On the Inversion of Mel-Frequency Cepstral Coefficients for Speech Enhancement Applications

Laura E. Boucheron and Phillip L. De Leon
Klipsch School of Electrical and Computer Engineering
New Mexico State University
P.O. Box 30001, MSC 3-O
Las Cruces, NM, USA 88003-8001
e-mail: {lboucher, pdeleon}@nmsu.edu

***Abstract*—The use of Mel-frequency cepstral coefficients (MFCCs) is well established in the fields of speech processing, particularly for speaker modeling within a Gaussian mixture model (GMM) speaker recognition system. The use of GMMs for speech enhancement applications has only recently been proposed in the literature; the concept of direct inversion of the MFCCs, however, has not been studied. In this paper we present a means to invert MFCCs for use in speech enhancement applications. Results for cepstral inversion is evaluated on the TIMIT speech corpus using perceptual evaluation of speech quality (PESQ).**

I. Introduction

The cepstral analysis of speech signals is homomorphic signal processing to separate convolutional aspects of the speech production process [1]. The glottal pulse and formant structure of speech contains information important for characterizing individual speakers [2]–[4]; cepstral analysis allows these components to be easily elucidated. As such, cepstral coefficients (CCs) are common features in speaker recognition (SR) research; in particular, Mel-frequency cepstral coefficients (MFCCs), with basis in human pitch perception, are perhaps more common, e.g., [3]–[7]. Additionally, CCs, particularly MFCCs, have shown promise in phonetic recognition applications, e.g., [8]–[10]; this lends credence to the hypothesis that enhanced MFCCs could be used to reconstruct/synthesize clean speech from noise-corrupted speech.

Simultaneously, Gaussian mixture models (GMMs) have been used now for over a decade in SR systems [3], [4]. Due to the non-deterministic aspect of speech (i.e., the actual sound produced for the same sound class will vary from instance to instance), it is desirable to model each sound class as a probability density function [4]. Since Gaussian mixtures can model arbitrary distributions [3], they are well suited to modeling speech for SR systems, whereby each sound class is modeled by one Gaussian component.

The use of cepstral- or GMM-based systems for speech enhancement has only recently been investigated, however. Kundu et al. [11] used a minimum mean-square estimate (MMSE) of clean speech frames given noisy speech frames within a GMM framework. They did not extract any features from the speech, but rather directly used the time-domain speech frames as input to the GMM. Mouchtaris et al. [12] applied MMSE spectral conversion methods to speech enhancement, modeling speech and noise as jointly Gaussian within a GMM framework. The enhanced cepstral coefficients are used within two linear filtering frameworks; i.e., the cepstral feature vectors are not directly used to synthesize the enhanced speech with either an iterative Kalman or iterative Wiener filter. Deng et al. [13] developed closed form solutions (with several assumptions and simplifications) for MMSE of cepstral feature vectors from noisy speech. They use a GMM model of clean speech and use the enhanced cepstral feature vectors in speech recognition rather than for speech enhancement.

The subject of inversion of MFCCs for speech reconstruction has also recently been investigated, specifically within the framework of distributed speech recognition for speech communication in mobile devices which proposes a restricted set of MFCC-based speech features [14]. As such, Shao and Milner [14] worked with MFCCs to reconstruct a smoothed log spectral envelope, and combined this with phase and pitch estimation within a sinusoidal speech synthesis system. No objective measures of speech quality were presented.

In this paper we consider the direct inversion of MFCCs for ultimate use in a GMM-based speech enhancement. Our focus in this paper is the inversion process, not the enhancement aspects, nor any speech communication aspects. Our research differs from previous work in that we look to use cepstral coefficients to directly reconstruct/synthesize the speech. Section II presents the theoretical basis for inversion of CCs and MFCCs, while Section III discusses the perceptual artifacts introduced by the inversion. Section IV presents considerations for use of CCs/MFCCs in speech enhancement and Section V concludes.

II. Inversion of Cepstral Coefficients

We deal with a short-time Fourier transform (STFT) framework, wherein the signal $x(n)$ is windowed

$$\tilde{\mathbf{x}}_\mathbf{k} = \mathbf{x}_\mathbf{k} \cdot \mathbf{w} \tag{1}$$

where $\mathbf{x_k}$ is the k-th frame of signal $x(n)$, $\mathbf{w}$ is the length-K window (often Hamming), and $\cdot$ denotes an element-wise multiplication. These frames may be stacked in a $K \times N$ matrix $\tilde{\mathbf{x}}$, where N is the total number of frames in signal $x(n)$.

A. Cepstrum

We define the cepstrum of signal $x(n)$ as

$$\mathbf{C} = \mathcal{DCT}\left\{\log|\mathcal{DFT}\left[\tilde{\mathbf{x}}\right]|^2\right\} \tag{2}$$

where the discrete cosine transform (DCT) and discrete Fourier transform (DFT) are applied to each column of $\tilde{\mathbf{x}}$. In general, the length of window $\mathbf{w}$, the DFT, and DCT may be different lengths, but we choose (without loss of generality) the same length K for $\mathbf{w}$ and the DFT, and length $K/2+1$ for the DCT (considering only a symmetric half of the DFT).

The inversion of the cepstral coefficients $\mathbf{C}$ is straightforward, since the DCT, DFT, log, and square operations are invertible. $x(n)$ can be reconstructed from $\tilde{\mathbf{x}}$ using the overlap-add (OLA) method [1]. The use of the magnitude of the Fourier transform introduces a complication, however, since the phase information of $x(n)$ is discarded. This issue will be considered in detail in Section II-C.

B. Mel Cepstrum

The computation of the Mel cepstrum applies a weighting to $\tilde{\mathbf{X}} = \log|\mathcal{DFT}\{\tilde{\mathbf{x}}\}|^2$ prior to the DCT operation. This weighting is based on an model of human perception of pitch and is most commonly implemented in the form of a filterbank of triangular filters [5]. For a given number of filters J, the center frequencies of the first $J/2$ filters are linearly spaced from 0 to 1000 Hz, while the remaining are logarithmically spaced in the remainder of the bandwidth. The filters taper to zero at the previous and subsequent center frequencies, and the amplitudes are normalized to have unit energy.

The output of the J Mel filters ϕ_j can be expressed in matrix form as

$$\mathbf{E} = \mathbf{\Phi}\tilde{\mathbf{X}} \tag{3}$$

where $\mathbf{\Phi}$ is $J \times K$ whose j-th row is ϕ_j and $\tilde{\mathbf{X}}$ is $K \times N$. Thus the Mel cepstrum is computed as

$$\mathbf{MC} = \mathcal{DCT}\left\{\mathbf{\Phi}\log|\mathcal{DFT}\left[\tilde{\mathbf{x}}\right]|^2\right\} = \mathcal{DCT}\left\{\mathbf{\Phi}\tilde{\mathbf{X}}\right\}. \tag{4}$$

We choose (again without loss of generality) length K for $\mathbf{w}$ and the DFT, and length J for the DCT.

To invert the Mel weighting, we look for $\mathbf{\Phi}'$ such that

$$\hat{\mathbf{X}} = \mathbf{\Phi}'\mathbf{E} = \mathbf{\Phi}'\mathbf{\Phi}\tilde{\mathbf{X}} \approx \tilde{\mathbf{X}} \tag{5}$$

Defining $\mathbf{\Phi}'$ as the Moore-Penrose pseudoinverse $\mathbf{\Phi}^\dagger$ ($\mathbf{\Phi}^\dagger = \left(\mathbf{\Phi}^\mathsf{T}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\mathsf{T}$ for full rank $\mathbf{\Phi}$), we assure that $\hat{\mathbf{X}}$ is the solution of minimal Euclidean norm [15]. The remaining operations can be inverted without loss as in the straight cepstrum. It is important to note, however, that $\hat{\mathbf{X}}$ will not necessarily be a valid STFT in the sense of having the required constraints of a STFT [1], and the reconstruction of $x(n)$ via inverse DFT and OLA may contain artifacts; these artifacts are directly related to the underconstrained (generally $J < K$) nature of the pseudoinverse $\mathbf{\Phi}^\dagger$.

C. Inversion Considerations for Modified STFTs

The modification of cepstral coefficients results in a modification of the STFT $\tilde{\mathbf{X}}$. As such, the direct inversion of the DFT and OLA reconstruction is no longer valid. The "closest" valid STFT $\tilde{\mathbf{X}}_\mathbf{e}$, in terms of least squared error (LSE) can be obtained via the methods presented in [16]. The first method, inverse STFT (LSE-ISTFT), assumes that the STFT has been modified, but that some valid estimate of the phase exists; the second method, inverse STFT magnitude (LSE-ISTFTM), operates on a modified STFT magnitude and iteratively develops an estimate of a valid STFT including an estimate of phase. For the following experiments, we save the phase information of the STFT prior to the magnitude operation; thus the phase information can be used for the STFT inversion process.

III. Artifactual Effects of the Inversion Process

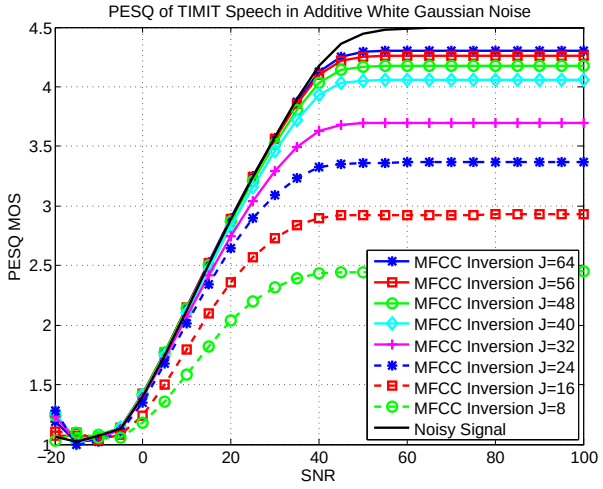
It is expected that the inversion of Mel-frequency cepstral coefficients will introduce some distortion to the speech signal, since the computation of $\tilde{\mathbf{X}}$ from $\mathbf{E}$ is an underconstrained problem. We wish to quantify the perceptual quality artifacts that this inversion process may introduce. With the ultimate goal of leveraging the MFCCs in a speech enhancement scenario, we are interested in the performance of the inversion process for a variety of signal-to-noise ratios (SNRs). To this end, we compute the perceptual evaluation of speech quality (PESQ) metric [17] of various reconstructed signals, using the implementation provided in [18].

We use 10 randomly selected speakers from the TIMIT speech corpus (5 male and 5 female) for evaluation. For each of these speakers, we use 24 seconds of speech utterances. White Gaussian noise is added at various SNRs and the cepstrum of the noisy signal is computed and inverted. For these experiments $K = 320$, corresponding to 20 ms at the 16 kHz sample rate, and a Hamming window with 50% overlap is used. The PESQ mean opinion score (MOS) is computed for the reconstructed signal.

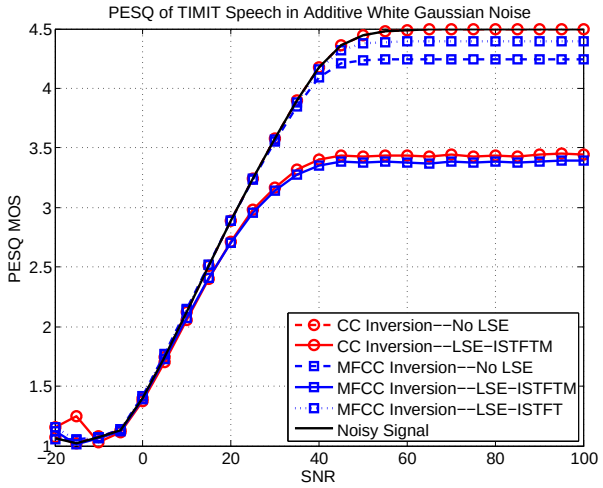
Figure 1 displays the MOS values for MFCC inverted signals for three different reconstructions: (a) directly inverting the STFT without an LSE method (i.e., assume that $\hat{\mathbf{X}}$ is a valid STFT and using the original phase), (b) LSE-ISTFTM, and (c) LSE-ISTFT (using the original phase).

From Figure 1 (a), we note that the underconstrained nature of the Mel cepstrum inversion introduces a degradation of $\sim$ 0.2 MOS points at high SNR for reasonably large J, but these artifacts become masked by the noise below about 30 dB SNR.

In Figure 1 (b), we show PESQ score for LSE estimations of the STFT. The estimation of phase (LSE-ISTFTM) introduces significant perceptual degradation ($\sim$ 1.2 MOS points) which is not masked by the noise until very low levels of SNR ($\sim$ 5 dB). The LSE-ISTFT method using the original phase, on the other hand, improves the PESQ MOS of MFCC inversion to almost the level attainable by the lossless inversion of straight CCs, and the artifacts become masked by the noise below $\sim$40 dB SNR. Thus, with the use of MFCC inversion and LSE-ISTFT with the noisy phase and within a reasonable



(a) Effect of MFCC inversion using the original phase for the inverse DFT, and assuming a valid STFT $\hat{\mathbf{X}}$. Note the degradation of ~ 0.2 MOS points at high levels of SNR for large J; the artifacts become masked by the noise below about 30 dB SNR.



(b) Effect of LSE-ISTFT estimation and LSE-ISTFTM estimation. The LSE-ISTFTM method introduces significant perceptual degradation (~ 1.2 MOS points). LSE-ISTFT, however, increases the MOS to nearly the score attainable with a straight cepstral inversion. $J = 56$ for the MFCC results here, and $K/2 + 1 = 161$ for the CC results.

Fig. 1. Artifacts of the cepstral inversion process, in terms of PESQ mean opinion score (MOS). These results are tallied for 10 randomly chosen TIMIT speakers (5 male and 5 female), and for additive white Gaussian noise at different levels of signal-to-noise ratio (SNR).

operating range of -10 to 40 dB SNR for speech enhancement applications, the artifacts due to the inversion of the Mel cepstrum are negligible.

IV. Use of Cepstral Coefficients in Speech Enhancement

A. On the Use of the Zeroth Coefficient

In SR systems using cepstral coefficients as features for a GMM-based modeling, the zeroth coefficient is generally discarded [3]. In the straight cepstrum the zeroth coefficient is a measure of energy in the signal, which could lend an

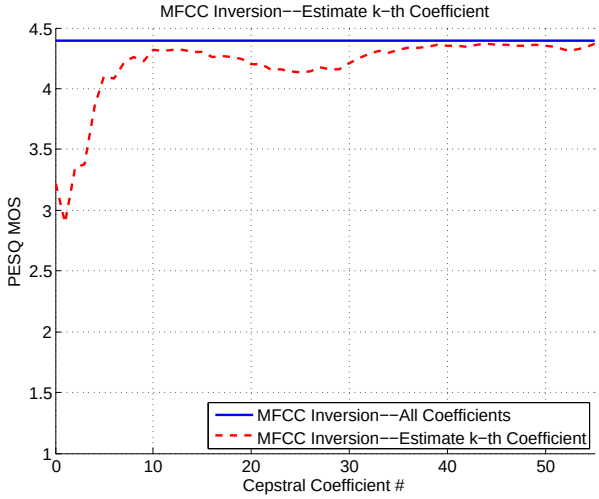


Fig. 2. Relative importance of individual MFCCs, $J = 56$. For these experiments, the k-th cepstral coefficient for each frame is replaced by the mean of that coefficient across all speech frames in the signal. These results are averaged for the same 10 randomly chosen TIMIT speakers. LSE-ISTFT is used for reconstruction.

undesirable bias in the speaker model. However, many GMM-based SR systems append the log energy to the cepstral vector, e.g., [19], which would seem to provide opportunity for an energy-based bias. Even in cepstral-based phoneme recognition systems [9], [10] and distributed speech recognition standards [14] the zeroth coefficient is often discarded.

For speech enhancement systems, where the speech will be reconstructed from the cepstral coefficients, the zeroth coefficient is of utmost importance to the perceptual quality. As an illustration of this effect, we inverted MFCCs for clean speech of the same 10 randomly chosen TIMIT speakers, while estimating the zeroth coefficient as the mean across all the frames. Thus, for each frame of the input signal the zeroth cepstral coefficient is replaced with the mean of the same coefficient as computed across all the speech frames. Reconstruction is achieved via the LSE-ISTFT method. This introduces a degradation of about 1.2 MOS points according to PESQ, as can be seen in Figure 2. While not shown here, similar conclusions hold for the use of the zeroth CC.

It is important to note that we are not simply discarding the zeroth MFCC, but instead are replacing it with the optimal estimate of its value (assuming that we have no observation of the frame-to-frame variation in the coefficient) [20]. Even with this reasonable estimate of the coefficient value, there is significant degradation in PESQ MOS.

B. Relative Importance of Individual Coefficients

We can conduct similar tests replacing other single MFCCs with their mean prior to reconstruction. These results are also plotted in Figure 2. We see a significant degradation in PESQ MOS when we estimate any of the first several (~ 10) MFCCs; less significant degradation occurs for coefficients in the approximate ranges of 20-30 and 50-56. This is not unexpected given the direct correspondence of the initial

cepstral coefficients to formant structure. The source of the smaller degradations are most likely due to the appearance of pitch period (i.e., vocal excitation) information in the range of 60-400 Hz [1], or ∼3-17 ms. For a window of 20 ms as used here, the first dip in PESQ performance is most likely due to the 5 female speakers used in the experiment, while the second dip is due to the male speakers. Overall, it appears the the most important MFCCs for perceptual quality of the reconstructed speech are the first 10 coefficients which are directly related to the glottal pulse and formant structure of the speech [1]. While not plotted here, similar conclusions hold for straight CCs as well.

When viewed in light of a speech enhancement application, it is clear that the first several MFCCs must be appropriately estimated on a frame-by-frame basis if good quality speech is to be reconstructed.

C. Considerations for GMM-based Speech Enhancement

Within the typical expectation maximization (EM) framework for computation of a GMM, the dimensionality of the feature set is of utmost importance for the convergence of the EM algorithm. The actual convergence is dependent on many factors, including the dimensionality, the initialization of the GMM, and the data itself, but it appears that the feature dimensionality has the most effect. We have found that feature dimensions much greater than 64 (with diagonal covariance matrices) tend to cause divergence in the EM algorithm. Thus, for any GMM-based speech enhancement algorith, we are restricted to the use of a feature set of 64 dimensions or less.

As MFCCs have demonstrated great use for characterization of vocal tract configurations, and also provide a convenient means to reduce the dimensionality of the STFT of a speech utterance, it is expected that MFCCs will have great promise for GMM-based speech enhancement.

D. Cepstral Feature Estimation for Speech Enhancement

Within a speech enhancement framework, the goal is to take MFCC feature vectors of noisy speech and estimate the MFCCs of the underlying clean speech while seeking to minimize the perceptual artifacts introduced by the estimation procedure. Since the inversion of MFCCs does not introduce significant artifacts, the problem of speech enhancement with MFCCs becomes solely a problem of estimation.

V. Conclusions

In this paper we have discussed the inversion of Mel-frequency cepstral coefficients. Perceptual artifacts due to this inversion were quantified through the use of the PESQ objective speech quality metric. We have demonstrated that the perceptual artifacts of the MFCC inversion process are negligible in the range of SNR for typical speech enhancement applications. We have also discussed the relative importance of the various individual cepstral coefficients as quantified by PESQ. This has implications for the proper frame-by-frame estimation of cepstrum for reconstruction of enhanced speech. We have also considered the implications of using cepstral coefficients in a GMM-based speech enhancement framework, particularly the dimensionality limitations of the feature set. Continuing work is looking at the process of cepstral feature estimation for speech enhancement.

Acknowledgments

This work was supported through a grant from Rosettex Technology and Ventures Group.

References

[1] T. F. Quatieri, *Discrete Time Speech Signal Processing*. Prentice Hall, 2002.

[2] R. P. Ramachandran, K. R. Farrell, R. Ramachandran, and R. J. Mammone, "Speaker recognition–general classifier approaches and data fusion methods," *Pattern Recognition*, vol. 35, pp. 2801–2821, 2002.

[3] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[4] D. A. Reynolds, "Automatic speaker recognition using Gaussian mixture speaker models," *The Lincoln Laboratory Journal*, vol. 8, no. 2, pp. 173–192, 1995.

[5] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[6] G. Zweig and S. J. Russell, "Speech recognition with dynamic bayesian networks," in *Proc AAAI*, 1998, pp. 173–180.

[7] R. Saeidi, H. R. S. Mohammadi, R. D. Rodman, and T. Kinnunen, "A new segmentation algorithm combined with transient frames power for text independent speaker verification," in *Proc. ICASSP*, vol. IV, 2007, pp. 305–308.

[8] P. Schwarz, P. Matějka, and J. Černocký, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. ICASSP*, vol. I, 2006, pp. 325–328.

[9] M. T. Johnson, R. J. Povinelli, A. C. Lindgren, J. Ye, X. Liu, and K. M. Indrebo, "Time-domain isolated phoneme classification using reconstructed phase spaces," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 458–466, Jul. 2005.

[10] J. Zeng and Z.-Q. Liu, "Type-2 fuzzy hidden Markov models and their application to speech recognition," *IEEE Transactions on Fuzzy Systems*, vol. 14, no. 3, pp. 454–467, Jun. 2006.

[11] A. Kundu, S. Chatterjee, A. S. Murthy, and T. V. Sreenivas, "GMM based Bayesian approach to speech enhancement in signal/transform domain," in *Proc. ICASSP*, 2008, pp. 4893–4896.

[12] A. Mouchtaris, J. Van der Spiegel, P. Mueller, and P. Tsakalides, "A spectral conversion approach to single-channel speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processin*, vol. 15, no. 4, pp. 1180–1193, May 2007.

[13] 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 Transactions on Speech and Audio Processing*, vol. 12, no. 3, pp. 218–233, May 2004.

[14] X. Shao and B. Milner, "Clean speech reconstruction from noisy mel-frequency cepstral coefficients using a sinusoidal model," in *Proc. ICASSP*, vol. I, 2003, pp. 704–707.

[15] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*. SIAM, 2000.

[16] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

[17] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," International Telecommunication Union, Telecommunication Standardization Sector, ITU-T Recommendation P.862, 2001.

[18] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC, 2007.

[19] R. Auckenthaler, E. S. Parris, and M. J. Carey, "Improving a GMM speaker verification system by phonetic weighting," in *Proc. ICASSP*, vol. I, 1999, pp. 313–316.

[20] A. H. Sayed, *Fundamentals of Adaptive Filtering*. John Wiley & Sons, Inc., 2003.