional entropy of each sub-band [7]. In this work, we propose a weighting factor based on HMM entropy which is defined in following.

2-1. HMM ENTROPY FOR LIKELIHOOD COMBINATION

The distribution of likelihoods at the output of the HMM contains information on the reliability of input observation vector to HMM. If an HMM shows a very high likelihood and all other HMMs have a low likelihood, this indicates a reliable input observation vector. On the contrary, when all HMMs have almost the equal likelihood, the input observation vector is very unreliable. This information can be obtained by computing the entropy of estimated likelihood for input observation vector to all HMMs. The entropy can be computed as follows:

$$H(O) = -\sum_{i=1}^{K} P(O \mid M_i) \log_2 P(O \mid M_i) \quad (2)$$

where:

O: observation vector in frame or word level

P (O|M$_i$): the generated likelihood for observation vector O by i-th HMM

K: the number of all HMMs

H (O): the entropy of observation vector O for all HMMs

Because the above entropy indicates the reliability of an observation vector, we can find a relationship between the entropy and weighting factor $\alpha_{b,M}$ in equation (1). The lower entropy value signifies the reliability of observation, whereas the higher entropy value is a sign of unreliable observation vector. Due to this property, we can use the inverse value of the entropy as weighting factor $\alpha_{b,M}$. So, if we rewrite equation (1) as follows:

$$H_b(O) = -\sum_{i=1}^{K} P(O \mid M_i, b) \log_2 P(O \mid M_i, b) \quad (3)$$

we can compute $\alpha_{b,M}$ like this:

$$\alpha_{b,M} = \frac{1}{H_b(O)} \quad (4)$$

3. WEIGHTED PROJECTION MEASURE

The theory behind WPM is based on the observation of Mansour and Juang [14] that the norms of cepstral vectors are reduced by additive white noise. From using 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 expanded 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 improve significantly 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:

$$b_{j,i}(c_t) = N(c_t, \lambda_{j,i,t}, \mu_{j,i}, C_{j,i}) = \frac{\exp\left(-\frac{1}{2}(c_t - \lambda_{j,i,t}\mu_{j,i})^T C_{j,i}^{-1}(c_t - \lambda_{j,i,t}\mu_{j,i}) \right)}{(2\pi)^{\frac{n}{2}} \mid C_{j,i} \mid^{\frac{1}{2}}} \quad (5)$$

where

c_t : observation vector for frame t

n: dimensionality of observation vector

$\mu_{j,i}$: mean vector of j-th Gaussian mixture in state i

$C_{j,i}$: covariance matrix of j-th Gaussian mixture in state i

$\lambda_{j,i,t}$: scale factor for frame t in j-th Gaussian mixture in state i

$b_{j,i}(c_t)$: 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:

$$\begin{aligned} \log b_{j,i}(c_t) = {} & (c_t - \lambda_{j,i,t}\mu_{j,i})^T C_{j,i}^{-1}(c_t - \lambda_{j,i,t}\mu_{j,i}) \\ & + \log \mid C_{j,i} \mid + N \log(2\pi) \end{aligned} \quad (6)$$

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

$$\lambda_{j,i,t} = \frac{c_t^T C_{j,i}^{-1} \mu_{j,i}}{\mu_{j,i}^T C_{j,i}^{-1} \mu_{j,i}} \quad (7)$$

With this value of $\lambda_{j,i,t}$, the above log likelihood becomes what is called as WPM.

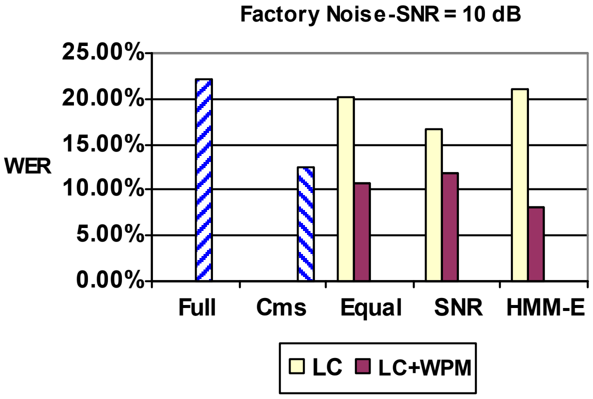
3-1. WPM AND HMM ENTROPY

When we apply WPM to CDHMM for noisy environments, the estimated likelihood by corresponding HMM for a noisy observation vector is maximized. Consequently, significant improvement in recognition rate is obtained [12]. This improvement is the logical result of increasing disparity between likelihood of different HMMs for noisy observation vector. An increase in difference between likelihoods of HMMs, leads to lower entropy value for observation vector in equation (2). Thus, we can say that applying WPM to CDHMM, decreases the HMM entropy defined by equation (2). This means an improvement in HMM entropy and HMM performance in noisy environments. Hence, when we apply WPM to HMM corresponding to b-th sub-band, we have a decrease in H_b value in equation (3). This means that we have a more reliable recognition result in b-th sub-band. Thus, the weighting factor $\alpha_{b,M}$ for this sub-band must increase. This can be supported by equation (4).

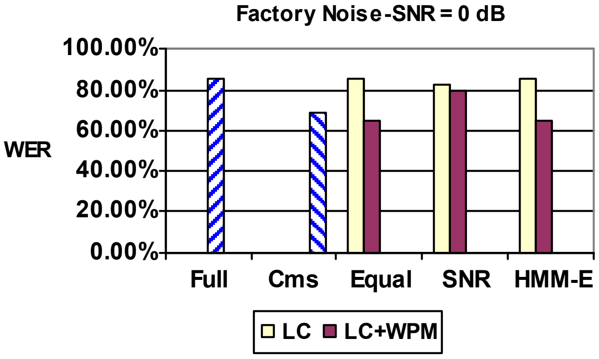
4. EXPERIMENTS AND RESULTS

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. 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: pink 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 pervious 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 MFCC and 3 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 MFCC and 12 delta-MFCC features and so its length is 24.

Fig. 1 shows results of likelihood combination (LC) and the effect of improving HMM entropy on it, in presence of factory 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. The word "Cms" indicates the conventional cepstral mean subtraction method for noise robustness. The three abbreviations "Equal", "SNR" and "HMM-E" show three different weighting methods based on equal weights, sub-band signal to no ise ratio and HMM entropy, respectively. As Fig. 1 shows, in SNR value of 10 dB, all three weighting methods in likelihood combination perform better than full-band system. In this case, SNR weighting method has the highest performance among them. When we apply WPM to LC in order to improve HMM entropy, the performances of all methods improve, especially in case of HMM-E weighting method. In SNR value of 10 dB, the improvements in performance are almost 9.4%, 5% and 13% for equal, SNR and HMM-E weighting methods, respectively. These results show that by improving HMM entropy, HMM-E has the highest performance in case of SNR value of 10 dB. When SNR value is equal to 0 dB, all three weighting methods perform a little higher than the full-band system. But, by applying WPM to LC, their performances increase significantly. In this case, decreases in word error rates are 21.2%, 3.2% and 20% for equal, SNR and HMM-E weighting method, respectively. It can be seen that in SNR value of 0 dB, the same performance is achieved for HMM-E and equal weighting methods by improving HMM entropy which both are higher than SNR weighting method. In addition, HMM-E has a better performance than conventional cepstral mean subtraction method.



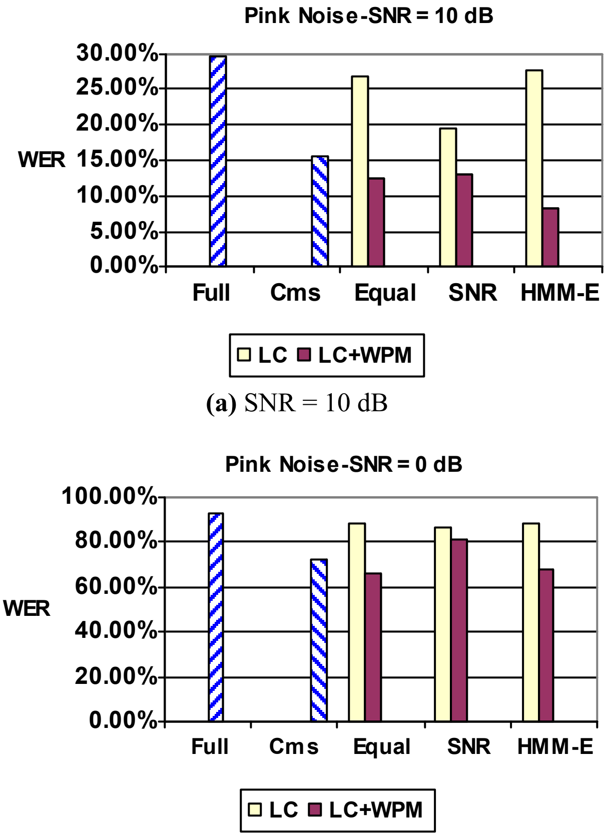
(a) SNR = 10 dB



(b) SNR = 0 dB

Fig. 1. Word error rates in presence of *factory noise* for SNR values of 10 and 0 dB

Fig. 2 illustrates the evaluation results of using likelihood combination and WPM in presence of pink noise. As Fig. 2 displays, in SNR value of 10 dB, higher performance is achieved, when we use likelihood combination. In this case, SNR weighting method has the lowest word error rate. Instead, when we apply WPM to LC, the highest performance belongs to HMM-E weighting method and the LC performance also improves significantly for all other weighting methods. This can be seen from the word error rates decrease. These decreases are almost 14.3%, 6.5% and 19.3% for equal, SNR and HMM-E weighting methods, respectively. In SNR value of 0 dB, the LC performance is higher than the full-band system. This performance improves when we apply WPM to LC. The improvements for equal, SNR and HMM-E weighting methods are almost 21.8%, 5.3%, 19.9%, respectively. In this case, both HMM-E and equal weighting methods have almost the same performance which is higher than SNR weighting method performance. Furthermore, HMM-E outperforms conventional cepstral mean subtraction.



(a) SNR = 10 dB

(b) SNR = 0 dB

Fig. 2. Word error rates in presence of *Pink noise* for SNR values of 10 and 0 dB

5. CONCLUSION

In this paper, we proposed a weighting method based on HMM entropy for likelihood recombination in robust sub-band speech recognition. That is a measure for reliability of observation vector and as a result, HMM classifier. Then, we used an HMM adaptation method named WPM to improve HMM entropy and therefore, our proposed weighting method. Our results show that using WPM improves performance of likelihood combination and especially, likelihood combination based on HMM entropy. As future work, we plan to define other new weighting methods for likelihood combination based on both the HMM entropy and the sub-band reliability. In addition, we want to use other HMM adaptation methods and support vector machines (SVM) to improve HMM entropy and as a consequence, likelihood combination performance.

6. REFERENCES

[1] B. Nasersharif, A. Akbari, "Application of wavelet transform and wavelet thresholding in robust sub-band speech recognition ", *European signal processing Conference*, vol.1, pp. 345-348, 2004.

[2] S. Ikbal, H. Misra, H. Bourlard, "Phase autocorrelation derived robust speech features", *IEEE International Conference on Acoustic, Speech, and Signal processing,* vol. 2, pp. 133-136, 2003.

[3] A. Hagen, *Robust speech recognition based on multi-stream processing,* Ph.D. thesis, Switzerland, 2001.

[4] C. Cerisara, D. Fohr, "Multi-band automatic speech recognition", *Computer Speech and Language*, vol.15, Issue. 2, pp. 151-174, April 2001

[5] X. Huang, A.Acero, H. Hon, *Spoken Language processing*, Prentice Hall, 2001.

[6] N. Mirghafoori, "A multi-band approach to automatic speech recognition", PhD thesis, ICSI, Berkeley, 1999.

[7] S. Okawa, E. Boochieri, A.Potamianos "Multi-band speech recognition in noisy environment", *IEEE International Conference on Acoustic, Speech, and Signal processing,* vol. 2, pp. 641-644, 1998.

[8] F. Berthommier, H. Glotin, E. Tessier, H. Bourlard, "Interfacing of CASA and partial recognition based on a multi-stream technique", *International Conference on Spoken Language Processing,* Australia,1998.

[9] M.J.F.Gales, S.J.Young, "Robust continuous speech recognition using parallel model combination", *IEEE Trans. on Speech and Audio Processing,* vol. 4, no. 5 , pp.352-359, September 1996.

[10] D.L.Donoho, "Denoising by soft thresholding", *IEEE Trans. on Information Theory ,* vol.41, no.3, pp. 613-627, May 1995.

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

[12] B.A. Carlson, M.A. Clements, "A projection-based likelihood measure for speech recognition in noise", *IEEE Trans. on Speech and Audio Processing,* vol. 2, no. 1, pp. 97-102, January 1994.

[13] H. Hermansky, N. Morgan, "Rasta processing of speech", *IEEE Trans. on Speech and Audio Processing,* vol.2, no. 4, pp. 578-589, 1994.

[14] D. Mansour, B. Juang, "A family of distortion measure based upon projection operation for robust speech recognition", *IEEE Trans. on Acoustic, Speech and signal processing,* vol. 37, pp.1659-1671, 1989