Speech Enhancement of Spectral Magnitude Bin Trajectories using Gaussian Mixture-Model based Minimum Mean-Square Error Estimators

James G. O'Connell [1], *Kuldip K. Paliwal* [1], *Kamil Wójcicki* [2]

[1] Griffith University, QLD, Australia

[2] The University of Texas at Dallas, U.S.A.

james.oconnell@griffithuni.edu.au, k.paliwal@griffith.edu.au, kamil.wojcicki@ieee.org

Abstract

Gaussian mixture-model based minimum mean-square error estimators have been applied to speech enhancement in the temporal, transform (e.g., discrete cosine transform), and subspace domains. In this paper, we propose a method for applying a GMM-based MMSE estimator to spectral magnitude-bin trajectories. In addition, methods for incorporating speech presence uncertainty into the proposed system to improve performance are discussed. The proposed system outperforms previously published GMM-based estimators, and the well-known Ephraim and Malah estimator for 8 kHz telephone-quality speech.

Index Terms: Speech enhancement, minimum mean-square error (MMSE) estimator, Gaussian mixture-model (GMM)

1. Introduction

Speech enhancement aims to improve the quality of noisy speech. This is generally accomplished by suppression or altering the characteristics of the noise in such a way that it is less unpleasant to the listener, while minimising the speech distortion introduced in the enhancement process. This paper considers single-channel speech enhancement, in which a single microphone is used to capture sound. Single-channel enhancement is particularly relevant to mobile speech communication, where a single microphone is often employed due to cost and size requirements.

Many popular single-channel speech enhancement methods employ the analysis-modification-synthesis (AMS) framework [1, 2, 3, 4, 5, 6]. The AMS framework consists of three stages: 1) the analysis stage, where short-time Fourier transform (STFT) analysis is applied to the input signal; 2) the modification stage, where the noisy spectrum is modified; and 3) the synthesis stage, where the output signal is reconstructed using the inverse STFT and overlap-add synthesis.

The proposed Gaussian mixture-model (GMM) based enhancement method is a statistical estimator, and all discussions in this paper assume this class of enhancement algorithm. During the modification stage, statistical speech enhancement methods assume a specific probability distribution for the signal and the disturbance when estimating the clean spectrum. Statistical models applied to speech enhancement include Rayleigh, Gaussian and Laplace distributions. However, any distribution capable of adequately modelling the speech signal can be used.

In automatic speech and speaker recognition, GMM are often used to model the features extracted from the speech signal. Their application in the speech enhancement field has, however, been limited. In this paper GMM will be used to model speech, for application to speech enhancement.

Given a sufficient number of mixtures, GMM are capable of approximating any continuous probability distribution. As a result, no initial assumptions need to be made about the distribution of the speech except that it is continuous. The disadvantage is that GMM require more computation than simpler models such as Rayleigh and Gaussian distributions.

One issue with statistical methods of speech enhancement is that the disturbance, either by pure chance or as a result of being speech-like, can generate high likelihood outputs from the speech model. In this case the disturbance will not be suppressed, resulting in degradation of the enhanced speech. In addition, the base estimators don't differentiate between periods of speech and silence, and attempt to enhance regions of the signal where speech is absent. In order to address these shortcomings, a speech-presence uncertainty (SPU) estimator is commonly incorporated.

An SPU algorithm provides an indicator of the presence of speech. Depending on the algorithm employed, the output can take the form of a continuous probability, or a dichotomy in which speech is either present or not. Generally, the outputs from the SPU algorithm and statistical estimator are multiplied to provide an improved estimate of the clean signal. The overall effect is to suppress the enhanced signal where there is a low probability of speech, and retain the enhanced signal in regions where the probability of speech is high.

This paper will focus on two main topics. First, speech enhancement by application of GMM-based MMSE estimators to spectral magnitude bin trajectories (SMBT) will be examined. And second, methods for improving the quality of the enhanced signal using speech-presence uncertainty will be discussed. The proposed methods will be applied to utterances from the TIMIT corpus [7] that have been corrupted with additive noise, and the enhanced speech will be evaluated using both objective and subjective measures. Section 2 provides a general overview of speech enhancement using GMM. Sections 3 and 4 discuss applying GMM-based enhancement to SMBT, and the inclusion of SPU, respectively. Sections 5 and 6 describe the experiments and results obtained.

2. GMM-Based Speech Enhancement

The GMM MMSE estimator described by Kundu *et al.* [8, 9, 10] is given by:

$$\hat{\mathbf{S}} = \sum_{m=1}^{M} \omega_m \ \left[\mu_m + \mathbf{\Sigma}_m(\mathbf{\Sigma}_m + \mathbf{\Sigma}_D)^{-1}(\mathbf{X} - \mu_D - \mu_m)\right] \tag{1}$$

where μ_m and $\mathbf{\Sigma}_m$ are the mean and covariance for mixture m respectively, μ_D and $\mathbf{\Sigma}_D$ are the mean and covariance of the disturbance, and ω_m is given by:

$$\omega_m(\mathbf{X}) = \frac{\alpha_m \, \mathcal{N}(\mathbf{X}; \mu_m + \mu_D, \mathbf{\Sigma}_m + \mathbf{\Sigma}_D)}{\sum_{k=1}^{M} \alpha_k \, \mathcal{N}(\mathbf{X}; \mu_k + \mu_D, \mathbf{\Sigma}_k + \mathbf{\Sigma}_D)} \quad (2)$$

where $\mathcal{N}$ represents a multivariate Gaussian probability density function and α is a vector of mixture weights.

The estimator given in Equation (1) is a special case of Gaussian mixture regression (GMR), in which the source and distortion are independent.[1] Equation (1) differs slightly from that in [8], in which a linear transform applied to the noisy signal is included. In practice, however, it is more efficient to calculate the complex spectral representation using a fast Fourier transform prior to applying the estimator—rather than applying the estimator to the noisy signal with a linear transform. In addition, speech enhancement is generally performed in the spectral magnitude domain, which is not a direct linear transform of the signal. In later papers, Kundu *et al.* describe the estimator without the linear transform included [9, 10].

GMM are limited in dimensionality by the computational complexity of the training process, and the tendency to produce mixtures with zero variance due to the appearance of outliers when separability is improved. This imposes limitations on the length of the frame that can be processed in the time domain. As a consequence, Kundu *et al.* limited frames to between 10 and 60 samples long (1.25–7.5 ms at 8 kHz) when processing in the time-domain [9], which is far shorter than the 10–40 ms generally applied in speech enhancement [12, 13, 14]. In addition, GMM-based speech enhancement in the time domain results in relatively unintelligible speech with high levels of background static.

In order to address these deficiencies, Kundu *et al.* proposed applying GMM-based enhancement in the discrete cosine transform (DCT) domain [9]. The DCT was chosen over the STFT as the resulting coefficients are generally assumed to be non-correlated for speech signals. Consequently, longer frame sizes can be processed, with the resultant DCT split into sub-vectors and each sub-vector enhanced individually using a separate GMM model. In [9], Kundu *et al.* chose frames of 32 ms (256 samples) duration, then split the DCT into 8 sub-vectors of 32 coefficients each. Thus eight GMM are required and feature vectors are of 32-dimensions. This approach greatly improves the intelligibility of the enhanced speech in comparison to the time-domain method, and the residual noise is smoother and less harsh in nature.

3. GMM-Based Speech Enhancement Applied to Spectral Bin Time Trajectories

Figure 1 outlines the proposed GMM-based enhancement applied to spectral magnitude bin trajectories. Due to overlap of SMBT frames, two nested applications of the AMS framework were required. First, STFT analysis was performed on the waveform data; i.e., the time-domain speech signal was processed framewise using a STFT with a Hamming window. The resulting complex spectra were divided into magnitude and phase spectra, with the magnitude spectrum undergoing further processing while the phase spectrum was left unchanged for synthesis.

The time-trajectory of each frequency bin of the short-time acoustic magnitude spectrum (i.e., spectral magnitude bin trajectory) was then divided into frames and used as input into the GMM estimator. Since no further frequency analysis was performed on the SMBT frames, a rectangular window was sufficient during this second analysis stage.

[1] An introduction to GMR can be found in [11].

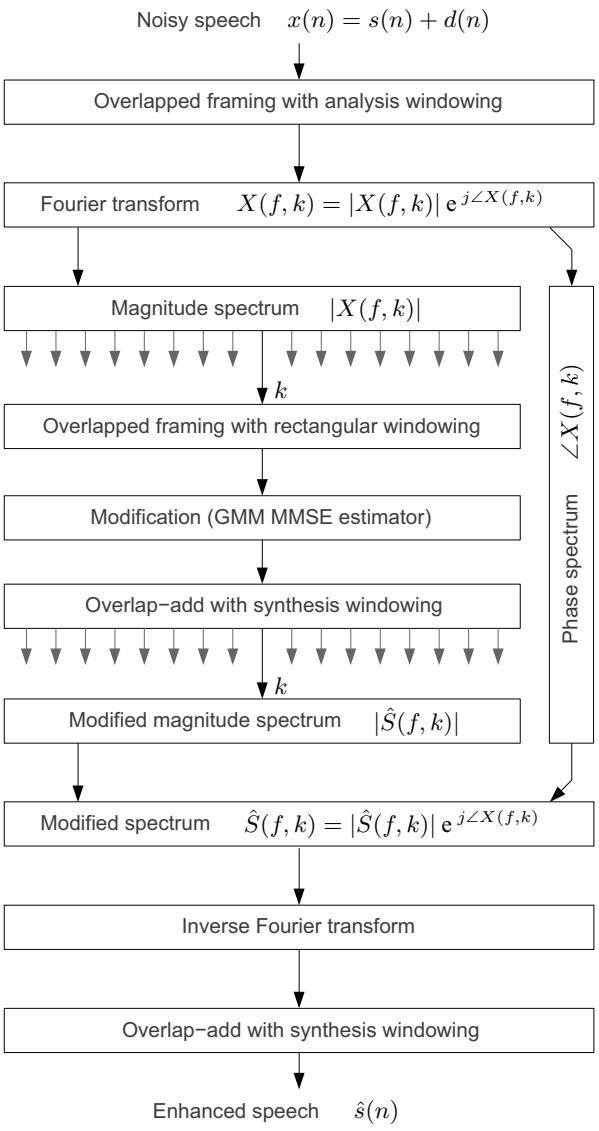

Figure 1: *Process diagram for GMM-based speech enhancement applied to spectral magnitude bin trajectories. The* (f, k) *indices refer to frame number and frequency bin respectively.*

The SMBT frames then underwent modification using the estimator given in Equation (1), as well as separate GMM speech models for each SMBT. This was followed by overlap-add synthesis to reconstruct each SMBT. From these a modified magnitude spectrum was constructed, which was then combined with the noisy phase spectrum to provide the estimated complex spectrum. The inverse STFT was then taken, and overlap-add synthesis applied to produce the estimated clean speech signal.

4. Incorporating Speech-Presence Uncertainty

Speech enhanced using the proposed system described thus far suffers from residual musical noise distortion for a number of

reasons. First, variances in the speech model are relatively large due to unavoidable variance in the spectral estimate. As a result, numerous mixtures will still generally output high likelihoods for background noise or silence regions.

Second, several models will necessarily model unvoiced speech. Since white noise has statistical properties more similar to unvoiced speech than silence, it is less likely to be suppressed. This can be somewhat mitigated by using SMBT frames that are longer than unvoiced speech segments. However, further mixtures are then required in the model to account for the increased combinations of acoustic events possible in longer frames.

To remove residual noise, further suppression of the signal during regions of silence is required. Speech-presence uncertainty provides a simple solution to this artefact. Speech-presence uncertainty is a metric that estimates the probability that speech is present in a given feature vector. The proposed speech enhancement algorithm attempts to enhance any supplied feature vector, whether speech is present or not. SPU can be used to scale, and subsequently deemphasise, regions of the signal where speech is unlikely to be present, and thus augments the speech enhancement process.

Since GMM-based speech enhancement necessarily employs statistical models of speech and noise, these can be used to estimate the probability of speech presence. The method employed in the present work is that described by Loizou [14, p. 271]:

$$p(H_s|\mathbf{x}) = \frac{P_s\ p(\mathbf{x}|H_s)}{P_s\ p(\mathbf{x}|H_s) + P_n\ p(\mathbf{x}|H_n)} \tag{3}$$

where $p(H_s|\mathbf{x})$ is the probability of speech given the input feature vector $\mathbf{x}$, P_s and P_n are the *a-priori* probabilities of speech and silence (only noise present) respectively, and H_s and H_n are the statistical models for speech and noise. The *a-posteriori* probability of speech output by the SPU is applied as a gain, during modification, to the estimated clean frame. Thus, frames that are likely to be noise are suppressed, while those highly likely to represent speech are left relatively untouched.

Originally, the likelihood outputs from the disturbance and speech models applied in enhancement were used in Equation (3). In order to achieve significant reductions in background noise, however, the *a-priori* probability of noise in Equation (3) had to be set unrealistically high (in some cases higher than 0.99). This is due to the speech model necessarily modelling silence, and the different distributions used to model the disturbance and speech. Acoustic events such as closures and other regions of silence need to be accurately estimated by the enhancement method, and thus need to be represented in the model. The use of GMM as the speech model results in feature vectors generally only scoring on a small number of mixtures. This means the likelihoods output by the speech model are lower than those output by the single Gaussian of the disturbance model. As a result, separate models to those used for enhancement were employed. In addition, using separate models for the SPU provides additional information in the enhancement process.

4.1. Rayleigh Distribution based SPU

If the real and imaginary components of the complex spectrum of a signal are assumed to be independently Gaussian distributed, then the magnitude spectrum is described by a Rayleigh distribution. Given normalised signals, the Rayleigh parameters for each spectral bin of speech can be determined during the training phase. During testing, the magnitude spectrum extracted from the first 500 ms of the noisy signal is used to determine the Rayleigh parameters for the noise in each spectral bin.

4.2. SPU based on SNR

Unlike the Rayleigh distribution, the SNR doesn't directly provide a likelihood estimate. The ratio can be used, however, to provide a measure of the proportion of signal energy attributable to speech. This proportion can be used in place of $p(H_s|\mathbf{x})$ in Equation (3) to apply gain to the estimated clean signal during modification.

5. Experiments

5.1. Speech Corpus

This study employed the TIMIT database, consisting of spoken sentences sampled at 16 kHz with 10 utterances each from 630 speakers [15]. Utterances were low-pass filtered to 3.4 kHz, to mimic telephone bandwidth, using a 7-th order Butterworth filter, and resampled to 8 kHz. The energy of the initial few samples of each utterance was calculated, and 500 ms of silence consisting of AWGN at the same energy prefixed. Very low-level AWGN was used in the silence region to avoid issues that may arise with some common implementations when a frame consists entirely of zeros.

Models were trained using the *test* subset of 168 speakers. The *si** and *sx** utterances for each speaker were used for training (8 utterances per speaker), for a combined total of approximately 66 minutes of training speech. For testing purposes, the utterances were degraded by AWGN (additive white Gaussian noise) at several SNR. Randomly chosen utterances from the *train* subset of TIMIT (462 speakers) were used for subjective and objective testing.

5.2. Processing Parameters

During tuning of the parameters, it was noted that the performance of the enhancement system was relatively invariant to acoustic frame length and overlap. As a result, 16 ms frames with 8 ms overlap were chosen for analysis to be consistent with the settings chosen by Kundu *et al.* [8], and to reduce computation required.[2] Each spectral magnitude bin trajectory signal was split, using a rectangular window, into 8 sample frames with 1 sample shift. Each SMBT frame thus corresponded to 72 ms of the original signal.

5.3. Performance Metrics

A mixture of objective and informal subjective tests were used to tune the performance of the system. The effectiveness of the proposed system was assessed using formal subjective tests.

5.3.1. Formal subjective listening test procedure

Formal subjective testing took the form of AB listening tests comprising all possible stimuli pairs, including the reverse cases. The subjective experiment was conducted in a quiet room, with stimuli pairs played back in random order over closed circumaural headphones at a comfortable listening level.

Participants were first familiarised with the task by a short practice session consisting of five randomly chosen stimuli pair. Subjects were presented with three labelled options for each

[2] Longer frames, however, performed slightly better.

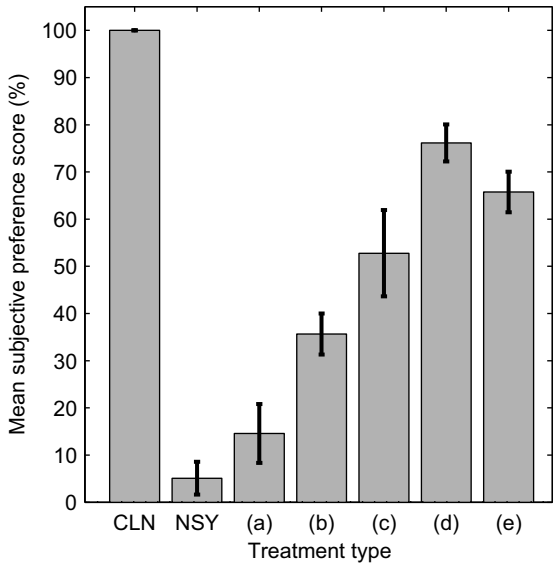

Figure 2: *Preference by treatment. Treatments are: (a) Kundu et al. [8] time-domain method; (b) Kundu et al. [9] DCT-domain method; (c) Ephraim and Malah [16] MMSE estimator with SPU; (d) Proposed method with Rayleigh distribution based SPU; (e) Proposed method with SPU based on SNR estimate.*

stimuli pair. The first and second options indicated a preference for one or the other stimulus, while the third option indicated that the subject preferred both stimuli equally. Participants could replay stimuli as many times as required before making a decision.

Pairwise scoring was used, with +1 and 0 points being assigned to the preferred and non-preferred stimulus respectively. If no preference could be assigned, each stimulus was awarded +0.5 points. The results for a subject were discarded as unreliable if they couldn't consistently indicate either a preference, or neutrality, for the clean signal. Ten native English speaking listeners, with no reported hearing defects, participated in the study, with one subject discarded as unreliable.

6. Results and Discussion

Figure 2 shows the mean preference, averaged across participants, for each treatment. The preference score indicates the percentage of comparisons in which the treatment was preferred. The error bars represent one standard deviation, calculated from the preference scores from all participants, from the mean.

For statistical analysis of the results, the null hypotheses was that methods being compared were equally preferred. The probability that our null hypothesis was true was calculated using a two-tailed test based on the binomial distribution, with an *a-priori* preference of 0.5 – that is, both equally preferred.

The results indicate a preference for GMM-based enhancement applied to SMBT frames (using either SPU) over the methods proposed by Kundu *et al.* ($p < 0.001$ for both the time-domain [8] and DCT-domain [9] methods when considering all comparisons), and Ephraim and Malah's MMSE-estimator with SPU ($p < 0.002$ when compared to the SPU based on SNR).

Although estimating distortion based on frame energy is slightly more computationally efficient, the incorporation of speech-presence uncertainty probabilities calculated from the Rayleigh distribution was preferred in subjective listening tests ($p < 0.002$). While the proposed method will run in real-time on a single-core of a modern processor, it is more computationally expensive than other common methods, and is thus more suited to offline processing.

7. Conclusions

In this paper, we applied MMSE estimators based on GMM distributions to telephone quality, single-channel speech enhancement. In the proposed method, the MMSE estimator was used to modify the spectral magnitude bin trajectories, and an SPU incorporated. The proposed method was found to significantly outperform similar, previously proposed GMM-based estimators, and Ephraim and Malah's MMSE estimator with SPU.

8. References

[1] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," vol. ASSP-25, no. 3, pp. 235–238, Jun 1977.

[2] J. Allen and L. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," vol. 65, no. 11, pp. 1558–1564, 1977.

[3] R. Crochiere, "A weighted overlap-add method of short-time Fourier analysis / synthesis," vol. ASSP-28, no. 1, pp. 99–102, Feb 1980.

[4] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," vol. ASSP-32, no. 2, pp. 236–243, 1984.

[5] M. Portnoff, "Short-time Fourier analysis of sampled speech," vol. ASSP-29, no. 3, pp. 364–373, 1981.

[6] T. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Upper Saddle River, NJ: Prentice Hall, 2002.

[7] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993.

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

[9] ——, "Speech enhancement using intra-frame dependency in DCT domain," in *ICASSP*, 2008, pp. 4893–4896.

[10] A. Kundu, S. Chatterjee, and T. V. Sreenivas, "Subspace based speech enhancement using Gaussian mixture model," in *INTERSPEECH*, 2008, pp. 395–398.

[11] H. G. Sung, "Gaussian mixture regression and classification," Ph.D. dissertation, Rice University, Houston, Texas, USA, 2004.

[12] J. Picone, "Signal modeling techniques in speech recognition," vol. 81, no. 9, pp. 1215–1247, Sep 1993.

[13] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: a guide to theory, algorithm, and system development*. Upper Saddle River, New Jersey: Prentice-Hall, 2001.

[14] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: Taylor and Francis, 2007.

[15] W. Fisher, G. Doddington, and K. Goudie-Marshall, "The DARPA speech recognition research database : Specifications and status," in *Proc. DARPA Workshop on Speech Recognition*, feb 1986, pp. 93 – 99.

[16] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," vol. ASSP-32, no. 6, pp. 1109–1121, Dec 1984.