APPLICATION OF WAVELET TRANSFORM AND WAVELET THRESHOLDING IN ROBUST SUB-BAND SPEECH RECOGNITION

Babak Nasersharif, Ahmad Akbari

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

ABSTRACT

Sub-band speech recognition approaches have been proposed for robust speech recognition, where full band speech are divided into several sub-bands and then likelihood or cepstral vectors of the sub-bands are merged depending on their reliability. In this paper, we use wavelet transform for splitting of input speech into sub-bands in two ways: equal bandwidths for sub-bands and dyadic bandwidths for sub-bands. We also use wavelet thresholding techniques to determine a criterion for reliability of sub-bands and use it as weight in sub-band recombination.

1. INTRODUCTION

It is well known that current ASR systems don't work as well as human perception. Fletcher and his colleagues [1] suggested that in human auditory perception, the linguistic message gets decoded independently in different frequency sub-bands and the final decoding decision is based on merging the decisions from the sub-bands. ASR machine could benefit if it had the human ability to de-emphasize the unreliable frequency sub-bands. Toward this end, many ways recently proposed to utilize information of sub-bands [3,4].

Two modes of sub-band approaches have been applied: Likelihood recombination (LC) and feature recombination (FC) [2]. In LC, each sub-band is modeled independently. During recognition process, different speech classifiers are applied independently to each sub-band and each classifier provides a set of likelihood scores. Then all classifier outputs are combined to obtain global recognition likelihood. Various kinds of recombination modules such as linear combination and neural combination were considered for recombination of likelihood scores in each sub-band [3,4].

In FC a single feature vector is composed by joining the sub-band feature vectors together. However, results show that both FC and LC don't perform well for clean speech. They cause further degradation recognition for clean speech [3,8], mainly because correlations across the sub-bands are lost in both approaches.

From a psycho-acoustic point of view, one can expect that some frequency bands and some models are more robust to noise than others. When recombination is done in clean speech, it can not make use of this intrinsic robustness of the bands and of the models. The only way for the system to learn which bands and models are the most robust and to exploit this information, is to be confronted with noisy speech. This can guide us to use sub-band approaches for robust speech recognition.

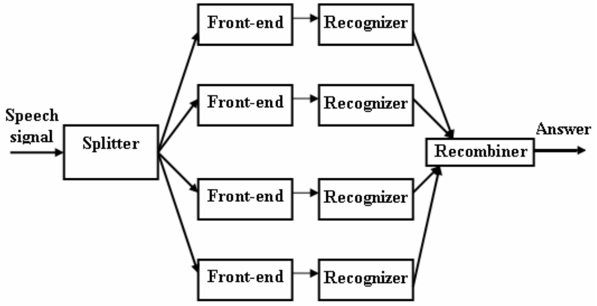

Figure 1. General architecture of likelihood recombination

In this paper, we use a multi-band speech recognition system as Fig. 1 [3]. The speech signal is divided into four frequency bands and independent processing is applied in each sub-band. Speech signal is first passed to a wavelet transform based filter bank which splits it into four sub-bands. The signal in each sub-band is encoded into a stream of acoustic vectors which are passed to a HMM based recognition system. The generated likelihoods by HMMs are given to a recombination module that delivers a unique answer to the recognition task. In this paper, we propose a new technique for likelihood recombination based on wavelet thresholding technique.

In section 2, we describe our speech decomposition method using wavelet transform. In the following sections, our proposed method for likelihood recombination are presented and compared for different noisy conditions.

2. WAVELET TRANSFORM AND SPEECH DECOMPOSITION

Wavelet transform has been intensively used in various fields of signal processing. It has the advantage of using variable size time-windows for different frequency bands. This results in a high frequency resolution in low bands and low frequency resolution in high bands.

Consequently, wavelet transform is a powerful tool for modeling non-stationary signals like speech that exhibit slow temporal variations in low frequency and abrupt temporal changes in high frequency. These properties can guide us to use wavelet transform in sub-band speech recognition. In this way, we select discrete wavelet transform to avoid complex design of filters that perform multi-band analysis.

A typical multi-level analysis for discrete wavelet transform (DWT) is depicted in Fig. 2. The decomposition filters, DL and DH, were adopted to split the input speech signal into two overlapped and equally spaced frequency bands in first level, where DL is low-passed and DH is high-passed. We then decimate each band by a factor 2, such that spectrum of each band is expanded to fill up the full frequency scale [9].

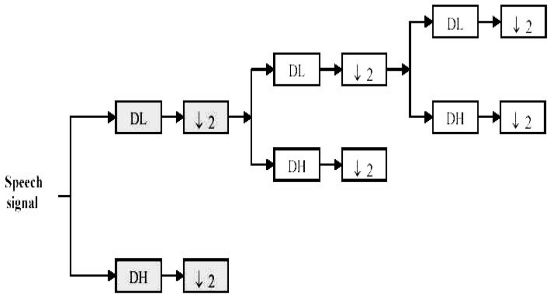

Figure 2. Block diagram of multi-level discrete wavelet analysis

One important problem in sub-band speech recognition is definition of frequency sub-bands and the frequency range spanned by each sub-band [3,4]. Narrow sub-bands may allow greater flexibility in isolating frequency-localized degradation, but the class discrimination within the sub-band decreases with decreasing amount of information in narrower sub-bands [4]. If we use multi-level wavelet transform for speech decomposition, we have a high frequency resolution in low bands and low frequency resolution in high bands as Fig. 2 shows. In this way, we obtain four sub-bands with dyadic bandwidths.

Fig. 3 shows the block diagram of another method for splitting input speech into four frequency bands. We use single-level discrete wavelet transform(SDWT) to split input speech signal into two frequency bands and use SDWT again to split each frequency bands into two sub-bands. At last, we obtain four sub-bands with equal bandwidths.

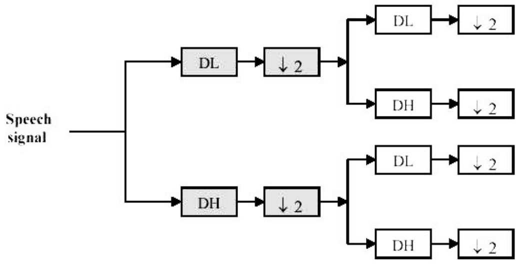

Figure 3. Block diagram of splitting method into 4 bands- equal bandwidths

2.1. Wavelet Thresholding

Removing noise components by thresholding the wavelet coefficients is based on observation that in many signals(like speech), energy is mostly concentrated in a small number of dimensions. The coefficients of these dimensions are relatively large compared to other dimensions or to any other signal (specially noise) that has its energy spread over a large number of coefficients. Hence, by setting smaller coefficients to zero, one can eliminate noise while preserving the important information of original signal [10]. Consequently, wavelet coefficients are compared to a threshold and it is determined which coefficients must be set to zero. The proper value of threshold can be determined in many ways. Donoho [10] has suggested the following relation for determining threshold value:

$$T = \sigma \sqrt{2 \log(N)} \quad (1)$$

where T is the threshold value and N is the length of noisy signal.

Thresholding can be performed as *Hard* or *Soft* thresholding that are defined as follows respectively*:*

$$THR_H (Y, T) = \begin{cases} Y & |Y| \succ T \\ 0 & |Y| \prec T \end{cases} \quad (2)$$

$$THR_S (Y, T) = \begin{cases} sign(Y)\ (|Y| - T) & |Y| \succ T \\ 0 & |Y| \prec T \end{cases} \quad (3)$$

where Y is wavelet coefficients of noisy signal.

3. LIKELIHOOD RECOMBINATION

Likelihoods returned by HMMs can be recombined using sub-band weighting as the following equation [3]:

$$S(x, M) = \sum_{b=1}^{B} \alpha_{b,M} P(x \mid M, b) \quad (4)$$

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. B is number of sub-bands.

One problem in sub-band weighting approach is the estimation of weighting factors $\alpha_{b,M}$. Many researchers use neural networks for non-linear estimation of weights [2,3]. They don't use any information of sub-bands in weighting directly. The weighting factor can be set if the reliability of the sub-bands is known. When characteristic of the noise signal is unknown, determining of this reliability is important. SNR can be an important factor for determining reliability of a sub-band under noise conditions [2,4]. Good SNR estimation is not simple and this problem can limit its application in a real world condition. In another method, entropy at the output of the sub-band recognizer (hybrid ANN/HMM), is computed and used as a measure of confidence to give weight to a recognizer output [6]. This method is very complicated for continuous density hidden Markov model (CDHMM), because entropy computation of CDHMM is very difficult [7].

We also propose a novel sub-band weighting based on wavelet thresholding to obtain a substitute for SNR, in determining reliability of sub-bands**.** We use Donoho's suggested wavelet thresholding technique in each sub-band to compute number of wavelet coefficients that are greater than threshold value. The ratio of this number to number of all coefficients in sub-band, can be used as a criterion for determining reliability of sub-band. We name this ratio as *WTH rate* and use it as $\alpha_{b,M}$ in equation (4).

We don't change wavelet coefficients in sub-bands by thresholding and only compute WTH rate. In this way, the kind of thresholding (soft or hard) is not important in computing of this ratio. But WTH rate can be affected by threshold determination method.

4. EXPERIMENTS AND RESULTS

Our experimental environment is as follows. The sampling rate of speech signal is 16 KHz. The test speech includes Persian numbers 1 to 10 which are recorded in a clean environment. There are 15 utterances for each number. Our recognition system is CDHMM with 6 state and 2 Gaussian mixtures per each state. We choose 3 type of additive noise: white, pink and Volvo noises from NOISEX92 database. We use 30 ms frames and 15 ms overlap and Hanning window in each sub-band. Our feature vector contains 12 MFCC coefficients and their first order derivative and logarithm of energy and its first order derivative. Hence, length of feature vector is 26. Under these conditions, results of full band noisy speech recognition are shown in Table 1.

	SNR=30	SNR=10	SNR=0
Pink noise	98.7%	36%	18%
Volvo Noise	99.3%	98%	86%
White Noise	97.3%	28.7%	14.7%

Table 1- Recognition rate for *full band* noisy speech under different kind of noise and noisy conditions

As indicated before, we use wavelet transforms for splitting input speech into 4 sub-bands in two ways: equal bandwidths for sub-bands (0-2 kHz, 2- 4 kHz, 4-6 kHz, 6-8 kHz) and dyadic bandwidths for sub-bands (0-1 kHz , 1-2 kHz, 2-4 kHz, 4-8 kHz). 5'Th order Daubechies wavelet is used as wavelet decomposition filter because of its smoothness and compact support. We use 3 methods for sub-band weighting: equal weighting (mean of likelihood scores as global likelihood score) and WTH rate (as indicated in section 3) and SNR of sub-bands as their weights.

Results of sub-band speech recognition under pink noise are shown in Tables 2, 3. Table 2 shows recognition results for dyadic bandwidths and Table 3 shows recognition results for equal bandwidths. In case of SNR and WTH weighting method ,as shown in these tables, dyadic bandwidths outperforms equal bandwidths. This result adapt to human ear nature. WTH weighting method shows good and acceptable performance in comparing to SNR weighting method. it may be have better performance with another wavelet thresholding techniques.

	SNR=30	SNR=10	SNR=0
Equal weighting	92%	29.3%	20%
SNR weighting	94%	62.7%	14%
WTH weighting	94.7%	53.3%	14.7%

Table 2- Recognition rate for noisy speech (*Pink noise)* -dyadic bandwidths

	SNR=30	SNR=10	SNR=0
Equal weighting	89.3%	44%	9.3%
SNR weighting	96%	52%	10%
WTH weighting	91.3%	46%	10.7%

Table 3- Recognition rate for noisy speech *(Pink noise)* - equal bandwidths

Tables 4,5 shows results of sub-band speech recognition under Volvo noise. Recognition results for dyadic bandwidths are shown in Table 4 and recognition results for equal bandwidths are shown in Table 5. Similar to pink noise, dyadic bandwidths have better performance than equal bandwidths. In this case, WTH weighting method has same performance of SNR weighting method.

Many researchers have used sub-band speech recognitions is presence of white noise [2,4] and reported results . Tables 6,7 shows results of sub-band speech recognition under white noise. Recognition results for dyadic bandwidths are shown in Table 6 and recognition results for equal bandwidths are shown in Table 7. As these tables show,

	SNR=30	SNR=10	SNR=0
Equal weighting	96.7%	94.7%	90.7%
SNR weighting	96.7%	96.7%	95.3%
WTH weighting	97.3%	96.7%	95.3%

Table 4- Recognition rate for noisy speech *(Volvo noise)* -dyadic bandwidths

	SNR=30	SNR=10	SNR=0
Equal weighting	96%	94.7%	88%
SNR weighting	96.7%	94%	92%
WTH weighting	97.3%	96%	92.7%

Table 5- Recognition rate for noisy speech *(Volvo noise)* -equal bandwidths

dyadic bandwidths have very better performance than equal bandwidths in case of SNR and WTH weighting. In our experiments, sub-band speech recognition, shows a good robustness in presence of white noise. WTH weighting method also shows an acceptable performance.

	SNR=30	SNR=10	SNR=0
Equal weighting	86%	20%	20%
SNR weighting	94.7%	80.7%	40%
WTH weighting	92%	62%	33.3%

Table 6- Recognition rate for noisy speech *(White noise)* -dyadic bandwidths

	SNR=30	SNR=10	SNR=0
Equal weighting	87.3%	35.3%	16%
SNR weighting	96.7%	59.3%	16%
WTH weighting	90%	52%	15.3%

Table 7- Recognition rate for noisy speech *(White noise)* -equal bandwidths

Fig. 4 displays results of sub-band recombination methods for different kind of noise, where SNR is 10 db and dyadic bandwidths is used. It can be seen that WTH weighting method has a performance near to SNR weighting method. Fig. 5 displays effect of dyadic bandwidths and equal bandwidths in sub-band speech recognition, where noise is pink and WTH weighting method is used.

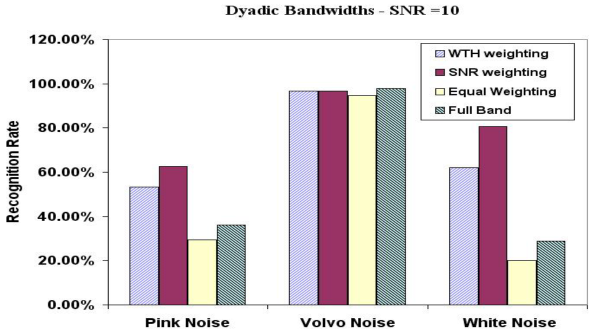

Figure 4. Comparison of full band and sub-band recombination for different kind of noise

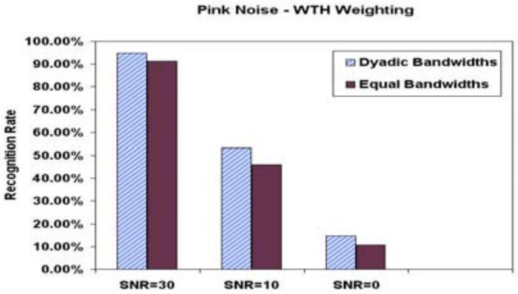

Figure 5. Comparison of equal bandwidths and dyadic bandwidths for Pink noise and WTH weighting

5. CONCLUSION

We used discrete wavelet transform and its multi-resolution property for robust sub-band speech recognition to split input speech into four sub-bands. Our recognition results under different kind of noise and noisy conditions, show that choosing dyadic bandwidths have better performance than choosing equal bandwidths in sub-band recombination. This result adapts to way which human ear recognizes speech and shows a useful benefit of dyadic nature of multi-level wavelet transform for sub-band speech recognition. We also defined a new weighting factor based on wavelet thresholding: WTH rate. In most of cases, it has acceptable performance in comparing to SNR weighting factor. It can be improved by choosing a more proper threshold value. In this way, we can obtain a reliable substitute for SNR in determining reliability of sub-bands.

6. REFERENCES

[1] Allen, J.B, "How do human process and recognize speech?" IEEE Trans. on Speech and Audio Processing, vol. 2, no. 4, pp.567-577, 1994.

[2] Okawa, S., Boochieri, E., Potamianos, A. "Multi-band speech recognition in noisy environment", ICASSP, vol. 2, pp. 641-644, 1998.

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

[4] Tibrewala, S., Hermansky, H. "Sub-band based recognition of noisy speech", ICASSP, pp.1255-1258, 1997.

[5] Berthommier, F., Glotin, H., Tessier, E. Bourlard, H., "Interfacing of CASA and partial recognition based on a multi-stream technique", ICSLP'98, Australia.

[6] Misra, H., Morris, A. "Confusion matrix based entropy correction in multi-stream combination", Eurospeech 2003, pp. 1817-1820, 2003.

[7] Ephraim,Y. Merhav, N. "Hidden Markov process" ,IEEE Trans. on Information theory, vol. 48, no. 6, June 2002.

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

[9] Mallat, S.G. " A theory for multi-resolution signal decomposition: the wavelet representation", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 11, no. 7, July 1989

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