Robust Feature Extraction for Speech Recognition by Enhancing Auditory Spectrum

Md Jahangir Alam\(^1\), Patrick Kenny\(^2\), Douglas O’Shaughnessy\(^3\)

\(^1\)INRS-EMT, University of Quebec, Montreal, Quebec, Canada
\(^2\,\(^3\)CRIM, Montreal, Quebec, Canada

alam@emt.inrs.ca, Patrick.kenny@crim.ca, dougo@emt.inrs.ca

Abstract

The goal of this work is to improve the robustness of speech recognition systems under additive noise and reverberation. Various research has been done in the literature to improve the robustness of speech recognition systems under additive noise and reverberation. The methods to compensate for the effects of environmental mismatch can be implemented at the front-end (or feature extractor) or at the back-end or both. Robust feature extractors are usually obtained either by appending a pre-processing step, like speech enhancement [8, 9, 14], or by incorporating algorithms in an MFCC or PLP computation framework such as PNCC [3], frequency masking [9], or by adding a post-processing step, like feature normalization techniques [7, 11], (e.g., cepstral mean normalization (CMN)) or by combining any two or all of the above mentioned steps [3, 4]. Most of the front-ends use, in addition to other techniques for environmental mismatch compensation, a feature normalization technique, at the least CMN, as a post-processing scheme.

Additive noise reduction approaches usually have a tradeoff between the amount of noise reduction and speech distortion induced due to processing of a speech signal. At very low SNR the intensity of this induced distortion is high, thereby deteriorating the performance of the speech recognition systems. Compensation of reverberant noise is usually done by dereverberation, which can be obtained by inverse filtering the impulse response of the room [12, 13]. However, room impulse response is dependent on the distance between the speaker and the microphone, and the conditions of the room. Therefore, extracting a common set of robust features, which can perform well at low SNRs and also can handle various room impulses, is a difficult and challenging task.

To deal with additive noise distortion various speech enhancement methods have been proposed in [8, 9, 14]. In [10, 12, 13] several approaches have been proposed for handling convolutional noise distortions. The ETSI-AFE, described in [4], uses a two-stage Wiener filter and blind equalization technique, which is based on the comparison to a flat spectrum and the application of the LMS algorithm, for improving robustness of ASR systems against additive noise distortions and channel effects. The PNCC technique, proposed in [3], includes the use of a gammatone filter-bank (GTFB) and power law nonlinearity in place of the Mel filter-bank and log nonlinearity, used in conventional MFCCs framework, a medium duration power bias subtraction technique, for noise reduction, based on the arithmetic mean (AM)-geometric mean (GM) ratio and cepstral mean normalization as a post-processing scheme for DC offset removal, for robust feature extraction.

In this work, for robust feature extraction, we propose to enhance the speech auditory spectrum using a weighting rule based on the subband a posteriori signal-to-noise ratio (SNR). In order to allow a realistic and controllable frequency-domain asymmetry and to model most of the level dependency observed in place of the Mel filter-bank and log nonlinearity, used in conventional MFCCs framework, a medium duration power bias subtraction technique, for noise reduction, based on the arithmetic mean (AM)-geometric mean (GM) ratio and cepstral mean normalization as a post-processing scheme for DC offset removal, for robust feature extraction.

1. Introduction

Speech intelligibility as well as the performance of speech recognition systems degrades in practical environments due to a variety of signal variabilities. Additive noise and reverberation are the important causes of signal variabilities. Additive noise from interfering noise sources and convolutive noise arising from acoustic environment mainly cause a reduction of speech recognition performance. The acoustic features most commonly used in speech recognition systems are Mel Frequency Cepstral Coefficients (MFCC) [1] and Perceptual Linear Prediction (PLP) [2] features. Both MFCC and PLP front-ends perform well in matched environments, where speech data are collected from reasonably clean environments. But their performance degrades when testing environment is different from the training environment. Degradation of performance due to mismatched environments (such as channel, handset, additive background noise and reverberant environments) has been a barrier for deployment of speech recognition technologies. It is necessary to address this problem to enable the deployment of recognition systems in real world conditions.
in basilar membrane (BM) filtering, the proposed method includes the use of a compressive gammachirp filter-bank (cGCFB) [16] for auditory spectral analysis. We use power function nonlinearity as it has been found in [3] that it is more robust than the logarithmic nonlinearity used in conventional MFCC framework. As a post-processing scheme, for the normalization of the features, we use short-time cepstral mean and scale normalization (STCMNSN) technique, proposed in [7]. Feature normalization is normally performed over the whole utterance with the assumption that the channel effect is constant over the entire utterance, such as CMN or CMV. Also, normalizing a feature vector over the entire utterance is not a feasible solution in real-time applications as it causes unnecessarily long processing delay. To relax this assumption and to reduce the processing delay, cepstral features in the proposed method are normalized over a sliding window of 1.5s duration. The feature vector to be normalized is located at the centre of the sliding window.

2. Proposed feature extractor

Fig. 1 (a) presents the complete diagram of the proposed feature extractor for robust speech recognition. In the proposed method processing of a speech signal begins with pre-processing (including DC removal and pre-emphasis, typically using a first-order high-pass filter). Short-time Fourier Transform (STFT) analysis is performed using a finite duration (25 ms) Hamming window to estimate the power spectrum of the signal. cGCFB integration is performed on both speech and noise power spectra for auditory spectral analysis

2.1. Computation of cGCFB weights

The compressive gammachirp filter bank (cGCFB), introduced by Irino et al in [16], is a generalization of the GTFB. The frequency response of a gammatone filter is very nearly symmetric, which is not a good match to the auditory data. The cGCFB allows a realistic and controllable frequency-domain asymmetry and has also been shown to model most of the level dependency observed in the BM filtering [16]. Fig. 1 (b) shows how cGCFB weights are computed from the GTFB weights. The frequency response of the cGCFB is given by:

\[ H_{\text{cGCFB}} = a_0 \left[ H_{\text{GT}}(f) \right] e^{i\theta_1} \cdot e^{i\theta_2} \]

where \( \theta_1 = \tan^{-1} \left( \frac{f-f_1}{hB(f_1)} \right) \), \( \theta_2 = \tan^{-1} \left( \frac{f-f_2}{hB(f_2)} \right) \), \( f_1 \), and \( f_2 \) are the center frequencies of the low-pass asymmetric function (LP-AF) \( e^{i\theta_1} \) and high-pass asymmetric function (HP-AF) \( e^{i\theta_2} \), respectively. \( a_0 \) is an amplitude normalization factor, \( f_1 \) is defined in term of a frequency ratio \( f_{\text{ratio}} \) as [16]

\[ f_1 = f_{\text{ratio}} \cdot f_{\text{ref}}, \quad \text{with} f_{\text{ref}} = f_1 + \frac{c_1 B(f_1)}{\mu} \]  

the peak frequency of the passive gammachirp filter. So, the frequency response of the cGCFB can be obtained from the GTFB by multiplying the frequency response \( H_{\text{GT}}(f) \) of the GT filter with HP- and LP-AFs. The high-pass asymmetric function makes the pass-band of the cGCFB more symmetric at lower levels. The number of filters used in this work is 64.

2.2. Estimation of noise power spectrum

For noise power spectrum estimation we use a soft speech presence probability (SPP)-based noise estimation approach, proposed in [17]. In this method, initial estimate of the noise power spectrum is computed by averaging the first ten frames of the speech spectrum. The advantage of this method is that it does not require a bias correction term as required by a MMSE-based noise spectrum estimation method; it also results in less overestimation of noise power and is computationally less expensive.
where $k$ is the subband index, $m$ is the frame index, $\tau$ is a parameter that controls the lower limit of the weighting function and $\vartheta(k,m)$ is the instantaneous SNR (in dB) defined as:

$$\vartheta(k,m) = \gamma_{sb}(k,m) - 4.5,$$

where

$$\gamma_{sb}(k,m) = \max \left\{ 10 \log_{10} \left( \frac{S_{mm}(k,m)}{N_{mm}(k,m)} \right) , -4.0 \right\}.$$  

$N_{mm}(k,m)$ is the noise power spectrum mapped onto the auditory frequency axis. Here, $\tau = 4.5$ is chosen experimentally. In order to remove the outliers from the weighting function due to noise variability we use a two-dimensional median filter and a two-dimensional moving average filter is used for smoothing the decision regions. In order to achieve optimal performance it is desirable to have a feature extractor that is well suited both for clean and adverse acoustic conditions. It is observed from Figs. 2 & 3 that the proposed weighting rule enhances the noisy auditory spectrum and does not introduce any distortion to the clean auditory spectrum.

### 2.4. Feature normalization using STCMSN

The static features, obtained after applying power function nonlinearity, using a coefficient of 0.07, and discrete cosine transform (DCT), are normalized using the STCMSN approach [7]. In STCMSN approach the $n$th feature space and $m$th frame $C(n,m)$ are normalized as

$$C_{st}(n,m) = \frac{C(n,m) - \mu_{st}(n,m)}{d_{st}(n,m)},$$

where $\mu_{st}(n,m)$ and $d_{st}(n,m)$ are the short-time mean and short-time difference between the upper and lower bound, respectively, defined, for a short-time window of $L=150$ frames, as:

$$
\mu_{st}(n,m) = \frac{1}{L} \sum_{j=(m-1)/2}^{(m+1)/2} C(n,j),
$$

$$
d_{st}(n,m) = \max_{(m-1)/2 \leq j \leq (m+1)/2} \left( C(n,j) - \min_{(m-1)/2 \leq j \leq (m+1)/2} C(n,j) \right).
$$

The main idea behind STCMSN (or CMSN) technique is that under mismatched conditions a difference of log spectrum between the training and test environments is removed by adjusting the short-time scale and short-time mean. The advantage of short-time feature normalization technique is that it relaxes the constant channel assumption, used in the full utterance-based feature normalization method, and reduces the unnecessary long processing delay.

### 3. Performance Evaluation

#### 3.1 Experimental setup

The AURORA-2 [5] and AURORA-5 [6] corpora are used for comparing the performances of the proposed feature extractor to the conventional MFCC, PLP (HTK version of PLP), ETSI-AFE [4], and PNCC [3] features, in the context of speech recognition. For the performance evaluation of our feature extractors, we have chosen mismatched conditions. Features extracted from the clean training data are used for training the recognizer. For testing we have used the following ten noise scenarios of the AURORA-2 corpus at six different SNRs (clean (SNR > 30 dB), 20 dB, 15 dB, 10 dB, 5 dB, 0 dB): in test set A - subway, babble, car, exhibition hall; in test set B - restaurant, street, airport and train-station and in test set C - MIRS (modified intermediate reference system) filtered subway and street noise.

In order to demonstrate the performance of the proposed feature extractor in real-time reverberant environments, experiments were conducted on a subset of the AURORA-5 corpus. The corpus comprises real time recording in a meeting room, recorded in a hands-free mode at the ICSI in Berkeley [6]. The database consists of 2388 utterances from 24 speakers with 7800 digits in total. The speech was captured with four different microphones, labeled as mic 6, mic 7, mic E and mic F, placed at the middle of the table in the meeting room. The recording contain only a small amount of additive noise, providing the typical effect of hands-free recording in the reverberant room. For our experiments, we use 13 static features (including the 0th cepstral coefficient) augmented with their delta and double delta coefficients, making 39-dimensional feature vectors. The analysis frame length is 25 ms with a frame shift of 10 ms. The Delta and double features were calculated using a 3-frame and 2-frame window, respectively. For all the feature extractors, the features, before appending delta and double delta features, are

![Figure 2: Auditory spectra of a clean signal, (a) before applying weighting rule, (b) after applying weighting rule.](image1)

![Figure 3: Auditory spectra of noisy speech signal degraded with train-station noise with SNR = 3 dB, (a) before applying weighting rule, (b) after applying weighting rule.](image2)
normalized using the feature normalization technique. For the recognition task we use the HTK speech recognizer. In the experiments we use a simple HMM-based system with 16 states per word model, 3 Gaussian components per state.

3.2 Results

We use the percentage of word accuracy as a performance evaluation measure for comparing the recognition performances of the feature extractors considered in this paper.

Table 1 presents the average word accuracy (in %), averaged over all noise scenarios of all the three test sets A, B, and C of the AURORA-2 corpus. In additive noise conditions the proposed method provides comparable results to that of the PNCC and ETSI-AFE features. Table 2 presents a comparison of the performances in terms of word accuracy (in %) of the proposed feature extractor with the other considered feature extractors in a real-time reverberant environment on the AURORA-5 corpus. In convolutive noise conditions, the proposed method provides consistently better word accuracy on all microphones of the AURORA-5 corpus than all other methods. From table 2 it is evident that the ETSI-AFE front-end has the lowest word accuracy compared to all other feature extractors. In reverberant environment the ETSI-AFE front-end is not as effective as its performance in additive background noise distortions, which is consistent with the studies of [10]. The proposed method performs well both in additive noise distortions and reverberation. Computationally, the proposed method is less expensive the PNCC and ETSI-AFE.

Table 1: Word accuracy (%) obtained using different front-ends on the AURORA-2 corpus

<table>
<thead>
<tr>
<th></th>
<th>Clean</th>
<th>20dB</th>
<th>15dB</th>
<th>10dB</th>
<th>5dB</th>
<th>0dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>99.03</td>
<td>96.76</td>
<td>93.31</td>
<td>85.22</td>
<td>66.10</td>
<td>33.01</td>
</tr>
<tr>
<td>PLP</td>
<td>99.07</td>
<td>96.68</td>
<td>93.41</td>
<td>85.57</td>
<td>67.48</td>
<td>36.32</td>
</tr>
<tr>
<td>PNCC</td>
<td>99.20</td>
<td>98.48</td>
<td>97.39</td>
<td>93.34</td>
<td>82.26</td>
<td>54.33</td>
</tr>
<tr>
<td>ETSI-AFE</td>
<td>99.23</td>
<td>98.06</td>
<td>96.56</td>
<td>93.02</td>
<td>83.32</td>
<td>60.10</td>
</tr>
<tr>
<td>Proposed</td>
<td>99.34</td>
<td>98.39</td>
<td>97.39</td>
<td>93.36</td>
<td>82.67</td>
<td>53.97</td>
</tr>
</tbody>
</table>

Table 2: Word accuracy (%) obtained using different front-ends on the AURORA-5 corpus

<table>
<thead>
<tr>
<th></th>
<th>Mic 6</th>
<th>Mic 7</th>
<th>Mic E</th>
<th>Mic F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>81.41</td>
<td>76.12</td>
<td>75.22</td>
<td>79.33</td>
</tr>
<tr>
<td>PLP</td>
<td>79.55</td>
<td>73.98</td>
<td>73.10</td>
<td>77.51</td>
</tr>
<tr>
<td>PNCC</td>
<td>88.12</td>
<td>83.40</td>
<td>80.20</td>
<td>84.51</td>
</tr>
<tr>
<td>ETSI-AFE</td>
<td>64.32</td>
<td>47.66</td>
<td>58.12</td>
<td>62.72</td>
</tr>
<tr>
<td>Proposed</td>
<td>89.16</td>
<td>85.27</td>
<td>82.82</td>
<td>86.21</td>
</tr>
</tbody>
</table>

4. Conclusions

In this paper we presented a robust feature extractor that includes the use of an asymmetric level dependent cGCFS, a simple way to enhance the auditory spectrum, a soft speech presence probability (SPP)-based noise spectrum method [17], and a short-time feature normalization method called STCMSN [7], for speech recognition. Experimental results on the AURORA-2 and AURORA-5 corpora show that the proposed method provides comparable results to that of the ETSI-AFE and PNCC front-ends under additive noise distortions and performed better under reverberant environments. For future study we would like to apply the proposed feature extractor on a large vocabulary continuous speech recognition (LVCSR) task.

5. References

[4] ETSI ES 202 050, Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; advanced front-end feature extraction algorithm; Compression algorithms; 2003.