ENERGY-CONSTRAINED MINIMUM VARIANCE RESPONSE FILTER FOR ROBUST VOWEL SPECTRAL ESTIMATION

Colin Vaz, Andreas Tsiartas, and Shrikanth Narayanan

Ming Hsieh Department of Electrical Engineering
University of Southern California, Los Angeles, CA 90089
<cvaz,tsiartas>@usc.edu, shri@sipi.usc.edu

ABSTRACT

We propose the energy-constrained minimum-variance response (ECMVR) filter to perform robust spectral estimation of vowels. We modify the distortionless constraint of the minimum-variance distortionless response (MVDR) filter and add an energy constraint to its formulation to mitigate the influence of noise on the speech spectrum. We test our ECMVR filter on a vowel classification task with different background noises at various SNR levels. Results show that vowels are classified more accurately in certain noises using MFCC and PLP features extracted from the ECMVR spectrum compared to using features extracted from the FFT and MVDR spectra.

Index Terms: frequency estimation, MVDR, robust signal processing, spectral estimation.

1. INTRODUCTION

Spectral modeling is a fundamental tool in speech processing and is a key building block in many applications, ranging from phonetic speech analysis to speech coding and automatic speech recognition (ASR). Good modeling of the speech spectrum allows one to capture properties of speech, such as pitch (fundamental frequency of the glottis) and formant frequencies (resonant frequencies of the vocal tract) across various speech sounds and speaking conditions. The classical approaches are based on (short time) Fourier analysis or Linear Prediction (LP) models. These methods, however, are sensitive to noise in the signal; real-world speech processing often has to contend with noise corruption.

The computation of the (Fourier) speech spectrum is an essential step in the early part of the processing pipeline in contemporary ASR systems, notably in the calculation of the popular Mel frequency cepstral coefficients (MFCCs) [1]. Limitations in the estimation of the (noisy) speech spectrum can influence the quality of the features, such as MFCCs, that are estimated from it. Likewise, Linear Prediction, a powerful speech modeling approach at the heart of many speech coders, approximates the speech spectrum with the frequency response of an all-pole filter [2]. However, the filter tends to overshoot the peaks of the formant frequencies, especially for female and child speech. Furthermore, LP-based cepstra are sensitive to noise, thus degrading the performance of an ASR that relies on LPC for spectral estimation [3]. Finally, perceptual linear prediction (PLP), yet another popular speech feature extraction method for ASR, relies on both a Fourier analysis step followed by linear prediction modeling [4]. As with MFCC and LPC features, PLP features can also benefit from improved spectral estimation. In summary, given the fundamental nature of speech spectral estimation across various applications, it is desirable to make the basic underlying spectral estimation more robust to noise.

Kleiner et al. tackled the problem of robust spectral estimation of time-series data in [5]. Motivated by overcoming the sensitivity of second-order statistics to outliers, they developed two methods of spectral estimation to be used in conjunction with a prewhitening filter. The first method is a robust filtering algorithm using robust estimates of the autoregressive/linear predictor coefficients and an estimate of the innovation variance to obtain residual coefficients, which are used in the autoregressive model to compute an estimate of the data. The second approach is to fit an autoregressive model to the data and operate on the resulting residuals. They point out that outliers in the data, even minor ones, can significantly alter the spectrum because their effect may be large relative to the power in small peaks. This has important implications for spectral estimation of vowels because formant frequencies, which appear as peaks in the spectrum, distinguish one vowel from another. Therefore, noise in the speech may obscure the formant frequencies and hinder analysis, such as vowel classification, or degrade the performance of a system, such as an ASR, that relies on non-robust spectral estimation.

Murthi et al. proposed the minimum-variance distortionless response (MVDR) filter to improve the modeling of the spectral envelope of speech [6]. Its formulation prevents overestimation of spectral peaks, making it better for estimating the spectrum of medium- to high-pitched speech. As with LPC, noise affects proper spectral estimation of speech using MVDR. We propose a reformulation to MVDR that reduces the influence of noise in frequency regions outside of human speech and constrains the energy of the filter to reduce the effect of additive noise on the speech. This way, we draw on MVDR's ability of undistorted spectral estimation while reducing the effects of noise on spectral estimation of speech.

The paper is organized as follows. Section 2 describes the MVDR formulation and describes our proposed modifications to MVDR to create the energy-constrained minimum-variance response (ECMVR) filter. Section 3 shows results from an isolated vowel classification task when we subject the vowels to different noises at different SNR levels. In Section 4, we discuss the results of the classification task and point out the noises in which ECMVR works well and the ones where the performance is not satisfactory. Finally, we state our conclusions and future work in Section 5.

2. ENERGY-CONSTRAINED MINIMUM-VARIANCE RESPONSE

MVDR has its roots in array signal processing, where one forms a beam at a certain angle to receive signals without distortion at that angle and suppress signals arriving at the microphone array from other angles [7]. Murthi et al. applied this concept to model desired frequencies of a signal [6]. They created a filter that passes a certain frequency of interest without distortion and suppresses the other frequencies of the signal. This property is useful for finding the formant frequencies in a speech segment, and thus enables bet-

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

ter vowel recognition (speech sounds with a rich formant structure) when extracting features from the MVDR spectrum.

MVDR constrains an FIR filter $h[n]$ to have a unity gain at a frequency of interest ω_k. Namely,

$$\left|H\left(e^{j\omega_k}\right)\right| = \left|\sum_{n=0}^{N-1} h[n]e^{-j\omega_k n}\right| = 1. \tag{1}$$

This is the distortionless constraint, and it can be written in vector form as $\left|\boldsymbol{v}^H(\omega_k)\boldsymbol{h}\right| = 1$, where $\boldsymbol{h} = [h[0]\cdots h[N-1]]^T$ and $\boldsymbol{v}(\omega_k) = \left[1\ e^{j\omega_k}\cdots e^{j\omega_k(N-1)}\right]^T$. The MVDR filter with a unity gain at ω_k is obtained by solving

$$\hat{\boldsymbol{h}}_{\omega_k} = \underset{\boldsymbol{h}_{\omega_k}}{\arg\min}\ \boldsymbol{h}_{\omega_k}^H R\boldsymbol{h}_{\omega_k} \quad \text{subject to} \quad \left|\boldsymbol{v}^H(\omega_k)\boldsymbol{h}_{\omega_k}\right| = 1, \tag{2}$$

where R is the $N \times N$ Toeplitz autocorrelation matrix of the input signal. The solution to this constrained optimization problem is

$$\hat{\boldsymbol{h}}_{\omega_k} = \frac{R^{-1}\boldsymbol{v}(\omega_k)}{\boldsymbol{v}^H(\omega_k)R^{-1}\boldsymbol{v}(\omega_k)}. \tag{3}$$

With this $\hat{\boldsymbol{h}}_{\omega_k}$, one can compute the MVDR power spectrum at ω_k as

$$\begin{aligned} P(\omega_k) &= \hat{\boldsymbol{h}}_{\omega_k}^H R\hat{\boldsymbol{h}}_{\omega_k} \qquad (4)\\ &= \frac{1}{\boldsymbol{v}^H(\omega_k)R^{-1}\boldsymbol{v}(\omega_k)}. \end{aligned}$$

We note that it is not necessary to calculate a different $\hat{\boldsymbol{h}}_{\omega_k}$ for each ω_k. Instead, we form a $N \times N$ matrix $V = [\boldsymbol{v}(\omega_0)\ \boldsymbol{v}(\omega_1)\cdots\boldsymbol{v}(\omega_{N-1})]$ and compute the MVDR power spectrum for all frequencies as

$$P(\omega) = \frac{1}{\text{diag}(V^H R^{-1} V)}, \tag{5}$$

where $\text{diag}(X)$ returns a vector of the diagonal elements of a square matrix X, and the division is element-wise.

Noise adds extra energy to the signal and introduces frequency components outside what is deemed as meaningful range for speech intelligibility, typically covering the first three formants. Schafer and Rabiner show that the first three formants range from 200 Hz to 3000 Hz [8]. To combat noise outside of this frequency range, we changed the distortionless constraint from an all-pass filter to a band-pass filter. This way, the MVDR filter passes frequencies of interest undistorted and attenuates out-of-band frequencies. Thus, we modify the distortionless constraint to $|\boldsymbol{v}^H(\omega_k)\boldsymbol{h}_{\omega_k}| = |A(\omega_k)|$, where $A(\omega)$ is a band-pass filter with a passband from 200 Hz to 4000 Hz. The formulation is general; the values can be adjusted based on the frequency range of interest. To deal with noise within the frequency band, we impose an energy constraint on the MVDR filter to prevent noise from adding too much energy to the spectrum. We formulate the energy constraint as $\boldsymbol{g}^H\boldsymbol{h}_{\omega_k} = \beta_{\omega_k}$, where $\boldsymbol{g}$ is a vector of coefficients that tries to approximate the MVDR filter coefficients $\boldsymbol{h}_{\omega_k}$. We use this constraint as an approximation for the true energy of the filter $\boldsymbol{h}_{\omega_k}^H\boldsymbol{h}_{\omega_k} = \beta_{\omega_k}$ because we want to keep the MVDR formulation linear in $\boldsymbol{h}_{\omega_k}$. We solve for β_{ω_k} as follows:

$$\begin{aligned} \beta_{\omega_k} &= \boldsymbol{g}^H\boldsymbol{h}_{\omega_k} &(6)\\ &= \frac{1}{N}\boldsymbol{g}^H\boldsymbol{v}(\omega_k)\boldsymbol{v}^H(\omega_k)\boldsymbol{h}_{\omega_k} &(7)\\ &= \frac{1}{N}\boldsymbol{g}^H\boldsymbol{v}(\omega_k)A(\omega_k) &(8)\\ &= \frac{1}{N}|A(\omega_k)|^2 &(9) \end{aligned}$$

where N is the length of the filter. We use Parseval's theorem in Equation 7 and substitute the constraint $\boldsymbol{v}^H(\omega_k)\boldsymbol{h}_{\omega_k} = A(\omega_k)$ in Equation 8. We approximate $\boldsymbol{g}^H\boldsymbol{v}(\omega_k)$ with $A^*(\omega_k)$ in Equation 9 because we want $\boldsymbol{g}$ to approximate $\boldsymbol{h}_{\omega_k}$ as closely as possible. Hence, we set $\boldsymbol{g}$ to be the impulse response of $A(\omega)$. In our experiments, $A(\omega)$ is a least-squares FIR filter with a passband of 200 Hz to 4000 Hz and a filter length of $N = 24$. In summary, the band-pass constraint tries to assign a certain amount of energy to each frequency while the total energy constraint limits the amount of energy given to all frequencies. With these constraints, we solve for the ECMVR filter to get

$$\hat{\boldsymbol{h}}_{\omega_k} = \frac{(\alpha\text{gRg} - \beta_{\omega_k}\text{vRg})R^{-1}\boldsymbol{v}(\omega_k) - (\beta_{\omega_k}\text{vRv} - \alpha\text{gRv})R^{-1}\boldsymbol{g}}{\text{gRg} \times \text{vRv} - \text{gRv} \times \text{vRg}} \tag{10}$$

where $\alpha = |A(\omega_k)|$, $\text{gRg} = \boldsymbol{g}^H R^{-1}\boldsymbol{g}$, $\text{vRg} = \boldsymbol{v}^H(\omega_k)R^{-1}\boldsymbol{g}$, $\text{vRv} = \boldsymbol{v}^H(\omega_k)R^{-1}\boldsymbol{v}(\omega_k)$, and $\text{gRv} = \boldsymbol{g}^H R^{-1}\boldsymbol{v}(\omega_k)$. We solve for the ECMVR power spectrum at ω_k to get

$$\begin{aligned} P(\omega_k) &= \hat{\boldsymbol{h}}_{\omega_k}^H R\hat{\boldsymbol{h}}_{\omega_k} \qquad (11)\\ &\propto \frac{1}{|\text{gRg} \times \text{vRv} - \text{gRv} \times \text{vRg}|^2}. \end{aligned}$$

For comparative illustration of the ECMVR spectrum with the FFT and MVDR spectra, we calculated these spectra for one 20-ms frame in the middle of three vowels: "aa", "iy", and "uw". The left column of Figure 1 shows these spectra for the clean vowels and the right column shows them for the same vowels with additive 0 dB speech babble. Speech babble has energy in a similar frequency range as speech and also has formant frequencies. Thus, speech babble can considerably alter the speech spectrum and obscure important details about the speech, such as the formants. When comparing the clean and noisy spectra in Figure 1, one can see that the FFT and MVDR spectra are affected by the speech babble, whereas the ECMVR spectrum is affected relatively less. Therefore, features of noisy signals extracted from the ECMVR spectrum, such as MFCC or PLP, will better match the features extracted from the clean signals. This property could lead to better ASR performance. Also, one can see that the formant frequencies are more clearly shown in the MVDR and ECMVR spectra than in the FFT spectrum. This means that these spectra can be useful in speech analysis, such as analyzing formants and classifying vowels.

3. VOWEL CLASSIFICATION EXPERIMENT

To test and analyze the performance of ECMVR, we ran an isolated vowel classification experiment. Since the focus of the proposed model is spectral modeling, analysis of vowels provides an ideal test study. We obtained vowels from the TIMIT database. We chose the TIMIT database because it is phonetically balanced, provides phoneme-level ARPAbet annotations, includes time stamps for the onset and offset of the phonemes, and is widely used for phoneme classification experiments. Moreover, the data is clean, offering good references for systematic comparisons against various noisy versions. There are 15 monophthong and diphthong vowels in the (American English) ARPAbet. We extracted these vowels from the TIMIT sentences. We windowed each vowel segment with a 20 ms Hamming window with 10 ms shift and calculated the FFT, MVDR, and ECMVR spectra of each windowed frame. From these spectra, we extracted 13-dimensional MFCC and PLP features and calculated the delta and delta-delta features. We did not use energy for the first MFCC coefficient. For comparison purposes, we also extracted the LP coefficients. We normalized the features using a diagonal covariance matrix.

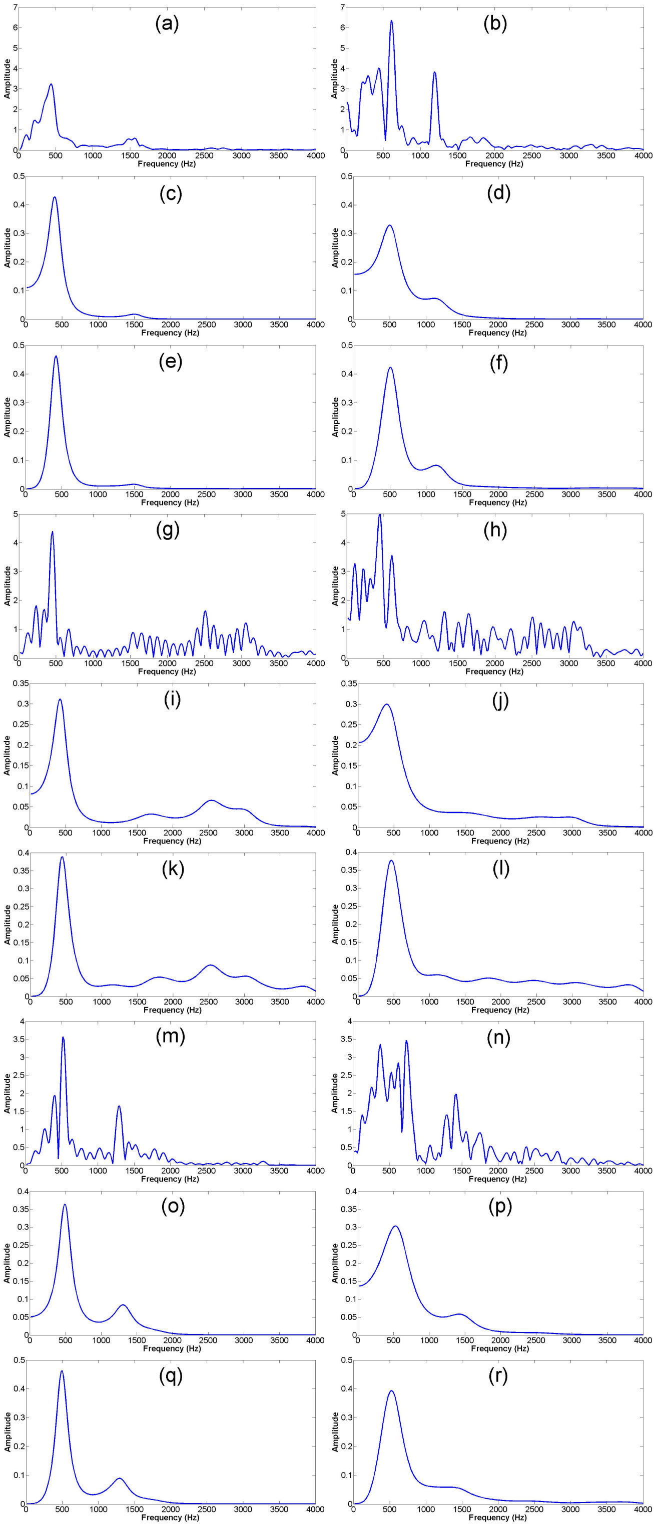
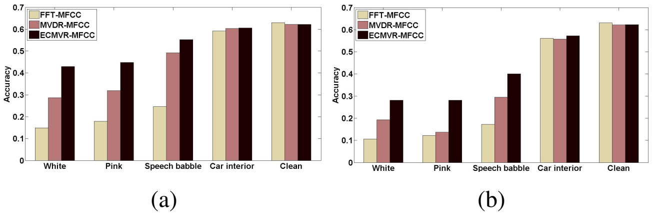


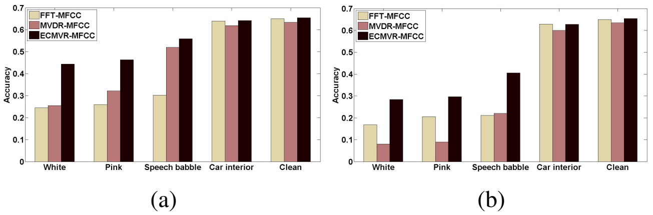
Fig. 1: Male speaking vowels "aa" (a–f), "iy" (g–l), and "uw" (m–r) in clean condition (left column) and 0 dB speech babble (right column). Each vowel is shown in the FFT spectrum (a,b,g,h,m,n), MVDR spectrum (c,d,i,j,o,p), and ECMVR spectrum (e,f,k,l,q,r).

We classified the features using K Nearest Neighbors (KNN). We extracted training features from clean vowels in TIMIT's training set. We extracted testing features from vowels in TIMIT's testing set, to which we added white noise, pink noise, car interior noise, and speech babble from the NOISEX database at 5 dB and 0 dB SNR levels. Thus, we performed classification with mismatched features because of the different noise conditions between training and testing vowels and because the same speaker does not appear in both the train and test sets. We performed a 10-fold cross-validation on the test set, using a 80%/20% split for each phoneme in the test set. In each fold, we found the K that maximized the weighted classification accuracy of the features from ECMVR, and we used this K to determine the weighted accuracy on the held-out set. For comparison purposes, we also computed the accuracy for clean vowels in the test set (matched condition). Figure 2 shows the mean weighted vowel classification accuracy for female speakers in different noise conditions using MFCC features. Figure 3 shows the same plots for male speakers. Figure 4 shows the mean weighted classification accuracy for female speakers when extracting PLP features from the different spectra. Figure 5 shows the same information for male speakers.



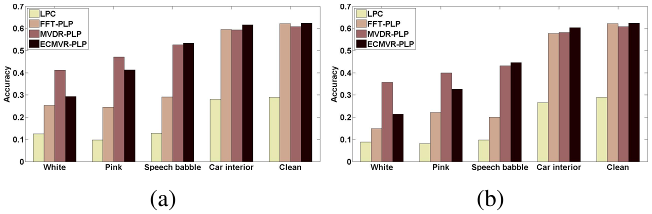
(a) (b)

Fig. 2: Weighted classification accuracy with MFCC features for female speakers in different noises and SNRs of (a) 5 dB and (b) 0 dB.



(a) (b)

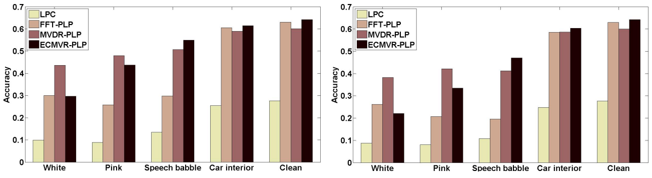
Fig. 3: Weighted classification accuracy with MFCC features for male speakers in different noises and SNRs of (a) 5 dB and (b) 0 dB.



(a) (b)

Fig. 4: Weighted classification accuracy with PLP features for female speakers in different noises and SNRs of (a) 5 dB and (b) 0 dB.

Unlike [9] and [10], we do not do dimension reduction on the features, use the context of surrounding phonemes, or perform speaker normalization in this classification experiment. We also do



(a) (b)

Fig. 5: Weighted classification accuracy with PLP features for male speakers in different noises and SNRs of (a) 5 dB and (b) 0 dB.

not use a language model to constrain the vowel classification, as is done in [11]. These techniques will most likely improve classification performance, but we did not do these because the focus of this paper is on robust spectral estimation, not phoneme classification. However, we intend to follow up on this work with large scale ASR experiments that take into account acoustic and language context.

We used the Wilcoxon rank-sum statistical test, a non-parametric version of the Student's T-test, to determine if the accuracy results for ECMVR are statistically significantly better than the results from the other spectra. For the MFCC features, the accuracy for ECMVR is significantly better at the 95% level than the accuracies for the other spectra in white, pink, and babble noises. For the PLP features, ECMVR performed significantly better than MVDR in speech babble and car interior noise but worse than MVDR in white and pink noises, both at the 95% level.

4. DISCUSSION

The idea behind doing the vowel classification experiment with mismatched conditions was to determine how well the ECMVR spectrum can reduce feature mismatch. The more similar the estimated spectra of a given phoneme in clean and noisy conditions, the more similar the extracted features will be, thereby reducing feature mismatch and improving classification performance. For MFCC features, one can see in Figures 2 and 3 that ECMVR boosted the classification accuracy over the other spectra. Moreover, ECMVR performed consistently well for a wide range of noise types: wideband, such as white and pink noises; narrowband, such as car interior noise; and non-stationary, such as speech babble. This suggests that ECMVR can improve spectral estimation of speech in a variety of noisy situations.

For PLP features, one can see in Figures 4 and 5 that the classification accuracies of vowels from MVDR-PLP features were higher than LPC features, especially for female speakers. This result corroborates the claims in [3] and [6] that MVDR models the speech spectrum better than LPC, particularly for high-pitched speech. ECMVR boosted the classification accuracies for speech babble and car interior noise with PLP features but decreased the classification performance in white and pink noises. White and pink noises have energy at all frequencies in the spectrum. To meet the energy constraint, the ECMVR filter sometimes adds the noise energy into passband region of the ECMVR spectrum if the energy in the speech signal is too low (remember that the modified distortionless constraint created a band-pass filter from 200 Hz to 4 kHz). This addition shows up as ripples in the passband region of the spectrum. Figure 6 shows an example of the ripples in the ECMVR spectrum due to white noise. Unlike MFCC, PLP does equal-loudness preemphasis that boosts frequencies from 400 Hz to 1200 Hz [4]. The equal-loudness preemphasis further exaggerates the ripples. This results in a mismatch between the clean and noisy spectra, especially in the higher frequencies, leading to poorer classification performance. This effect is more pronounced in female speakers because the higher fundamental frequency of the speech shows up as harmonic peaks in the spectrum, which exaggerates the ripples even more.

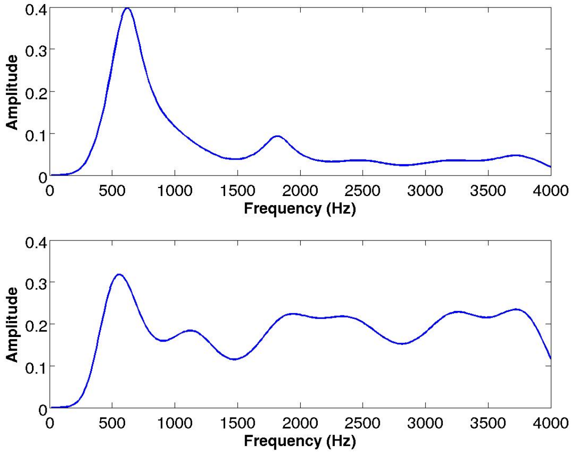


Fig. 6: ECMVR spectra for a female speaking vowel "ae" in clean condition (top) and 0 dB white noise (bottom). The bottom figure shows the ripples that white noise introduces into the ECMVR spectrum, causing a deviation between the clean and noisy spectra and increasing feature mismatch.

5. CONCLUSION

We have presented the ECMVR filter for robust spectral estimation of speech in the presence of noise. We evaluated its performance on modeling and classifying vowels from noisy audio. We modified the distortionless constraint of the MVDR filter into a band-pass filter to handle noise at frequencies outside the range of typical human speech and added an energy constraint to handle noise at frequencies within this range. ECMVR produces spectra of noisy speech that closely matches the clean spectra. Using ECMVR in the front-end of an ASR system can improve overall ASR accuracy by reducing the mismatch between features in the noisy test set and the clean train set. Preliminary experiments on isolated vowel classification show that features extracted from the ECMVR spectrum classify vowels better than FFT-based MFCC. For PLP features, the performance was improved for certain noise types while degraded for others; this is attributed to the nonlinear transformations of the signal in the PLP processing.

To further improve our method, we will investigate filtering the autocorrelation matrices R to see if information in the time domain can improve ECMVR. We will reformulate our energy constraint to deal with low SNR speech in broadband noise. We will explore using a perceptually-motivated filter in place of the band-pass filter constraint, like the equal-loudness filter used in PLP. Additionally, we will fine tune our approach for continuous phoneme classification and apply it to a full-fledged ASR system.

6. REFERENCES

[1] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," in *Proc. Joint Workshop on Pattern Recognition and Artificial Intelligence*, Hyannis, MA, 1976, pp. 374–388.

[2] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoustical Society of America*, vol. 50, no. 2B, pp. 637–655, Aug. 1971.

[3] S. Dharanipragada and B. D. Rao, "MVDR based feature extraction for robust speech recognition," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, 2001, pp. 309–312.

[4] H. Hermansky, "Perceptual linear prediction (PLP) analysis of speech," *J. Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[5] B. Kleiner, R. D. Martin, and D. J. Thomson, "Robust estimation of power spectra," *J. Royal Statistical Society*, vol. 41, no. 3, pp. 313–351, 1979.

[6] M. N. Murthi and B. D. Rao, "Minimum Variance Distortionless Response (MVDR) Modeling of Voiced Speech," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997, pp. 1687–1690.

[7] J. Capon, "High-Resolution Frequency-Wavenumber Spectrum Analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[8] R. W. Schafer and L. R. Rabiner, "System for Automatic Formant Analysis of Voiced Speech," *J. Acoustical Society of America*, vol. 47, no. 2B, pp. 634–648, 1970.

[9] H. M. Meng and V. W. Zue, "Signal representation comparison for phonetic classification," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Toronto, Canada, 1991, pp. 285–288.

[10] S. A. Zahorian, P. Silsbee, and X. Wang, "Phone classification with segmental features and a binary-pair partitioned neural network classifier," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997, pp. 1011–1014.

[11] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.