Regularized Minimum Variance Distortionless Response-Based Cepstral Features for Robust Continuous Speech Recognition

Md Jahangir Alam\textsuperscript{1}, Patrick Kenny\textsuperscript{2}, Douglas O’Shaughnessy\textsuperscript{3}

\textsuperscript{1,3}INRS-EMT, University of Quebec, Montreal, Quebec, Canada
\textsuperscript{1,2}CRIM, Montreal, Quebec, Canada
alam@emt.inrs.ca, Patrick.kenny@crim.ca, dougo@emt.inrs.ca

Abstract

In this paper, we present robust feature extractors that incorporate 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, to estimate the speech power spectrum. Direct spectrum estimators, e.g., single tapered periodogram, have high variance and they perform poorly under noisy and adverse conditions. To reduce this performance drop we propose to increase the robustness of speech recognition systems by extracting features that are more robust based on the regularized MVDR technique. The RMVDR spectrum estimator has low spectral variance and is robust to mismatch conditions. Based on the RMVDR spectrum estimator, robust acoustic front-ends, namely, are regularized MVDR-based cepstral coefficients (RMCC), robust RMVDR cepstral coefficients (RRMCC) and normalized RMVDR cepstral coefficients (NRMCC). In addition to the RMVDR spectrum estimator, RRMCC and NRMCC also utilize auditory domain spectrum enhancement methods, auditory spectrum enhancement (ASE) and medium duration power bias subtraction (MDPBS) techniques, respectively, to improve the robustness of the feature extraction method. Experimental speech recognition results are conducted on the AURORA-4 large vocabulary continuous speech recognition corpus and performances are compared with the Mel frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP), MVDR spectrum estimator-based MFCC, perceptual MVDR (PMVDR), cochlear filterbank cepstral coefficients (CFCC), power normalized cepstral coefficients (PNCC), ETSI advancement front-end (ETSI-AFE), 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 error rate on the AURORA-4 large vocabulary continuous speech recognition (LVCSR) task under clean and multi-condition training conditions. In clean training conditions, on average, the RRMCC and NRMCC provide significant reductions in word error rate over the rest of the front-ends. In multi-condition training, the RMCC, RRMCC, and NRMCC perform slightly better in terms of the average word error rate than the rest of the front-ends used in this work.

Keywords: Speech recognition, robust feature extraction, regularized MVDR, ASE, MDPBS, feature normalization, multi-condition training

1. Introduction

Mel-frequency cepstral coefficients (MFCC) [1], which have proven to be one of the most effective feature sets for speech and speaker recognition tasks, are frequently used as a low-dimensional set of features to represent short-time speech signals. MFCC 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. MFCC and perceptual linear prediction (PLP) [10]-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 MFCC (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, 9, 56, 59], etc. In MFCC [1] and PLP [10] front-ends, and in most of the robust feature extractors the features are computed from a windowed (e.g., Hamming) direct spectrum estimate (the squared magnitude of the Fourier transform of the short-time windowed observed signal) that has a high spectral variance. The variances of these features are greatly influenced by the variances of the spectral estimates of the observed speech signal. Variance in the feature vectors has a
direct bearing to the variance of Gaussians modeling the speech classes. Reduction in the variance of the feature vector increases class separability and improved class separability can potentially increase recognition accuracy and decrease search speed [11]. Although direct spectrum estimators (also known as non-parametric 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 mismatched training/test conditions. Among the parametric spectrum estimators, the linear predictive coding (LPC) based all-pole spectrum estimator is most widely used [12]. It has been noted in speech modeling literature that the LP-based all-pole models do not provide good models of the spectral envelope for medium and high pitch voiced speech [11]. Also, the LP-based cepstra are known to be very sensitive to noise. They tend to overestimate or overemphasize sparsely spaced harmonic peaks [13]. The standard feature extractors used for speech recognition are based on either DFT, e.g., MFCC or linear prediction, e.g., PLP. The MFCC feature extractor is not robust and therefore shows poor performance under noisy and adverse conditions. On the other hand, the PLP front-end is ill-suited for reliable estimation of the spectra of speech signals, which is true for all methods using linear prediction envelopes [13]. In order to overcome the problems associated with linear prediction, namely, over-estimation of spectral power at the harmonics of voiced speech, the MVDR method was proposed in [14]. It is also known as Capon's method [12] for all pole modeling of speech.

In this paper, we propose to incorporate a regularized minimum variance distortion-less response (RMVDR) spectrum estimator, in place of the DFT-based direct spectrum estimator, into the traditionally used feature extraction framework, e.g., MFCC, for speech recognition task. Based on RMVDR spectrum estimation method we also propose robust feature extractors, dubbed as robust regularized MVDR cepstral coefficients (RRMCC) and normalized RMVDR cepstral coefficients (NRMCC), that include the use of sigmoid-shape auditory domain spectrum enhancement (ASE) [6] and medium duration power bias subtraction [5] techniques, respectively, to improve the robustness of speech recognition systems in adverse conditions while having little performance reduction in matched train/test conditions. The advantages of a RMVDR spectrum estimator are:
a) It overcomes the problems apparent in linear prediction spectral estimation,
b) The regularization parameter helps to penalize rapid changes in all-pole spectral envelopes thereby producing smooth spectra without affecting the formant positions [15, 16], and
c) It provides robust spectral estimates under noisy environments [17, 18, 19].

The MVDR spectral estimator has already been applied in speech recognition [11] and speaker identification [13] tasks. An extension of the MVDR method was proposed in [20] by warping the frequency axis with the bilinear transformation prior to MVDR spectral estimation. In [21], a perceptual MVDR-based cepstral coefficients (PMCC) approach is proposed where perceptual information is directly incorporated into the spectrum estimation. The perceptually motivated MVDR (PMVDR) front-end, proposed in [22], completely eliminates the auditory filterbank processing step and directly performs warping on the DFT power spectrum.

In order to compare the performance of the proposed front-ends, the following conventional and robust front-ends were chosen: MFCC [1], PLP [10], MVDR-based MFCC [11], PMVDR [22], ETSI-AFE [4], power normalized cepstral coefficients (PNCC) [5], cochlear filterbank cepstral coefficients (CFCC) [23], and the robust feature extractor (RFE) proposed in [6]. 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 (Least Mean Squares) algorithm, for improving robustness of ASR systems against additive noise distortions and channel effects. The PNCC technique, proposed in [5], includes the use of a gammatone filter-bank (GTFB) and power law nonlinearity, a medium duration power bias subtraction technique based on the arithmetic mean (AM)-geometric mean (GM) ratio for noise reduction, and cepstral mean normalization as a post-processing scheme for DC offset removal, for robust feature extraction. The robust feature extractor of [6] includes the use of an asymmetric and level-dependent compressive gammachirp filter-bank (cGCFB) [24] for auditory spectral analysis, power function nonlinearity, a sigmoid-shape weighting rule based on the subband a posteriori signal-to-noise ratio (SNR) to enhance the speech auditory spectrum and, as a post-processing scheme, a short-time cepstral mean and scale normalization (STCMSN) technique [25] to normalize the features [59].
Most of the robust 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. In the proposed robust feature extractors RRMCC and NRMCC, to normalize features we use short-time cepstral mean and scale normalization and cepstral mean normalization, respectively.

To evaluate and compare the performance of the proposed feature extractors, RMCC, RRMCC and NRMCC, speech recognition experiments are performed on the AURORA-4 [26] LVCSR task. The reported speech recognition results show that the proposed robust front-ends, RRMCC and NRMCC, outperform all other front-ends considered here for comparison of performances. RMCC features are also found robust with respect to the MFCC, PLP, MVDR and PMVDR features.

The rest of the paper is organized as follows: section 2 presents a brief overview of the various methods used for spectral analysis of speech signals in a speech recognition task. Section 3 describes the regularized minimum variance distortionless response (RMVDR) spectrum estimator and section 4 provides a detailed description of the RMVDR-based robust feature extractors proposed in this work for robust speech recognition. The performances of the robust feature extractors are evaluated and compared with the conventional MFCC, PLP other robust front-ends in section 5 and concluding remarks are drawn in section 6.

2. Background Spectral Analysis

Spectral estimation can be loosely defined as any process in which the frequency content of a signal is determined automatically. Applications of spectral analysis include speech analysis, music analysis, communications, radar, sonar, and experimental sciences. Spectrum estimators are classified as parametric and nonparametric (or direct). The Discrete Fourier transform (DFT)-based periodogram is an example of a nonparametric (or direct) spectrum estimator and the LPC-based spectrum estimator is a parametric method.
2.1. DFT- and LP-based spectrum estimators

MFCC (and PLP) features are computed from discrete Fourier transform (DFT)-based windowed periodogram estimates given by

$$\hat{S}_{\text{DFT}}(f) = \left| \sum_{j=0}^{N-1} w(j)s(j)e^{\frac{j2\pi f}{N}} \right|^2,$$

where \(f\) denotes the discrete frequency index, \(N\) is the frame length, \(j \in \{0,1,...,N-1\}\) is the sample index, \(s(j)\) is the time domain speech signal and \(w(j)\) denotes the time domain window function, e.g., Hamming.

In the LPC (linear predictive coding) analysis the current value of the speech sample \(s(n)\) is obtained as a weighted sum of its \(p\) past samples as follows [3]:

$$s(n) = -\sum_{q=1}^{p} a_q s(n-q) + e(n),$$

where \(p\) is the model order, \(a = \{a_q, q=1,...,p\}\) are the predictor coefficients, and \(e(n)\) is the prediction error or residual. The spectrum of the LP method is then given by:

$$S_{\text{LP}}(f) = \frac{1}{1 + \sum_{q=1}^{p} a_q e^{-j2\pi fq}}.$$

In the autocorrelation method the predictor coefficients \(a = \{a_q, q=1,...,p\}\) are expressed as a solution of (2) as [27]:

$$a_{\text{opt}}^{\text{LP}} = -R_{\text{LP}}^{-1}r_{\text{LP}},$$

where \(R_{\text{LP}}\) represents the Toeplitz autocorrelation matrix and \(r_{\text{LP}} = \left[ r_1, r_2, ..., r_p \right]^T\) is the autocorrelation vector, and \(r_q = E[s(n)s(n-q)]\) with \(q = 1,2,...,p\). Eqn. 4 is usually solved through the computationally efficient Levinson-Durbin recursion algorithm [48-49].
2.2. MVDR spectrum estimation

The Minimum Variance Distortionless Response (MVDR) spectrum estimator, introduced by Capon [12], is mostly used in array signal processing applications, and has also been investigated in relation to other applications such as speech modeling [14], robust speech recognition [11], and speaker recognition [13] systems. The MVDR method defines a filter that leaves the signal undistorted at frequencies of interest while suppressing other frequencies in an optimal way. The MVDR spectrum is given by:

\[ S_{MVDR}(f) = \frac{1}{\mathbf{v}^H(f) \mathbf{R}_{p+1}^{-1} \mathbf{v}(f)} \]  \hspace{0.5cm} (5)

where \( \mathbf{R}_{p+1} \) is the autocorrelation matrix, \( \mathbf{v}(f) = [1, e^{-i2\pi f}, e^{-i4\pi f}, ..., e^{-i2\pi p}] \) is a frequency tuning vector and \((\cdot)^H\) denotes the Hermitian transpose operator. The model order \( p \) corresponds to the largest correlation lag in the autocorrelation matrix. Eqn. (5) shows that 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 spectral estimator variance [11, 27]. In [11], a more detailed description about the bias and variance reduction by the MVDR spectrum estimator can be found.

The MVDR spectrum has an affinity with nonparametric filterbank spectral analysis methods. In the filterbank interpretation of DFT-based spectral analysis methods, the spectrum at any given frequency can be viewed as the power at the output of a bandpass filter. In this case, the bank of bandpass filters is data-independent and its characteristics are defined by the length and choice of the analysis window, arranged along an equally spaced frequency grid. Moreover, the frequency characteristics of the individual filters are such that they are also frequency independent among themselves [47]. Similarly to DFT-based methods, the MVDR spectral estimate can be conceptually viewed as an output of a bank of filters, with each filter centered at one of the analysis frequencies. However, in contrast to DFT-based methods, the bandpass filters of the MVDR bank are both data and frequency dependent, and information that is captured by the signal autocorrelation matrix \( \mathbf{R}_{p+1} \) appears in the definition of the MVDR spectrum in eqn. (5).
In particular, the MVDR spectral estimate at frequency \( f \) utilizes a specially designed FIR bandpass filter with impulse response \( h = \{ h(n), n = 0,1,...,p \} \). The passband characteristics of this filter are designed so that it would pass, without distortion, the signal components subject to the constraint that its response at the frequency of interest \( f \) has unity gain. This constraint is known as a distortionless constraint and can be expressed mathematically as:

\[
H(f) = \sum_{k=0}^{p} h(i)e^{-j2\pi kf} = v^H(f)h = 1.
\]

The distortionless filter \( h(n) \) is obtained by solving the following constrained optimization problem: \( \min_h \{ h^H R_{p+1} h \} \) subject to \( v^H(f)h = 1 \). In other words, the MVDR filter designed for a frequency of interest will let the input signal component at that frequency pass undistorted. It minimizes the total output power of the filter, which ensures that the undesired frequency components of the signal are suppressed in an optimal manner.

The \( p \)-th order MVDR spectral estimate can be parametrically obtained from the LP (linear prediction) coefficients \( a_q \) in a very interesting and computationally efficient way as [11]:

\[
S_{\text{MVDR}}(f) = \frac{1}{\sum_{k=-p}^{k=p} \mu(k)e^{-j2\pi kf}}, \tag{6}
\]

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

\[
\mu(k) = \begin{cases} 
\frac{1}{\sigma} \sum_{q=0}^{p-k} (p+1-k-2q)a_q a_{q+k}^*, & \text{for } k \geq 0 \\
\mu^*(-k), & \text{for } k < 0,
\end{cases} \tag{7}
\]

where \( \sigma \) is the residual variance (i.e., variance of the prediction error signal) and \((.)^*\) denotes the complex conjugate operator.
3. Regularized MVDR (RMVDR) spectrum estimation

All-pole spectral envelope estimates based on LP for speech signals often exhibit unnaturally sharp peaks, especially for speakers with high pitch frequency [14, 28]. The LP method fails to separate the short-term dependency (the envelope) from the long-term dependency (the pitch) and the resulting envelope is contaminated with harmonics. These peaky spectral envelopes can cause problems in speech modification [28]. Because of the inherent smoothing properties of the MVDR spectrum estimator, MVDR spectral estimates obtained using the LP coefficients have less rapid variations than the LP spectral estimates (see fig. 3), but MVDR estimates may still be affected.

In [15, 28] 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 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) [17-19].

Similar to the MVDR spectrum estimator, the $p$th order RMVDR spectral estimate can be parametrically written as:

$$S_{RMVDR}(f) = \frac{1}{\sum_{k=-p}^{k=p} \mu_r(k)e^{-i2\pi fk}},$$

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

$$\mu_r(k) = \begin{cases} \frac{1}{\sigma^2_e} \sum_{q=0}^{p-k} (p+1-k-2q)a_q^\ell a_q^{\ell+*}, & \text{for } k \geq 0 \\ \mu_r(-k), & \text{for } k < 0. \end{cases}$$

In the RLP method, the predictor coefficients $a_q^\ell$ 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 minimizes that modified objective
function of the following form [15, 11, 28]:

$$\sum_n \left( s(n) + \sum_{q=1}^{r} a_q s(n-q) \right)^2 + \lambda \psi(a),$$

where regularization constant $\lambda > 0$ controls the trade-off between the fit of speech
spectrum and smoothness measure. The RLP method penalizes the rapid changes in the
all-pole spectral envelope and therefore produces a smooth spectral estimate, keeping the
formant positions unaffected. In [15, 16, 28] the authors use the following penalty
measure

$$\psi(a) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{A'(e^{i\omega})}{W(\omega)} \right|^2 d\omega,$$

where $1/|W(\omega)|^2$ serves as a coarse approximation of the spectral envelope
$1/|A(e^{i\omega})|^2$ and $A'(e^{i\omega})$ is the frequency derivative of the inverse filter
$A(e^{i\omega}) = \sum_{q=0}^{r} a_q e^{-i\omega q}$ with $a_0 = 1$. The merit of this penalty measure is that a closed form
and computationally efficient non-iterative solution exists [16, 28].

It has been shown in [15, 16, 28] that the above penalty measure (eqn. (11)) can be
approximated as:

$$\hat{\psi}(a) = a^T DFDa,$$

where $a = [a_1, ..., a_p]^T$ are the predictor coefficients, $D$ is a diagonal matrix in which each
diagonal element consists of the row number, $F$ is a Toeplitz matrix corresponding to the
windowed autocorrelation sequence $f(\tau) = r(\tau) w'(\tau) = \sum_{j=0}^{N-1} s(j)s(j-\tau)w'(\tau)$,
$\tau = 0, 1, ..., p-1$, of a frame of the speech signal, and $w'(\tau)$ is a window function. Here
$\tau$ is the time lag. The matrix $F$ represents the denominator term of eqn. (11). The coarse
spectral envelope $1/|W(\omega)|^2$ was derived from the windowed autocorrelation sequence
The matrix $F$ is equal to $R_{LP}$ if $w'(\tau)$ is chosen as a rectangular window. The compact solution obtained from the modified cost function (eqn. (10)) is [15, 28]:

$$a_{opt} = a^r = -(R_{LP} + \lambda DFD)^{-1} r_{LP},$$

(13)

Various tapers (or windows) were considered in [15-16, 28] to compute the windowed autocorrelation matrix $F$. It has been reported in [28, 33] that the double autocorrelation (DAC) sequence can be used for robust estimation of spectral envelopes in adverse environments. In this paper, similar to [28], we use a DAC sequence, i.e.,

$$f(\tau) = \sum_{j=0}^{p-1} r(j) r(j-\tau), \tau = 0, 1, ..., p-1,$$

for the computation of $F$ instead of the windowed autocorrelation (WAC) sequence. It was found in [16] that computation of $F$ from the WAC sequence does not improve recognition accuracy.

Figs. 1 & 2 present a comparison of the estimated spectra of a frame of (a) clean speech, and (b) noisy speech (corrupted with babble noise at a signal-to-noise ratio of 0 dB) signals obtained by the various spectrum estimators described in this paper. Figs. 1 & 2 demonstrate that both MVDR and RMVDR provide robust spectral estimates compared to the DFT- and LP-based spectrum estimators. Compared to LP and MVDR methods, RMVDR provides smoothed spectral peaks.

![Figure 1](image)

**Figure 1.** Comparison of the estimated speech spectra of a frame of speech signal (a female speaker uttering the vowel phoneme /ae/) obtained using various spectrum estimators when the speech signal is (a) clean, and (b) corrupted by 0 dB babble noise. The
spectra in each plot have been shifted for better visualization. The model order used is $p = 100$. The value for the regularization parameter $\lambda$ used for the regularized MVDR (RMVDR) estimator is $10^{-9}$ [12, 13].

Figure 2. Comparison of the estimated speech spectra of a frame of speech signal (a male speaker uttering the vowel phoneme /ae/) obtained using various spectrum estimators when the speech signal is (a) clean, and (b) corrupted by 0 dB babble noise. For better visualization the spectra in each plot have been shifted. The model order used is $p = 100$. The value for the regularization parameter $\lambda$ used for the regularized MVDR (RMVDR) estimator is $10^{-9}$ [12, 13].

Figure 3. Comparison of the estimated short-term spectra of a speech signal (taken from the AURORA-4 corpus) frame using various spectrum estimators. For better
visualization the spectra in this plot are shifted. The model order used is $p = 100$. The value for the regularization parameter $\lambda$ used for the regularized MVDR (RMVDR) estimator is $10^{-9}$ [12, 13].

**Figure 4.** Comparison of the estimated running short-term spectra of the speech signal (a) (vowel /æ/ uttered by a female speaker) using the MVDR and RMVDR spectrum estimators. The model order used is $p = 100$. The value for the
regularization parameter $\lambda$ used for the regularized MVDR (RMVDR) estimator is $10^{-9}$ [12, 13].

It is observed from fig. 3 that, compared to the Hamming-windowed periodogram, LP and the MVDR spectrum estimators, the RMVDR method provides a smooth spectral estimate and therefore results in reduced spectral variance. Fig. 4 (b) provides a comparison of estimated short-term running spectra (all frames of a speech (vowel /ae/) signal, frame index increases from left to right and top to bottom) obtained by the MVDR and RMVDR spectrum estimators. It is observed from this figure that the RMVDR provides better and smoothed spectral estimates than the MVDR spectral estimates.

4. Regularized MVDR (RMVDR)-based Robust Front-ends

In addition to the RMVDR-based cepstral coefficients feature, dubbed as RMCC, we also propose two robust feature extractors, namely, robust RMVDR cepstral coefficients (RRMCC) and normalized RMVDR cepstral coefficients (NRMCC). These front-ends incorporate auditory domain enhancement methods to enhance the speech spectrum. The various steps of the conventional MFCC and PLP feature extraction process are shown in fig. 5. Here, we use an HTK version of PLP [29], where, instead of the Bark scale filterbank, a Mel scale filterbank is used for auditory frequency analysis of the speech spectrum. The only difference between the RMCC and MFCC feature extraction process is in the spectrum estimation. MFCC features are computed from the DFT-based direct spectral estimates, whereas, in the RMCC extraction method, an RMVDR spectrum estimator is used for the estimation of speech power spectra. A combined block diagram of the proposed RMVDR-based robust feature extractors RRMCC and NRMCC is shown in figure 6. The stages that belong to only RRMCC or NRMCC feature extraction processes are enclosed by dotted lines. In the combined block diagram, 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 both front-ends are computed from the RMVDR spectral estimates. In the following sub-sections we provide a detailed description of the proposed robust front-ends.
Figure 5. Schematic diagram showing the various steps of the conventional MFCC and PLP (as implemented in HTK) feature extraction process.

4.1. RRMCC Front-end

In this case, the Mel filterbank auditory spectrum is passed through an auditory domain spectrum enhancement method, introduced in [6], that utilizes a sigmoid-shape weighting rule based on the subband a posteriori signal-to-noise ratio (SNR). A power function nonlinearity with a coefficient of 0.07 ($\approx 1/15$) is applied on the enhanced auditory spectrum. Static cepstral features, obtained in this front-end, are normalized using a short-time cepstral mean and scale (or gain) normalization technique (STCMSN) [25] 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 [29].

4.1.1. Noise Spectrum Estimation

The estimation of a noise power spectrum from the noisy speech signal plays a very important rule in noise reduction/speech enhancement algorithms. For relatively stationary noises, an accurate estimation of the noise spectrum can be done using
minimum statistics (MS)-based approaches [30-31]. These algorithms lead to less satisfying results for rapidly changing noises. In this paper, for accurate estimation of noise power spectra, we employ a minimum mean square error (MMSE) - soft speech presence probability (SPP) (MMSE-SPP)-based noise estimation approach, proposed in [32]. In this method, the 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 [32]; it also results in less overestimation of noise power and is computationally less expensive.

4.1.2. Auditory Spectrum Enhancement (ASE)

In order to enhance the auditory spectrum $S_{a}(m,k)$ we chose a sigmoid-shape weighting function $\tilde{W}(m,k)$ based on the subband a posteriori SNR (in dB) $\gamma_{sb}(m,k)$ as follows:

$$\tilde{W}(m,k) = \frac{1}{1 + e^{-\frac{\vartheta(m,k)}{\eta}}}$$

(14)

where $k=1,2,\ldots,K$ is the subband index, $m=1,2,\ldots,M$ is the frame index, $\eta$ is a parameter that controls the lower limit of the weighting function and $\vartheta(m,k)$ is the instantaneous SNR (in dB) defined as:

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

(15)

where $\gamma_{sb}(m,k) = \max \left( \log_{10} \left( \frac{S_{a}(m,k)}{N_{a}(m,k)} \right) - 4.0 \right)$. $N_{a}(m,k)$ is the noise power spectrum mapped onto the auditory frequency axis. Here, $\eta = 4.5$ is chosen experimentally.

In order to remove the outliers from the weighting function $\tilde{W}(m,k)$ due to noise variability, we use a $3 \times 5$ two-dimensional median filter. For smoothing the decision regions, a $2 \times 2$ two-dimensional moving average filter is also applied. The enhanced
auditory spectrum $S_{\text{ext}}(m,k)$ can be obtained using the smoothed weighting function $\tilde{W}_s(k,m)$ as:

$$S_{\text{ext}}(m,k) = \tilde{W}_s(k,m) \cdot S_{\text{ext}}(m,k).$$ \hspace{1cm} (16)

**Figure 6.** Schematic diagram showing the various steps of the proposed feature extractors RRMCC (when 1 is connected to 2) and NRMCC (when 1 and 3 are connected). Both feature extractors incorporate a regularized MVDR spectrum estimator for spectrum estimation, auditory domain spectrum enhancement methods, and power law nonlinearity in the conventional MFCC feature extraction framework.

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. The proposed ASE technique helps to reduce noise from noisy speech signals and does not introduce much distortion to
clean speech signals. Speech spectrograms presented in figures 12-14 (in section 5) confirmed the effectiveness of this enhancement method.

4.1.3. Power function Nonlinearity

While the auditory filtering (e.g., Mel-scale filterbank) approximates the nonlinear characteristics of the human auditory system in frequency, the natural logarithmic nonlinearity or power function nonlinearity deals with the loudness nonlinearity. It approximates the relationship between a human’s perception of loudness and the sound intensity [34]. The logarithmic nonlinearity is found to be sensitive to noise, whereas the power-law nonlinearity is more immune to noise [5, 35-36]. One of the reasons why MFCC features have poor performance under noisy environments can be attributed to the logarithmic nonlinearity [35-36]. This nonlinearity gives large negative values to inputs close to zero, which leads to a spreading of energy after the discrete cosine transform (DCT) [36]. This problem is normally avoided by appropriately flooring the power values before applying the logarithm. Power function nonlinearity or root compression (which can be expressed as $y_i = y^a$, $0 < a < 1$) followed by DCT leads to better compaction of energy [29]. In this paper we use a power function nonlinearity with a coefficient of $a = 0.07$ to approximate the loudness nonlinearity of human perception.

4.1.4. Feature Normalization (FN)

Feature normalization strategies are employed in speech (and speaker) recognition systems to compensate for the effects of environmental mismatch. These techniques are preferred because a priori knowledge and adaptation are not required under any environment. Most of the normalization techniques are applied as a post-processing scheme on the Mel-frequency cepstral coefficient (MFCC) speech features. Normalization techniques can be classified as model-based or data distribution-based techniques. In model-based normalization techniques, certain statistical properties of speech such as mean, variance, and moments, are normalized to reduce the residual mismatch in feature vectors, e.g., STMVN (short-term mean and variance normalization)
[25]. Data distribution-based techniques aim at normalizing the feature distribution to the reference, such as short-time Gaussianization (STG) [38]. Some feature normalization techniques have been proposed in the past for speech and speaker recognition systems, including feature warping [37], STG [38], cepstral mean normalization (CMN) [39-40], cepstral variance normalization (CVN) [41], histogram equalization [42], Quantile-based Histogram Equalization [57], and RASTA filtering [43-44]. Almost all the feature extractors include a feature normalization technique as a post-processing scheme. 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 MVN (mean and variance normalization) [39, 41]. 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 MFCC features are normalized over a sliding window of more than 1s duration. The feature vector to be normalized is located at the centre of the sliding window.
Figure 7: The zero-th cepstral coefficients ($c_0$) of clean and noisy speech (street noise, SNR = 5 dB) when (a) no normalization technique is applied, and (b) normalized by the STCMSN technique. The number of frames in the center-aligned window to compute the mean and difference between upper and lower bound is $L = 150$.

In the proposed feature extractor, in order to relax the constant channel assumption as used in conventional full utterance-based feature normalization methods, the static features are normalized using the short-time cepstral mean and scale (STCMSN) approach [25]. In the STCMSN approach, the $n$-th feature space and $m$-th frame $C(m,n)$ are normalized as

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

(17)

where $\mu_{st}(m,n)$ and $d_{st}(m,n)$ 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:

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

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

The main idea behind STCMSN (or cepstral mean and scale normalization (CMSN)) technique is that under mismatched conditions a difference of cepstra between the training and test environments is removed by adjusting the short-time scale and short-time mean. The advantage of the 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. Fig. 7 presents the zero-th cepstral coefficient feature ($c_0$) of a clean speech signal and that of a noisy speech signal, having a signal-to-noise ratio (SNR) of 5 dB, (a) before applying feature normalization and (b) after normalizing by the STCMSN method. It is observed from fig. 7 (a) that the effects of additive noise on a clean speech signal are:

i. The minimum values of the cepstral features are elevated.
ii. The valleys of the cepstra are affected by the additive noise energy while the peaks remain almost unaffected.

The larger difference in valleys leads to a mismatch between clean and noisy speech. It is clear from fig. 7 (b) that the STCMSN method is able to reduce the mismatch between the clean and noisy speech features.

4.2 NRMCC Front-end

In this front-end, after estimating the speech spectra using the RMVDR spectrum estimator, Mel filterbank (MelFB) integration is performed on the speech spectra for auditory spectral analysis. 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 before and 2 after) power bias subtraction (MDPBS) that is based on maximization of the sharpness of the power distribution. A power flooring is applied in the MDPBS method to reduce the spectral distortion between the training and test data. This is important because the regions of the spectro-temporal speech segments that exhibit the lowest power are the most vulnerable to additive noise [5].

4.2.1 Medium Duration Power Bias Subtraction (MDPBS)

In MDPBS processing, as introduced in [5], a medium duration (5 speech frames, the current plus 2 frames before and after) power $\tilde{S}_{as}(m,k)$ for the $m$-th frame and $k$-th subband is computed. This is done in the Mel filterbank domain using the following equation:

$$
\tilde{S}_{as}(m,k) = \frac{1}{2M'+1} \sum_{q=m-M'}^{q=m+M'} S_{as}(q,k)
$$

where $M'$ is the lag size of the medium duration window, $S_{as}(m,k)$ is the Mel filterbank integrated RMVDR power spectrum. Similar to [5], in this paper we have used $M' = 2$. 

From the medium duration power $\tilde{S}_{\text{as}}(m,k)$, a power bias level $p_b$ is selected as that level above which the signal power distribution is sharpest. The ratio of the arithmetic mean (AM) to the geometric mean (GM) of medium-duration power is used to measure the sharpness of signal power distribution. The motivation behind this idea is that the human hearing system is more sensitive to changes in signal power over time and frequency than to relatively constant background excitation [5]. In order to enhance robustness a power flooring $p_{flr}$ is applied while subtracting this bias level. Power flooring means imposing a lower bound on power spectral values. Spectro-temporal speech segments that exhibit small power are the most susceptible to additive background noise. Use of power flooring can reduce the spectral distortion between the training and test sets for those vulnerable regions [5]. Similar to [5], in this work we use $p_{flr} = 0.01$.

For each $k$-th channel, the bias-subtracted auditory spectrum $\tilde{S}_{\text{bias}}(m,k)$ can be expressed as [5]

$$\tilde{S}_{\text{bias}}(m,k) = \max\{\tilde{S}_{\text{as}}(m,k) - p_b, p_{flr}\}.$$  \hfill (19)

For a detailed description of the power bias subtraction method, bias level selection, and for the matlab code we refer to [5].

Now, the weighting rule for the $m$-th frame and $k$-th subband is given by

$$H(m,k) = \frac{\tilde{S}_{\text{bias}}(m,k)}{S_{\text{as}}(m,k)}.$$ \hfill (20)

In order to smooth the decision regions of the weighting function, a 1-D moving average filter is applied across the channel. The smoothed weighting factor for the $k$-th channel and $m$-th frame is given by:

$$H_s(m,k) = \frac{1}{\min(k + N, K) - \max(k - N, 1) + 1} \sum_{q=\max(k-N,1)}^{\min(k+N,K)} H(m,q),$$ \hfill (21)

where $N$ is the lag size of the moving average window. Similar to [5], we use $N = 4$.

The enhanced auditory power spectra can now be obtained as:

$$S_{\text{eas}}(m,k) = H_s(m,k) \cdot S_{\text{as}}(m,k).$$ \hfill (22)
Fig. 8 shows the various steps of the MDPBS method to enhance the auditory spectrum. A power function nonlinearity with a coefficient of 0.07 (≈1/15) is applied on the enhanced auditory spectrum to approximate the loudness nonlinearity of human perception. A 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].

5. Performance Evaluation

The performances of the proposed feature extractors, RRMCC and NRMCC, as presented in fig. 6, are evaluated and compared with other competing feature extraction approaches in a wide variety of noise environments. The front-ends that are selected for comparison
purposes are: conventional MFCC [1], PLP [10], MVDR spectrum estimation based MFCC [11], perceptual MVDR (PMVDR) [22], Cochlear filterbank cepstral coefficient (CFCC) [23], PNCC [5], ETSI-AFE [4], and the robust feature extractor (RFE) proposed in [6]. Word error rate (WER) is used as a performance evaluation measure for comparing the recognition performances on the AURORA-4 LVCSR task of the proposed front-ends to that of the baseline feature extractors. Both the clean condition and multi-condition training modes are used for performance evaluation.

In the following sub-sections we provide a description of the speech corpus that is chosen for performance evaluation, experimental setup and some discussion on the reported experimental speech recognition results.

5.1. Speech corpus

The AURORA-4 [26] continuous speech recognition corpus is derived from the Wall Street Journal (WSJ0) corpus. It is divided into 3 sets, namely, training (clean condition and multi-condition training data), development (dev test) and evaluation (eval or test) sets. This task is often referred to as the 5k closed vocabulary task, i.e., there are no out-of-vocabulary words (OOVs) in the evaluation set. The training set contains 7138 utterances from 83 speakers, totaling 14 hours of speech data. The multi-condition training set contains 7138 utterances. Half of the utterances were recorded with the primary Sennheiser microphone and the other half were recorded using one of several different secondary microphones. Both halves include a combination of clean speech and speech corrupted by one of six different noises - street traffic, train station, car, babble, restaurant, airport at 10-20 dB SNR. 14 evaluation sets were defined in order to study the degradations in speech recognition performance due to microphone conditions, filtering and noisy environments. Each of the filtered versions of the evaluation set recorded with a Sennheiser microphone and a secondary microphone was selected to form the two evaluation sets. The remaining 12 subsets were defined by randomly adding each of the 6 noise types (car, babble, restaurant, street traffic, airport, and train-station noises) at a randomly chosen SNR between 5 and 15 dB for each of the microphone types as mentioned above. The goal was to have an equal distribution of each of the 6 noise types
and the SNR with an average SNR of 10 dB [26]. Each of the test sets contains 166 utterances from 8 speakers, totaling 20.69 minutes of speech data. The 14 test sets were grouped into the following 4 families [26, 45]:

a. Test set A - clean/multi-condition speech in training and clean speech in test, same channel (set 1),

b. Test set B - clean/multi-condition speech in training and noisy speech in test, same channel (sets 2-7),

c. Test set C - clean/multi-condition speech in training and clean speech in test, different channel (set 8), and

d. Test set D - clean/multi-condition speech in training and noisy speech in test, different channel (sets 9-14).

The number inside the parentheses represents the test set number defined in the AURORA-4 corpus. In the case of clean condition training, test set A represents the matched training/test condition whereas test sets B, C, and D represent the mismatched training/test conditions due to various additive background noise and channel distortions. On the other hand for multi-condition training, test set A represents the mismatched training/test condition whereas test sets B, C, and D represent the matched training/test conditions.

5.2. Experimental setup

For the continuous speech recognition task on the AURORA-4 corpus, all experiments employed state-tied crossword speaker-independent triphone acoustic models with 16 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 [26, 45].

For our experiments, we use 13 cepstral 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 delta features were calculated using the regression formula with a 5-frame window. All the front-ends, considered in this paper including the proposed methods,
incorporate feature normalization as a post-processing scheme. MFCC [1], PLP [10], MVDR [11], PMVDR [22], and the proposed RMVDR-based cepstral coefficient (RMCC) front-ends use conventional full-utterance based mean and variance (MVN) + normalization. In the CFCC [23] and the proposed RRMCC front-ends, features are normalized using short-time cepstral mean and scale normalization (SMCMSN) over a sliding window of 1.5 s duration. PNCC [5] and the proposed NRMCC utilize classical cepstral mean normalization (CMN) over the whole utterance. Some of the baseline feature extractors (i.e., MFCC, PLP) considered in this paper are implemented using the rastamat toolbox [46].

5.3. Experimental Results and Discussion

In order to verify the effectiveness of the proposed RMCC, RRMCC and NRMCC feature extractors, speech recognition experiments are conducted on the AURORA-4 large vocabulary continuous speech recognition (LVCSR) corpus. Word error rate is used as an evaluation metric for comparing the recognition performances of the proposed methods to that of the baseline feature extractors.

![Figure 9. Influence of the model order p on the speech spectrum for the regularized MVDR method. The value for the regularization parameter $\lambda = 10^{-9}$. The spectra in each plot has been shifted for better visualization.](image)
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. Fig. 9 illustrates the influence of the model order on the spectral estimate of the speech signal. It is observed from fig. 9 that a higher model order provides more detail of the fine structure of the spectrum and represents the first harmonic (or fundamental frequency), whereas a low model order results in a reduction of influence of the excitation and is more or less a representation of the vocal tract transfer function [13].

Fig. 10 presents speech recognition word accuracy $WAcc$ ($WAcc = 100 - WER$) obtained using RMCC (regularized RMVDR based cepstral coefficient) front-end for model orders between 20 and 120 with a step size of 10. It is observed from fig. 10 that the recognition accuracy is high for $p = 20$ and it decreases with increase in $p$ till $p = 60$. For $p > 60$ recognition accuracy again begins to increase linearly with $p$ until a model order of $p = 100$ is reached and after that accuracy becomes almost constant. So, for our work, the optimal model order is chosen as $p = 100$.

Fig. 11 demonstrates the recognition word accuracy $WAcc$ of the RMCC front-ends for various values of the regularization constant $\lambda$. The accuracy increases with decrease in $\lambda$ and the highest performance is obtained for $\lambda \leq 10^{-9}$. So, in this work, the optimal value for the regularization constant is chosen as $\lambda = 10^{-9}$.

![Figure 10](image.png)

**Figure 10.** Variation of Word Accuracy (WAcc) of the RMVDR-based feature extractor with the model order $p$. Model order $p$ was varied from $p = 20$ to $p = 120$. 
Figure 11. Variation of Word recognition accuracy (WAcc) of the RMVDR-based feature extractor with the logarithm of the inverse of regularization constant $\lambda$. $\lambda$ was varied from $10^{-4}$ to $10^{-10}$.

Compared to the DFT and LP spectrum estimator-based feature robustness of the RMVDR spectrum estimator-based proposed feature extractors, RMCC, RRMCC and NRMCC are also shown in figures 13-14 by presenting speech auditory spectrograms of noisy speech signals. The speech signals used in figures 13-14 were corrupted with babble and street noise.

Fig. 12 presents the auditory spectrograms of a clean speech signal obtained by the various feature extractors. It is observed from this figure that the ASE and MDPBS processing did not introduce much distortion to the spectrum.

Fig. 13 and fig. 14 present the auditory spectrograms of noisy speech signals, corrupted by babble noise and street noise with a signal-to-noise ratio (SNR) of 5 dB, respectively, obtained by the various feature extractors. It is observed from these figures that the proposed RMCC is robust compared to the MFCC, PLP, MVDR and PMVDR front-ends. The RMVDR-based proposed front-ends RRMCC and NRMCC provide more robust features than the conventional DFT and LP spectrum estimator-based features.

Table 1 depicts the recognition word error rate obtained by the different feature extractors on various test sets of the AURORA-4 LVCSR corpus. The results reported here were obtained with clean condition training. The RMCC demonstrated lower recognition WER than the MFCC, PLP, MVDR, and PMVDR front-ends in both matched and mismatched
conditions. Relative WER reductions obtained by the RMCC front-end are presented in fig. 15. An average relative error reduction of approximately 6.4% is achieved over the MFCC, PLP, MVDR and PMVDR front-ends. The proposed robust feature extraction methods, RRMCC and NRMCC, outperformed the other feature extractors, considered here for comparison, in terms of the recognition WER. Figures 16 and 17 depict the relative WER reduction achieved by the RRMCC and NRMCC front-ends, respectively, in clean condition training over the rest of the feature extractors included in this work for comparison and evaluation purposes. NRMCC provided an average relative WER reduction of approximately 23.6%, whereas an average relative WER reduction obtained by the RRMCC front-end is approximately 20.5%. NRMCC showed an average relative WER reduction of 3.3% over the RRMCC. The NRMCC front-end performed the best in terms of the recognition WER.

When a multi-condition training condition (i.e., training on (clean plus noisy) data and testing on all test data) is used, the WERs achieved by all the front-ends are presented in table 2. The goal of multi-condition training is to create matched training/test environments. Although it is expensive to obtain enough representative noisy data to cover a wide range of noise types and signal-to-noise ratios, it is an effective method for mismatch compensation. From table 2 it is evident that the NRMCC, RRMCC, RMCC, and ETSI-AFE yielded lower WERs. In multi-condition training mode the performance of the RMCC is comparable to that of the RRMCC, NRMCC, and ETSI-AFE. Among all the features the NRMCC performed the best (provided lowest WER) both in tables 1 & 2. Comparing the performances of the RFE [6] and that of the RRMCC from both tables, it can be said that use of a triangular-shaped Mel scale filterbank and a RMVDR spectrum estimator instead of a compressive gammachirp filterbank and a DFT (discrete Fourier transform) based direct spectrum estimator provided reduced WER both in clean and multi-condition training modes. Proposed robust feature extractors helped to reduce WER both in clean and multi-condition training modes, though in the multi-condition training mode the reduction in WER is not as huge as observed in the clean training mode. This because multi-condition training data of the AURORA-4 corpus uses the same six noise channels as used in the test sets (as described in section 5.1). There it creates matched training/test environments. Although multi-condition training is an effective method for
mismatch compensation, it has three major problems. Firstly, it is computationally expensive to obtain enough representative noisy data that can cover a wide range of adverse acoustic conditions and SNR levels. Secondly, because of different signal conditions between train and test data, such different noise types and different value of SNRs, there will still be mismatch between train and test conditions. Thirdly, if the training data includes speech with very diverse characteristics the model trained from the training data will not be sharp enough to discriminate various sound classes [58]. This is demonstrated by the higher WER (table 2) obtained by the MFCC front-end on the test set A when multi-condition training is used.
Figure 12: Auditory spectra of a clean speech signal, obtained for (a) DFT spectrum-based MFCC, (b) LP spectrum-based MFCC, (c) MVDR spectrum-based MFCC, (d) RMVDR spectrum-based MFCC, i.e., RMCC, (e) RRMCC, and (f) NRMCC front-ends.
Figure 13: Auditory spectra of a noisy speech signal (babble noise, SNR = 5 dB), obtained for (a) DFT spectrum-based MFCC, (b) LP spectrum-based MFCC, (c) MVDR spectrum-based MFCC, (d) RMVDR spectrum-based MFCC, i.e., RMCC, (e) RRMCC, and (f) NRMCC front-ends.
Figure 14: Auditory spectra of a noisy speech signal (street noise, SNR = 5 dB), obtained for (a) DFT spectrum-based MFCC, (b) LP spectrum-based MFCC, (c) MVDR spectrum-based MFCC, (d) RMVDR spectrum-based MFCC, i.e., RMCC, (e) RRMCC, and (f) NRMCC front-ends.
Table 1. Word error rate (WERs) obtained by the various feature extractors in clean training condition (recognizer is trained on clean data and evaluated on clean as well as noisy data) on the AURORA-4 corpus. The lower the word error rate the better is the performance of the feature extractor. The lowest WERs are highlighted in boldface

<table>
<thead>
<tr>
<th>Feature Extractor</th>
<th>WER (%)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>9.98</td>
<td>50.81</td>
<td>28.88</td>
<td>64.56</td>
</tr>
<tr>
<td>PLP(HTK)</td>
<td>10.28</td>
<td>49.59</td>
<td>25.56</td>
<td>60.36</td>
</tr>
<tr>
<td>MVDR</td>
<td>10.53</td>
<td>47.9</td>
<td>25.49</td>
<td>60.40</td>
</tr>
<tr>
<td>PMVDR</td>
<td>11.31</td>
<td>47.31</td>
<td>23.35</td>
<td>58.92</td>
</tr>
<tr>
<td>RMCC</td>
<td>9.94</td>
<td>45.75</td>
<td>21.77</td>
<td>59.37</td>
</tr>
<tr>
<td>CFCC</td>
<td>13.66</td>
<td>36.95</td>
<td>21.40</td>
<td>45.30</td>
</tr>
<tr>
<td>ETSI-AFE</td>
<td>11.41</td>
<td>30.42</td>
<td>20.48</td>
<td>38.49</td>
</tr>
<tr>
<td>PNCC</td>
<td>11.36</td>
<td>30.15</td>
<td>18.93</td>
<td>40</td>
</tr>
<tr>
<td>RRMCC</td>
<td>10.71</td>
<td>28.80</td>
<td>18.12</td>
<td>37.12</td>
</tr>
<tr>
<td>NRMCC</td>
<td>11.38</td>
<td><strong>26.78</strong></td>
<td><strong>17.53</strong></td>
<td><strong>35.88</strong></td>
</tr>
</tbody>
</table>

Figure 15: Relative error reduction obtained by the proposed regularized RMVDR cepstral coefficients (RMCC) features with respect to the MFCC, PLP, MVDR and PMVDR feature extractors under clean condition training mode on the AURORA-4 LVCSR task.
Figure 16: Relative reduction word error rate obtained by the proposed robust regularized RMVDR cepstral coefficients (RRMCC) features over the other feature extractors considered in this paper under clean condition training mode on the AURORA-4 LVCSR task.

Figure 17: Relative reduction of word error rate obtained by the proposed normalized regularized RMVDR cepstral coefficients (NRMCC) features over the other feature extractors considered in this paper under clean condition training mode on the AURORA-4 LVCSR task.
Table 2. Word error rate (WERs) obtained by the various feature extractors in multi-condition training condition (i.e., recognizer is trained on (clean + noisy) data and evaluated on clean as well as noisy data) on the AURORA-4 corpus. The lowest WERs are highlighted in boldface. The lower the word error rate the better is the performance of the feature extractor.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>14.62</td>
<td>23.84</td>
<td>19.19</td>
<td>31.47</td>
<td>22.28</td>
</tr>
<tr>
<td>PLP(HTK)</td>
<td>16.10</td>
<td>24.98</td>
<td>18.27</td>
<td>30.23</td>
<td>22.40</td>
</tr>
<tr>
<td>MVDR</td>
<td>14.25</td>
<td>23.58</td>
<td>17.86</td>
<td>30.78</td>
<td>21.62</td>
</tr>
<tr>
<td>PMVDR</td>
<td>15.65</td>
<td>23.94</td>
<td>18.78</td>
<td>29.76</td>
<td>22.03</td>
</tr>
<tr>
<td>RMCC</td>
<td>13.81</td>
<td>23.81</td>
<td>17.53</td>
<td>30.45</td>
<td>21.40</td>
</tr>
<tr>
<td>CFCC</td>
<td>18.45</td>
<td>29.53</td>
<td>20.77</td>
<td>35.06</td>
<td>25.95</td>
</tr>
<tr>
<td>ETSI-AFE</td>
<td>14.55</td>
<td>23.32</td>
<td>18.31</td>
<td>29.68</td>
<td>21.47</td>
</tr>
<tr>
<td>PNCC</td>
<td>15.29</td>
<td>25.81</td>
<td>17.86</td>
<td>31.33</td>
<td>22.57</td>
</tr>
<tr>
<td>RRMCC</td>
<td>13.37</td>
<td>23.16</td>
<td>18.27</td>
<td>30.62</td>
<td>21.36</td>
</tr>
<tr>
<td>NRMCC</td>
<td>15.58</td>
<td>22.28</td>
<td>17.94</td>
<td>28.23</td>
<td>21.01</td>
</tr>
</tbody>
</table>

6. Conclusions

New acoustic front-ends, namely, RMCC, RRMCC and NRMCC, are proposed for robust large vocabulary continuous speech recognition tasks. The RMCC front-end incorporates a regularized MVDR spectrum estimator, instead of the DFT-based direct spectrum estimator, in the MFCC feature extraction framework. The regularization constant $\lambda$ used in the RMVDR spectrum estimation method helps to penalize rapid changes in all-pole spectral envelopes, thereby producing better spectral estimates over time without affecting the formant positions. In addition to the RMVDR spectrum estimator, RRMCC and NRMCC front-ends also incorporate auditory domain spectrum enhancement methods, auditory spectrum enhancement (ASE) and medium duration power bias subtraction (MDPBS) techniques, respectively, for robust feature extraction. Experimental speech recognition results on the AURORA-4 LVCSR task under clean as well as multi-condition training mode, plotted spectra and spectrograms of the noisy speech signals showed that the proposed feature extraction techniques provided
significant reduction in word error rate and increased robustness with respect to some classical and recent robust front-ends, specifically, when there is mismatch between training and test conditions. On the average both RRMCC and NRMCC front-ends demonstrated lower word error rates than the other considered methods in clean condition training. Proposed robust feature extractors RMCC, RRMCC, and NRMCC helped to reduce WER both in clean and multi-condition training modes, though in the multi-condition training mode the reduction in WER is not as large as observed in the clean training mode. In terms of the average WER, the NRMCC features performed the best among all the feature extractors in both training modes on the AURORA-4 corpus. Our future work will be to evaluate the performance of feature extractors used in this work by using the i-vector-based speaker adaptation and DNN-HMM (Deep Neural Networks - Hidden Markov Model) hybrid architecture [50] using the KALDI recognizer [54] on the REVERB Challenge [55] and the AURORA-4 [26] corpora. It has been shown in [50-53] that, compared to the GMM-HMM recognizer, the DNN-HMM system provides an absolute reduction of WER by 6-8%.

Acknowledgement

The authors would like to thank the reviewers for their relevant and constructive remarks, which have enabled us to significantly improve the quality of the paper.

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.


