AN EFFICIENT FRONT-END FOR AUTOMATIC SPEECH RECOGNITION

S.M. Ahadi[1,3], *H. Sheikhzadeh*[2], *R.L. Brennan*[2] *& G.H. Freeman*[3]

[1]EE Dept., Amirkabir University of Technology, Tehran, Iran
[2]Dspfactory Ltd., Waterloo, Ont., Canada
[3]E&CE Dept., University of Waterloo, Waterloo, Ont., Canada

Emails: smahadi@ece.uwaterloo.ca, {hsheikh, rob.brennan}@dspfactory.com, freeman@pce.uwaterloo.ca

ABSTRACT

This paper deals with the introduction of an efficient speech front-end for automatic speech recognition. This front-end not only performs well, in comparison to the traditional and widely used MFCC, but is also efficiently implemented in a low-resource system. Furthermore, due to its desirable characteristics that allow near-perfect reconstruction of the speech signal, this front-end can directly be used for speech enhancement purposes before the recognition is carried out. Experimental results show that the new front-end is capable of speech recognition with comparable or superior results to MFCC both in clean and noisy conditions. Similar results were also obtained in sub-band speech recognition experiments.

1. INTRODUCTION

Front-end signal processing plays a crucial role in the realization of speech recognition systems. This is the result of the fact that better signal feature extraction leads to better recognition performance. Hence, intensive efforts have been carried out to achieve a high performance front-end. These efforts have led to the introduction of highly successful speech features including: Linear Predictive Cepstral Coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Predictive (PLP) coefficients [1, 2]. Some of these techniques have been used extensively in speech recognition experiments and are the standard choice in many applications. MFCC feature extraction, for example, has been introduced as part of the ETSI standard front-end [3].

In this paper, we will discuss the implementation of a Weighted OverLap Add (WOLA) approach [4] for speech feature extraction. This approach has been found very useful in different speech processing applications, including speech enhancement, for its flexibility, high performance, low cost and complexity [5, 6]. Due to its special characteristics, this approach leads to very low cost implementations of speech processing algorithms that are suitable for computation sensitive applications such as embedded real-time implementations.

In this paper, we will show that these desirable features can also be used advantageously in a speech recognition system. Speech enhancement as a first step, for example, is known to lead to better recognition performance. However, many speech enhancement systems use different speech features, compared to speech recognizers. A unified feature set, leads to considerable savings in computations and a higher performance speech recognizer.

Another approach could be the integrated use of a speech enhancement system, e.g. in a hearing aid or a similar system, and a speech recognizer, which can simultaneously transcribe the enhanced speech signal for real-time or future use. The simplified block diagram of such a system is shown in Figure 1.

The feature extraction section of the speech recognition system, in this approach, uses a linearly spaced set of power complementary filters, which are later grouped to form a mel-warped filter bank. The cepstral parameters are derived using this group of filter outputs. The filters are formed and distributed in a way to constitute a flat energy output. The recognition performance of the system is evaluated under clean and noisy conditions and using different sub-band based speech recognition configurations. The implemented system has either outperformed or come very close to similar MFCC-based system under different test conditions.

2. WOLA FILTERBANK ANALYSIS

The WOLA is a highly efficient implementation of an over-sampled Generalized DFT (GDFT) filterbank, offering a low-delay, computationally cost effective, perfect/near-perfect reconstruction system [4, 5]. A simplified block diagram of the WOLA analysis stage is shown in Figure 2. Here, the input signal is shifted R samples at a time into the input buffer and analyzed. The buffer length (L) is the analysis frame size, while the

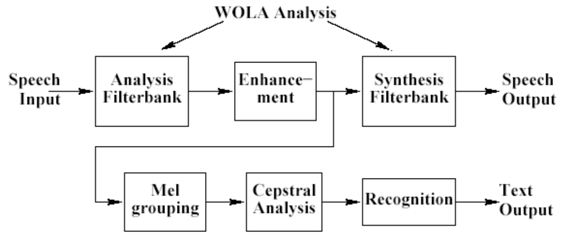

Figure 1. Example of integrated implementation of an enhancement system and a recognizer, using the same speech front-end.

FFT size is N, which could be different from the frame size. Hence, according to the L/N ratio, an order of saving in FFT computations can be obtained. For every R new input samples (a block), there are N/2 unique complex sub-bands, the other N/2 are complex conjugate. Since each complex band requires two numbers, an over-sampling factor of OS=N/R is achieved. High uniform oversampling ratios of 2 and more are often used to simplify the analysis filter design to achieve low processing delay, low aliasing and low reconstruction errors. Efficient WOLA synthesis is similarly implemented [4, 5].

3. RECOGNIZER IMPLEMENTATION

The speech recognition system was implemented using an isolated word recognition task. This was realized using TI-20 section of the TI 46 words speech corpus [7]. For all speech coding and recognition purposes, except for the WOLA front-end coding, the HTK speech recognition toolkit was used [8].

The data was originally recorded using 12500 Hz sampling frequency. Sixteen speakers (8 male and 8 female) uttered 20 English words 26 times. Ten of these utterances per speaker per word were designated as training material and the rest as test material. The corpus was designed to be used as a speaker dependent set. However, in order to increase the speech diversity for our tests, we have used it as a multi-speaker system, i.e. all the training material per word were used to build a single model for that word.

3.1. Basic Recognition System

The recognition system was built as an isolated word recognizer using 7 states per model (9 states including the two non-emitting states used in HTK). The MFCC coding, for comparison purposes, was carried out using HTK HCopy tool. Here, initially, the speech signal was pre-emphasized with a coefficient of 0.97. A frame size of 25 msec. with a frame shift of 10 msec. was used and the hamming window applied. The filter bank, in this case, consisted of 24 triangular-shaped half-overlapped filters spaced linearly over the mel frequency scale. 12 cepstral parameters were calculated from the filter bank outputs and weighted. The cepstral parameters for each utterance were mean normalized over the whole utterance. The normalized log energy was also appended to form a 13-component vector, whose size was later increased to 39 by appending the dynamic parameters.

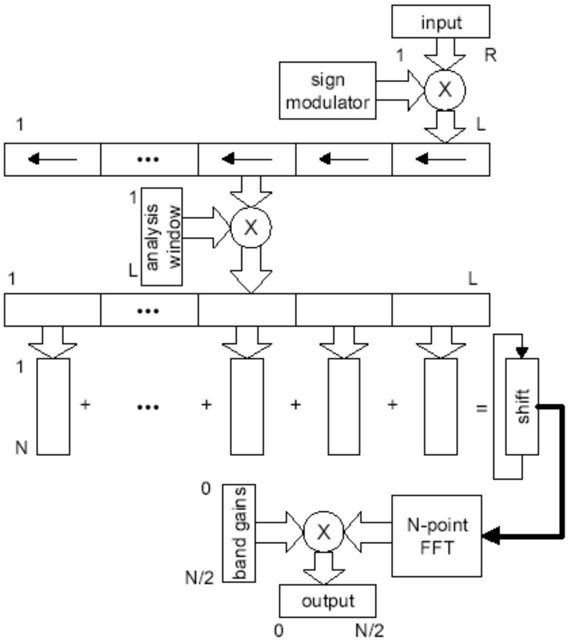

Figure 2. WOLA analysis block diagram.

The WOLA-based front-end was designed in a way to have the most possible similarities to the MFCC front-end. However, this is different to the approach taken in [9], where a WOLA front-end was used to exactly replicate a similar MFCC, to benefit from the computation efficiency of WOLA on a low resource platform. Here, similar to the MFCC front-end, the process started with a pre-emphasis step. The efficient implementation of WOLA, however, limited the L, N and R parameters to be powers of 2. In order to have specifications comparable to those of the MFCC, the values of L and R were chosen to be 256 and 128 respectively. This led to a frame size of 20.48 msec. and a frame shift of 10.24 msec. Special time-domain windows were designed so that the overall energy response of WOLA filters would be flat. The frequency responses of a few such filters are shown in Figure 3.

To obtain 24 mel-spaced filters, 128 linearly spaced WOLA filters were combined together in groups of 2 to 18. This resulted in a set of 24, almost similar, energy complementary filters distributed evenly on the mel scale. The DCT was then applied to the outputs of these filters to extract 12 cepstral coefficients per frame. A Juang lifter was then applied to these parameters and cepstral mean normalization carried out. Furthermore, the normalized raw log energy was appended to this basic cepstral vector. Later, the delta and acceleration parameters were added to make up a 39-element vector per frame.

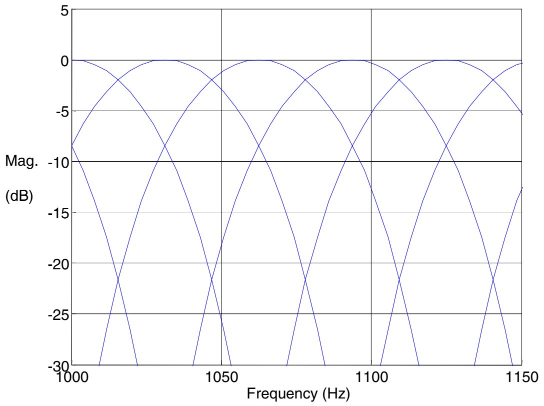

Figure 3. Frequency responses of WOLA analysis filters.

The Training phase, in both cases, was started by applying the HTK tool HInit to a predefined simple prototype model. This tool uses all the available training data and utilizing the Viterbi alignment repeatedly, tries to provide initial estimates of HMM parameters. Later, the HRest tool was used to provide more accurate parameter estimates using the Baum-Welch algorithm. The recognition and results analysis phases were carried out later using the appropriate tools.

For noisy speech recognition, the NATO RSG-10 noise data [10] were used to contaminate the TI-20 clean test data. These were first downsampled from the original 19.98 KHz to 12.5 KHz and then added to the clean data at various SNR levels. White, pink and babble noises were used in these experiments.

Table 1 displays the recognition results for clean and noisy speech using both the MFCC and WOLA front-ends. The results are percent recognition rate over all available test utterances. As depicted, although the results for the MFCC and WOLA front-ends are very close, especially in the clean speech case, in most of the noisy speech cases, the WOLA front-end slightly outperforms the widely used MFCC front-end. The main reason for the WOLA's better performance, especially in noisy speech cases can be attributed to its better representation of the speech signal due to its energy complementary filters.

3.2. Sub-band Speech Recognition

In recent years, sub-band speech recognition is found useful in dealing with speech signals contaminated by band-limited noise [11, 12]. We decided to evaluate the performance of WOLA-based front-end in sub-band ASR. In order to perform this evaluation, the results for sub-band ASR were needed from both MFCC-based and WOLA-based recognizers. As a more straightforward implementation of sub-band ASR, the Full-reCombination (FC, also called feature concatenation) approach was taken [13, 14]. Furthermore, this approach has been found to perform better in certain conditions [13].

Table 1. MFCC and WOLA front-end performances for clean and noisy speech recognition under different noises and SNRs. Figures indicate % recognition rates.

Noise	SNR	MFCC	WOLA
Clean		98.86	99.29
White	15 dB	57.10	65.17
	5 dB	31.92	28.77
Pink	15 dB	60.53	75.27
	5 dB	28.61	29.50
Babble	15 dB	78.28	76.13
	5 dB	45.26	50.87

For the MFCC case, modifications to the HTK source were necessary. This was done by modifying the HCopy tool and some of its library files. The modifications consisted of equally dividing the mel scale to several sub-bands, deriving the cepstral parameters from these sub-band groups and constructing the final basic cepstral vector by concatenating these individual cepstral sub-vectors. The energy, delta and acceleration parameters were calculated and appended to this vector as before.

Use of 24 mel filters and 12 cepstral parameters, that are multiples of 2, 3, 4 and 6, made this task simpler for these sub-band counts. The liftering process was slightly modified to use different liftering coefficients for different number of sub-bands. Similar modifications were carried out in the WOLA front-end to obtain WOLA-based sub-band parameters.

The results of sub-band MFCC and WOLA analyses with different numbers of sub-bands for clean and contaminated speech with different noises are shown in Table 2. The results indicate that the WOLA is performing better than the MFCC method in most of the situations, especially for the cases of sub-band noisy speech recognition. This can once again be attributed to the possibly better representation of the speech signal dynamics by the WOLA filters.

4. CONCLUSIONS AND FUTURE WORK

An efficient filter bank technique, WOLA, is used as the front-end for a speech recognition system. This technique has shown a good performance in speech recognition in comparison to the well-known and widely used MFCC. Furthermore, there are more benefits in using WOLA in a speech recognition system:

- As already shown, WOLA can be very efficiently implemented [5] and, due to its computation and resource usage efficiencies, it is a desirable candidate for implementation on low-resource systems.

Table 2. Comparison of MFCC-based and WOLA-based frontend performances in sub-band speech recognition under clean and noisy conditions. Figures indicate % recognition rates.

Sub-bands		2		3		4		6	
Noise	SNR	MFCC	WOLA	MFCC	WOLA	MFCC	WOLA	MFCC	WOLA
Clean		99.23	98.74	99.35	98.74	99.11	99.00	98.23	97.58
White	15 dB	52.26	68.67	60.29	73.44	48.09	63.93	58.13	62.32
	5 dB	28.71	33.49	34.65	50.79	29.93	34.65	38.00	29.79
Pink	15 dB	55.65	77.13	60.53	82.17	52.70	72.75	57.36	72.04
	5 dB	23.30	37.50	28.45	55.86	25.32	38.25	29.61	31.54
Babble	15 dB	73.46	85.87	83.06	90.85	70.80	87.15	69.11	81.50
	5 dB	39.57	57.26	53.86	70.37	39.65	51.18	45.28	54.55

- The proposed front-end easily allows speech enhancement algorithms to be directly applied before the recognition process. Specifically, multi-microphone processing schemes such as sub-band adaptive filters (SAF), beamforming, and echo cancellation have been efficiently implemented in sub-band domain and on the oversampled filterbanks [6]. The proposed front-end allows the integration, in frequency-domain, of such sub-band processing schemes with the speech recognition system.
- Due to its perfect to near-perfect reconstruction property, the proposed front-end allows integrated implementations of speech enhancement (with time-domain synthesized output), and recognition systems, using the same front-end.

The above advantages can play an important role in the realization of speech recognizers. The first one can be considered as an important step toward the realization of speech recognition in portable and low resource systems. The second and third can help in providing cleaner signal for recognition in real environments, without imposing significant computation cost through the implementation of different feature extraction algorithms appropriate for enhancement and recognition.

Our future work consists of joint implementation of an enhancement and a recognition system using the WOLA front-end described above. The ETSI Aurora 2 database will be used for this purpose.

5. REFERENCES

[1] S.B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”, *IEEE Trans. Acoust., Speech, Sig. Proc.*, Vol.28, pp. 357-366, Aug. 1980.

[2] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech”, *J. Acoust. Soc. Amer.*, pp. 1738-1752, 1990.

[3] “ETSI ES 201 108 v1.1.2 Distributed speech recognition; front–end feature extraction algorithm; compression Algorithm”, April 2000.

[4] R.E. Crochiere and L. R. Rabiner, *Multirate digital signal processing*, Prentice-Hall, 1983.

[5] R. Brennan and T. Schneider, “A Flexible Filterbank Structure for Extensive Signal Manipulations in Digital Hearing Aids”, *Proc. IEEE Int. Symp. Circuits and Systems*, pp.569-572, 1998.

[6] K. Tam, H. Sheikhzadeh and T. Schneider, “Highly oversampled subband adaptive filters for noise cancellation on a low-resource DSP system”, *Proc. Int. Conf. Spoken Lang. Proc.,* 2002.

[7] TI 46-word speech database, available from Linguistic Data Consortium, University of Pennsylvania, Catalog No. LDC93S9.

[8] Hidden Markov model Toolkit ver. 3.2, available from http://htk.eng.cam.ac.uk.

[9] E. Cornu and H. Sheikhzadeh, “A low-resource miniature implementation of the ETSI distributed speech recognition front-end”, *Proc. Int. Conf. Spoken Lang. Proc.,* 2002.

[10] NATO RSG-10 Noise data, available from http://spib.rice.edu/spib/select_noise.html.

[11] H. Hermansky, M. Pavel and S. Tibrewala, “Towards ASR using partially corrupted speech”, In *Proc. Intl. Conf. Spoken Lang. Proc.*, Philadelphia, 1996.

[12] H. Bourlard and S. Dupont: “A new ASR approach based on independent processing and re-combination of partial frequency bands,” In *Proc. Intl. Conf. Spoken Lang. Proc.,* Philadelphia, October 1996.

[13] S. Okawa, E. Bocchieri, and A. Potamianos, “Multi-band speech recognition in noisy environments,” in *Proc. ICASSP*, vol. II, May 1998, pp. 641–644.

[14] B. Mak, “A mathematical relationship between full-band and multiband mel-frequency cepstral coefficients”, *IEEE signal Proc. Letters*, Vol.9, No.8, August 2002, pp. 241-244.