Regularized MVDR Spectrum Estimation-based Robust Feature Extractors for Speech Recognition

Md Jahangir Alam¹,², Patrick Kenny², Douglas O'Shaughnessy ¹

¹ INRS-EMT, University of Quebec, Montreal (QC), Canada
² CRIM, Montreal (QC), Canada

{alam, dougo}@emt.inra.ca, patrick.kenny@crim.ca

Abstract

In this paper, we present two robust feature extractors that use a regularized minimum variance distortionless response (RMVDR) spectrum estimator instead of the discrete Fourier transform-based direct spectrum estimator, used in many front-ends including the conventional MFCC, for estimating the speech power spectrum. Direct spectrum estimators, e.g., single tapered periodogram, have high variance and they perform poorly under noisy and adverse conditions. RMVDR spectrum estimator has low spectral variance and are robust to mismatch conditions. Based on RMVDR spectrum estimator two robust feature extractors, robust RMVDR cepstral coefficients (RRMCC) and normalized RMVDR cepstral coefficients (NRMCC), are proposed that incorporate an auditory domain spectrum enhancement (ASE) method and a medium duration power bias subtraction (MDPBS) technique, respectively, for enhancement of the speech spectrum. Experimental speech recognition results are conducted on the AURORA-4 corpus and performances are compared with the MFCC, PLP, MVDR-MFCC, RMVDR-MFCC, PMVDR, ETSI advancement front-end (ETSI-AFE), PNCC, CFCC, and the robust feature extractor (RFE) of [6]. Experimental results demonstrate that the proposed robust feature extractors outperformed the other robust front-ends in terms of percentage word accuracy on the AURORA-4 large vocabulary continuous speech recognition (LVCSR) task under different mismatch conditions.

Index Terms: speech recognition, feature extractor, regularized MVDR, auditory spectrum enhancement, MDPBS

1. Introduction

Mel-frequency cepstral coefficients (MFCCs) [1], have proven to be one of the most effective features for speech and speaker recognition tasks, are frequently used as a low-dimensional set of features to represent short-time speech signal. MFCCs are usually computed by integrating a triangular-shaped Mel-scaled filterbank (MelFB) either to the DFT-based short-time spectrum or to the linear predictive coding (LPC)-based spectrum. MFCCs-based speech recognizers perform well under matched training/test conditions but the performance gap between automatic speech recognizers (ASRs) and human listeners in real world settings is significant [2, 3]. Different operating conditions during signal acquisition (e.g., channel response, handset type, additive background noise, reverberation, etc.) lead to feature mismatch across training and testing and thereby degrade the performance of the MFCCs (and PLP)-based speech recognition systems. To tackle this problem various robust feature extractors are employed in speech recognition tasks such as the ETSI advanced front-end (ETSI-AFE) [4], power normalized cepstral coefficients (PNCC) [5], and the robust feature extractors proposed in [6, 7, 8], etc. In MFCC and PLP front-ends, and in most of the robust feature extractors the features are computed from a windowed (e.g., Hamming) direct spectrum estimates (the squared magnitude of the Fourier transform of the observed signal) that has a high spectral variance. The variance of these features are greatly influenced by the variance of the spectral estimate of the observed speech signal. Variance in the feature vectors has a direct bearing to the separability of Gaussians modeling the speech classes. Reduction in variance of the feature vector increases class separability and improved class separability can potentially increase recognition accuracy and decrease search speed [9]. Although direct spectrum estimators are entirely independent of data and therefore do not suffer from problems arising from modeling deficiencies, these methods are not robust to noise and hence they perform poorly under mismatch training/test conditions.

Here, we present two robust feature extractors, dubbed as robust RMVDR cepstral coefficients (RRMCC) and normalized RMVDR cepstral coefficients (NRMCC), that include the use of a regularized minimum variance distortionless response (RMVDR) spectral estimator instead of the windowed direct spectrum estimator for the estimation of speech power spectrum. The advantages of RMVDR spectrum estimator are that the regularization parameter helps to penalize rapid changes in all-pole spectral envelopes thereby producing smooth spectra without affecting the formant positions [10, 11] and it provides robust spectral estimate under noisy environments [12, 13]. In order to enhance the estimated spectrum in the auditory domain, RRMCC front-end uses a sigmoid-shaped auditory spectrum enhancement (ASE) technique proposed in [6], whereas, in NRMCC, a medium duration power bias subtraction (MDPBS) technique [5] is utilized.

The MVDR spectral estimator has already been applied in speech recognition [9] and speaker identification [14] tasks. An extension of the MVDR method was proposed in [14, 19] by warping the frequency axis with the bilinear transformation prior to MVDR spectral estimation. The perceptually motivated MVDR (PMVDR) front-end, proposed in [15], directly performs warping on the DFT power spectrum and eliminates the auditory filterbank processing step.

To compare the performances of the RRMCC and NRMCC front-ends speech recognition experiments are performed on the AURORA-4 [16] LVCSR task. For comparison purposes following front-ends are used: MFCC, PLP [18], MVDR-MFCC, PMVDR [15], ETSI-AFE [4], power normalized cepstral coefficients (PNCC) [5], cochlear filterbank cepstral coefficients (CFCC) [17], and the robust feature extractor (RFE) proposed in [6]. Reported speech recognition results show that the presented front-ends outperform all other front-ends considered here for comparison of performances.
2. RMVDR-based front-ends

A combined block diagram for the RRMCC and NRMCC feature extractors (or front-ends) are shown in figure 1. RRMCC features are obtained when point 1 is connected to point 2 and when points 1 and 3 are connected then NRMCC features are obtained. Features of the both front-ends are computed from a RMVDR spectral estimates.

2.1. MVDR and RMVDR spectrum estimation

The Minimum Variance Distortionless Response (MVDR) spectrum estimator, introduced by Capon [21], is mostly used in array signal processing applications, and has also been investigated in relation to other applications such as speech modeling [22], robust speech recognition [9], and speaker recognition [14] systems. MVDR method defines a filter that leaves the signal undistorted at frequency of interest while suppressing the other frequencies in an optimal way. In MVDR method the $p$th order MVDR spectral estimate can be parametrically obtained from the LP (linear prediction) coefficients $a_q$ as:

$$S_{\text{MVDR}}(f) = \frac{1}{\sum_{k=p}^{\infty} \lambda(k)e^{-i2\pi f_k}},$$

where the parameter $\lambda(k)$ of the MVDR method can be directly obtained using a non-iterative computation from the LP coefficients $a_q$ as:

$$\lambda(k) = \left\{ \begin{array}{ll}
\frac{1}{\sigma_r^2} \sum_{k=0}^{\infty} (p+1-k-2q)a_q a_r^*, & \text{for } k \geq 0 \\
\lambda(-k), & \text{for } k < 0,
\end{array} \right.$$  \hspace{1cm} (2)

where $\sigma_r^2$ is the residual variance. In MVDR, the power is obtained by averaging several samples at the output of the optimum constrained filter. This averaging results in a reduction of the spectrum estimator variance [9, 14].

All-pole spectral envelope estimates based on LP for speech signals often exhibit unnaturally sharp peaks, especially for speakers with high pitch frequency [22, 27]. LP method fails to separate the short-term dependency from the long-term dependency and the resulting envelope is contaminated with harmonics. These peaky spectral envelopes can cause problems in speech modification [27]. Because of the inherent smoothing properties of MVDR spectrum estimator, MVDR spectral estimates obtained using the LP coefficients has less rapid variations than the LP spectral estimates (see fig. 2), but MVDR estimates are still affected as it is computed from the LP coefficients. In [10, 27], regularization is introduced to the objective function of the LP method to penalize rapid changes in the spectral envelope, which helps to improve the all-pole spectral envelope estimate. For a robust and improved spectral estimates, in this paper, we propose to compute the MVDR spectral estimates from the regularized LP (RLP) coefficients.

We denote this method as regularized MVDR (RMVDR) [12, 13]. Similar to the MVDR spectrum estimator, the $p$th order RMVDR spectral estimate can be parametrically written as:

$$S_{\text{RMVDR}}(f) = \frac{1}{\sum_{k=p}^{\infty} \mu(k)e^{-i2\pi f_k}},$$

where the parameter $\mu(k)$ of the regularized MVDR method can be obtained from a non-iterative computation using the regularized LP (RLP) coefficients $a_q^*$ and the prediction error variance $\sigma_r^2$ as:

$$\mu(k) = \left\{ \begin{array}{ll}
\frac{1}{\sigma_r^2} \sum_{k=0}^{\infty} (p+1-k-2q)a_q^* a_r^*, & \text{for } k \geq 0 \\
\mu(-k), & \text{for } k < 0.
\end{array} \right.$$  \hspace{1cm} (4)

In RLP method, the predictor coefficients $a_q^*$ are computed by adding a penalty measure $\psi(a')$, which is a function of the unknown predictor coefficients $a'$, to the objective function of the LP method and therefore minimizing that modified objective function of the following form [10, 11, 27]

$$\sum_n (s(n) + a_q^* s(n-q))^2 + \lambda \psi(a'),$$

where regularization constant $\lambda > 0$ controls the smoothness of the all-pole spectral envelope. RLP method penalizes the rapid changes in all-pole spectral envelope and therefore, produces a smooth spectral estimate keeping the formant positions unaffected. In [10, 27] the authors show that the penalty measure can be approximated as:

$$\psi(a') = a'^T D F D a',$$  \hspace{1cm} (6)

where $a'=[a_0, ..., a_p]^T$ are the predictor coefficients, $D$ is a diagonal matrix in which each diagonal element consists of the row number, $F$ is Toeplitz autocovariance matrix for a windowed autocorrelation sequence of a frame of the speech signal. The compact solution obtained from the composite cost function is [10, 27]:

$$a_n^* = a' = -(R + \lambda D F D)^{-1} r,$$  \hspace{1cm} (7)

where $R$ is the autocovariance matrix of the speech frame and $r$ is the vector of speech autocovariance values $[r(t), r(2), ..., r(p)]^T$.

Fig. 2 presents a comparison of the estimated spectra obtained by the various spectrum estimators described in this paper. It is observed from this fig. that compared to the periodogram, LP and the MVDR spectrum estimators, RMVDR method provides smooth spectral estimate and therefore, results in reduced spectral variance.

2.2. Spectrum enhancement

After estimating the speech spectra using the RMVDR spectrum estimator, noise spectra are estimated with the help of a soft speech presence probability (SPP)-based MMSE noise estimation approach, proposed in [20]. Mel filterbank (MelFB) integration is performed on both the speech and noise spectra for auditory spectral analysis. In order to enhance the MelFB-subband spectra we adopt two enhancement methods, presented in [6] and [5], and we denote the obtained features as RRMCC (robust RMVDR cepstral coefficients) and NRMCC (normalized RMVDR cepstral coefficients), respectively. Before taking the discrete cosine transform (DCT) we use root compression (i.e., power function nonlinearity) with a coefficient of 0.07 to the enhanced
auditory spectra of both of the front-ends as it is found to be more robust than the conventional log compression [5].

2.2.1 RRMCC front-end

In this case, the Mel filterbank auditory spectrum is passed through an auditory domain spectrum enhancement method, introduced in [6, 13], that utilizes a sigmoid-shaped weighting rule based on the subband a posteriori signal-to-noise ratio (SNR). Static cepstral features, obtained in this front-end, are normalized using a short-time cepstral mean and gain (or scale) normalization technique (STCMGN or STCMSN) [24, 6, 13] with a sliding window of 1.5 sec duration. Delta and double delta features are computed with a 5-frame window using the regression formula [26].

2.2.2. NRMCC front-end

In this front-end, the auditory spectra are first normalized by the 95th percentile power across all frames and channels [5]. Normalized auditory spectra are then processed using a medium duration (5 speech frames by taking 2 frames from the left and 2 from the right) power bias subtraction (MDPBS) that is based on maximization of the sharpness of the power distribution [5]. A power flooring is applied in MDPBS method to reduce the spectral distortion between the training and test data for the region where spectro-temporal segments representing speech that exhibit the smallest power are the most vulnerable to additive noise [5]. Cepstral mean subtraction (CMS) technique is used to normalize static features and delta features are computed with a 5-frame window using the regression formula [26].
represents the test set number defined in the AURORA-4 corpus.

For the continuous speech recognition task on the AURORA-4 corpus, all experiments employed state-tied crossword speaker-independent triphone acoustic models with 4 Gaussian mixtures per state. A single-pass Viterbi beam-search-based decoder was used along with a standard 5K lexicon and bigram language model with a prune width of 250 [16, 23].

For our experiments, we use 13 Mel-frequency cepstral coefficient (MFCC) features (including the 0th cepstral coefficient) augmented with their delta and double delta coefficients, making 39-dimensional MFCC 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 5-frame window. For all methods except the RRMCC, extracted features are normalized using the normalization technique over the whole utterance.

3.2. Results and discussion

Percentage word accuracy is used as a performance evaluation measure for comparing the recognition performances of the proposed front-ends to that of the baseline feature extractors. The optimal model order \( p \) for the MVDR and regularized MVDR methods is adjusted to allow for highest speech recognition accuracy on the development test set of the AURORA-4 corpus. Similar to [12, 13] in this work the optimal model order is set \( p = 100 \).

We chose the optimal value for the regularization constant \( \lambda \) of the RMVDR method that provided the highest word accuracy on the dev-test set of the same corpus. The optimal value for the regularization constant is set to \( \lambda_{opt} = 10^{-9} \) [12, 13]. Robustness of the RMVDR spectrum estimator was shown in [12] by presenting speech spectrograms of a noisy speech obtained by the various spectrum estimators. Fig. 3 presents the auditory spectra of a noisy speech signal, corrupted with the babble noise, obtained by the various feature extractors. It is observed from this figure that compared to the MFCC and RMVDR front-ends, both the RRMCC and NRMCC methods result in a reduction of the noise. Table 1 depicts the word accuracies obtained by the different features on the various test sets, as described in section 3.1, of the AURORA-4 LVCSR corpus. Both RRMCC and NRMCC feature extractors outperformed the other feature extractors, considered here for comparison, in terms of the recognition word accuracy. The average relative improvements obtained by the RRMCC in recognition word accuracy over the ETSI-AFE, PNCC and RFE of [6] are 6.0%, 5.6% and 7.1%, respectively and the NRMCC provided an average relative improvements of 9.1%, 8.8% and 10.2% over the ETSI-AFE, PNCC and RFE of [6], respectively. An average relative improvement of 3.3% is obtained by NRMCC over the RRMCC. NRMCC front-end performed the best in terms of the recognition accuracy.

<table>
<thead>
<tr>
<th>Word Accuracy (%)</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>90.02</td>
<td>49.19</td>
<td>71.12</td>
<td>35.44</td>
<td>61.44</td>
</tr>
<tr>
<td>PLP(HTK)</td>
<td>89.72</td>
<td>50.41</td>
<td>74.44</td>
<td>39.64</td>
<td>63.55</td>
</tr>
<tr>
<td>MVDR</td>
<td>89.47</td>
<td>52.10</td>
<td>74.51</td>
<td>39.60</td>
<td>63.92</td>
</tr>
<tr>
<td>PMVDR</td>
<td>88.69</td>
<td>52.69</td>
<td>76.65</td>
<td>41.08</td>
<td>64.78</td>
</tr>
<tr>
<td>RMVDR</td>
<td>90.06</td>
<td>54.25</td>
<td>78.23</td>
<td>40.63</td>
<td>65.79</td>
</tr>
<tr>
<td>CFCC</td>
<td>86.34</td>
<td>63.05</td>
<td>78.00</td>
<td>54.70</td>
<td>70.67</td>
</tr>
<tr>
<td>ETSI-AFE</td>
<td>88.59</td>
<td>69.58</td>
<td>79.52</td>
<td>61.51</td>
<td>74.80</td>
</tr>
<tr>
<td>PNCC</td>
<td>88.64</td>
<td>69.85</td>
<td>81.07</td>
<td>60.00</td>
<td>74.89</td>
</tr>
<tr>
<td>RFE [4]</td>
<td>88.90</td>
<td>68.87</td>
<td>80.94</td>
<td>59.25</td>
<td>74.49</td>
</tr>
<tr>
<td>RRMCC</td>
<td>89.29</td>
<td>71.20</td>
<td>81.88</td>
<td>62.88</td>
<td>76.31</td>
</tr>
<tr>
<td>NRMCC</td>
<td>88.62</td>
<td>73.22</td>
<td>82.47</td>
<td>64.12</td>
<td>77.11</td>
</tr>
</tbody>
</table>

4. Conclusions

Two robust front-ends, namely robust RMVDR cepstral coefficients (RRMCC) and normalized RMVDR cepstral coefficients (NRMCC) are introduced. Performances of the proposed front-ends are evaluated on the AURORA-4 LVCSR task and compared with some latest robust front-ends. Experimental results showed that the proposed methods are more robust than the other features extractors when there is mismatch between training and test conditions. On the average both front-ends demonstrated better recognition word accuracy than the other considered methods and the NRMCC features performed the best among all the feature extractors.
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.


