View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by epublications@Marquette

Marquette University

e-Publications@Marquette

Electrical and Computer Engineering Faculty Research and Publications

Electrical and Computer Engineering, Department of

10-3-2008

Minimum Mean-Squared Error Estimation of Mel-Frequency Cepstral Coefficients Using a Novel Distortion Model

Kevin M. Indrebo
Crabel Capital Management

Richard J. Povinelli
Marquette University, richard.povinelli@marquette.edu

Michael T. Johnson
Marquette University, mike.johnson@marquette.edu

Follow this and additional works at: https://epublications.marquette.edu/electric_fac

 Part of the Computer Engineering Commons, and the Electrical and Computer Engineering Commons

Recommended Citation

Indrebo, Kevin M.; Povinelli, Richard J.; and Johnson, Michael T., "Minimum Mean-Squared Error Estimation of Mel-Frequency Cepstral Coefficients Using a Novel Distortion Model" (2008). *Electrical and Computer Engineering Faculty Research and Publications*. 106.
https://epublications.marquette.edu/electric_fac/106

Marquette University

e-Publications@Marquette

Electrical and Computer Engineering Faculty Research and Publications/College of Engineering

This paper is NOT THE PUBLISHED VERSION; **but the author's final, peer-reviewed manuscript.** The published version may be accessed by following the link in the citation below.

IEEE Transactions on Audio, Speech, and Language Processing, Vol. 16, No. 8 (November 2008): 1654 - 1661. DOI. This article is © The Institute of Electrical and Electronics Engineers and permission has been granted for this version to appear in e-Publications@Marquette. The Institute of Electrical and Electronics Engineers does not grant permission for this article to be further copied/distributed or hosted elsewhere without the express permission from The Institute of Electrical and Electronics Engineers.

Minimum Mean-Squared Error Estimation of Mel-Frequency Cepstral Coefficients Using a Novel Distortion Model

Kevin M. Indrebo
Department of Electrical and Computer Engineering, Marquette University, Milwaukee, WI
Richard J. Povinelli
Department of Electrical and Computer Engineering, Marquette University, Milwaukee, WI
Michael T. Johnson
Department of Electrical and Computer Engineering, Marquette University, Milwaukee, WI

Abstract:

In this paper, a new method for statistical estimation of Mel-frequency cepstral coefficients (MFCCs) in noisy speech signals is proposed. Previous research has shown that model-based feature domain enhancement of

speech signals for use in robust speech recognition can improve recognition accuracy significantly. These methods, which typically work in the log spectral or cepstral domain, must face the high complexity of distortion models caused by the nonlinear interaction of speech and noise in these domains. In this paper, an additive cepstral distortion model (ACDM) is developed, and used with a minimum mean-squared error (MMSE) estimator for recovery of MFCC features corrupted by additive noise. The proposed ACDM-MMSE estimation algorithm is evaluated on the Aurora2 database, and is shown to provide significant improvement in word recognition accuracy over the baseline.

SECTION I. Introduction

Robustness to additive noise remains a largely unsolved problem in automatic speech recognition research today. Various approaches to combating degradation of recognition performance due to noise distortion have been studied [1]–[2][3][4][5], with some level of success. Many of the approaches to building noise-robust recognition systems can be classified into one of three primary categories: back-end adaptation techniques, front-end enhancement algorithms, and alternative feature approaches. The first of these classes focuses on adapting acoustic model parameters to better match the environmental conditions present. The other approaches concentrate the effort on signal parameterization. Enhancement algorithms attempt to remove the noise distortion either from the acoustic signals directly or from the features extracted from the signals. The well-known Ephraim–Malah filter [6] is an example of such an algorithm, as are Bayesian cepstral estimation models [2]. Systems that take the third approach attempt to extract features that are affected less by the noise than traditional features such as Mel-frequency cepstral coefficients (MFCCs). Often the novel features are used in conjunction with the standard feature set. Examples of features studied include frequency subband features and coefficients derived from the phase of the signals [7]–[8][9].

Some of the more successful approaches taken to date have attempted to estimate true clean speech features given a noisy speech signal, often in the log spectral domain. In this domain, the interaction between the speech and noise signals is nonlinear, resulting in high complexity of compensation models even when the speech and noise signals are assumed independent. A common method for dealing with this issue involves the use of a Taylor series expansion to make the compensation algorithm tractable [3], [10], [11]. As a result, the reliability of the estimator depends on the choice of an expansion point. Because the optimal expansion point is not known *a priori*, the algorithm may become iterative. A method for finding a reasonable initial expansion point is still required.

In this paper, these issues are addressed with the introduction of a minimum mean-square error (MMSE) estimator of Mel-frequency cepstral coefficients (MFCCs). The estimation procedure is noniterative and requires no Taylor series approximation. Additionally, the estimator works entirely in the cepstral domain, without the need for an inversion of the discrete cosine transform (DCT). The estimator is developed using a novel approach to modeling the interaction between speech and noise. As a result, the new method models the noise distortion as additive in the cepstral domain, leading to a closed-form solution to the estimation problem. The model is developed using filter bank energy coefficients of the speech and noise signals to match the computation of MFCC features. These coefficients are assumed to be Gamma distributed. In addition, the distortion of the cepstral coefficients is assumed to have a Gaussian distribution. These assumptions, which are discussed in the following section, lead to a tractable solution for the estimator. The proposed estimator performs as the front-end parameterizer to a speech recognition system. Recognition experiments run over speech signals corrupted by various nonstationary noises at multiple signal-to-noise (SNR) ratios are used to demonstrate the efficacy of the proposed approach. The proposed estimator is compared to traditional baselines, in which no noise removal is implemented, and to the well-known Vector Taylor Series (VTS) algorithm [11]. A theoretical comparison of

the proposed estimator and the VTS algorithm is given in the Appendix, highlighting the differences between the two front-ends.

The rest of this paper is structured as follows. In SectionII, the new distortion model is presented, and the MMSE estimator for the MFCC's is derived. SectionIII discusses the practical issues of the algorithm used for robust recognition, followed by a presentation of experimental validation of the given method. In the final section, a discussion of the new noise compensator appears, along with comments on the future directions of this work.

SECTION II. MMSE Estimation of MFCC Features

The MMSE estimator is found using the mean of the conditional distribution of the clean (desired) cepstral coefficients given the distorted values, as

$$\hat{\mathbf{c}} = E[\mathbf{c}|\mathbf{d}]$$

(1)

where $\mathbf{c}$ is the vector of clean cepstral coefficients and $\mathbf{d}$ is the vector of distorted coefficients. Using the definition of the mean and Bayes' theorem, this can be computed by

$$\hat{\mathbf{c}} = \frac{\int_{-\infty}^{\infty} \mathbf{c} p(\mathbf{d}|\mathbf{c}) p(\mathbf{c}) d\mathbf{c}}{p(\mathbf{d})}.$$

(2)

Previous research has used a Gaussian mixture model (GMM) to represent the prior distribution, $p(\mathbf{c})$, [2] and that approach is used in this work as well. This GMM is built by training over a large set of clean speech and helps mitigate any undesired distortion to the features caused by poor estimates of the noise signal present in the corrupted speech signals. A new distortion model is proposed here to represent the conditional distribution.

A. Novel Statistical Distortion Model

The proposed additive cepstral distortion model (ACDM) is derived by representing the true speech spectral (filter bank) coefficients as a function of the distorted spectral coefficients and a gain vector, i.e.,

$$\mathbf{x} = \mathbf{g} \cdot \mathbf{y}$$

(3)

where $\mathbf{x}$ and $\mathbf{y}$ are the clean and distorted speech filter bank energy coefficient vectors for a frame of speech, $\mathbf{g}$ is the appropriate gain vector, and • represents element-wise multiplication. In the log domain, the relationship becomes

$$\ln(\mathbf{x}) = \ln(\mathbf{g}) + \ln(\mathbf{y})$$

(4)

where the log operation of a vector is given by

$$\ln(\mathbf{z}) = \begin{bmatrix} \ln(z_0) \\ \ln(z_1) \\ \vdots \\ \ln(z_n) \end{bmatrix}.$$

(5)

Multiplication of both sides of (4) by a discrete cosine transform matrix, $\mathbf{\Lambda}$, results in

$$\mathbf{\Lambda}\ln(\mathbf{x}) = \mathbf{\Lambda}\ln(\mathbf{g}) + \mathbf{\Lambda}\ln(\mathbf{y}).$$

(6)

Since $\mathbf{\Lambda}\ln(\mathbf{x})$ and $\mathbf{\Lambda}\ln(\mathbf{y})$ are, by definition, equivalent to $\mathbf{c}$ and $\mathbf{d}$, respectively, substitution of these terms and rearrangement gives

$$\mathbf{d} = \mathbf{c} - \mathbf{\Lambda}\ln(\mathbf{g})$$

(7)

in which $\mathbf{\Lambda}\ln(\mathbf{g})$ represents the additive distortion in the cepstral domain. The gain variable, $\mathbf{g}$, is treated as a random vector, allowing the form for the conditional distribution $p(\mathbf{d}|\mathbf{c})$ to be found, provided the distribution of $\mathbf{g}$ is known. To ensure that the MMSE estimator of (2) has a closed-form solution, $p(\mathbf{d}|\mathbf{c})$ is assumed to be Gaussian. This assumption can be justified with the use of the central limit theorem [12], as $p(\mathbf{d}|\mathbf{c})$ is formed as a linear combination of random variables that are exponentially beta distributed. The number of variables in the summation that produces the conditional distribution is 23, the size of the filter bank, which is a value that is generally sufficient to produce distributions that are very close to Gaussian [12]. The mean and variance of the conditional can be computed as

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{d}|\mathbf{c}} &= \mathbf{c} - E[\mathbf{\Lambda}\ln(\mathbf{g})] \\ \Sigma_{\mathbf{d}|\mathbf{c}} &= E[\mathbf{\Lambda}\ln(\mathbf{g})^2] - E[\mathbf{\Lambda}\ln(\mathbf{g})]^2 \end{aligned}$$

(8)

where $\Sigma_{\mathbf{d}|\mathbf{c}}$ is a diagonal matrix, since the gain variables are assumed to be independent across frequency bins. The conditional distribution of the gain variables is determined by using a linear MMSE estimator, the Wiener filter, to represent the gain

$$g_k = \frac{x_k}{x_k + n_k}$$

(9)

where x and n are the kth filter bank energy coefficients of the speech and noise signals, respectively. Ideally, x_k is a known quantity, since c is given. However, the values for x cannot be fully recovered from c, as the discrete cosine transform used is not necessarily invertible. A least-squares fit transform of the coefficients back into the log spectral domain is possible, but inclusion of that transform into the estimator would require that the algorithm become iterative. Therefore, as an approximation, x_k and x_k are both treated as random variables, and are assumed to be gamma distributed. Gamma distributions have often been used to model speech time samples and spectra in prior work [13]–[14][15] and an empirical goodness-of-fit test over clean condition training data in the filter bank energy domain confirms that the gamma distribution has a chi-squared

statistic an order of magnitude better than a normal, uniform, exponential, or Rayleigh distribution.
If x_k and x_k are assumed to have independent gamma distributions with parameters $(\alpha_{x,k}, \beta)$ and $(\alpha_{n,k}, \beta)$, respectively, g_k can be shown to be beta distributed, with

$$p(g_k) = \frac{\Gamma(\alpha_{x,k} + \alpha_{n,k})}{\Gamma(\alpha_{x,k})\Gamma(\alpha_{x,k})}(1 - g_k)^{\alpha_{n,k}-1} g_k^{\alpha_{x,k}-1}$$

(10)

where $\alpha_{x,k,}$, and $\alpha_{n,k}$ are parameters derived from the distributions of x and n. Note that the beta value must be equivalent for the two gamma distributions. The distribution for $\ln(g_k)$ is known as the exponential beta distribution, and the mean and variance can be computed by [16]

$$\begin{aligned} \mu_{\ln g_k} &= \psi_0(\alpha_{x,k}) - \psi_0(\alpha_{x,k} + \alpha_{n,k}) \\ \sigma^2_{\ln g_k} &= \psi_1(\alpha_{x,k}) - \psi_1(\alpha_{x,k} + \alpha_{n,k}) \end{aligned}$$

(11)(12)

where ψ_0 and ψ_1 are the digamma and trigamma functions, respectively [17]. The final form of $p(\mathbf{d}|\mathbf{c})$ is then given by

$$p(\mathbf{d}|\mathbf{c}) = N(\mathbf{d}; \mathbf{c} - \mathbf{\Lambda}\mu^g, |\mathbf{\Lambda}|\Sigma^g)$$

(13)

where $\boldsymbol{\mu}^g$ and Σ^g are the mean and variance vectors computed using (11) and (12). Inserting (13) into (2), and using the same procedure found in [2], the MMSE estimator can be fit into a standard quadratic form, and a closed form solution for the estimator can be found as

$$\begin{aligned} \hat{\mathbf{c}} &= \sum_{m=1}^{M} \gamma_m [\mathbf{W}_1(m)\boldsymbol{\mu}_m^c + \mathbf{W}_2(m)(\mathbf{d} + \mathbf{\Lambda}\boldsymbol{\mu}^g)] \\ \mathbf{W}_1(m) &= |\mathbf{\Lambda}|\Sigma^g(\Sigma_m^c + |\mathbf{\Lambda}|\Sigma^g)^{-1} \\ \mathbf{W}_2(m) &= \Sigma_m^c(\Sigma_m^c + |\mathbf{\Lambda}|\Sigma^g)^{-1} \end{aligned}$$

(14)

where $\boldsymbol{\mu}_m^c$ and Σ_m^c are the mean vector and covariance matrix of the GMM used for the prior model $p(\mathbf{c})$ and

$$\gamma_m = \frac{w_m p(\mathbf{d}|m)}{\sum_{m=1}^{M} w_m p(\mathbf{d}|m)}.$$

(15)

As in [2], $p(\mathbf{d}|m)$ is computed by

$$p(\mathbf{d}|m) = N(\mathbf{d}; \boldsymbol{\mu}_m^c + \mathbf{\Lambda}\boldsymbol{\mu}^g, \Sigma_m^c + |\mathbf{\Lambda}|\Sigma^g).$$

(16)

Examination of (14) shows that the final ACDM-MMSE estimator is essentially a weighted average between the mean of the distortion compensation factor $\mathbf{d} + \mathbf{\Lambda}\boldsymbol{\mu}^g$ and the mean of each component in the prior model, where the weighting is determined by the ratio of variances of the conditional and prior distributions. If the prior model was assumed to be uniform, the estimator would essentially be equivalent to applying a Wiener filter, though the computation is performed in the cepstral domain. The inclusion of the prior in the estimator forces the estimate to more closely match a pattern of actual speech.

Because the trigamma function is monotonically decreasing for positive numbers, for a given α_x, the variance computed in (12) increases along with α_n. Thus, as the estimated signal-to-noise ratio decreases, this variance increases, and more weight is given to the prior model in the estimation of the features. This is desirable, as it is expected that the accuracy of our distortion compensation factor will be worse for lower SNR values. The use of the prior model then becomes especially important, so that the negative effects caused by the inaccuracy of the distortion factor are not as detrimental to the estimate of the features and subsequently the robustness of the recognition system.

SECTION III. Algorithm Implementation

The ACDM-MMSE estimator derived in the previous section requires knowledge of the parameters of the distributions of the speech and noise filter bank energy coefficients in order to compute the $\boldsymbol{\mu}^g$ and Σ^g (mean and variance) terms of (14). *A priori* estimates for the speech and noise power are generated using noise estimation and spectral estimation algorithms. The improved minima-controlled recursive averaging (IMCRA) method [18] is used for estimation of the noise power. For the spectral estimate, a decision-directed generalized Wiener filter [19] is implemented. The form of the filter is

$$H(\omega) = \frac{S_x(\omega)}{S_x(\omega) + \rho S_n(\omega)}$$

(17)

where ρ is a multiplier that controls additional noise suppression. The value $\rho = 4$ is chosen based on experiments run over a development set. Because of the increased noise suppression, this filter will sometimes significantly underestimate the speech power. Consequently, the filter in (17) is bounded, resulting in the modified form

$$H(\omega) = \frac{S_x(\omega)}{S_x(\omega) + min\{\rho S_n(\omega), S_y(\omega)\}}$$

(18)

where $S_y(\omega)$ is the spectral power value computed from the distorted speech signal. Additionally, the spectral estimate obtained using this filter is smoothed in frequency using a normalized window. The Wiener filtering and smoothing process is defined by

$$\hat{X}(\omega) = \sum_{i=-l}^{l} b(i)[H(\omega)S_y(\omega)]$$

(19)

where the b coefficients must sum to unity. The value used for l is one.

Once the *a priori* speech and noise estimates are generated, the parameters α_x and α_n can be computed. The values for the noise and speech estimates, which are obtained by application of the IMCRA algorithm and the modified Wiener filter, respectively, are first converted from spectral coefficients to filter bank energy coefficients by applying a Mel-spaced triangular filter bank. The resulting values for the kth filter banks of the speech and noise are treated as the means of the gamma distributions for x_k and x_k. Using the definitions of the mean and variance for a gamma distribution, the alpha parameters can be computed by

$$\alpha_{x,k} = \frac{\hat{x}_k}{\beta} \alpha_{n,k} = \frac{\hat{n}_k}{\beta}$$

(20)

where $\hat{x}_k$ and $\hat{n}_k$ are the *a priori* estimates of the speech and noise filter bank energy coefficients, and β is treated as a free parameter. Once the alpha values are computed, the values for the mean vector $\boldsymbol{\mu}^g$ and the variance matrix Σ^g can then be found using (11) and (12). These mean and variance measures, and subsequently the estimated values for the MFCC features, are affected by the choice of β. It has been observed that the choice of an appropriate β is important for success of the estimation algorithm, and that the computation of the mean is adversely affected by a poor choice of β more so than the variance. Because of this sensitivity, in the implementation of the algorithm the computation of the mean from (11) is replaced by

$$\mu_{\ln g_k} = \log\left(\frac{\hat{x}_k}{\hat{x}_k + \hat{n}_k}\right).$$

(21)

This approach can be viewed as treating the speech and noise *a priori* estimates as deterministic instead of stochastic for the purpose of estimation of the mean of the conditional distribution. However, the variance of the conditional is still derived using the statistical assumptions developed in the previous sections. Empirical observations have indicated that it is beneficial to bound the variance computed in (14) to prevent impact from occasional outliers. Values for β and the upper and lower bounds of the variance of the conditional distribution in the MMSE estimator are chosen to optimize recognition accuracy over a development set, resulting in $\beta = 9000$, and bounds of [1.1, 4.5].

The estimator is implemented to estimate the static cepstral coefficients, including C0. The first and second derivative coefficients are then computed from the estimated features. While it is possible to compute all parameters for (14), including first and second derivatives, it has been observed that doing so provides no benefit in terms of recognition accuracy over the approach of estimating static coefficients only.

While the IMCRA algorithm and decision-directed generalized Wiener filter are used for estimating the noise and speech components, other methods could easily be used, such as minimum statistics [20] or Ephraim–Malah filtering [6]. The proposed estimator is independent of the *a priori* estimators and allows for the inclusion of spectral estimation in a feature domain compensation scheme.

SECTION IV. Speech Recognition Experiments

Table I Average Word Accuracies for Proposed Estimator and Baseline Front-Ends Using Clean-Condition Trained Acoustic Models on Aurora2

Front-end	Set A	Set B	Set C	Overall

ACDM-MMSE	81.64%	83.05%	82.03%	82.24%
No enhancement (baseline 1)	56.56%	52.98%	66.77%	58.77%
CMS enhancement only (baseline 1)	65.17%	70.67%	66.39%	67.41%
VTS (256 mixture prior)	78.88%	80.11%	78.68%	79.22%
VTS (16 mixture prior)	67.00%	70.26%	68.03%	68.43%

The proposed ACDM-MMSE estimator is tested using the Aurora2 database [21]. Aurora2 is a speaker independent database of connected digits, zero through nine, plus “oh.” The data was originally collected under a clean environment, but has been corrupted by various real-world noises at multiple SNR levels. The data has also been dowsampled to 8 kHz, and filtered with either a G712 or MIRS characteristic, depending on the set. Two training sets, clean-condition and multicondition, and three test sets, labeled A, B, and C, are provided. The clean-condition set is left undistorted, while the multicondition set is corrupted with subway, babble, car, and exhibition hall noises, matching the noises in test set A. Test set B is corrupted by restaurant, street, airport, and train station noises. Both training sets and test sets A and B are filtered with the G712 characteristic. Test set C is corrupted with the subway and street noises, but is filtered with the MIRS characteristic to allow for the study of channel distortion. For the experiments presented in this paper, the range of SNR levels used is 0–20 dB.

An HMM is built for each word, each with 16 states and three mixtures per state. A three-state silence model with six mixtures per state is also trained. This results in a total of 163 states and 498 mixtures. The training procedure matches that of the script provided by the Aurora2 database. The speech feature set in all experiments consists of a 39-element vector containing 13 static MFCCs, including C0, along with first and second derivative features. The proposed estimation system is used as a front-end to a standard speech recognition system, which is implemented using Sphinx-4 [22]. The static feature vector estimates are produced by first running the IMCRA noise estimation algorithm and the decision-directed generalized Wiener filter to give *a priori* estimates for the speech and noise filter bank energy components, followed by application of (14). First and second derivative coefficients are then computed from the estimated static coefficients in the standard manner.

Results for two sets of experiments are presented, based on the training set used to build the acoustic models. In the clean-condition trained experiments, all acoustic models, as well as the prior model used in the estimator are trained using the clean-condition training set. In the multicondition training experiments, the prior model is first learned over the clean-condition training data. The proposed estimator is applied to the multicondition training data, resulting in an “enhanced” training set. This data is then used to train the acoustic models for use in recognition experiments for the proposed system. The baseline system is built by training the acoustic models directly on the multicondition training data. Configuration of algorithm parameters described in the previous section is executed using a development set based on the multicondition training set, with all models trained on the clean-condition set.

A summary of the clean-condition training experimental results is found in TableI, along with baseline comparisons. The VTS method [11] is also evaluated and compared to the proposed method. Like the ACDM-MMSE estimator, VTS uses a prior distribution model trained over clean speech, but the feature estimation is done in the log-spectral domain as opposed to the cepstral domain. As is the case for the proposed estimator, the IMCRA algorithm is used to obtain the noise estimate for use with the VTS algorithm. Cepstral mean subtraction (CMS) is applied to the features produced by the ACDM-MMSE front-end as a postprocessing step,

as well as the VTS front-end. These results are compared to two baselines, in which no explicit noise modeling or removal is applied, one with CMS postprocessing and one without.

Based on parameterization tuning experiments on a development set, the number of mixtures used for the ACDM-MMSE prior is 16. Two versions of the VTS method are employed: one with 256 mixtures to match [11] and one with 16 mixtures to compare more closely with the proposed method. Because of the difference in number of mixtures, the ACDM-MMSE estimation algorithm is significantly faster than the 256 mixture VTS algorithm. Recognition experiments for the 256 mixture VTS method run in approximately 2.7× real time as compared to around 0.5× real time for the proposed estimation method. The experimental time for the 16 mixture VTS system is comparable to the ACDM-MMSE system.

The proposed estimator outperforms all baselines, including both versions of VTS. Inspection of the results for each of the test subsets (by noise type and SNR) for clean-condition training indicates that the proposed system gives superior performance over the both the standard feature set and VTS estimated features in all SNR levels 15 dB or lower and nearly equal numbers at the 20-dB SNR level. To see the effect of the prior distribution, recognition is also run using the modified Wiener filter described in (18) as the front-end. The overall accuracy for this system is 77.33%, showing that inclusion of the prior model results in an absolute error reduction of 4.91%.

Each algorithm is also tested on clean data. The accuracy for the ACDM-MMSE estimation method is 98.50%, compared to 98.97% for the VTS and a baseline of 99.12%. While the proposed method causes some degradation in accuracy on clean data, the amount is relatively small.

Results for the multicondition experiments are presented in TableII. The relative improvement seen in these experiments is smaller than that of the clean-condition experiments, but the improvement seen is still consistent. A modified Wiener front-end system gives an overall accuracy of 89.27% here, showing that inclusion of the GMM prior model results in an absolute error reduction of 0.47%. The VTS algorithm does not perform well on this task, actually decreasing the word accuracy in comparison to the baseline.

Table II Average Word Accuracies for Proposed Estimator and Baseline Front-Ends Using Multicondition Trained Acoustic Models on Aurora2

Front-end	Set A	Set B	Set C	Overall
ACDM-MMSE	89.80%	89.54%	89.87%	89.74%
No enhancement (baseline 1)	87.28%	85.75%	84.79%	85.94%
CMS enhancement only (baseline 1)	88.63%	88.66%	89.24%	88.84%
VTS (256 mixture prior)	85.04%	84.73%	85.42%	85.06%
VTS (16 mixture prior)	81.33%	82.22%	81.99%	81.85%

As stated in the previous section, the ACDM-MMSE estimator has a free parameter β which controls the scaling of the conditional variance. To study the sensitivity of the algorithm to this parameter, a series of clean-condition recognition experiments are run, varying the value for β. A range of values from 10 to 50 000 is used, spaced logarithmically. The minimum and maximum accuracies, averaged over all test sets, are 79.91% and 82.80%. The lowest accuracy is a result of $\beta = 50\,000$, and all other accuracies are within 0.4% of the maximum. This indicates that, provided the β value is not excessively large, the proposed estimator is robust to variances in the actual value.

In addition to the recognition experiments presented, analysis of the error of the MFCC estimates is executed. A relative mean squared error (MSE) is computed for the static coefficients for each frame in all test utterance between the estimated and clean features. The baseline error is computed directly between the corrupted and

original clean features. No CMS is performed in the error computations. A relative MSE value is computed for each SNR in test sets A, B, and C and transformed into log scale by

$$\log \frac{\sum_t \sum_i (c_i - \hat{c}_i)^2}{\sum_t \sum_i c_i^2}$$

(22)

where t is the frame index, c_i is the ith clean cepstral coefficient, and $\hat{c}_i$ is the ith estimated or corrupted coefficient. Figs. 1–3 show the error trends for the baseline and proposed front-ends for test sets A, B, and C, respectively. The ACDM-MMSE front-end MSE is lower in every case, and the relative improvement is quite consistent.

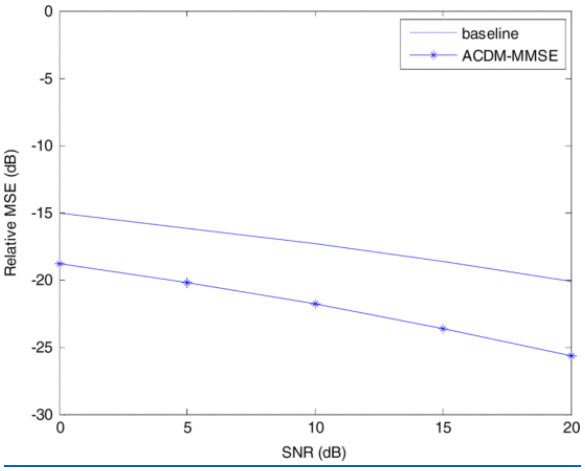

Fig. 1. Relative log mean-squared error of static MFCC features for baseline and proposed front-ends on Test Set A by SNR level.

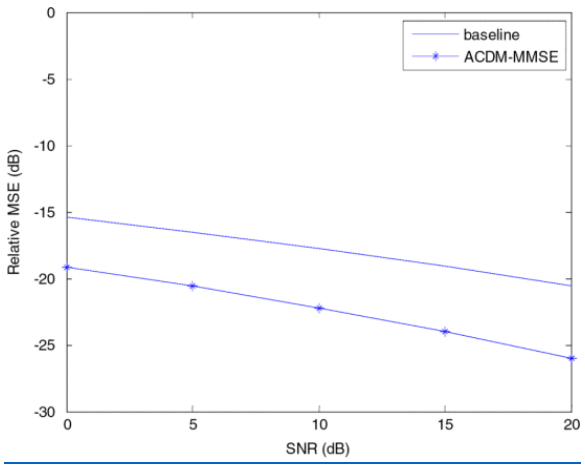

Fig. 2. Relative log mean-squared error of static MFCC features for baseline and proposed front-ends on Test Set B by SNR level.

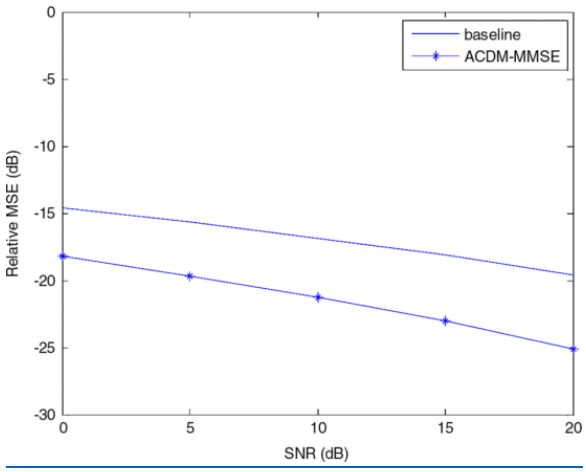

Fig. 3. Relative log mean-squared error of static MFCC features for baseline and proposed front-ends on Test Set C by SNR level.

SECTION V. Discussion

A new method for estimation of MFCC features for use in robust speech recognition has been proposed. This approach models the noise distortion as additive in the cepstral domain, and makes use of assumptions of the statistical distribution of the speech and noise in the spectral domain to derive the MMSE estimator. Unlike some previous approaches to estimation of speech features, the algorithm used is not iterative. Additionally, the estimation is performed entirely in the cepstral domain. Experimental results show significant improvement in word recognition accuracy in noisy connected digit utterances over a baseline system with no feature enhancement.

The success of the estimation algorithm depends primarily on the quality of three components: the *a priori* noise power estimates, the *a priori* speech power estimates, and the cepstral prior model. The IMCRA algorithm is used for the noise estimate, and a generalized Wiener filter is used for estimation of speech. Improvement in these estimation algorithms is likely to lead to improvement in recognition accuracy using ACDM-MMSE estimator.

The prior model used is a simple GMM trained over a large set of clean speech. Its major contribution is to ensure that the enhanced cepstral values are reasonable (i.e., they resemble actual speech). However, the prior model does not differentiate between different classes of phonemes, such as vowels and fricatives. Instead, all frames of speech use the same prior model, which is a conglomeration of different classes of phonemes. If the prior model could be made more specific for each individual frame of speech (i.e., a vowel model used for frames that are likely vowels, a fricative model for frames that are likely fricatives, etc.), it is likely the estimator would produce yet more accurate features. This is the focus of our continuing work.

In this Appendix, a derivation of the VTS-1 estimation equation in the cepstral domain is presented, with the objective of deriving a result for comparison with the proposed ACDM-MMSE estimator. We start with the well-known nonlinear acoustic distortion model

$$\mathbf{y}' = \mathbf{x}' + g(\mathbf{x}', \mathbf{n}'),$$
$$g(\mathbf{x}', \mathbf{n}') = \log\left(\mathbf{i} + e^{\mathbf{n}'-\mathbf{x}'}\right).$$

(23)

Here, $\mathbf{x}'$, $\mathbf{n}'$, and $\mathbf{y}'$ are the clean speech, noise, and corrupted speech log filter bank coefficient vectors, respectively, and i is the identity vector. Equation (23) is expanded around an initial point $\mathbf{x}'_0$ with a first-order Taylor series expansion, using $\mathbf{n}' = \mathbf{n}'_0$, to give

$$\mathbf{y}' = (I + \nabla_{\mathbf{x}'} g(\mathbf{x}'_0, \mathbf{n}'_0))\mathbf{x}' + g(\mathbf{x}'_0, \mathbf{n}'_0) - \nabla_{x'} g(\mathbf{x}'_0, \mathbf{n}'_0)\mathbf{x}'_0.$$

(24)

If both sides of (24) are multiplied by a DCT matrix, Λ, we have (after splitting the first term)

$$\Lambda\mathbf{y}' = \Lambda(I + \nabla_{\mathbf{x}'} g(\mathbf{x}'_0, \mathbf{n}'_0) - I)\mathbf{x}' + \Lambda\mathbf{x}' + \Lambda g(\mathbf{x}'_0, \mathbf{n}'_0) - \Lambda\nabla_{\mathbf{x}'} g(\mathbf{x}'_0, \mathbf{n}'_0)\mathbf{x}'_0.(25)$$

(25)

Which, using $\mathbf{d} = \Lambda\mathbf{y}'$ and $\mathbf{c} = \Lambda\mathbf{x}'$, can be rewritten as

$$\mathbf{d} = \mathbf{c} + \Lambda(\nabla_{\mathbf{x}'} g(\mathbf{x}'_0, \mathbf{n}_0))\mathbf{x}' + \Lambda(g(\mathbf{x}'_0, \mathbf{n}'_0) - \nabla_{\mathbf{x}'} g(\mathbf{x}'_0, \mathbf{n}'_0)\mathbf{x}'_0).$$

(26)

The second term on the right side of (26) can be rewritten as

$$\Lambda(\nabla_{\mathbf{x}'} g(\mathbf{x}'_0, \mathbf{n}'_0))\mathbf{x}' = \Lambda(\nabla_{\mathbf{x}'} g(\mathbf{x}'_0, \mathbf{n}'_0))(\Lambda^{-1}\Lambda)\mathbf{x}'.$$

(27)

We can then represent (26) as

$$\begin{aligned} \mathbf{c} &= \mathbf{d} - \mathbf{A}\mathbf{c} + \mathbf{b} \\ \mathbf{A} &= \Lambda(\nabla_{\mathbf{x}} g(\mathbf{x}_0, \mathbf{n}_0))\Lambda^{-1} \\ \mathbf{b} &= \Lambda(\nabla_{\mathbf{x}} g(\mathbf{x}_0, \mathbf{n}_0)x_0 - g(\mathbf{x}_0, \mathbf{n}_0)). \end{aligned}$$

(28)

The MMSE estimator for c is found by

$$\begin{aligned} \hat{\mathbf{c}}_{\mathrm{MMSE}} &= \int_C \mathbf{c}p(\mathbf{c}|\mathbf{d})d\mathbf{c} \\ &= \int_{\mathbf{C}} \mathbf{d} - \mathbf{A}\mathbf{c} + \mathbf{b}\, p(\mathbf{c}|\mathbf{d})d\mathbf{c} \\ &= \sum_{m=0}^{M} p[m|\mathbf{d}] \int_C \mathbf{d} - \mathbf{A}\mathbf{c} + \mathbf{b}p(\mathbf{c}|\mathbf{d}, m)d\mathbf{c} \end{aligned}$$

(29)(30)

where m is the index of a mixture in a GMM prior model of clean speech. The integral can be split and terms can be rearranged to give

$$\hat{\mathbf{c}} = \sum_{m=0}^{M} P[m|\mathbf{d}] \{-\mathbf{A} \int_{C} \mathbf{c} p(\mathbf{c}|\mathbf{d}, m) d\mathbf{c} + \mathbf{d} \int_{\mathbf{C}} p(c|\mathbf{d}, m) + \mathbf{b} \int_{C} p(\mathbf{c}|\mathbf{d}, m) d\mathbf{c}\}. \quad (31)$$

(31)

Substituting $\int_{\mathbf{C}} \mathbf{c} p(\mathbf{c}|\mathbf{d}, m) d\mathbf{c} = \boldsymbol{\mu}_{\mathbf{c},m}$ and $\int_{\mathbf{C}} p(\mathbf{c}|\mathbf{d}, m) d\mathbf{c} = 1$, (31) can be transformed into

$$\begin{aligned} \hat{\mathbf{c}} &= \sum_{m=0}^{M} \gamma_m \{\mathbf{W}_1 \boldsymbol{\mu}_{\mathbf{c},m} + [\mathbf{d} + \mathbf{f}_0]\} \\ \mathbf{W}_1 &= -\mathbf{A} = (-\Lambda \nabla_{\mathbf{x}'} g(\mathbf{x}'_0, \mathbf{n}'_0) \Lambda^{-1}) \\ \mathbf{f}_0 &= \mathbf{b} = \Lambda (g(\mathbf{x}'_0, \mathbf{n}'_0) - \nabla_{\mathbf{x}'} g(\mathbf{x}'_0, \mathbf{n}'_0) \mathbf{x}'_0) \\ \gamma_m &= P[m|\mathbf{d}]. \end{aligned} \quad (32)$$

(32)

By comparing (14) and (32), we can see that, although the form is similar, the weights on the two components for each mixture m, the prior mean and the enhanced value, are not the same. In the ACDM-MMSE estimator, they will always sum to unity and are based on the relative variances of the two Gaussians (prior and conditional). In the VTS equation, the weight for the prior mean is not based on the variance of the prior or conditional Gaussian and the weights will never sum to unity, since the weight on the enhanced value is already 1. Also, the enhanced values $\mathbf{f}_0$ and $\boldsymbol{\mu}^g$ are computed differently.

References

1. A. Morris, A. Hagen and H. Bourlard, "The full-combination sub-bands approach to noise robust HMM/ANN base ASR", *Eurospeech*, 1999.

2. L. Deng, J. Droppo and A. Acero, "Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features", *IEEE Trans. Speech and Audio Processing*, vol. 12, pp. 218-233, 2004.

3. L. Deng, J. Droppo and A. Acero, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition", *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 568-580, Nov. 2003.

4. C.-H. Lee, "On stochastic feature and model compensation approaches to robust speech recognition", *Speech Commun.*, vol. 25, pp. 29-47, 1998.

5. A. Acero and R. Stern, "Environmental robustness in automatic speech recognition", *ICASSP*, 1990.

6. Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator", *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-32, no. 6, pp. 1109-1121, Dec. 1984.

7. D. Zhu and K. K. Paliwal, "Product of power spectrum and group delay function for speech recognition", *Int. Conf. Acoust. Speech Signal Process. (ICASSP 04)*, 2004.

8. A. Morris, A. Hagen, H. Glotin and H. Bourlard, "Multi-steam adaptive evidence combination for noise robust ASR", *Speech Commun.*, vol. 34, pp. 25-40, 2001.

9.H. Bourlard and S. Dupont, "Subband-based speech recognition", *Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 1997.
10.M. Afify and O. Siohan, "Sequential noise estimation with optimal forgetting for robust speech recognition", *ICASSP*, 2001.
11.P. J. Moreno, *Speech recognition in noisy environments*, 1996.
12.A. Papoulis, Probability Random Variables and Stochastic Processes, New York:McGraw-Hill, 1991.
13.R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors", *Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2002.
14.S. Gazor, "Speech probability distribution", *IEEE Signal Process. Lett.*, vol. 10, no. 7, pp. 204-207, Jul. 2003.
15.J. W. Shin, J.-H. Chang and N. S. Kim, "Statistical modeling of speech signals based on generalized gamma distribution", *IEEE Signal Process. Lett.*, vol. 12, no. 3, pp. 258-261, Mar. 2005.
16.A. K. Gupta and S. Nadarajah, Handbook of Beta Distribution and Its Applications, New York:Marcel Dekker, 2004.
17.J. L. Spouge, "Computation of the gamma digamma and trigamma functions", *SIAM J. Numer. Anal.*, vol. 31, pp. 931-944, 1994.
18.I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging", *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466-475, Sep. 2004.
19.L. Arslan, A. McCree and V. Viswanathan, "New methods for adaptive noise suppression", *ICASSP*, 1995.
20.R. Martin, "Spectral subtraction based on minimum statistics", *Eur. Signal Process. Conf.*, 1994.
21.D. Pearce and H. Hirsch, *"The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions"*, 2000.
22.*"Sphinx-4"*.