A Perceptually Approach for Speech Enhancement Based on Mmse Error Estimators and Masking in an Auditory System

T.Raghavendra Vishnu, P.S.Brahamanandam, Raghava Yathiraju, Naga Himaja Alla, K.PhaniSrinivas, B.T.P.Madhav, Radhika Chinaboina, Usha Mallaparapu

Department of ECE, K L University, Guntur DT, AP, India
Department of ECE, ST.Mary’s college of Engg &Tech, Hyderabad, AP, India
Email: tadivakavishnu@gmail.com, raghava.yathiraju@gmail.com

ABSTRACT

Speech processing is used widely in every application that most people take for granted, such as network wire lines, cellular telephony, telephony system and telephone answering machines. Due to its popularity and increasing of demand, engineers are trying various approaches of improving the process. One of the methods for improving the process is MMSE based Wiener filter. The objective of speech enhancement is to improve the quality of a speech signal, often degraded by some type of distortion (for example communication channel distortion, additive noise, convolution filtering operation, etc.) The quality of a speech signal is judged, depending on the application, by one or more of the following factors intelligibility, perceptual quality, listener fatigue, signal-to-noise ratio (SNR), speech distortion, and (occasionally) recognition accuracy of an automatic speech recognizer. Numerous schemes have been proposed and implemented that perform speech enhancement under various constraints/assumptions and deal with different issues and applications. An application of these estimators in an auditory enhancement scheme using the masking threshold of the human auditory system is formulated, resulting in the GMMSE auditory masking threshold (AMT) enhancement method. Finally, a detailed evaluation of the proposed algorithms is performed over the audio archive using subjective and objective speech quality measures.

Keywords: Auditory masked threshold (AMT), denoising, generalized minimum mean square error (GMMSE) AMT, log minimum mean square error (MMSE), noise suppression, speech enhancement, Weiner filter.

1. INTRODUCTION

The objective of speech enhancement is to improve the quality of a speech signal, often degraded by some type of distortion (for example communication channel distortion, additive noise, convolution filtering operation, etc.). The quality of a speech signal is judged, depending on the application, by one or more of the following factors intelligibility, perceptual quality, listener fatigue, signal-to-noise ratio (SNR), speech distortion, and (occasionally) recognition accuracy of an automatic speech recognizer. Numerous schemes have been proposed and implemented that perform speech enhancement under various constraints/assumptions and deal with different issues and applications.

For example, the problem of dual-channel speech enhancement assumes the availability of two signals one a noisy speech signal (called the primary signal) and the other (called the reference signal) comprised only of noise that is correlated to the noise in the primary signal but not to the speech. The problem of enhancing speech becomes much harder for single-channel speech enhancement where there is no reference input available to estimate the noise. A log-MMSE-based estimator, or maximum probability of error maximum likelihood (ML) and maximum a posteriori (MAP) estimators. Another closely related family of enhancement schemes is based on short-term Wiener filtering and its variations (iterative Wiener filtering and constrained iterative Wiener filtering. In particular, the spectral amplitude MMSE estimator proposed by Ephraim and Malah was shown to have good musical noise suppression characteristics. In this paper, a general family of MMSE estimators is derived as

\[ \hat{X}_p = \left( E(X_\alpha|Y_p) \right)^{1/\alpha}, \alpha > 0 \]

Where \( X_p \) and \( Y_p \) are the power spectra of the clean and noisy speech, respectively. The noise suppression and musical noise elimination mechanisms of these generalized MMSE estimators are also studied. It is shown that these estimators offer more flexibility in optimizing noise reduction and musical artifacts over the EM estimator.

2. WIENER FILTER

The Wiener filter is a popular technique that has been used in many signal enhancement methods. The basic principle of the Wiener filter is to obtain a clean signal from that corrupted by additive noise. It is required estimate an optimal filter for the noisy input speech by minimizing the Mean Square Error
(MSE) between the desired signal \( s(n) \) and the estimated signal \( \hat{s}(n) \). The frequency domain solution to this optimization. The drawback of the Wiener filter is the fixed frequency response at all frequencies and the requirement to estimate the power spectral density of the clean signal and noise prior to filtering.

2.1. Process of estimation

Wiener filtering, spectral subtraction, subspace methods and Kalman filtering are popularly used approaches for noise reduction. In the following subsections, we discuss these methods, and study their differences and similarities.

Consider a \( K \times K \) linear estimator \( H \) that results in an estimate \( \hat{x} = Hy \) of the clean speech from the noisy speech. The estimation error can be written as

\[
\varepsilon = Hy - x = (H - I)x + Rw
\]

where \( I \) is the \( K \times K \) identity matrix. The mean squared estimation error is given by \( \text{tr} E[\varepsilon \varepsilon^T] \), where \( \text{tr} \) denotes the matrix trace. From the independence assumption, the cross terms in the mean squared error vanish and we have

\[
\text{tr} E[\varepsilon \varepsilon^T] = \text{tr} E[\varepsilon_x \varepsilon_x^T] + \text{tr} E[\varepsilon_w \varepsilon_w^T]
\]

where \( \text{tr} (H - I)R_x(H - I)^T + \text{tr} HR_wH^T \),

\[
R_x = E[\varepsilon_x \varepsilon_x^T], \quad R_w = E[\varepsilon_w \varepsilon_w^T]
\]

are the covariance matrices of speech and noise respectively.

\[
H^* = \text{arg min}_H \left( \text{tr} (H - I)R_x(H - I)^T + \text{tr} HR_wH^T \right)
\]

\[
= R_x^{-1} + R_w^{-1}.
\]

The above estimator can be efficiently implemented in the frequency domain. Under the assumption of large \( K \), the covariance matrices \( R_x \) and \( R_w \) can be approximated as circulant and are hence diagonalized by the DFT, i.e.,

\[
R_x = F^*Px F, \quad R_w = F^*Pw F,
\]

Where \( Px = \text{diag}(Px(0) Px(1) : : : Px(K - 1)) \) is a diagonal matrix containing the PSD of \( x \), \( F \) is the DFT matrix, and the superscript * denotes complex conjugate transpose. Similarly, \( R_w = F^*Pw F \),

Where \( Pw = \text{diag}(Pw(0)Pw(1) : : : Pw(K-1)) \) is a diagonal matrix containing the PSD of \( w \). With the above digitalization, the Wiener filter can be rewritten in the frequency domain as

\[
H_{\text{Wiener}}(k) = \frac{P_x(k) + P_w(k)}{P_x(k)} = P_x^{-1}.
\]

However, \( P_x \) is not known, and, in practice, an estimate \( \hat{P}_x \) of \( P_x \) is used. This estimate is commonly obtained in a subtractive fashion using \( H \) and an estimate \( \hat{P}_w \) of \( P_w \), and negative values are set to zero since the PSD cannot be negative (negative values may arise since \( \hat{P}_w \) is only an estimate of the noise PSD).

\[
\hat{P}_x(k) = \max(\hat{P}_x(k) - \hat{P}_w(k), 0), \quad k = 0,1,\ldots,K - 1,
\]

So that the clean speech spectrum is then estimated according to

\[
\hat{X}(k) = \frac{\max(\hat{P}_x(k) - \hat{P}_w(k), 0)}{P_x(k)} Y(k), \quad k = 0,1,\ldots,K - 1.
\]

2.2 Single-channel speech enhancement

Single-channel speech enhancement systems obtain the input signal using only one microphone. This is in contrast to multi-channel systems where the presence of two or more microphones enables both spatial and temporal processing. Single-channel approaches are relevant due to cost and size factors. They achieve noise reduction by exploiting the spectral diversity between the speech and noise signals. Since the frequency spectra of speech and noise often overlap, single-channel methods generally achieve noise reduction at the expense of speech distortion. The reduction of background noise using single-channel methods requires an estimate of the noise statistics. Early approaches were based on voice activity detectors (VAD), and noise estimates were updated during periods of speech inactivity. Accuracy deteriorates with decreasing signal-to-noise ratios (SNR) and in non-stationary noise. Soft-decision VADs, update the noise statistics even during speech activity.

The segmentation is performed using a sliding window of finite support. The windowed signal (assuming it is absolute summable) is transformed to the frequency domain using the discrete short-time Fourier transform (STFT)

\[
X_m(k) = \frac{1}{\sqrt{N}} \sum_{n=-N}^{N} x(n)h(n-m) \exp \left(-\frac{j2\pi kn}{N}\right), \quad k = 0,1,\ldots,K - 1
\]

Where \( x(n) \) is the sampled speech signal, \( h(n) \) is the analysis window that is non-zero only in the interval \( [0; K - 1] \), \( m \) is the index to the current windowed segment, and \( k \) is the discrete frequency index.2 While it is not customary to include the normalization by \( PK \) in the definition, we do so for convenience in notation introduced later in the thesis. To obtain \( X_{m+1}(k) \), the window is shifted by one sample from its previous position. In practice, the sequence of frames is sub sampled by a factor \( L \), resulting in \( X_m L(k) \), which is equivalent to a larger frame-shift.

Typical values at a sampling frequency of 8 kHz are \( K = 256 \) (32 ms) and a frame-shift of \( L = 128 \) (50% overlap). For a given window length \( K \) over which the analysis window is non-zero, to ensure invariability of the discrete STFT, we must have \( L \leq K \). In practice, (9) is implemented by buffering \( K \) samples of the signal, applying a smooth window, followed by a \( K \)-point discrete
(fast) Fourier transform (DFT). We refer to $X(k)$ as the (complex) spectrum of the signal and to $jX(k)j$ as the magnitude spectrum. The quantity $jX(k)j^2$ denotes the periodogram. For stationary signals, as $K - 1$, the expected value of the periodogram can be shown to be the power spectral density (PSD), $P_x(k) = E[jX(k)j^2]$. The PSD and the autocorrelation function of the signal form a Fourier transform pair.

We consider an additive noise model

$$y(n) = x(n) + w(n)$$

Where $y(n)$ represents the sampled noisy speech. The speech and noise signals are modeled as independent random processes. Let $x = [x(0), x(K-1)]^T$ denote a segment of length $K$ of the clean speech signal. $y$ and $w$ are defined analogously. In the noise reduction problem, the additive signal model defined above applies to all the single-channel algorithms described in this project. The additive model can be expressed in the frequency domain as

$$Y(k) = X(k) + W(k)$$

Where $Y(k), X(k)$ and $W(k)$ are obtained by applying the DFT to the time domain entities $y, x$ and $w$ respectively. Since the speech and noise signals are independent, the following relation holds between the corresponding PSDs:

$$P_x(k) = P_x(k) + P_w(k)$$

A block diagram of a generic frequency domain single-channel speech enhancement system is shown in Fig. below. It is common to modify only the spectral amplitude and use the noisy phase an estimate of the noise PSD is obtained from the noisy speech. Any available prior knowledge about the noise signal may be exploited. Again, prior knowledge about the speech signal or about the human auditory system can be exploited. In some systems (e.g., the systems described in papers A and B), the speech and noise PSD are jointly estimated, as indicated by the bidirectional arrow in the figure.

3. DEVELOPMENT OF ATM USING WIENER FILTER TO SPEECH

Algorithm Development:

The noisy speech is broken down into frames and shaped using a Hamming window and a time-to-frequency transformation is applied using an FFT. An estimate of the clean spectrum is found using the GMMSE algorithm. The Auditory Masking Threshold

Calculation of the AMT can be broken down into 4 steps

1. Do sub-band processing with the center of each band equal to the center frequency of each auditory filter. Auditory filters are represented using their equivalent rectangular bandwidth (ERB).
2. Calculation the excitation pattern for each sub-band using the Moore-Glasberg spreading functions.
3. Calculate the offset term, based on the tonality of the speech waveform for each sub-band.
4. Subtract the offset from the excitation pattern and compare with the absolute threshold of hearing.

Next, the noisy power spectrum is compared with the AMT. If the noisy power spectrum is greater than the AMT, the signal is enhanced using a Wiener filtering operation.

4. RESULTS

Figure 1: Input Speech PSNR Value: 57.4887
Figure 2: Total computational coefficients 100
Figure 3: Output Speech PSNR Value: 63.2569
Figure 4: Output Speech applied with Masking

Speech Samples signal channel wise outputs:
6. CONCLUSIONS

In this paper, an implementation of employing wiener filtering to speech processing had been developed. As has been previously mentioned, the purpose of this approach is to reconstruct an output speech signal by making use of the accurate estimating ability of the wiener filter.

A general class of MMSE spectral estimators was presented in this paper, resulting in the GMMSE-AMT. It was shown that these estimators allowed greater flexibility in trading off noise suppression versus speech distortion and musical noise, over the EM algorithm. A sensitivity analysis was performed to study the performance when the statistical model parameters (a priori SNR) were not accurately known. A detailed performance analysis was done on 8-kHz sampled using segmental SNRs. In our study; we proposed the integration of the AMT within the generalized MMSE estimators. Future studies could consider integrating AMT within a log-MMSE. Other advances could include investigating noise spectral update rates and how these might influence quality improvement. Finally, the use of an integrated AMT within the GMMSE enhancement method could be employed further for individuals with hearing loss.

REFERENCES


