Low-SNR, Speaker-Dependent Speech Enhancement using GMMs and MFCCs

Laura E. Boucheron and Phillip L. De Leon

New Mexico State University, Klipsch School of Elect. and Comp. Eng., Las Cruces, N.M., U.S.A.
{lboucher, pdeleon}@nmsu.edu

Abstract

In this paper, we propose a two-stage speech enhancement technique. In the training stage, a Gaussian Mixture Model (GMM) of the mel-frequency cepstral coefficients (MFCCs) of a user's clean speech is computed wherein the component densities of the GMM serve to model the user's "acoustic classes." In the enhancement stage, MFCCs from a noisy speech signal are computed and the underlying clean acoustic class is identified via a maximum *a posteriori* (MAP) decision and a novel mapping matrix. The associated GMM parameters are then used to estimate the MFCCs of the clean speech from the MFCCs of the noisy speech. Finally, the estimated MFCCs are transformed back to a time-domain waveform. Our results show that we can improve PESQ in environments as low as -10 dB SNR.

Index Terms: Speech enhancement, MFCC, GMM

1. Introduction

Enhancement of noisy speech remains an active area of research due to the difficulty of the problem. Standard methods such as spectral subtraction [1], Wiener filtering [2], minimum mean-square error (MMSE) estimation [3], and generalized subspace [4] can improve perceptual evaluation of speech quality (PESQ) scores but at the expense of other distortions such as musical artifacts. With all of these methods, PESQ can be improved by as much as 0.6 for speech with 10 to 30 dB input SNR. The effectiveness of these methods deteriorates rapidly below 5 dB input SNR.

Gaussian Mixture Models (GMMs) of a speaker's mel-frequency cepstral coefficients (MFCCs) have been successfully used in speaker recognition systems [5]. Due to non-deterministic aspects of speech, it is desirable to model each acoustic class with a Gaussian probability density function [6]. Since GMMs can model arbitrary distributions [5], they are well suited to modeling speech, whereby each acoustic class is modeled by a single component density.

The use of cepstral/GMM-based systems for speech enhancement has only recently been investigated. Compared to most algorithms which do not require clean speech signals for training [1–4], recent research assumes availability of a clean speech signal to build user-dependent models for enhancing noisy speech [7, 8].

In this paper, we propose a two-stage speech enhancement technique which leverages a user's clean speech. In Sections 2 and 3, we provide details of the training and enhancement stages, respectively. In Section 4, we describe the experimental evaluation and provide results and commentary. Finally, in Section 5, we conclude the paper and discuss areas of future research.

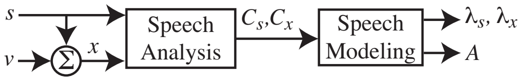


Figure 1: Training stage of proposed speech enhancement system.

2. Training

In the training stage (Fig. 1), we synthesize a noisy speech signal x from a clean speech signal s and a representative noise signal v as

$$x = s + v. \tag{1}$$

In synthesizing x, the noise type (e.g., white) and SNR should be chosen according to the known/anticipated operational environment. Estimation of noise type and SNR can be achieved through analysis of the non-speech portions of the acquired noisy speech signal. In a real-time application, one could create a family of synthesized noisy speech training signals using different noise types and SNRs and select the appropriate noisy speech model based on enhancement performance.

2.1. Speech Analysis

The cepstral analysis of speech signals is homomorphic signal processing to separate convolutional aspects of the speech production process; mel-frequency cepstral analysis has a basis in human pitch perception and is perhaps more common. The glottal pulse (pitch) and formant structure of speech contains information important for characterizing individual speakers [5, 6], as well as for characterizing individual acoustic classes contained in the speech.

We use a 20 ms Hamming window with 50% overlap to compute a 62-dimensional vector of MFCCs C_s, C_x from s, x, respectively. The MFCCs are based on an DFT length of 320 (the window length) and a DCT of length 62 (the number of mel-filters). The mel-scale filters are 20 triangular weighting functions linearly-spaced from 0-1 kHz, 40 triangular weighting functions logarithmically-spaced from 1-8 kHz, and two "half-triangle" weighting functions centered at 0 and 8 kHz. The two "half-triangle" weighting functions improve the quality of the enhanced speech signal by improving the accuracy in the inversion of the MFCC vector to a time-domain waveform.

Algorithm 1 Computation of ACMM A

```
Initialize $A = 0$
for each MFCC vector $C_s$ and $C_x$ do
    $j = \arg\max_i p(i|C_s, \lambda_s)$
    $k = \arg\max_i p(i|C_x, \lambda_x)$
    $A_{j,k} \leftarrow A_{j,k} + 1$
end for
$A_{j,k} \leftarrow A_{j,k} / \sum_{i=1}^{M} A_{i,k}$ for $1 \leq j, k \leq M$
```

2.2. Signal Modeling

The time-aligned sequences of MFCC vectors C_s and C_x are modeled by a GMM:

$$p(C|\lambda) = \sum_{i=1}^{M} w_i p_i(C) \tag{2}$$

where M is the number of component densities, C is the 62-dimensional vector of MFCCs, w_i are the weights, and $p_i(C)$ is the i-th component density

$$p_i(C) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(C-\mu_i)^T\Sigma_i^{-1}(C-\mu_i)\right\} \tag{3}$$

where $D = 62$ is the dimensionality of the MFCC vector, μ_i is the mean vector, and Σ_i is the diagonal covariance matrix. Each GMM is parametrized by $\lambda = \{w_i, \mu_i, \Sigma_i\}$, $1 \leq i \leq M$ and we denote the GMMs for C_s, C_x by λ_s, λ_x respectively. The GMM parameters are computed via the Expectation Maximization (EM) algorithm [5]. As in [6], we use a GMM to model the distribution of MFCC vectors and use individual component densities as models of distinctive acoustic classes for more specialized enhancement over the acoustic classes.

2.3. Acoustic Class Mapping Matrix (ACMM)

In the EM computation of the GMM parameters, there is no guarantee that the j-th component density in λ_s models the same acoustic class as the j-th density in λ_x. Thus, for each acoustic class, we must link corresponding component densities in λ_s and λ_x.

This mapping from clean acoustic class to noisy acoustic class can be ascertained from the MFCC vectors. We can identify which acoustic class C_s, C_x belongs to, given the GMM λ_s, λ_x respectively by computing the *a posteriori* probabilities for the acoustic classes and identifying the acoustic class which has the maximum [5]

$$\begin{aligned} j &= \arg\max_i p(i|C, \lambda) \\ &= \arg\max_i \frac{w_i p_i(C)}{p(C|\lambda)}. \end{aligned} \tag{4}$$

With sufficiently long and phonetically diverse time-aligned training signals, we can develop a probabilistic model which enables us to map each component density in λ_s to the component densities in λ_x. Algorithm 1 gives a procedure for computing the ACMM, A. The column-wise normalization of A provides a probabilistic mapping from noisy component density k (column of A) to clean component density j (row of A). Thus, each column of A (noisy acoustic class) contains probabilities of that noisy acoustic class having been perturbed from each of the possible clean acoustic classes (rows of A).

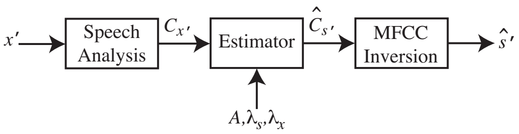


Figure 2: Speech enhancement stage.

Algorithm 1 above gives a procedure for computing the ACMM, A. The column-wise normalization of A provides a probabilistic mapping from noisy component density k (column of A) to clean component density j (row of A). Thus, each column of A contains probabilities of that noisy acoustic class having been perturbed from each of the possible clean acoustic classes (rows of A).

3. Enhancement

In this section we describe the enhancement stage illustrated in Fig. 2. We denote the noisy signal to be enhanced as x' and assume an additive noise model

$$x' = s' + v'. \tag{5}$$

We assume that s' is speech from the same speaker as s, v' is the same type of noise as v, and that x' is mixed from s' and v' at a SNR similar to that used in synthesizing x in the training stage.

As in the training stage, we compute the MFCC vector $C_{x'}$ from the noisy speech signal. The goal is to estimate $C_{s'}$ given $C_{x'}$, taking into account A, λ_s, and λ_x. We reconstruct the enhanced time-domain speech signal $\hat{s'}$ from the estimate $\hat{C}_{s'}$.

3.1. Speech Analysis

The parameters for speech analysis in the enhancement stage are identical to those of the training stage. A smaller frame advance, however, allows for slightly better performance in low-SNR due to added redundancy in the overlap-add and estimation processes.

3.2. Identifying the Underlying Clean Acoustic Class

The noisy acoustic class is identified from MFCC vector $C_{x'}$ via

$$k = \arg\max_{1\leq i\leq M} p(i|C_{x'}, \lambda_x). \tag{6}$$

Using the ACMM A, noisy acoustic class k can be probabilistically mapped to the underlying clean acoustic class j, by

$$\hat{j} = \arg\max_i A_{i,k}. \tag{7}$$

The clean acoustic class $\hat{j}$ is a probabilistic estimate of the true clean class identity for the particular speech frame.

3.3. Estimation of $C_{s'}$: "Phroming" Methods

The next step in enhancement is to "morph" the noisy MFCC vector to have characteristics of the desired clean MFCC vector. Since spectral→cepstral in the original cepstrum vocabulary, morphing→phroming. This cepstral phroming is more rigorously described as an estimation of the clean MFCC vector $C_{s'}$ based on the noisy MFCC vector $C_{x'}$, noisy acoustic class k, ACMM A, and GMMs λ_s and λ_x. We next present two phroming methods.

3.3.1. Phromed maximum method (PMAX)

Equation (7) returns the maximum-probability acoustic class $\hat{j}$ and this estimate is used as follows. Since the k-th component density in λ_x and the $\hat{j}$-th component density in λ_s are both Gaussian, a simple means of estimating $C_{s'}$ is to transform the (Gaussian) vector $C_{x'}$ into another (Gaussian) vector $\hat{C}_{s'}$:

$$\hat{C}_{s'} \quad = \quad \mu_{s,\hat{j}} + (\Sigma_{s,\hat{j}})^{1/2}(\Sigma_{x,k})^{-1/2}\,(C_{x'} - \mu_{x,k}) \quad (8)$$

where $\mu_{s,\hat{j}}$ and $\Sigma_{s,\hat{j}}$ are the mean vector and (diagonal) covariance matrix of the $\hat{j}$-th component density of λ_s, and $\mu_{x,k}$ and $\Sigma_{x,k}$ are similarly defined for λ_x. This method is referred to as phromed maximum (PMAX).

3.3.2. Phromed mixture method (PMIX)

Rather than using a single maximum probability acoustic class, we use a weighted mixture of (8) with $A_{j,k}$ as the weights

$$\hat{C}_{s'} \quad = \quad \sum_{j=1}^{M} A_{j,k}\left[\mu_{s,j} + \Sigma_{s,j}^{1/2}\Sigma_{x,k}^{-1/2}\,(C_{x'} - \mu_{x,k})\right].(9)$$

This phromed mixture (PMIX) method results in a superposition of the clean speech acoustic classes in the mel-cepstrum domain, with the weights determined based on the ACMM. Due to the added redundancy in the weighted average of the PMIX method, our research shows it consistently outperforms the PMAX method.

3.4. Inverse Transformation of MFCCs

The final step in the enhancement stage (Fig. 2) is to inverse transform $\hat{C}_{s'}$ and obtain the speech frame $\hat{s}'$. This is achieved with the direct cepstral inversion (DCI) method [9] summarized below, followed by a simple overlap-add reconstruction.

Denote the spectrum of the enhanced speech frame as $\hat{S}' = \mathcal{DFT}\left(\hat{s}'\right)$. We define the mel-frequency cepstrum as

$$\hat{C}_{s'} = \mathcal{DCT}\left\{\log\left[\Phi\left|\hat{S}'\right|^2\right]\right\} \quad (10)$$

where Φ is a bank of J mel-scale filters. In general, the speech frame, DFT, and DCT may be different lengths, but we choose (without loss of generality) length K for speech frame and the DFT, and length J for the DCT.

To invert the mel weighting, we find Φ' such that

$$\left|\tilde{S}'\right|^2 = \Phi'\Phi\left|\hat{S}'\right|^2 \approx \left|\hat{S}'\right|^2. \quad (11)$$

Defining Φ' as the Moore-Penrose pseudoinverse $\Phi^\dagger$ ($\Phi^\dagger = \left(\Phi^{\mathsf{T}}\Phi\right)^{-1}\Phi^{\mathsf{T}}$ for full rank Φ), we assure that $\left|\tilde{S}'\right|^2$ is the solution of minimal Euclidean norm. The remaining operations can be inverted without loss, since the DCT, DFT, log, and square operations are invertible, assuming that we use the noisy phase (i.e., the phase of x') for inversion of the DFT. It has been shown previously that the phase of the noisy signal is the MMSE estimate for the phase of the clean signal [3].

We have shown in [9] that the underconstrained nature of the mel cepstrum inversion introduces a degradation in PESQ of ~ 0.2 points at very high SNR (for $J > 52$), but these artifacts become masked by the noise below about 20 dB SNR.

3.5. Modeling Separate Formant and Pitch Information

We find significant speech enhancement improvement if the MFCC vector is partitioned into two subvectors

$$\begin{aligned} C^{\mathrm{f}} &= [C(0), \cdots, C(12)]^T \\ C^{\mathrm{p}} &= [C(13), \cdots, C(61)]^T \end{aligned} \quad (12)$$

where ‘f’ and ‘p’ refer to the formant and pitch subsets, respectively. Both formant (vocal tract configuration) and pitch (excitation) are important components to a total speech sound, but should be allowed to vary independently. The cutoff for the formant and pitch subsets is chosen based on the range of pitch periods expected for both males and females, translated into the mel-cepstrum domain.

We thus compute GMMs λ_s^{f}, λ_s^{p}, λ_x^{f}, λ_x^{p} based on MFCC subvectors C_s^{f}, C_s^{p}, C_x^{f}, C_x^{p} respectively. ACMMs A^{f}, A^{p} are computed with Algorithm 1 using $\{C_s^{\mathrm{f}}, C_x^{\mathrm{f}}\}$, $\{C_s^{\mathrm{p}}, C_x^{\mathrm{p}}\}$ respectively and $\hat{C}_{s'}^{\mathrm{f}}$, $\hat{C}_{s'}^{\mathrm{p}}$ are estimated using $\{C_{x'}^{\mathrm{f}}, \lambda_s^{\mathrm{f}}, \lambda_x^{\mathrm{f}}\}$, $\{C_{x'}^{\mathrm{p}}, \lambda_s^{\mathrm{p}}, \lambda_x^{\mathrm{p}}\}$ respectively. Finally, the estimate of the clean MFCC vector is formed as the concatenation of $\hat{C}_{s'}^{\mathrm{f}}$ and $\hat{C}_{s'}^{\mathrm{p}}$ followed by inversion of $\hat{C}_{s'}$ as described in the previous section.

We are separating the MFCCs into two subsets to better individually model formant and pitch information, rather than for computational reasons as in [7]. Both formant (vocal tract configuration) and pitch (excitation) are important components to a total speech sound, but should be allowed to vary independently.

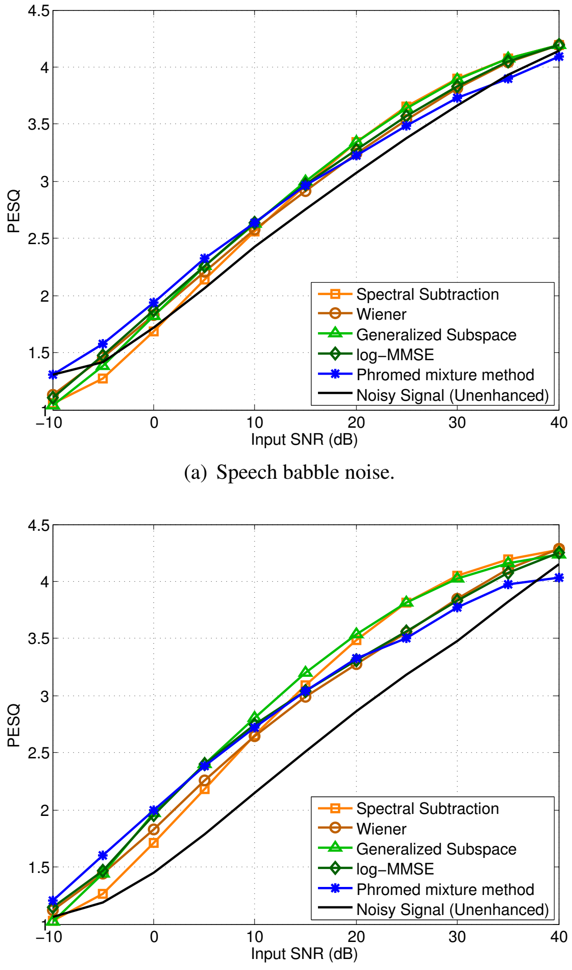
4. Results

The system described above has been implemented and simulations run to measure average performance using ten randomly-chosen speakers (five male and five female) from the TIMIT corpus and noise signals from the NOISEX-92 corpus. Speech frames are 320 samples, training signals are ∼24s long with a frame advance of 160 samples, and test signals are ∼6s long with a frame advance of 1 sample. Separate GMMs are used to model formant and pitch information and the number of GMM components M is 15. Results are presented in terms of PESQ versus input SNR; PESQ has been shown to have the highest correlation to overall signal quality [10].

4.1. Performance and Comparison to Other Methods

Fig. 3 shows the performance of the proposed method for babble and white noises. In addition, performance for spectral subtraction using oversubtraction [1], Wiener filtering using *a priori* SNR estimation [2], log-MMSE spectral amplitude estimator [3], and the generalized subspace method [4] are provided for comparison. These methods improve upon the respective standard methods.

For the proposed method, we see a maximum improvement in PESQ of 0.3–0.6 points over the unenhanced signal, depending on the noise type. In general, the proposed method has an input SNR operating range from -10 dB to +35 dB, with performance tapering off at the ends of the operating range. Phroming typically outperforms all of the compared methods for input SNRs below 10 dB. For further reference, the PESQ scores are shown in Table 1 for input SNRs between -10 and



(a) Speech babble noise.

(b) White noise.

Figure 3: Speech enhancement results (PESQ vs. input SNR).

Table 1: PESQ performance of various enhancement methods: spectral subtraction (SS) [1], Wiener (WA) [2], generalized subspace (GS) [4], log-MMSE (LM) [11], and the proposed method (PM). Bold entries correspond to the best performance.

(a) Speech babble noise.

SNR	Noisy	SS	WA	GS	LM	PM
15	2.75	2.96	2.92	3.00	**2.97**	2.96
10	2.43	2.56	2.58	2.63	2.63	**2.64**
5	2.07	2.14	2.20	2.25	2.26	**2.32**
0	1.72	1.69	1.83	1.82	1.87	**1.94**
-5	1.42	1.27	1.46	1.38	1.48	**1.58**
-10	1.31	1.06	1.13	1.04	1.11	1.31

(b) White noise.

SNR	Noisy	SS	WA	GS	LM	PM
15	2.51	3.09	2.99	**3.20**	3.04	3.04
10	2.15	2.65	2.65	**2.80**	2.75	2.72
5	1.79	2.19	2.25	**2.40**	**2.40**	2.39
0	1.45	1.71	1.83	1.97	1.95	**2.00**
-5	1.19	1.26	1.44	1.44	1.46	**1.60**
-10	1.06	1.03	1.13	1.02	1.15	**1.21**

15 dB. Subjective evaluation of the resulting enhanced waveforms reveals good noise reduction with minimal artifacts. In particular, the musical noise present in spectral subtraction and Wiener filtering is not apparent in the proposed method.

5. Conclusions and Future Research

We have proposed a two-stage speech enhancement technique which uses GMMs to model the MFCCs from clean and noisy speech. A novel acoustic class mapping matrix (ACMM) allows us to probabilistically map the identified acoustic class in the noisy speech to an acoustic class in the underlying clean speech. Finally, we use the identified acoustic classes to estimate the clean MFCC vector. Our results show that we can improve PESQ in environments as low as -10 dB input SNR.

6. References

[1] M. Berouti, M. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1979, pp. 208–211.

[2] P. Scalart and J. Filho, "Speech enhancement based on a priori signal-to-noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1996, pp. 629–632.

[3] 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.

[4] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 334–341, July 2003.

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

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

[7] A. Kundu, S. Chatterjee, and T. V. Sreenivas, "Speech enhancement using intra-frame dependency in DCT domain," in *Proc. European Signal Process. Conf. (EUSIPCO)*, 2008.

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

[9] L. E. Boucheron and P. L. De Leon, "On the inversion of mel-frequency cepstral coefficients for speech enhancement applications," in *Proc. Int. Conf. Signals and Electronic Systems (ICSES)*, Sept. 2008.

[10] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.