SPEECH RECOGNITION USING REGULARIZED MINIMUM VARIANCE DISTORTIONLESS RESPONSE SPECTRUM ESTIMATION-BASED CEPSTRAL FEATURES

Md Jahangir Alam[1, 2], *Patrick Kenny*[2], *Douglas O'Shaughnessy*[1]

[1]INRS-EMT, University of Quebec, Montreal, Canada
[2]CRIM, Montreal, Canada

ABSTRACT

This paper presents regularized minimum variance distortion-less response (MVDR)-based cepstral features for robust continuous speech recognition. The mel-frequency cepstral coefficient (MFCC) features, widely used in speech recognition tasks, are usually computed from a direct spectrum estimate, that is, the squared magnitude of the discrete Fourier transform (DFT) of speech frames. Direct spectrum estimation methods (also known as nonparametric estimators) perform poorly under noisy and adverse conditions. To reduce this performance drop we propose to increase robustness of the speech recognition system by extracting more robust features based on the regularized MVDR technique. The proposed method, when evaluated on the AURORA-4 speech recognition task, provides an average relative improvement in word accuracy of 11.3%, 6.1%, and 5.2% over the conventional MFCC, PLP, MVDR and PMVDR-based MFCC features, respectively.

Index Terms— Speech recognition, spectrum estimation, regularized MVDR, linear prediction

1. INTRODUCTION

Speech spectrum estimation is a key first step in most feature extraction methods, e.g., MFCC or PLP, for speech recognition. The Mel-frequency cepstral coefficients (MFCCs) computed from a short-time direct spectrum estimate are the widely used feature set and have been empirically observed to be most effective for speech recognition, specifically, under controlled environments [1]. Parametric as well as nonparametric methods of spectrum estimation have been studied for modeling speech signals. Nonparametric spectrum estimators, such as a discrete Fourier transform (DFT)-based periodogram or modified periodogram, are attractive as these estimators are completely independent of data and therefore do not suffer from problems arising from modeling deficiencies. However, these methods are not robust and therefore show poor performance in noisy and adverse conditions. Among the parametric spectrum estimators, the LPC (linear predictive coding)-based all-pole spectrum estimator is most widely used [1]. However, the LP-based cepstra are known to be very sensitive to noise. They tend to overestimate or overemphasize sparsely spaced harmonic peaks [7]. The standard feature extractors used for speech recognition are based on either DFT, e.g., MFCC or linear prediction, e.g., PLP. Both of the feature extractors are either not robust and therefore show poor performance under noisy and adverse conditions, such as MFCC, or ill-suited for the reliable estimation of the spectra of the speech signals, which is true for all methods using linear prediction envelopes [7]. 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 [2]. It is also known as Capon's method [1], for all pole modeling of speech.

In this paper we propose to replace the traditionally used feature extractors by a feature extractor that is based on the regularized minimum variance distortion-less response (MVDR) spectral estimator. The MVDR spectral estimation overcomes the problems apparent in linear prediction spectral estimation and a regularization parameter penalizes rapid changes in all-pole spectral envelopes, thereby producing smooth spectra without affecting the formant positions [5, 6]. The MVDR spectral estimator has already been applied in speech recognition [4] and speaker identification [7] tasks. An extension of the MVDR method was proposed in [3] by warping the frequency axis with the bilinear transformation prior to MVDR spectral estimation. The perceptually motivated MVDR (PMVDR) front-end, proposed in [15], completely eliminates the auditory filterbank processing step and directly performs warping on the DFT power spectrum. In [5], the regularized LP (RLP) has been applied for speaker recognition. To the best of our knowledge, regularized MVDR has not been applied to any recognition tasks. Experimental results on the AURORA-4 continuous speech recognition task show that the regularized MVDR-based MFCC features outperform the MFCC, PLP and MVDR-based MFCC features.

2. SPECTRUM ESTIMATION

Spectrum estimators are classified as parametric and nonparametric. The Discrete Fourier transform (DFT)-based

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

periodogram is an example of a nonparametric estimator and the LPC-based spectrum estimator is a parametric method.

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

$$\hat{S}_{DFT}(f) = \left|\sum_{j=0}^{N-1} w(j)s(j)e^{-\frac{i2\pi fk}{N}}\right|^2, \qquad (1)$$

where f denotes the discrete frequency index, N is the frame length, $j \in \{0,1,...,N\text{-}1\}$ 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 [13]:

$$s(n) = \sum_{q=1}^{p} a_q s(n-q) + e(n), \qquad (2)$$

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

$$S_{LP}(f) = \frac{1}{\left|1+\sum_{q=1}^{p} a_q e^{-i2\pi fq}\right|^2}. \qquad (3)$$

In the autocorrelation method the predictor coefficients $\{a_q\}_{q=1}^{p}$ is expressed as a solution of (2) as [14]:

$$\mathbf{a}_{opt}^{LP} = -\mathbf{R}_{\mathbf{LP}}^{\mathbf{-1}}\mathbf{r}_{\mathbf{LP}}, \qquad (4)$$

where $\mathbf{R}_{LP}$ and $\mathbf{r}_{\mathbf{LP}}$ represents the Toeplitz autocorrelation matrix and autocorrelation vector, respectively.

2.3. MVDR spectrum estimation

The Minimum Variance Distortionless Response (MVDR) spectrum estimator, introduced by Capon [1], is mostly used in array signal processing applications, and has also been investigated in relation to other applications such as speech modeling [2], robust speech recognition [4], and speaker recognition [5] systems. The MVDR spectrum is given by

$$S_{MVDR}(f) = \frac{1}{\mathbf{v}^{\mathbf{H}}(f)\mathbf{R}_{p+1}^{-1}\mathbf{v}(f)}, \qquad (5)$$

where $\mathbf{R}_{p+1}$ is the autocorrelation matrix, $\mathbf{v}(f)$=[1 $e^{-i2\pi f}$ $e^{-i4\pi f}$ ... $e^{-i2\pi pf}$] is a frequency tuning vector with $\mathbf{v}^{\mathbf{H}}(f)$ denoting its conjugate transpose. The model order p corresponds to the largest correlation lag in the autocorrelation matrix. Eqn. (5) represents the power obtained by averaging several samples at the output of the optimum constrained filter. This averaging results in a reduction of the spectral estimator variance [4, 14].

For computational purpose, the *p*th order MVDR spectral estimate can be parametrically written as

$$S_{MVDR}(f) = \frac{1}{\sum_{k=-p}^{k=p} \mu(k) e^{-i2\pi fk}}, \qquad (6)$$

where the parameter $\mu(k)$ of the MVDR method can be directly obtained using a non-iterative computation based on the LPC technique as:

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

where a_q is the LP coefficients and σ_e is the residual variance.

From (6), the MVDR spectral estimator can also be viewed as an all-pole model based spectrum estimator. The MVDR all-pole filter is stable and causal and can be used in a manner similar to the way in which LP filters are used in speech processing systems.

2.4. Regularized MVDR spectrum estimation

Similar to the MVDR spectrum estimator, the *p*th order regularized MVDR spectral estimate can be parametrically written as

$$S_{regMVDR}(f) = \frac{1}{\sum_{k=-p}^{k=p} \mu_{reg}(k) e^{-i2\pi fk}}, \qquad (7)$$

where the parameter $\mu_{reg}(k)$ of the regularized MVDR method can be obtained from a non-iterative computation using the regularized LP (RLP) coefficients a_q^{reg} and the prediction error variance σ_e^{reg} as:

$$\mu_{reg}(k) = \begin{cases} \frac{1}{\sigma_e^{reg}}\sum_{q=0}^{p-k}(p+1-k-2q)a_q^{reg} a_{q+k}^{reg*}, & \text{for } k \geq 0 \\ \mu_{reg}^*(-k), & \text{for } k < 0. \end{cases}$$

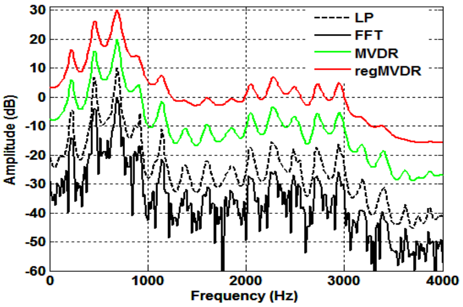


Figure 1 Comparison of the estimated short-term spectra of a noisy speech signal frame (street noise 5 dB) using various spectrum estimators. For better visualization the spectra in each plot is shifted by 10 dB. Model order used is p = 100. The value for the regularization parameter λ used for the regularized MVDR (regMVDR) estimator is 10^{-9}.

In RLP method, the predictor coefficients a_q^{reg} are computed by adding a penalty measure, which is a function of the unknown predictor coefficients a^r ($\psi(a^r)$), to the objective function of the LP method and therefore minimizing that modified objective function of the following form [5, 6]

$$\sum_{n}\left(s(n) + \sum_{q=1}^{p} a_q s(n-q) \right)^2 + \lambda \psi\left(a^r\right), \qquad (8)$$

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. For more detail about the RLP method please see [5, 6]. Fig. 1 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, regularized MVDR method provides smooth spectral estimate.

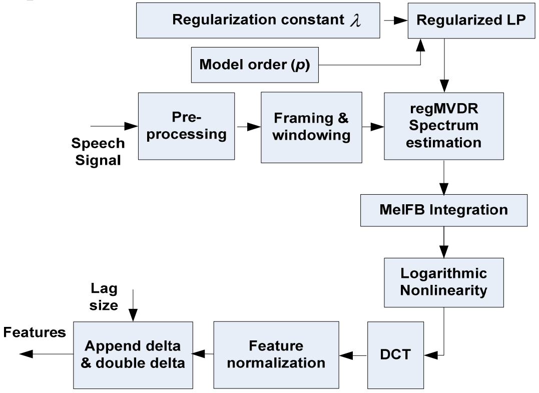


Figure 2 Block diagram of the regularized MVDR spectrum-based MFCC feature extraction process. Here we chose p = 100, and $\lambda_{opt} = 10^{-9}$ based on the recognition experiments.

3. EXPERIMENTS

The proposed feature extractor, as presented in fig. 2, is evaluated and compared with other feature extractors, namely, conventional MFCC, PLP and MVDR-based MFCC, on the AURORA-4 corpus in the context of speech recognition. Note that, for the extraction of PLP features, we have followed HTK-based processing [11], that is, for the auditory spectral analysis a Mel filterbank is used instead of a trapezoid-shaped bark filterbank. All feature extractors considered in this paper are implemented using the rastamat toolbox [10].

3.1. Experimental set-up

The AURORA-4 continuous speech recognition corpus, derived from the Wall Street Journal (WSJ0) corpus, is divided into 3 sets, namely, training, 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. 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 secondary microphone was selected to form the two eval 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 [8]. Each of the test sets contains 166 utterances from 8 speakers, totaling 20.69 minutes of speech data. The 14 test sets are grouped into the following 4 families [8, 9]: (a) Test set A - clean speech in training and test, same channel (set 1), (b) Test set B - clean speech in training and noisy speech in test, same channel (sets 2-7), (c) Test set C - clean speech in training and test, different channel (set 8), and (d) Test set D - clean speech in training and noisy speech in test, different channel (sets 9-14). The number inside the brackets 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 [8, 9].

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, extracted features are normalized using the conventional mean and variance (MVN) normalization technique over the whole utterance.

3.2. Results and discussion

In order to verify the effectiveness of the regularized MVDR-based feature extractor, speech recognition experiments are conducted on the AURORA-4 large vocabulary continuous speech recognition (LVCSR) corpus. Percentage word accuracy is used as a performance evaluation measure for comparing the recognition performances of the proposed method 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. Fig. 3 illustrates the influence of the model order on the spectral estimate of the speech signal. It is observed from fig. 3 that a higher model order provides more detail of the fine structure of the spectrum and represents the 1st 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 [7]. For our work the optimal model order is found to be p = 100.

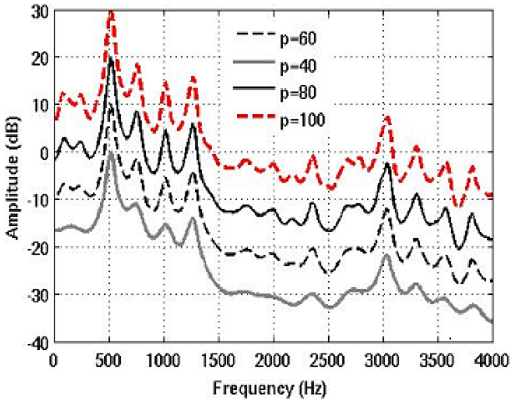
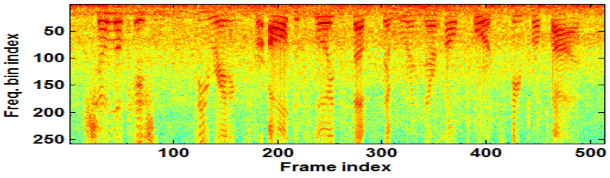
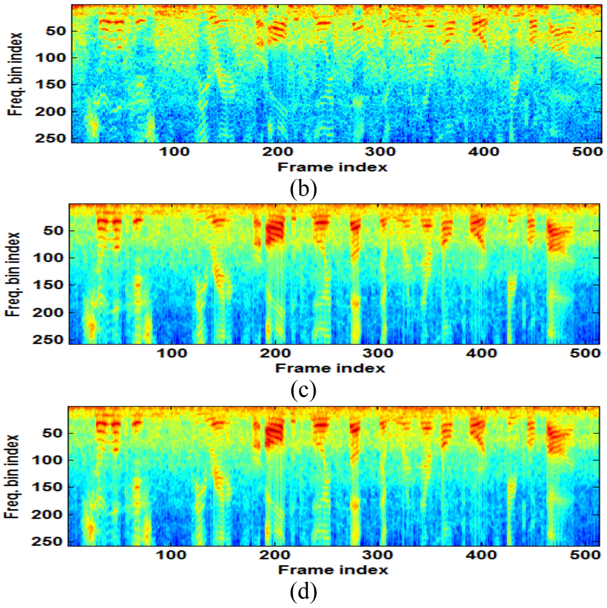


Figure 3 Influence of the model order p on the speech spectrum for the regularized MVDR method. The value for the regularization parameter λ is 10^{-9}.

Fig. 4 presents the speech spectrograms of a noisy speech signal, corrupted with the street noise (SNR = 5 dB), obtained by the various spectrum estimators. It is observed from this figure that compared to the other estimators, both the MVDR and regularized MVDR methods result in a reduction of the noise while preserving the formant structure. We chose the optimal value for the regularization constant λ of the regularized MVDR method that provided the highest word accuracy on the dev-test set of the same corpus. The optimal value for the regularization constant is found to be $\lambda_{opt} = 10^{-9}$. 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. The regularized MVDR spectral estimator-based feature extractor outperformed the other feature extractors in terms of the recognition word accuracy. The average relative improvements obtained by the proposed feature extractor in recognition word accuracy over the conventional MFCC, PLP, and MVDR-based MFCC feature extractors are 11.3%, 6.1%, and 5.2%, respectively.



(a)



(b)

(c)

(d)

Figure 4 Speech spectrograms, street noise, SNR = 5 dB, (a) DFT-based periodogram, (b) LP, (c) MVDR, and (d) regularized MVDR spectrum estimators.

Table 1 Word accuracies (%) obtained by the various feature extractors on the AURORA-4 corpus. The higher the word accuracy the better is the performance of the feature extractor.

	Word Accuracy (%)				
	A	**B**	**C**	**D**	**Avg.**
MFCC	90.02	49.19	71.12	35.44	61.44
PLP(HTK)	89.72	50.41	74.44	39.64	63.55
MVDR	89.47	52.10	74.51	39.60	63.92
regMVDR	**90.06**	**54.25**	**78.23**	40.63	**65.79**

4. CONCLUSION

A robust feature extraction method for the large vocabulary continuous speech recognition (LVCSR) is described. The method incorporates the regularized MVDR spectrum estimator in the MFCC feature extraction framework. A regularization parameter used in this method helps to penalize the rapid changes in all-pole spectral envelopes, thereby producing smooth spectra without affecting the formant positions. Experimental results on the AURORA-4 LVCSR corpus showed that the proposed feature extractor gave significant improvement in word accuracy over the baseline methods.

Our possible future works are:

- Adaptive selection of the regularization constant λ .
- Incorporation of this regularized MVDR spectrum estimator in the feature extraction framework of [12] and [16].

5. REFERENCES

[1] J. Capon, “High-resolution frequency - ‘wavenum-ber spectrum analysis,” Proc. IEEE, vol. 57, pp. 1408–1418, Aug. 1969.

[2] M.N. Murthi and B.D. Rao, “All-pole modeling of speech based on the minimum variance distortionless response spectrum,” IEEE Trans. Speech Audio Processing, vol. 8, no. 3, pp. 221–239, May 2000.

[3] M.C. Wolfel and J.W. McDonough, “Minimum variance distortionless response spectral estimation, review and refinements,” IEEE Signal Processing Magazine, vol. 22, no. 5, pp. 117–126, Sept. 2005.

[4] S. Dharanipragada, B. D. Rao, "MVDR based Feature Extraction for Robust Speech Recognition”, *Proc. ICASSP*, pp. 309-312, 2001.

[5] C. Hanilci, T. Kinnunen, F. Ertas, R. Saeidi, J. Pohjalainen, P. Alku, “Regularized All-Pole Models for Speaker Verification Under Noisy Environments”, *IEEE Signal Processing Letters* 19(3), 163--166, March 2012.

[6] M. N. Murthi and W. B. Kleijn, “Regularized linear prediction all-pole models,” in *IEEE Speech Coding Workshop*, 2000, pp. 96–98.

[7] M. Wolfel, Q. Yang, Q. Jin, T. Schultz, "Speaker Identification using Warped MVDR Cepstral Features," Proc. Interspeech, pp. 912-915, 2009.

[8] N. Parihar, J. Picone, D. Pearce, H.G. Hirsch, "Performance analysis of the Aurora large vocabulary baseline system," Proceedings of the European Signal Processing Conference, Vienna, Austria, 2004.

[9] S.-K. Au Yeung, M.-H. Siu, "Improved performance of Aurora-4 using HTK and unsupervised MLLR adaptation," Proceedings of the Int. Conference on Spoken Language Processing, Jeju, Korea, 2004.

[10] Daniel P. W. Ellis, " PLP and RASTA (and MFCC, and inversion) in Matlab," . online:http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/.

[11] S. J. Young et al., HTK Book, Entropic Cambridge Research Laboratory Ltd., 3.4 edition, 2006. online: http://htk.eng.cam.ac.uk/.

[12] M. J. Alam, P. Kenny, D. O'Shaughnessy, "Robust Feature Extraction for Speech Recognition by Enhancing Auditory Spectrum," Proc. INTERSPEECH, Portland Oregon September 2012.

[13] Douglas O'Shaughnessy, Speech communications - human and machine, 2. ed., IEEE Press, I-XXV, pp. 1-547, 2000.

[14] Djuric, P. M., Kay, S. M., Spectrum Estimation and Modeling, Digital Signal Processing Handbook, CRC Press LLC, 1999.

[15] Umit H. Yapanel, John H.L. Hansen, “A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition,” Speech Comm., Vol. 50, pp. 142-152, 2008.

[16] C. Kim and R. M. Stern., “Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring,” In IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 4574-4577, March 2010.