A Weighted Overlap Add-based Front-end for Speech Recognition

S.M. Ahadi$^{1,3}$  H. Sheikhzadeh$^{2,3}$  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

Abstract—Speech signal enhancement is frequently referred to as a preprocessing step to speech recognition. However, in practice, this cannot be easily accomplished since the front-end signal processing techniques and/or parameters used in these two frequently differ. We apply a signal processing technique successfully used in speech enhancement to speech recognition and show that it can perform equally well compared to well-known speech recognition front-ends such as MFCC. The technique, oversampled filterbank analysis/synthesis through weighted overlap add (WOLA), has been tested and performed satisfactorily on the TI-46 and Aurora tasks in both clean and noisy conditions and also in sub-band speech recognition. The results indicate the capability of this technique in reducing the front-end signal processing blocks of enhancement and recognition into a single block.

1. Introduction

Since the introduction of early speech processing systems, feature extraction has been paid significant attention as a main block in such systems. Several different feature sets extracted from the speech signal have been used in speech processing applications. In speech recognition, in particular, the evolution of speech signal processing has led to some highly successful speech features leading to high performance speech recognition systems. Among these, Linear Predictive Cepstral Coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Predictive (PLP) coefficients [1, 2] have been extensively used in speech recognition experiments. The high performance of speech recognition systems based on some of these features in different environmental conditions has been the main reason for using them as the standard choice in speech recognition. One such feature extraction technique, for example, is the MFCC which has been introduced as part of the ETSI standard front-end [3].

Speech enhancement consists of the set of techniques that try to improve speech quality. The idea of using a speech enhancement block before performing recognition has long been discussed as being of obvious help in improving the performance of a speech recognition system. Meanwhile, researchers in the two fields have used either different speech features or different parameters in feature extraction to achieve the best possible performance in their respective systems. Therefore, it has usually been found difficult to use a certain feature set for both enhancement and recognition simultaneously as, in each case, different features or parameters lead to better performance.

In this paper, we discuss the use of an oversampled filterbank analysis efficiently implemented by a Weighted OverLap Add (WOLA) approach [4] for speech feature extraction in a speech recognition system. This approach has been found very useful in different speech processing applications, including speech enhancement, because of its flexibility, high performance, low cost and low complexity [5, 6]. Due to its special characteristics, this approach leads to very low cost implementations of speech processing algorithms that are suitable for computationally sensitive applications such as embedded real-time systems. We show that these desirable features can also be used advantageously in a speech recognition system. Therefore, this new approach is believed to be able to pave the way to the application of a speech enhancement technique before performing recognition in a speech recognizer. Such a unified feature set leads to considerable savings in computations and a higher performance of speech recognizer.

A further application of such 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. Figure 1 displays a simplified block diagram of such a
system. 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 so as to jointly 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 the performance of a similar MFCC-based system under different test conditions.

Fig. 1: Example of a unified speech front-end used in integrated implementation of an enhancement system and a recognizer.

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 (and window) size, while the FFT size is N, which could be different from the frame size. Hence, according to the L/N ratio, an order of savings in FFT computations can be obtained.

For every R new real input samples (a block), there are N/2 unique complex sub-bands, the other N/2 are their complex conjugates due to the Hermitian symmetry. Since each complex band requires two real numbers, an over-sampling factor of OS=N/R is achieved. High uniform over-sampling ratios of 2 and more are often used to simplify the analysis prototype filter (analysis window) 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 initially implemented using an isolated word recognition task. This was realized using only the 20-word section of the TI 46 words speech corpus, i.e., the 26-alphabet section was not used [7]. However, in the rest of this paper, this 20-word section will still be referred to as TI-46. For all speech coding and recognition purposes, except for the WOLA front-end coding, the HTK speech recognition toolkit was used [8].

The data in this corpus was originally recorded using a 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 had originally been 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 was used to build a single model for that word and tested using all the available test data.

Another speech recognition system was also implemented based on the ETSI Aurora 2 noisy speech recognition task [9]. This task is based on the TIDIGIT speaker-independent connected digit database. Only the speech from the adult speakers of TIDIGIT is used. The data is down-sampled to 8 kHz to extract the spectrum between 0 and 4 kHz. Additive and convolutional distortions are artificially added to this clean data at different SNRs. The noise data, however, were collected in real environments.

Fig. 2: WOLA analysis block diagram.

3.1. Basic Recognition Systems

3.1.1. TI-46 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 the HTK). The MFCC coding, for comparison purposes, was carried out using the HTK HCopy tool. Here, initially, the speech signal was pre-emphasized with a coefficient of
A frame size of 25 msec. with a frame shift of 10 msec. was used and the hamming window was 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 were 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.

The WOLA-based front-end was designed in a way to have the closest similarities to the MFCC front-end. However, to benefit from the computation efficiency of WOLA on a low resource platform, this was differently done from the approach taken in [10], where a WOLA front-end was used to exactly replicate a similar MFCC. 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 filterbank analysis filters would be flat. The frequency responses of a few such filters are shown in Figure 3.

![Figure 3: Frequency responses of WOLA analysis filters.](image)

To obtain 24 mel-spaced filters, 128 linearly spaced uniform filterbank analysis 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 discrete cosine transform 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.

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 performance analysis phases were carried out later using the appropriate tools.

For noisy speech recognition, the NATO RSG-10 noise data [11] were used to contaminate the TI-46 clean test data. These were first down-sampled 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 possibly be attributed to its better representation of the speech signal due to its power complementary analysis filters.

<table>
<thead>
<tr>
<th>Noise</th>
<th>SNR</th>
<th>MFCC</th>
<th>WOLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>98.86</td>
<td>99.29</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>15 dB</td>
<td>57.1</td>
<td>65.17</td>
</tr>
<tr>
<td></td>
<td>5 dB</td>
<td>31.92</td>
<td>28.77</td>
</tr>
<tr>
<td>Pink</td>
<td>15 dB</td>
<td>60.53</td>
<td>75.27</td>
</tr>
<tr>
<td></td>
<td>5 dB</td>
<td>28.61</td>
<td>29.5</td>
</tr>
<tr>
<td>Babble</td>
<td>15 dB</td>
<td>78.28</td>
<td>76.13</td>
</tr>
<tr>
<td></td>
<td>5 dB</td>
<td>45.26</td>
<td>50.87</td>
</tr>
</tbody>
</table>

### 3.1.2. Aurora 2 System

The Aurora 2 task includes a set of shell scripts to enable the users to build standard recognizers based on the Aurora speech data and test them on its test sets. Therefore most of the recognition parameters are already defined. The word models, except those provided for silence and space, have 16 states with left-right topology, 3 Gaussians per mixture and diagonal covariance matrices. The recognizers are based on the HTK.
The feature extraction, in our implementation, was different from what was done in the TI-46 case. The changes were made to make our results directly comparable to those of the original Aurora 2 implementation. Major differences were as follows:

- No energy normalization was carried out.
- A different sampling frequency (8 KHz) led to different frame size and shift and different WOLA parameters in some cases.
- 23 mel filters were used in place of 24. This did not make much difference in filter allocations though, as in Aurora 2, a lower cut-off frequency of 64 Hz was also set.
- No cepstral mean normalization was carried out.

Use of L=256 in this case resulted in a frame size of 32 msec. with a frame shift of 10 msec. which was obtained by setting R to 16 and using one in every 5 frames for speech recognition. Triangular binning of the WOLA filters was carried out in this case in place of the previous rectangular binning. The rest of the feature extraction was performed in a manner similar to the TI-46 case.

For a similar reason, the training phase was also different and the same procedure set up in Aurora 2 was followed. In this case, the initialization was carried out by finding the global mean and variance parameters using all the available training data. Then, several stages of embedded re-estimation followed by mixture incrementing at each stage were carried out until the desired number of mixture components was obtained. A few further re-estimation steps led to the final models. The models were only trained using the clean speech data.

The Aurora tests were carried out only on test set A of Aurora 2 which consists of speech data contaminated with four types of additive noise at different SNRs. The results of the tests are reported in Table 2 for both MFCC and WOLA-based front-ends. Note that the average reported values are calculated over SNRs of 20dB to 0dB and the clean and -5dB results have not been used in averaging, as is usually done in reporting Aurora 2 results. The MFCC and WOLA results, in this case, are very close and the slight differences cannot be interpreted as an indication of higher performance of one technique. However, they indicate that WOLA can perform equally well, compared to MFCC, in different conditions.

### 3.2 Sub-band Speech Recognition

Sub-band speech recognition has been found useful in certain conditions, where the speech signal is contaminated by band-limited noise [12, 13]. The idea is to divide the signal into several sub-bands and try to either perform (usually part of) the recognition process on them individually or extract separate features from them and perform recognition on a set of concatenated features. Therefore, if the contamination is band-limited, it will not spread through the whole recognition process and will be contained in the originally contaminated sub-band(s).

We evaluated the performance of WOLA-based front-end in sub-band ASR. To do so, 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 [14, 15]. Furthermore, this approach has been found to perform better in certain conditions [14].

The experiments were carried out on TI-46 corpus. 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. 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 numbers of sub-bands. The energy, delta and acceleration parameters were calculated and appended to this vector as before. Similar modifications were carried out in the WOLA front-end to obtain WOLA-based sub-band parameters.

The recognition rates for sub-band MFCC and WOLA analyses for clean and contaminated speech with different noise types and with different numbers of sub-bands are compared in Figure 4. The results indicate that the WOLA is performing better than the MFCC 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 analysis process.

### 4. Conclusions

An efficient filter bank technique, WOLA, is used as the front-end for a speech recognition system. The basics of the new approach and the recognizer implementation have been discussed. The system implementation on two well-known recognition tasks has been explained and the results for both the MFCC and the WOLA front-end reported and compared.

The high performance of the WOLA front-end was
demonstrated on at least two cases, where it performed usually better than MFCC, especially in noisy cases, in TI-46 plain system and sub-band-based system experiments. A possible reason for this higher performance can be the better representation of the speech signal in the WOLA front-end due to its power

**Tab. 2** MFCC and WOLA front-end performances for clean and noisy speech recognition under different noises and SNRs on Aurora 2 task. Figures indicate % recognition rates.

<table>
<thead>
<tr>
<th></th>
<th>Clean</th>
<th>SNR20</th>
<th>SNR15</th>
<th>SNR10</th>
<th>SNR5</th>
<th>SNR0</th>
<th>SNR-5</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MFCC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subway</td>
<td>98.83</td>
<td>96.96</td>
<td>92.91</td>
<td>78.72</td>
<td>53.39</td>
<td>27.3</td>
<td>12.62</td>
<td>69.86</td>
</tr>
<tr>
<td>Babble</td>
<td>98.97</td>
<td>89.96</td>
<td>73.43</td>
<td>49.06</td>
<td>27.03</td>
<td>11.73</td>
<td>4.96</td>
<td>50.24</td>
</tr>
<tr>
<td>Car</td>
<td>98.81</td>
<td>96.84</td>
<td>89.53</td>
<td>66.24</td>
<td>33.49</td>
<td>13.27</td>
<td>8.35</td>
<td>59.87</td>
</tr>
<tr>
<td>Exhibition</td>
<td>99.14</td>
<td>96.2</td>
<td>91.85</td>
<td>75.1</td>
<td>43.51</td>
<td>15.98</td>
<td>7.65</td>
<td>64.53</td>
</tr>
<tr>
<td><strong>WOLA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subway</td>
<td>99.41</td>
<td>97.36</td>
<td>92.67</td>
<td>76.54</td>
<td>57.77</td>
<td>24.34</td>
<td>10.85</td>
<td>69.74</td>
</tr>
<tr>
<td>Babble</td>
<td>99.39</td>
<td>90.49</td>
<td>76.38</td>
<td>47.55</td>
<td>29.14</td>
<td>15.34</td>
<td>5.21</td>
<td>51.78</td>
</tr>
<tr>
<td>Car</td>
<td>99.39</td>
<td>96.94</td>
<td>92.05</td>
<td>71.87</td>
<td>37.31</td>
<td>13.15</td>
<td>7.65</td>
<td>62.26</td>
</tr>
<tr>
<td>Exhibition</td>
<td>98.8</td>
<td>96.08</td>
<td>89.16</td>
<td>73.49</td>
<td>44.88</td>
<td>16.87</td>
<td>8.73</td>
<td>64.1</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>99.25</td>
<td>95.22</td>
<td>87.56</td>
<td>67.36</td>
<td>42.27</td>
<td>17.43</td>
<td>8.11</td>
<td><strong>61.97</strong></td>
</tr>
</tbody>
</table>

**Fig. 4:** MFCC and WOLA front-end performances for clean and noisy sub-band speech recognition under different noises on noisy TI-46 task.
complementary analysis filters. For Aurora experiments, however, the results only indicate very slight overall improvement for the WOLA front-end. A possible reason for this might be that the Aurora front-end has been carefully tuned with the MFCC parameters for highest performance. Therefore, its adjustment to WOLA analysis parameters may need further efforts.

Further to the improved or equal performance of a speech recognition system based on the WOLA filterbank front-end analysis, in comparison to the widely-used MFCC front-end, there exist more benefits in using WOLA filterbank in a speech recognition system:

- As already shown, WOLA can be very efficiently implemented [5] and, due to its computational and resource-usage efficiencies, it is a desirable candidate for implementation on low-resource systems.
- 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), beam-forming, and echo cancellation have been efficiently implemented in the sub-band domain and with over-sampled filterbanks [6]. The proposed front-end allows the integration, in the frequency-domain, of such sub-band processing schemes (not to be confused with the above-mentioned sub-band speech recognition) with the speech recognition system.
- Due to its perfect or 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 a cleaner signal for recognition in real environments, without imposing the significant computational cost needed for the implementation of different feature extraction algorithms for enhancement and recognition.

References