



저작자표시-동일조건변경허락 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이차적 저작물을 작성할 수 있습니다.
- 이 저작물을 영리 목적으로 이용할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



동일조건변경허락. 귀하가 이 저작물을 개작, 변형 또는 가공했을 경우에는, 이 저작물과 동일한 이용허락조건하에서만 배포할 수 있습니다.

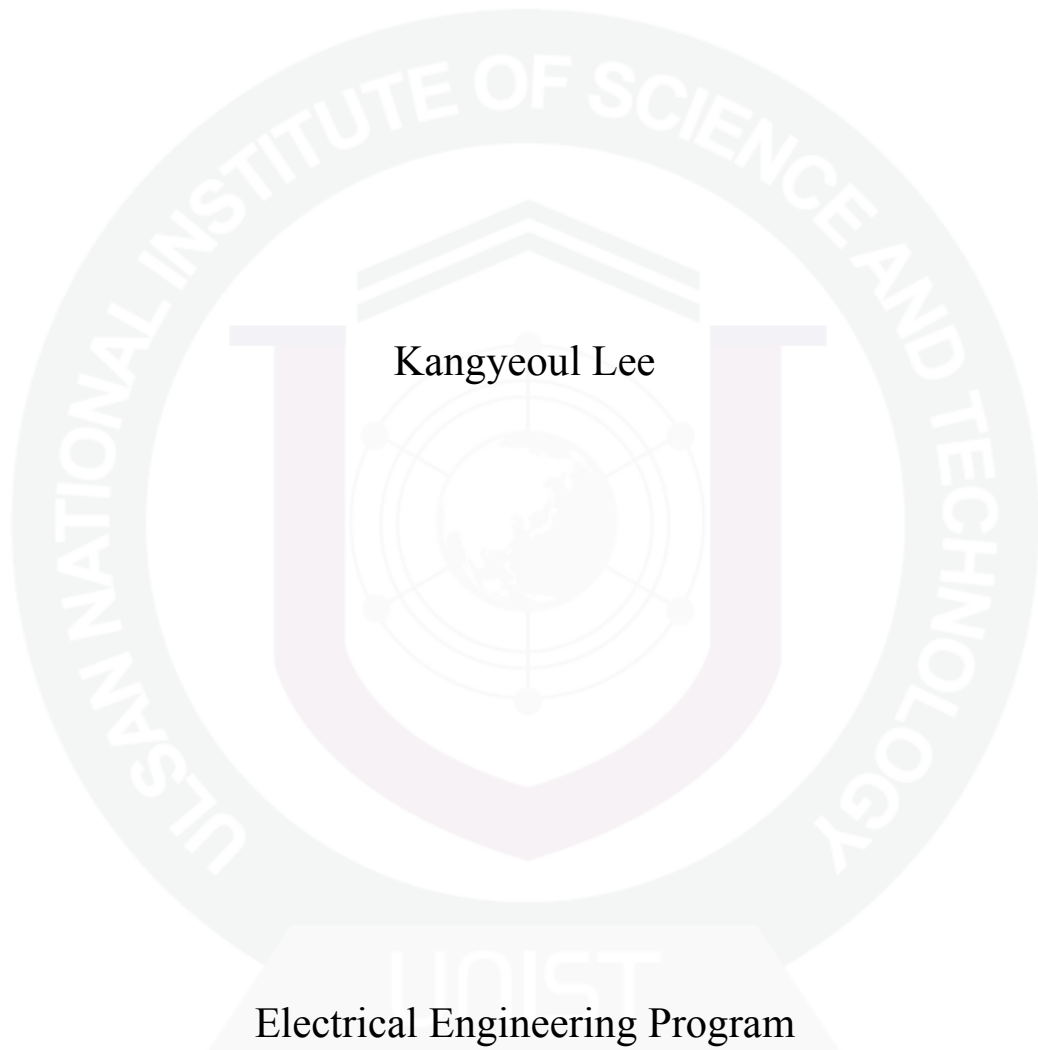
- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#) 

Efficient Noise Suppression for Robust Speech Recognition



Kangyeoul Lee

Electrical Engineering Program

Graduate School of UNIST

Efficient Noise Suppression for Robust Speech Recognition

A thesis
submitted to the Graduate School of UNIST
in partial fulfillment of the
requirements for the degree of
Master of Science

Kangyeoul Lee

. . 2013 of submission

Approved by



Major Advisor

Gil-Jin Jang

Efficient Noise Suppression for Robust Speech Recognition

Kangyeoul Lee

This certifies that the thesis of Kangyeoul Lee is approved.

. . . 2013. of submission

signature



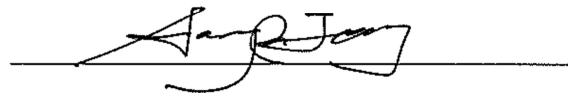
Thesis Supervisor: Gil-Jin Jang

signature



Jae-Young Sim: Thesis Committee Member #1

signature



Sangbae Jeong: Thesis Committee Member #2

Abstract

This thesis addresses the issues of single microphone based noise estimation technique for speech recognition in noise environments. A lot of researches have been performed on the environmental noise estimation, however most of them require voice activity detector (VAD) for accurate estimation of noise characteristics. I propose two approaches for efficient noise estimation without VAD. The first approach aims at improving the conventional quantile-based noise estimation (QBNE). I fostered the QBNE by adjusting the quantile level (QL) according to the relative amount of added noise to the target speech. Basically, we assign two different QLs, i.e., binary levels, according to the measured statistical moment of log scale power spectrum at each frequency. The second approach is applying dual mixture parametric model in computing likelihoods of speech and non-speech classes. I used dual Gaussian mixture model (GMM) and Rayleigh mixture model (RMM) for the likelihoods. From the assumption that speech is generally uncorrelated to the environmental noises, the noise power spectrum can be estimated by using each mixture model parameter of speech absence class.

I compared the proposed methods with the conventional QBNE and minimum statistics based method on a simple speech recognition task in various signal-to-noise ratio (SNR) levels. Based on the experimental results, the proposed methods are shown to be superior to the conventional methods.

Contents

1.	Introduction-----	1
2.	Single Microphone based Noise Suppression-----	3
2.1	Spectral subtraction-----	3
2.2	Wiener filter-----	6
3.	Noise Estimation-----	8
3.1	Minimum statistics based noise estimation-----	8
3.1.1	Principle of the minimum statistics method-----	8
3.1.2	Deriving optimal time-frequency dependent smoothing factor-----	9
3.1.3	Bias factor-----	10
3.2	Quantile based noise estimation-----	11
3.3	Histogram based noise estimation-----	12
4.	Proposed Method-----	14
4.1	Binary quantile level based noise estimation-----	14
4.1.1	Kurtosis based gaussianity estimation-----	18
4.1.2	Negentropy based gaussianity estimation-----	19
4.1.3	Extended infomax algorithm based gaussianity estimation-----	21
4.2	Dual mixture model based noise estimation-----	22
4.2.1	Dual Gaussian mixture model based noise estimation-----	23
4.2.2	Dual Rayleigh mixture model based noise estimation-----	25
5.	Experimental Results-----	28
6.	Conclusion-----	31

List of Figures

Figure 1.1 Single microphone based noise suppression system-----	1
Figure 2.1 Single microphone based noise suppression model-----	3
Figure 2.2 Half wave rectification for non-negative value-----	4
Figure 2.3 Definition of the error between clean and estimated speech-----	6
Figure 3.1 Quantiles of PSD at 300Hz, 1.5kHz and 3kHz in speech signal of the TIMIT corpus----	12
Figure 3.2 Speech signal, magnitude spectrum and histogram at 2 kHz according to SNR changes-	13
Figure 4.1 Histogram of log-scale PSD for various noises and clean speech, measured at 2 kHz----	15
Figure 4.2 Histogram of log-scale PSD for various SNR levels, measured at 2 kHz-----	16
Figure 4.3 Procedures of binary quantile level based noise reduction-----	17
Figure 4.4 Kurtosis changes according to the type of standard symmetric distribution-----	18
Figure 4.5 Non-quadratic function for measuring gaussianity-----	20
Figure 4.6 GMM based likelihoods for speech presence and absence-----	24
Figure 4.7 Dual GMM for 0 dB SNR noisy speech with spectral histogram at 2 kHz-----	25
Figure 4.8 Rayleigh distribution changes according to sigma parameter-----	26
Figure 4.9 Dual RMM for 0 dB SNR noisy speech with spectral histogram at 2 kHz-----	27
Figure 5.1 Recognition rate comparison binary QL based methods under 16 dB SNR-----	29
Figure 5.2 Recognition rate comparison biary QL based methods under 8 dB SNR-----	29

List of Tables

Table 4.1	Suitable quantile level according to various SNR conditions-----	14
Table 5.1	The utterance format of the SSC database-----	28
Table 5.2	Results of speech recognition experiment-1 (binary QL based methods)-----	28
Table 5.3	Results of speech recognition experiment-2 (mixture model based methods)-----	30

List of Abbreviations

ASR	Automatic Speech Recognition
DFT	Discrete Fourier Transform
EM	Expectation Maximization
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HTK	Hidden Markov model Toolkit
MFCC	Mel-Frequency Cepstral Coefficients
MLE	Maximum-Likelihood Estimator
MMSE	Minimum Mean-Squared Error
MS	Minimum Statistics
PSD	Power Spectral Density
QBNE	Quantile Based Noise Estimation
QL	Quantile Level
RHS	Right Hand Side
RMM	Rayleigh mixture model
SNR	Signal-to-Noise Ratio
STFT	Short-Time Fourier Transform
VAD	Voice Activity Detector
WSS	Wide Sense Stationary

1. Introduction

Due to recent advances in technology, the use of ASR is extremely increased by spread of using smart devices. Although there are several important factors to enhance speech recognition rate, purity of speech is regarded as most crucial one. Unfortunately, the speech is always exposed to numerous acoustic background noises in recording environment. The service suppliers such as wireless telecommunication companies or application providers for smart devices should take care of handling various background noises, since we commonly use the devices in outdoors.

So far, a lot of noise suppression algorithms are proposed, and they are mostly based on spectral subtraction that is first introduced by Boll in 1979 [1]. It assumes that additive noise changes slowly over time and uncorrelated with speech and approximates the PSD of the noise signal by an average in non-voice periods. Most conventional methods depend on VAD which can detect speech presence. However it is not easy to distinguish speech or noise, its performance varies a lot in accordance with types and the amount of additive noises.

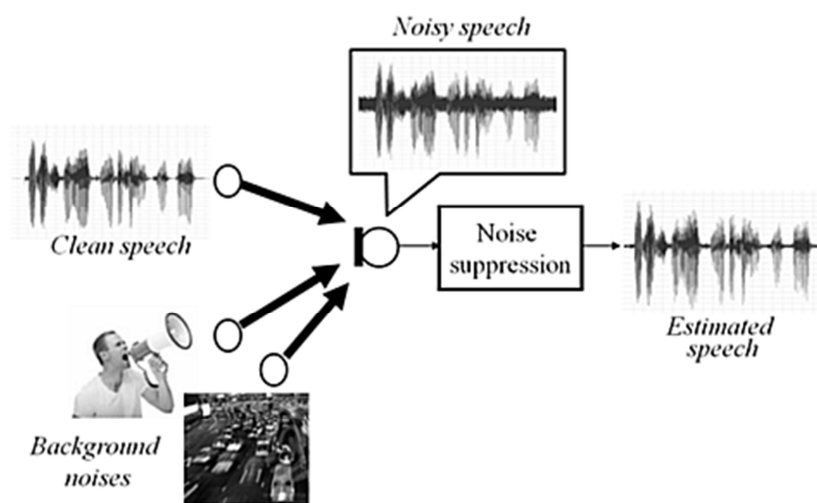


Figure 1.1 Single microphone based noise suppression system

So, severe methods based on spectral subtraction are proposed recently to eliminate VAD. MS based noise estimation [2] is notable for VAD independent method. It tracks noise power spectrum through taking minimum value of smoothed PSDs of noisy speech. QBNE [3] is another wide use method that is not required VAD. The basic concept of QBNE succeeded to MS based method but it assumes that the noise PSD is contained for significant percentage of noisy speech PSD instead taking minimum value. However, due to intrinsic assumptions of the noise characteristics, both methods suffer from performance variation in different noise conditions. Since minimum statistics based method is biased to minimum value of PSD, it cannot eliminate noise appropriately in highly noisy environment. On

the other hand, quantile based method estimates noise power spectrum using fixed QL in the distribution of the noisy signal. So it tends to suppress speech when the amount of additive noise is very small, i.e., high SNR conditions.

A novel of first approach is not only remedying the shortcomings of MS and QBNE, but also eliminating the need for VAD. Based on observed log PSD of stationary noise and speech, I can aware that the distribution of a stationary noise is close or peakier than Gaussian (super-Gaussian), while a speech signal is spreader than Gaussian (sub-Gaussian). Therefore, I impose binary quantile levels according to the measured statistical moments of the log PSD at each frequency.

A contrast function is used to decide the super-Gaussian (positive) and the sub-Gaussian (negative) by distance of the given distribution [4], and I adjusted higher quantile level when the distribution is Gaussian or super-Gaussian.

The second approach is started from speech presence likewise VAD problem. I estimated the noise power spectrum by likelihood of speech absence class. Dual GMM and RMM are used for likelihoods and low mean and sigma parameter are used for noise power estimation, respectively.

After the estimation, a time-domain Wiener filter suppressing the found noise PSD is derived from the noise estimate and applied to the input noisy speech signal. ASR experiments are carried out on speech separation challenge database [5] to verify improvement of proposed methods. The proposed method shows stable performance over various SNR conditions, while the conventional methods show degraded performance in high or low SNRs.

The deployment of this thesis is as follows; short explanation for single microphone based noise suppression techniques in Chapter 2 and Chapter 3 is for conventional noise estimation methods which are independent for VAD. Our proposed noise estimation methods are described in Chapter 4, and Chapter 5 summarizes experimental results of speech recognition. And finally, Chapter 6 concludes this thesis with future extensions.

2. Single microphone based noise suppression

Single microphone can be used to estimate and effectively suppress the stationary noise without compromising voice quality. In a single microphone environment, the same microphone will be used to capture voice and noise. The noise in signal is suppressed by severe algorithms that examine the frequency spectrum and segment it into many frames. Each frame is analyzed by its amplitude characteristics. So far, a lot of single microphone based noise reduction methods have been proposed in the field of noise reduction. Most of them are based on Boll's spectral subtraction [1], which is first introduced in 1979, because it has powerful advantages such as robustness and low complexity.

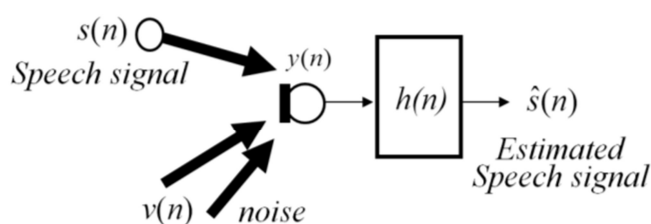


Figure 2.1 Single microphone based noise suppression model

2.1. Spectral subtraction

Basically, the spectral subtraction method is performed by deducting noise spectrum from noisy speech spectrum. The noise reduction problem with a single microphone input is formulated by

$$y(n) = s(n) + v(n), \quad (1)$$

where $y(n)$, $s(n)$ and $v(n)$ are noisy input, clean speech and additive noise signals respectively. And it is assumed that speech signal and noise are uncorrelated. So the autocorrelation $R_{yy}(\tau)$ of input signal can be expressed as

$$\begin{aligned} R_{yy}(\tau) &= E\{y(n)y(n+\tau)\} \\ &= E\{(s(n) + v(n))(s(n+\tau) + v(n+\tau))\} \\ &= E\{s(n)s(n+\tau)\} + E\{v(n)v(n+\tau)\} \\ &= R_{ss}(\tau) + R_{vv}(\tau) \end{aligned} \quad (2)$$

where, $R_{ss}(\tau)$ and $R_{vv}(\tau)$ are autocorrelation function of clean speech and noise. By forcing Fourier transform to both sides of equation (2), we can represent it in spectral domain, because

autocorrelation function and power spectrum are Fourier transform pair.

$$P_y(\omega) = P_s(\omega) + P_v(\omega) \quad (3)$$

$P_y(\omega)$, $P_s(\omega)$ and $P_v(\omega)$ are power spectrum of input signal, clean speech and noise, respectively.

From equation (3), we can obtain estimated speech power spectrum as

$$\hat{P}_s(\omega) = P_y(\omega) - \hat{P}_v(\omega) \quad (4)$$

where, $\hat{P}_s(\omega)$ and $\hat{P}_v(\omega)$ are estimated power spectrum of speech and noise. Now, equation (3) can be rewritten as equation (4) for the magnitudes of signal.

$$|Y(\omega)|^2 = |S(\omega)|^2 + |V(\omega)|^2 \quad (5)$$

$Y(\omega)$, $S(\omega)$ and $V(\omega)$ are Fourier transform of $y(n)$, $s(n)$ and $v(n)$, respectively. And we can also rewrite equation (4) as,

$$|\hat{S}(\omega)|^2 = \max\left(|Y(\omega)|^2 - |\hat{V}(\omega)|^2, 0\right). \quad (6)$$

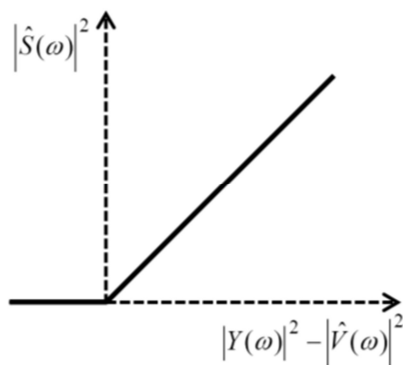


Figure 2.2 Half wave rectification for non-negative value

where, $|\hat{V}(\omega)|^2$ is the estimated power spectrum of $v(n)$. And it takes zero when estimated speech signal is less than 0 for half wave rectification. Finally, we can obtain noise suppressed speech $\hat{s}(n)$ from taking inverse DFT after square rooting RHS of equation (6).

$$s(n) = IDFT \left[\sqrt{|\hat{S}(\omega)|^2} \right] = IDFT \left[\sqrt{\max(|Y(\omega)|^2 - |\hat{V}(\omega)|^2, 0)} \right] \quad (7)$$

The phase of the noisy speech is not changed, since human does not sensitive with phase difference. Uppers can be expressed in time domain filter coefficient $h(n)$ as Figure 2.1. Estimated speech signal can be modeled in spectral domain as follows

$$\hat{S}(\omega) = H(\omega)Y(\omega) \quad (8)$$

where, $H(\omega)$ is spectral filter coefficient for spectral subtraction.

$$|H(\omega)|^2 = \frac{|\hat{S}(\omega)|^2}{|Y(\omega)|^2} = \frac{\max(|Y(\omega)|^2 - |\hat{V}(\omega)|^2, 0)}{|Y(\omega)|^2}, H(\omega) = \sqrt{\frac{\max(|Y(\omega)|^2 - |\hat{V}(\omega)|^2, 0)}{|Y(\omega)|^2}} \quad (9)$$

Taking inverse DFT equation (9), finally we are able to acquire impulse response $h(n)$.

$$h(n) = IDFT \left[\sqrt{\frac{\max(|Y(\omega)|^2 - |\hat{V}(\omega)|^2, 0)}{|Y(\omega)|^2}} \right] \quad (10)$$

Like this, the principle of spectral subtraction is very simple but it can apply only for stationary noises, as mentioned before. However most noises are generally non-stationary in real world, musical noises are often remained after the filtering. Hence we overestimate the noise power spectrum density occasionally to suppress more noise.

$$H(\omega) = \sqrt{\frac{\max(|Y(\omega)|^2 - \alpha|\hat{V}(\omega)|^2, 0)}{|Y(\omega)|^2}} \quad (11)$$

Where α is overestimation factor which is experimentally determined. However overestimation sometimes leads signal distortion because it not only suppresses noise, but also eliminates speech elements.

2.2. Wiener filter

In time domain filtering, we can obtain estimated speech as

$$\hat{s}(n) = \sum_{l=-\infty}^{\infty} h(l)y(n-l). \quad (12)$$

The definition of the error difference between clean speech signal and estimated speech signal is equivalent to equation (13) and it can represent graphically as Figure 2.3..

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{l=-\infty}^{\infty} h(l)y(n-l) \quad (13)$$

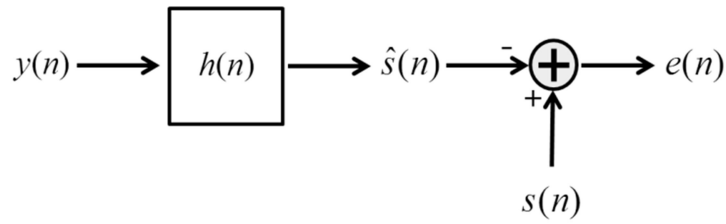


Figure 2.3 Definition of the error between clean and estimated speech

The impulse response $h(n)$ of the filter is derived in the MMSE sense by minimizing cost function J which is mean squared error.

$$J = E\left\{(s(n) - \hat{s}(n))^2\right\} = E\left\{\left(s(n) - \sum_{l=-\infty}^{\infty} h(l)y(n-l)\right)^2\right\} \quad (14)$$

To obtain optimal $h(n)$, we take partial differential to J with regarding to $h(\tau)$ as follows.

$$\frac{\partial J}{\partial h(\tau)} = -2E\left\{\left(s(n) - \sum_{l=-\infty}^{\infty} h(l)y(n-l)\right)y(n-\tau)\right\} = 0 \quad (15)$$

$$\sum_{l=-\infty}^{\infty} h(l)R_{yy}(\tau-l) = R_{sy}(\tau) \quad (16)$$

Where $R_{yy}(\tau)$ is correlation between clean and noisy speech. Equation (16) is called ‘‘Wiener-Hopf equation’’. As we assumed before, speech and noise are uncorrelated, and speech and noisy speech are WSS, the autocorrelation function of $y(n)$ can be rewrote as

$$\begin{aligned}
R_{yy}(\tau) &= E\{y(n)y(n-\tau)\} \\
&= E\{(s(n) + v(n))(s(n-\tau) + v(n-\tau))\} \\
&= E\{s(n)s(n-\tau)\} + E\{v(n)v(n-\tau)\} \\
&= R_{ss}(\tau) + R_{vv}(\tau)
\end{aligned} \tag{17}$$

In the same manner, correlation between clean and noisy speech can be changed as (18).

$$\begin{aligned}
R_{sy}(\tau) &= E\{s(n)y(n-\tau)\} \\
&= E\{s(n)(s(n-\tau) + v(n-\tau))\} \\
&= E\{s(n)s(n-\tau)\} \\
&= R_{ss}(\tau)
\end{aligned} \tag{18}$$

Substituting the result of equation (17) and (18) to equation (16), we can obtain,

$$\sum_{l=-\infty}^{\infty} h(l)(R_{ss}(\tau-l) + R_{vv}(\tau-l)) = R_{ss}(\tau). \tag{19}$$

Taking DFT to both sides of equation (19), the frequency response $H(\omega)$ can be acquired.

$$\begin{aligned}
H(\omega) &= \frac{P_s(\omega)}{P_s(\omega) + P_v(\omega)} \\
&= \max\left(1 - \frac{|\hat{V}(\omega)|^2}{|Y(\omega)|^2}, 0\right)
\end{aligned} \tag{20}$$

As spectral subtraction did, Wiener filter is prohibited for negative value and can also perform in time domain by taking inverse DFT on equation (20).

3. Noise estimation

Spectral subtraction based noise suppression techniques need good noise power spectrum estimator as we derived in equation (9), (20). Most conventional noise estimation methods use VAD to estimate background noise. The VAD updates estimated noise whenever speech is absent. However short pause detection is difficult and the performance of the VAD varies according to the kinds and conditions of noise [6]. On the other hand, [2], [3] and [8] proposed noise estimators which continuously track noise power spectrum for each frequency band regardless of existence of speech. These methods have advantage for reducing errors from VAD.

Usually the noisy signal $y(n)$ is processed frame by frame for STFT, and each frame length is proper to set 10-30ms for processing. It is assumed that for the duration of a frame, $s(n)$ and $v(n)$ can be considered to be WSS process. We can represent equation (1) to equation (21) in frequency domain.

$$Y(\omega, t) = S(\omega, t) + V(\omega, t) \quad (21)$$

Where t is current time index and $Y(\omega, t)$, $S(\omega, t)$ and $V(\omega, t)$ are STFT versions of $y(n)$, $s(n)$ and $v(n)$ respectively.

3.1 Minimum statistics based noise estimation

Martin proposed a noise power spectrum estimator based on minimum statistics and optimal power spectrum smoothing. This method was founded on two reasonable factors. One of them is independency between speech and noise. It means that summation of clean speech and noise power spectrum is equivalent to noisy speech power spectrum. And it can be represented as

$$|Y(\omega, t)|^2 = |S(\omega, t)|^2 + |V(\omega, t)|^2 \quad (22)$$

Another factor is that noisy speech power spectrum becomes noise power spectrum occasionally when speech is absent. Hence we can obtain estimated noise power spectrum by tracking the minimum of the noisy speech for each frequency component.

3.1.1 Principle of the minimum statistics method

As mentioned before, it is assumed that the power spectrum of noisy speech is summation of speech and noise power spectrum. So, noise variance was estimated by tracking the minimum of noisy speech power spectrum over a fixed buffer length. The buffer length has to be chosen enough for mixing speech and noise signal. It was experimentally found out that approximately 0.8-1.4s gave good results.

For searching the minimum a first-order recursive version of the noisy speech power spectrum was used:

$$P(\omega, t) = \beta P(\omega, t-1) + (1-\beta) |Y(\omega, t)|^2 \quad (23)$$

Where β is a constant smoothing constant which is typically set between 0.9 to 0.95. To enhance the performance of the minimum statistics based method following procedures were added.

1. Replacing the constant smoothing factor in equation (23) with time-frequency dependent smoothing factor.
2. Deriving a bias factor for the noise estimate since the minimum tracking was biased towards lower values.
3. Improving tracking speed of the algorithm for increasing noise levels.

3.1.2 Deriving optimal time-frequency dependent smoothing factor

The smoothing parameter used in equation (23), had to be low value to follow the noise faster. On the other hand, it had to be close to one to keep the power of the minimum tracking as small as possible. Hence time and frequency dependent smoothing factor is needed in place of a fixed factor.

This was derived for speech absent region. The requirement was that the smoothed power spectrum $P(\omega, t)$ had to be equal to the noise power $|V(\omega, t)|^2$ during speech pauses. Hence the smoothing parameter was derived by minimizing the conditional mean squared error between $|V(\omega, t)|^2$ and $P(\omega, t)$ as follows,

$$E \left\{ \left(P(\omega, t) - |V(\omega, t)|^2 \right)^2 | P(\omega, t-1) \right\} \quad (24)$$

where

$$P(\omega, t) = \beta(\omega, t) P(\omega, t-1) + (1-\beta(\omega, t)) |Y(\omega, t)|^2 \quad (25)$$

Note that in equation (25) time-frequency dependent smoothing factor $\beta(\omega, t)$ was used instead of fixed factor as defined in (23). Substituting equation (25) to (24) and setting the first derivative to zero gave the optimal value for $\beta(\omega, t)$:

$$\beta_{opt}(\omega, t) = \frac{1}{1 + \left(\frac{P(\omega, t-1)}{|\hat{V}(\omega, t)|^2} - 1 \right)^2} \quad (26)$$

But in real time implementation, the value of estimated noise variance $|\hat{V}(\omega, t)|^2$ lags behind true noise variance. Hence some correction factor $\beta_c(t)$ was calculated using the ratio of averaged smoothed periodogram to estimated noise power. The final smoothing factor with the correction parameter was given as

$$\beta_{opt}(\omega, t) = \frac{\beta_{max}\beta_c(t)}{1 + \left(\frac{P(\omega, t-1)}{|\hat{V}(\omega, t-1)|^2} - 1 \right)^2} \quad (27)$$

where β_{max} is typically 0.96.

3.1.3 Bias factor

Since minimum is biased to low values, the bias factor for compensating the minimum of noisy speech power spectrum was derived using the statistics of minimum of the correlated PSD estimates of noisy speech. It was stated that since the distribution of $P(\omega, t)$ was scaled by $|\hat{V}(\omega, t)|^2$, the minimum statistics of the short term estimates $P_{min}(\omega, t)$ was also scaled by $|\hat{V}(\omega, t)|^2$. Thus the bias term was derived by finding the mean of minimum PSD for some $|\hat{V}(\omega, t)|^2 = 1$ which after simplification gave

$$B_{min}(\omega, t) \approx 1 + (D-1) \frac{2}{\tilde{Q}_{eq}(\omega, t)} \quad (28)$$

where D is the window length over which the minimum is found and $\tilde{Q}_{eq}(\omega, t)$ called “equivalent degrees of freedom”, is function of smoothed periodogram, and the previous noise variance. The unbiased noise estimate is finally obtained as

$$|\hat{V}(\omega, t)|^2 = B_{\min}(\omega, t)P_{\min}(\omega, t) \quad (29)$$

3.2 Quantile based noise estimation

MS based algorithm has advantage for tracking noise power continuously. However the minimum is sensitive for outliers. For more reliable estimation, Stahl proposed QBNE which takes q -th quantile of the noisy speech spectrum. QBNE assumes that the noise power for each frequency band is contained for significant percentage of noisy speech segment.

In order to estimate noise power spectrum, firstly, observed power spectrum frames $|Y(\omega, t_m)|^2, m = 0, \dots, D$ are sorted for a frequency band ω as below equation where D is fixed window length.

$$|Y(\omega, t_0)|^2 \leq |Y(\omega, t_1)|^2 \leq \dots \leq |Y(\omega, t_D)|^2 \quad (30)$$

The noise power spectrum for each frequency band can be estimated by taking q -th quantile as follows.

$$|\hat{V}(\omega, t)|^2 = |Y(\omega, t_{[qD]})|^2 \quad (31)$$

For instance, $q = 1$ yields the maximum, $q = 0$ yields the minimum and $q = 0.5$ the median. Stahl *et al* experimentally found that $q \approx 0.5$ is optimal quantile level for estimating noise power.

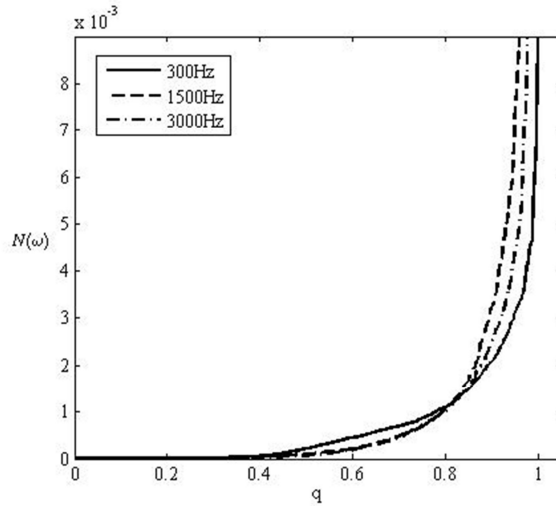


Figure 3.1 Quantiles of PSD at 300Hz, 1.5kHz and 3kHz in speech signal of the TIMIT corpus

The QBNE method has very simple concept and can track noise power spectrum without reference to speech presence as MS based method. However estimated noise power will be close to speech power component when little amount of noise is added. In that case, speech can be eliminated as well as additive noise when performing Wiener filtering. As a result, output signal will be distorted by overestimation of noise power spectrum and it leads low recognition rate.

3.3 Histogram based noise estimation

Hirsch proposed histogram based noise estimation approach [7]. It based on following observed statistical characteristics of distribute density function.

1. If SNR of noisy speech is low, magnitude spectrum is distributed to large value. On the other hand, magnitude spectrum is distributed to low value in high SNR environment.
2. If SNR of noisy speech is low, the distribution of magnitude spectrum gets broad. In other words, variance of distribution increases according to decreasing SNR.

From the observations, noise spectrum can be estimated by taking maximum frequency number of the histogram. This method is less affected by the signal SNR and doesn't need to identify whether speech or not, as MS and QBNE also did.

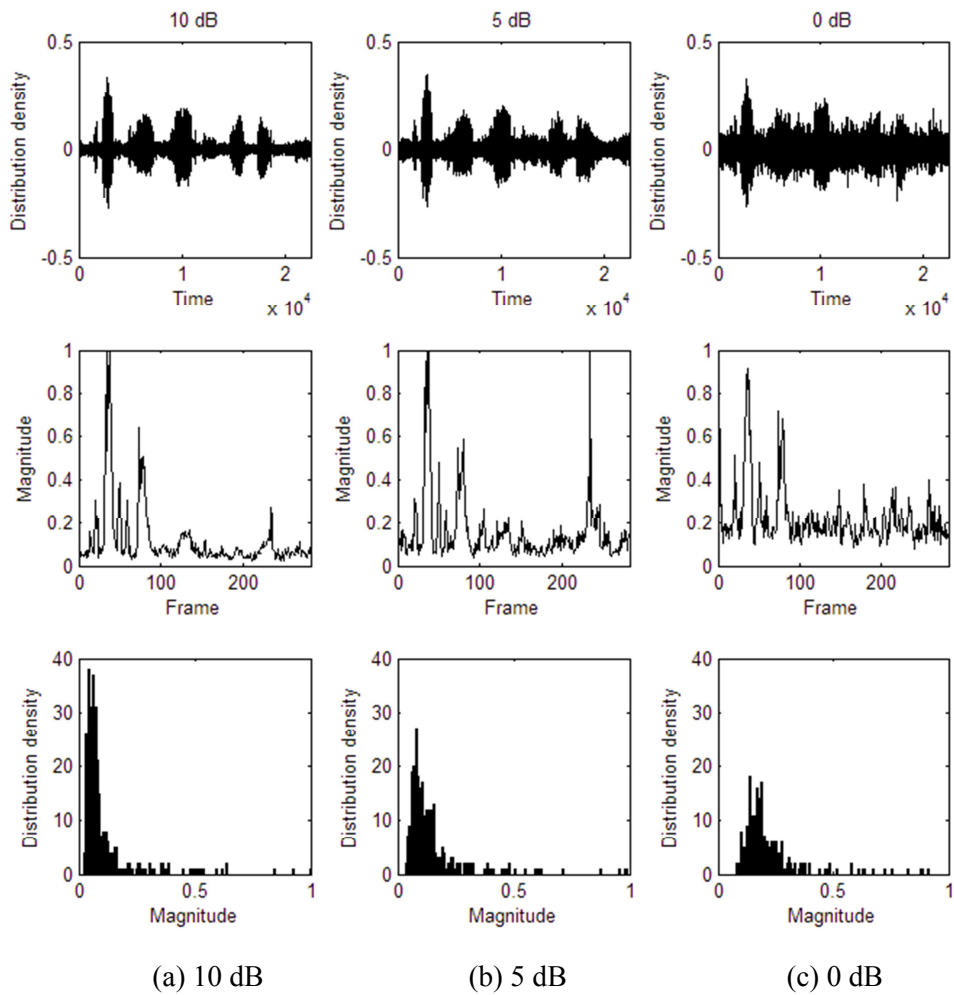


Figure 3.2 Speech signal, magnitude spectrum and histogram at 2 kHz according to change of SNR

We can practically find out those characteristics as Figure 3.2. The noisy speech data in Figure 3.2 is corrupted NOISEUS database by AURORA noise database. Figure 3.2-(a), (b) and (c) are speech signal, magnitude spectrum and the histogram of 10 dB, 5 dB and 0 dB, respectively. Comparing three different noisy speeches, we can easily know that the magnitude spectrum of noiseless speech is distributed around zero and the average of spectrum is less than others. To put it shortly, the less SNR of noisy speech, the more magnitude spectrum moves higher value. Finally, noise power estimated by taking maximum position of the histogram. And we can also observe that the variance gets bigger and bigger according to decreasing SNR.

4. Proposed methods

In this section, we introduce two different approaches for noise power estimation. The first approach remedies the weakness of QBNE in various SNR condition. The concept is that it adjusts QL according to estimated SNR instead using fixed QL. Another approach is regarding to dual mixture model based noise estimation. Each mixture is used for likelihood function of speech presence or absence class. And noise power can be estimated by taking long term average of the speech absence class. After the noise estimation, it is substituted to Wiener filter, i.e. equation (20), for noise reduction. Both approaches also take advantage that they can estimate noise power spectrum without using VAD like as conventional methods.

4.1 Binary quantile level based noise estimation

QBNE is good for tracking noise power spectrum without suffering outlier effect on the contrary to MS based method. Stahl *et al* proposed that taking median is the best way to estimate noise spectrum. However, in general cases, corruption level of noisy speech is different, applying fixed QL is not suitable for speech enhancement. Especially, in high SNR condition, speech power can be presumed as noise power and it leads to increase distortion of output signal. In practice, I simulated 5-different SNR levels from 0 dB to 20 dB noisy signal by adding AURORA2 noise sources [7] to speech signal from TIMIT corpus. And I experimentally found appropriate QLs for each SNR level by checking similarity between clean speech and filtered output. Table 4.1 shows the optimal quantile level for each SNR with a number of synthetic mixtures. There is significant correlation between appropriate QL and noise power.

SNR	q
0 dB	0.63
5 dB	0.44
10 dB	0.31
15 dB	0.22
20 dB	0.15

Table 4.1 Suitable quantile level according to various SNR conditions

One of the representative characteristics of stationary noise is that its power spectrum does not vary too much along time [8]. For example, Figure 4.1-(a), (b) and (c) show the histograms for 2 kHz log-scale power spectrum of the airport, babble and restaurant noise from AURORA-2 database,

respectively. As shapes of distributions, the histograms of noises are near to Gaussian distribution or super-Gaussian. On the other hand, as we discussed [8] [4], speech varies much faster than noise and power is spread broadly. This phenomenon makes a shape of log scale histogram deviate from Gaussian. Figure 4.1-(d) shows such shape of log-scale histogram which is much broader than Gaussian.

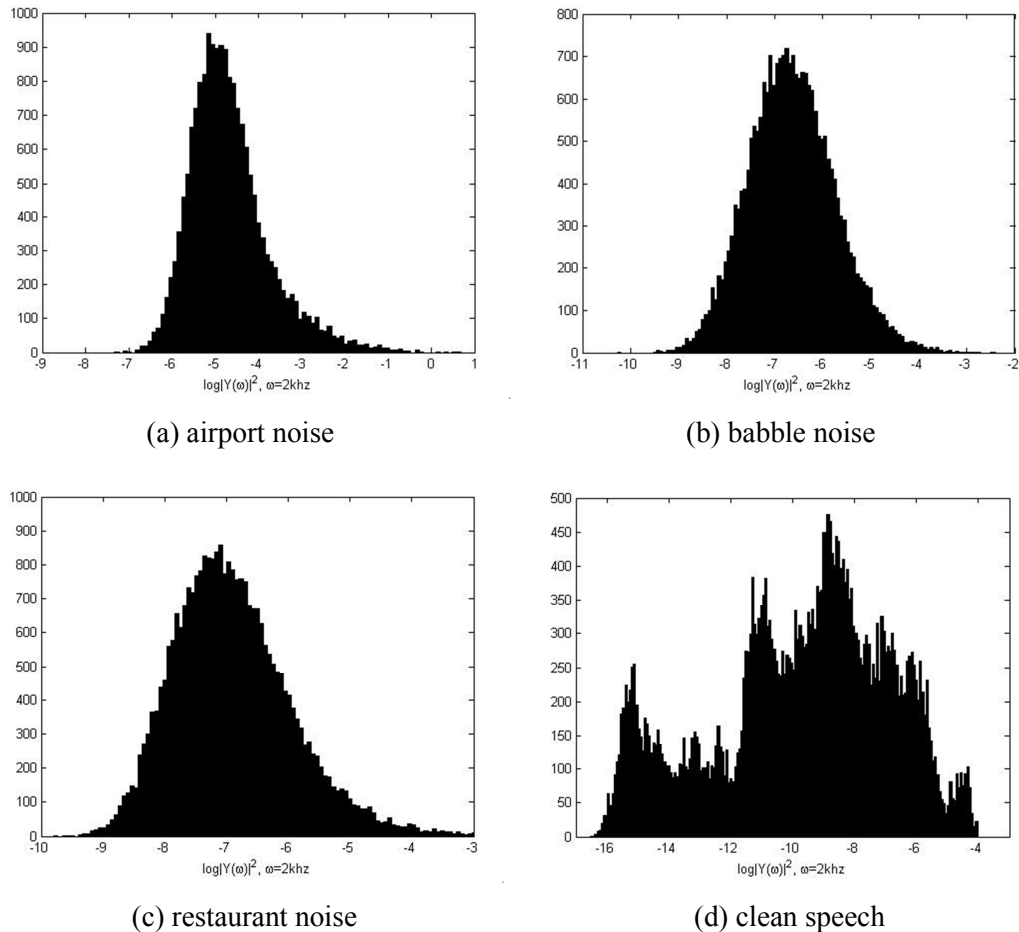


Figure 4.1 Histogram of log-scale PSD for various noises and clean speech, measured at 2 kHz

In noisy speech problem, I could expect that the more the environment is noisy, the more a distribution of logarithm power spectrum nears to super-Gaussian. Figure 4.2 demonstrates same histogram with varying SNR levels. The shape approaches to Gaussian as SNR decreases, and become broader as SNR increases.

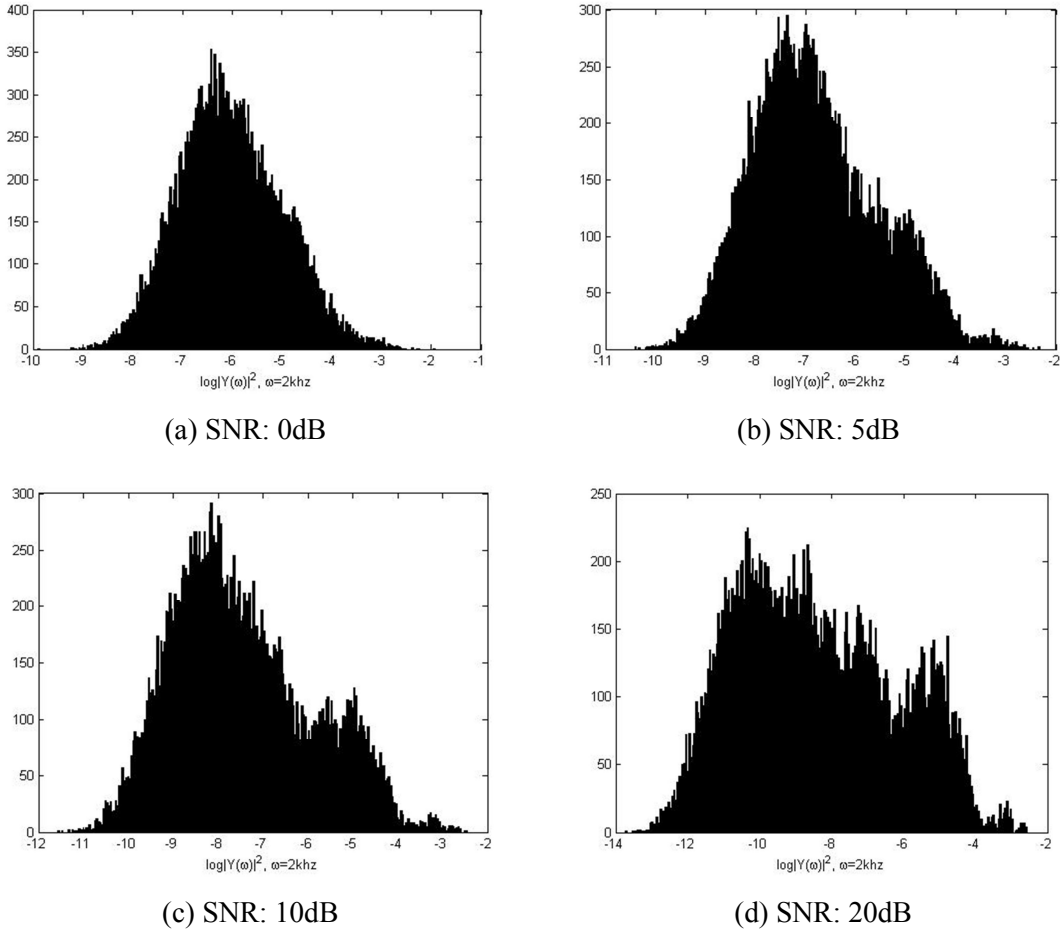


Figure 4.2 Histogram of log-scale PSD for various SNR levels, measured at 2 kHz

From those observations, a distribution is getting sharp, SNR of a frequency band is getting low. In this case, QL had to be adjusted to higher level. On the other hand, a distribution is getting obtuse means that SNR of input signal is getting high, so it needs to choose low QL. So gaussianity is the key factor for measuring degree of contamination of the noisy speech as well as selecting QL.

For example, let $\mathbf{b}(\omega, t)$ logarithm power spectrum buffer at current time t .

$$\mathbf{b}(\omega, t) = \left[\log\left(|Y(\omega, t-D)|^2\right), \log\left(|Y(\omega, t-D+1)|^2\right), \dots, \log\left(|Y(\omega, t)|^2\right) \right] \quad (32)$$

where D is buffer length. I define \hat{r} which represents gaussianity and segment SNR indirectly as equation (33). Input buffer $\mathbf{b}(\omega, t)$ needs to force zero mean with unit variance before function $f(\cdot)$ which estimates gaussianity of the distribution is performed.

$$\hat{r} = f\left(\frac{\mathbf{b}(\omega, t) - E(\mathbf{b}(\omega, t))}{\sigma_{\mathbf{b}}}\right) \quad (33)$$

Where σ_b is standard deviation of $\mathbf{b}(\omega, t)$. And then we need to map \hat{r} to $q_{\hat{r}}$ which is optimal quantile level for estimating noise power of segment $\mathbf{b}(\omega, t)$ according to the function $f(\cdot)$. Finally, noise power spectrum can be estimated as following equation.

$$|\hat{V}(\omega, t)|^2 = |Y(\omega, t_{[q, T]})|^2 \quad (34)$$

It is important to measure the gaussianity of the distribution, because it is highly correlated with amount of additive noise. So, appropriate quantile level can be selected after measuring the gaussianity.

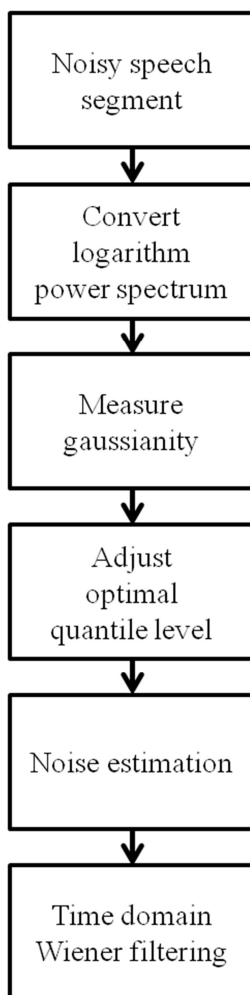


Figure 4.3 Procedures of binary quantile level based noise reduction

There are several methods for measuring the gaussianity, I employed 3 different measurements. The first measurement is kurtosis based gaussianity estimation which can classify sub-Gaussian,

Gaussian and super-Gaussian. The second one is negentropy which improved the disadvantage of kurtosis. However, it cannot provide information about super-Gaussian or sub-Gaussian. So, I applied extended infomax algorithm to overcome a drawback of negentropy.

4.1.1 Kurtosis based gaussianity estimation

One of the classical measurements of gaussianity is kurtosis or the fourth-order cumulant [9]. The kurtosis of random variable \mathbf{x} is defined by

$$\text{Kurt}(\mathbf{x}) = \frac{E\{\mathbf{x} - \mu\}^4}{\sigma^4}, \quad (35)$$

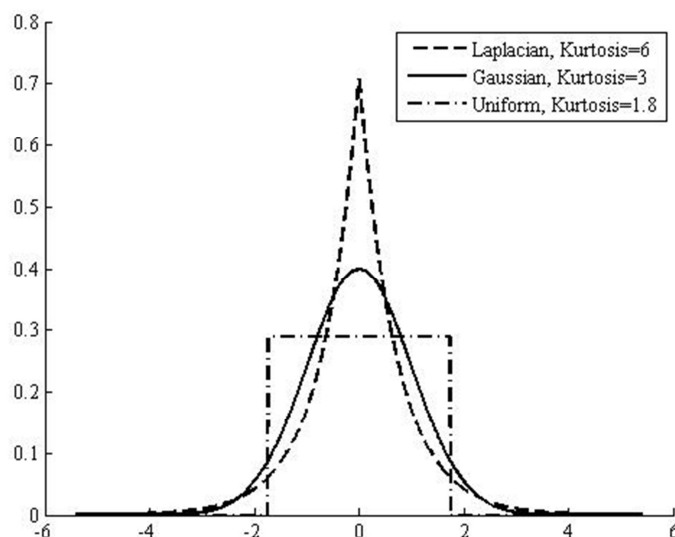


Figure 4.4 Kurtosis changes according to the type of standard symmetric distribution

where μ and σ are mean and standard deviation of \mathbf{x} , respectively. It can measure not only similarity with Gaussian but also distinguish sub-Gaussian or super-Gaussian. Kurtosis is equal to 3 for Gaussian random variable and also can be less than or larger than 3 for sub-Gaussian or super-Gaussian. Since the distribution of stationary noises follows super-Gaussian or Gaussian, I experimentally obtain binary quantile level as follows

$$q_K = \begin{cases} 0.5 & ; \text{Kurt}(\mathbf{x}) \geq 3 \\ 0.1 & ; \text{otherwise} \end{cases} \quad (36)$$

Finally, estimated noise power for each frequency band can be obtained as follows,

$$|\hat{V}(\omega, t)|^2 = |Y(\omega, t_{[q^k T]})|^2 \quad (37)$$

However, sometimes kurtosis cannot represent gaussianity precisely, because it is very sensitive for outlier [4].

4.1.2 Negentropy based gaussianity estimation

To measure Gaussianity more exactly, Hyvärinen proposed the negentropy based method which estimates nongaussianity of a random variable without suffering outlier effect [9].

The entropy of a random variable \mathbf{x} with probability density $f(\mathbf{x})$ is defined as

$$H(\mathbf{x}) = -\int f(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x} . \quad (38)$$

An important property of Gaussian distribution is that it has the maximum entropy among all distribution over the entire real axis $[-\infty, \infty]$. And uniform distribution has the maximum entropy among all distributions over a finite range. Based on this property, the negentropy is defined as

$$J(\mathbf{x}) = H(\mathbf{x}_G) - H(\mathbf{x}) \quad (39)$$

where \mathbf{x}_G is Gaussian random variable of the same mean and variance with \mathbf{x} . Since entropy of Gaussian is the largest over all random variable, negentropy is always greater than zero, and it is zero if and only if \mathbf{x} follows Gaussian random variable. The problem for using negentropy is that it is very difficult for computation hence approximations of negentropy are needed. Jones *et al* proposed approximated negentropy as follows

$$J(\mathbf{u}) \approx \frac{1}{12} E\{\mathbf{u}^3\}^2 + \frac{1}{48} \text{Kurt}(\mathbf{u})^2, \quad (40)$$

where \mathbf{u} is a random variable with zero mean and unit variance. However, this approximation is also suffers from the non-robustness due to kurtosis function. A better approximation is proposed by Hyvärinen as

$$\begin{aligned}
J(\mathbf{u}) &\approx \sum_{i=1}^p k_i [E\{G_i(\mathbf{u})\} - E\{G_i(\mathbf{g})\}]^2 \\
&\propto [E\{G(\mathbf{u})\} - E\{G(\mathbf{g})\}]^2
\end{aligned} \tag{41}$$

where k_i are some positive constants, and \mathbf{g} is a normal distribution. Although this approximation may be not accurate, equation (41) can be used to construct a measure of nongaussianity that is consistent in the sense that it is always non-negative, and equal to zero if \mathbf{u} follows Gaussian distribution. And G_i are some non-quadratic functions such as

$$G_1(\mathbf{u}) = \frac{1}{a} \log(\cosh(a\mathbf{u})), \quad G_2(\mathbf{u}) = -\exp\left(-\frac{\mathbf{u}^2}{2}\right) \tag{42}$$

where $1 \leq a \leq 2$ is suitable constant. Since results of $E\{G(\mathbf{u})\}$ indicates that how \mathbf{u} is close to Gaussian distribution, I used it to measure the gaussianity.

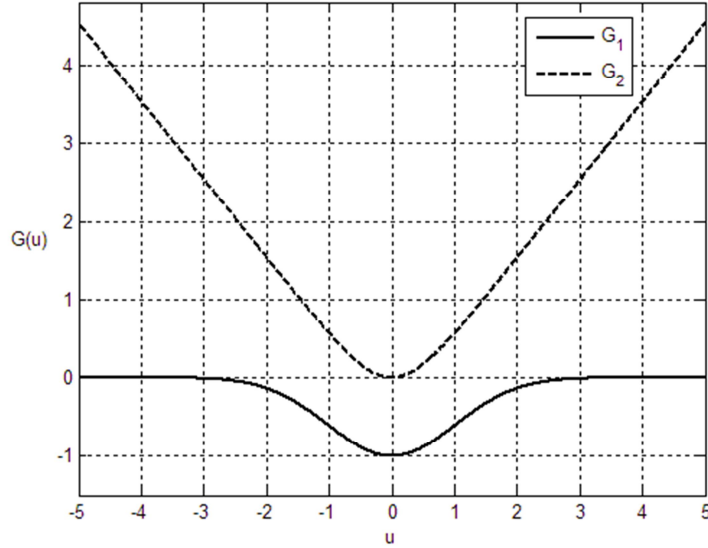


Figure 4.5 Non quadratic function for measuring gaussianity

In this manner, function $G(\cdot)$ is useful for adjusting QLs instead of kurtosis. If distribution \mathbf{u} is getting sharp, $E\{G(\mathbf{u})\}$ gets small. It means that SNR of a frequency band is getting low, high QLs are more suitable. While $E\{G(\mathbf{u})\}$ increased, distribution \mathbf{u} is getting obtuse. In other words, if SNR of input signal is getting high, we need to choose low QLs. We define $\hat{\tau}_i$ which can estimate gaussianity and segment SNR indirectly as follow,

$$\hat{r}_i = E \left\{ G_i \left(\frac{\mathbf{b}(\omega, t) - E\{\mathbf{b}(\omega, t)\}}{\sigma_{\mathbf{b}}} \right) \right\}, i = \{1, 2\} \quad (43)$$

where, i is a type of quadratic function and $\sigma_{\mathbf{b}}$ is standard deviation of $\mathbf{b}(\omega, t)$. Basically, QBNE works well in very noisy condition. As mentioned before, \hat{r}_i which is smaller than or equal to some threshold means that the distribution of log scale power spectrum is close to Gaussian, therefore the signal is low SNR. Optimal quantile level corresponding to \hat{r}_i is experimentally found as below equation.

$$q_N = \begin{cases} 0.5 & ; \hat{r}_1 < 0.447 \text{ or } \hat{r}_2 < -0.692 \\ 0.1 & ; \textit{otherwise} \end{cases} \quad (44)$$

Finally, estimated noise power for each frequency band can be obtained as follows,

$$|\hat{V}(\omega, t)|^2 = |Y(\omega, t_{\lfloor q_N T \rfloor})|^2. \quad (45)$$

4.1.3 Extended infomax algorithm based gaussianity estimation

Negentropy based algorithm has appeared to overcome the disadvantage of kurtosis. However negentropy cannot classify the type of Gaussian, such as sub-Gaussian or super-Gaussian, as compared with kurtosis. Evaluating type of Gaussian is important because the distribution gets close to super-Gaussian means very noisy condition. To complement the shortcoming, we employ extended infomax algorithm based method.

$$I(\mathbf{u}) = E\{\text{sech}^2(\mathbf{u})\}E\{\mathbf{u}^2\} - E\{\tanh(\mathbf{u})\mathbf{u}\} \quad (46)$$

$$k = \text{sign}(I(\mathbf{u})) = \begin{cases} 1 & : \text{super-Gaussian} \\ 0 & : \text{Gaussian} \\ -1 & : \text{sub-Gaussian} \end{cases} \quad (47)$$

where \mathbf{u} is a random distribution with zero mean and unit variance. The decision factor, k , determines the shape of the distribution of \mathbf{u} : 1, 0, and -1 for super-Gaussian, Gaussian, and sub-Gaussian, respectively. If distribution \mathbf{u} becomes peakier, the value of the function $I(\mathbf{u})$ moves along the positive

direction. Estimated gaussianity \hat{r}_I which based on extended infomax algorithm can be defined as following equation :

$$\hat{r}_I = I\left(\frac{\mathbf{b}(\omega, t) - E(\mathbf{b}(\omega, t))}{\sigma_{\mathbf{b}}}\right) \quad (48)$$

Optimal quantile levels corresponding to \hat{r}_I are found experimentally, which is obtained by the following equation:

$$q_I = \begin{cases} 0.5 & ; \hat{r}_I > 0 \\ 0.1 & ; \textit{otherwise} \end{cases} \quad (49)$$

Finally, estimated noise power for each frequency band can be obtained as follows,

$$|\hat{V}(\omega, t)|^2 = |Y(\omega, t_{[q_I, T]})|^2 \quad (50)$$

4.2 Dual mixture model based noise estimation

This section addresses another VAD-free noise estimation approach. As we assumed that speech and noise are uncorrelated, the power spectrum of noisy speech can be summation of speech and noise power spectrum the long-term average aspect, as represented equation (3). And some frames contain speech and noise power, but the others are equivalent to noise power because of the speech pause. Therefore we employ a method for detecting the activity of the speech, so that measuring how much noise component is contained in a frame. In all of the short-time analysis frames at frequency ω , they are classified into the following two classes:

$$\begin{aligned} C_0(\omega) &= \left\{ t \mid |Y(\omega, t)|^2 = |S(\omega, t)|^2 + |V(\omega, t)|^2 \right\} \\ C_1(\omega) &= \left\{ t \mid |Y(\omega, t)|^2 = |V(\omega, t)|^2 \right\} \end{aligned} \quad (51)$$

where $C_0(\omega)$ and $C_1(\omega)$ are classes for speech presence and absence at frequency ω , respectively, and $C_0(\omega)$ is the complementary of $C_1(\omega)$. Using Bayes's rule, *a posteriori* probability, which discerns presence of the speech, can be denoted by,

$$\begin{aligned}
P\left(C_j(\omega) \mid |Y(\omega, t)|^2\right) &= \frac{P\left(|Y(\omega, t)|^2 \mid C_j(\omega)\right)P\left(C_j(\omega)\right)}{P\left(|Y(\omega, t)|^2\right)} \\
&= \frac{P\left(|Y(\omega, t)|^2 \mid C_j(\omega)\right)P\left(C_j(\omega)\right)}{\sum_{j=0}^1 P\left(|Y(\omega, t)|^2 \mid C_j(\omega)\right)P\left(C_j(\omega)\right)}, \quad j = \{0,1\}
\end{aligned} \tag{52}$$

where $j = \{0,1\}$ and $|Y(\omega, t)|^2$ are type of class and marginal probability, respectively. And $P(C_0(\omega))$ is a *priori* probability for speech presence and $P(C_1(\omega)) = 1 - P(C_0(\omega))$ is for speech absence.

4.2.1 Dual Gaussian mixture model based noise estimation

Gaussian distribution is common and representative probability density function. Therefore we assumed that the likelihood of $|Y(\omega, t)|^2$ given $C_j(\omega)$ follows univariate Gaussian density function with different mean and variance as

$$P\left(|Y(\omega, t)|^2 \mid C_j(\omega)\right) = \frac{1}{\sqrt{2\pi\sigma_j^2(\omega)}} \exp\left(-\frac{\left(|Y(\omega, t)|^2 - \mu_j(\omega)\right)^2}{2\sigma_j^2(\omega)}\right), \quad j = \{0,1\}. \tag{53}$$

where $\mu_j(\omega)$ and $\sigma_j^2(\omega)$ are mean and variance of each likelihood function, respectively. To obtain optimal Gaussian parameters, such as $P(C_j(\omega))$, $\mu_j(\omega)$ and $\sigma_j^2(\omega)$, MLE is used as below equation,

$$\begin{aligned}
\hat{P}(C_j(\omega)) &= \frac{1}{N} \sum_{t=1}^N P\left(C_j(\omega) \mid |Y(\omega, t)|^2\right) \\
\hat{\mu}_j(\omega) &= \frac{\sum_{t=1}^N P\left(C_j(\omega) \mid |Y(\omega, t)|^2\right) |Y(\omega, t)|^2}{\sum_{t=1}^N P\left(C_j(\omega) \mid |Y(\omega, t)|^2\right)} \\
\hat{\sigma}_j^2(\omega) &= \frac{\sum_{t=1}^N P\left(C_j(\omega) \mid |Y(\omega, t)|^2\right) \left(|Y(\omega, t)|^2 - \mu_j(\omega)\right)^2}{\sum_{t=1}^N P\left(C_j(\omega) \mid |Y(\omega, t)|^2\right)}
\end{aligned} \tag{54}$$

where N is the number of samples. Generally, Gaussian parameters are iteratively updated by the EM algorithm. The important thing is that the mean $\mu_0(\omega)$ of likelihood $P(|Y(\omega, t)|^2 | C_0(\omega))$ is assumed to be always greater than or equal to $\mu_1(\omega)$ for all ω , because the power spectrum of noisy speech which is composed with sum of noise and speech power is larger than noise power spectrum.

$$\mu_1(\omega) \leq \mu_0(\omega) \quad (55)$$

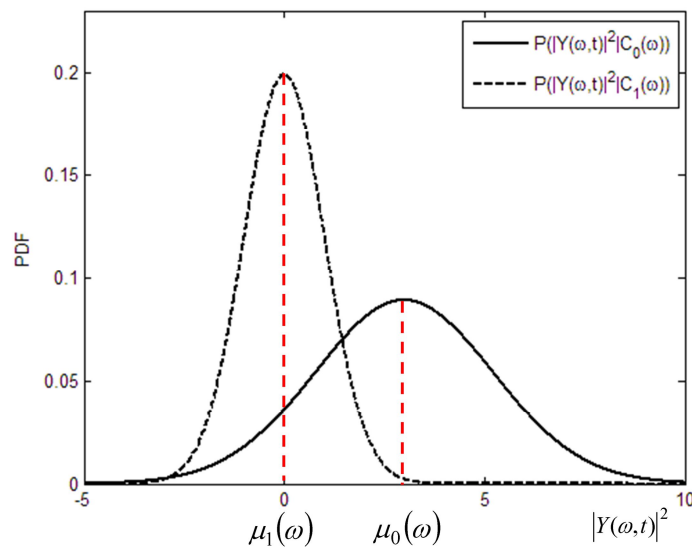


Figure 4.6 GMM based likelihoods for speech presence and absence

From this feature, we estimated the power spectrum of noise by taking long-term average of $P(|Y(\omega, t)|^2 | C_1(\omega))$ which is equivalent to $\hat{\mu}_1(\omega)$.

$$\begin{aligned} |\hat{V}(\omega, t)|^2 &= E \left\{ P \left(C_1(\omega) \middle| |Y(\omega, t)|^2 \right) \right\} \\ &= \hat{\mu}_1(\omega) \end{aligned} \quad (56)$$

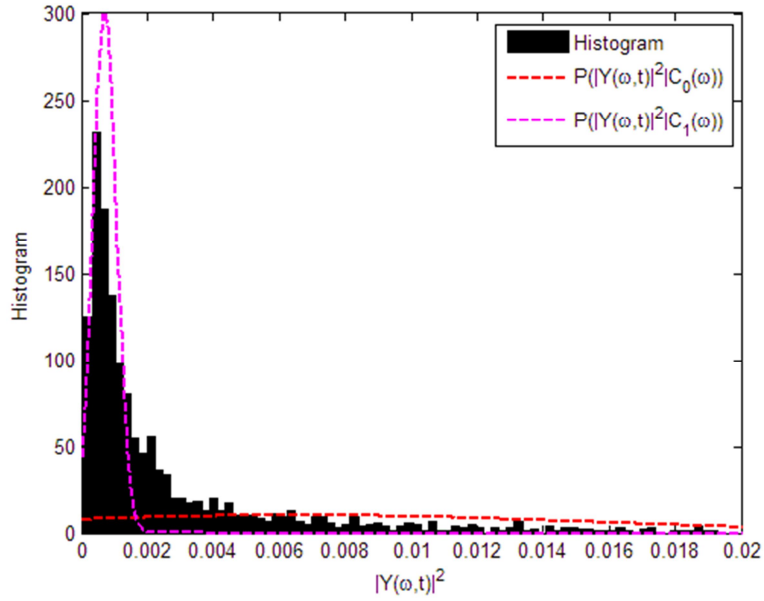


Figure 4.7 Dual GMM for 0 dB SNR noisy speech with spectral histogram at 2 kHz

4.2.2 Dual Rayleigh mixture model based noise estimation

It is common practice to apply Gaussian probability density function for likelihood, and GMM can represent any shape of distribution with less distortion. However, we restricted to dual mixture problem, sometimes GMM has a limitation that it cannot optimally approximate the histogram shapes. For estimating noise power spectra with little error, Rayleigh probability density function is more appropriate than Gaussian distribution. Because the power spectrum is non-negative value and Rayleigh distribution also defined to only positive values. Also, the spectral histogram of noisy speech is more close to Rayleigh distribution.

Rayleigh probability density function for random variable x is defined by

$$P(x) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right) \quad (57)$$

where σ is a sigma parameter which is different concept with standard deviation. The parameter σ can be estimated by MLE as follow

$$\hat{\sigma} \approx \sqrt{\frac{1}{2N} \sum_{i=1}^N x_i^2} \quad (58)$$

And the maximum value of the density function is equal to $1/\sigma\sqrt{e}$ and is reached when $x = \sigma$. The curve of the distribution is widely spread by increasing σ parameter as below Figure 4.8.

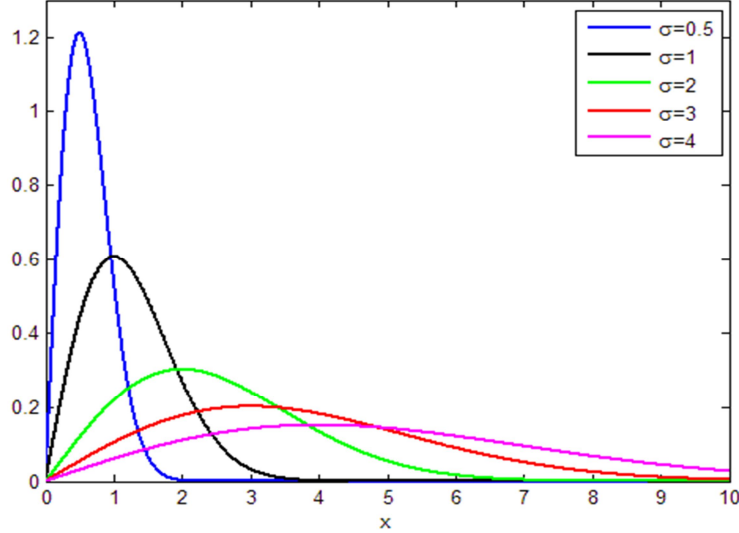


Figure 4.8 Rayleigh distribution changes according to sigma parameter

The likelihood of $|Y(\omega, t)|^2$ given $C_j(\omega)$ follows Rayleigh density function with different sigma parameter as

$$P\left(|Y(\omega, t)|^2 | C_j(\omega)\right) = \frac{|Y(\omega, t)|^2}{\sigma_j^2(\omega)} \exp\left(-\frac{|Y(\omega, t)|^4}{2\sigma_j^2(\omega)}\right), j = \{0, 1\}. \quad (59)$$

The optimal Rayleigh parameters, such as $P(C_j(\omega))$, $\sigma_j^2(\omega)$, can be updated by below equation,

$$\begin{aligned} \hat{P}(C_j(\omega)) &= \frac{1}{N} \sum_{t=1}^N P\left(C_j(\omega) | |Y(\omega, t)|^2\right) \\ \hat{\sigma}_j^2(\omega) &= \frac{\sum_{t=1}^N P\left(C_j(\omega) | |Y(\omega, t)|^2\right) |Y(\omega, t)|^4}{2 \sum_{t=1}^N P\left(C_j(\omega) | |Y(\omega, t)|^2\right)}. \end{aligned} \quad (60)$$

In the same manner with GMM, the sigma parameter $\sigma_0^2(\omega)$ of likelihood $P\left(|Y(\omega, t)|^2 | C_0(\omega)\right)$ is assumed to be always greater than or equal to $\sigma_1^2(\omega)$ for all ω , because larger sigma parameter indicate that the power spectrum distributed more widely.

$$\sigma_1^2(\omega) \leq \sigma_0^2(\omega) \quad (61)$$

From this feature, the noise power spectrum is estimated by taking the argument that maximize $P(|Y(\omega, t)|^2 | C_1(\omega))$ which is equivalent to $\sigma_1(\omega)$.

$$\begin{aligned} |\hat{V}(\omega, t)|^2 &= \arg \max_{|Y(\omega, t)|^2} P(C_1(\omega) | |Y(\omega, t)|^2) \\ &= \sqrt{\hat{\sigma}_1^2(\omega)} \end{aligned} \quad (62)$$

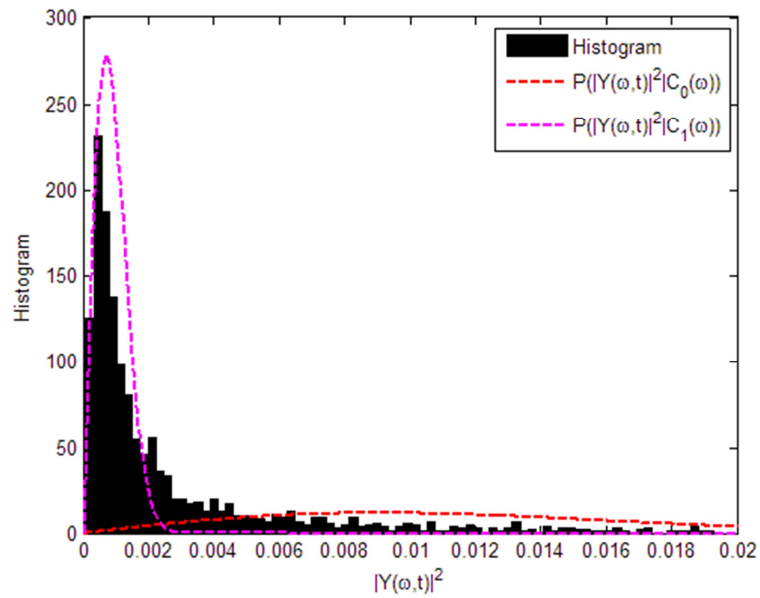


Figure 4.9 Dual RMM for 0 dB SNR noisy speech with spectral histogram at 2 kHz

5. Experimental results

To verify improvement, we compared the two types of proposed methods to the QBNE and MS based method separately, by ASR experiments on the Speech Separation Challenge (SSC) database [5]. The SSC database has clean speech set of 17,000 utterances for training, spoken by 34 different speakers. And training set is recorded in quiet condition without any background noise. Each utterance consists of 6 words in the format, such as “command-color-preposition-letter-number-adverb”. For example, “bin-blue-on-A-5-soon”.

Command	Color	Preposition	Letter	Number	Adverb
bin	white	at	A-Z (excluding W)	0-9 : *0 : zero	again
lay	blue	by			now
place	green	on			please
set	red	with			soon

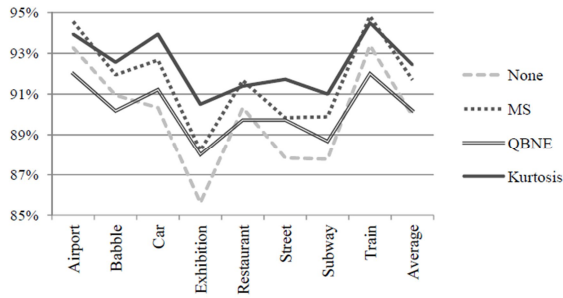
Table 5.1 The utterance format of the SSC database

The acoustic models of the words are built by the hidden Markov model toolkit (HTK). The features are extracted to 39 dimension vector which consists of 12 MFCCs plus log energy, plus their velocities and accelerations for every 10ms. We also employ separate testing set of 600 utterances which is exclusive with training set.

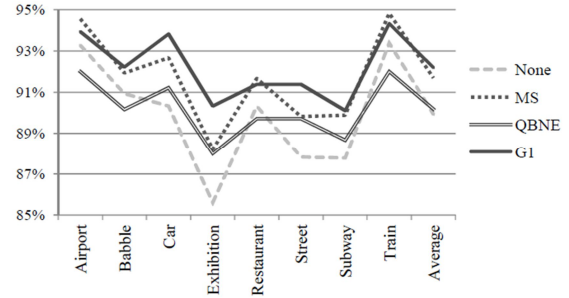
The proposed methods are evaluated on 8 different noise environments, “airport, babble, car, exhibition, restaurant, street, subway and train”, from the AURORA-2 database and noises are added to clean speeches. We simulated in 6 different SNR conditions, such as 20, 16, 12, 8, 4 dBs and clean speech. All hidden Markov Models (HMMs) are trained by clean speech to verify degree of noise reduction.

SNR	None	MS	QBNE	Proposed-1			
				Kurtosis	Neg-1	Neg-2	EI
Clean	97.56%	97.61%	95.39%	96.83%	96.00%	96.00%	96.61%
20dB	94.14%	94.83%	92.45%	94.74%	94.61%	94.48%	94.83%
16dB	89.94%	91.70%	90.19%	92.44%	92.19%	91.99%	92.68%
12dB	80.46%	84.26%	85.42%	87.36%	87.40%	87.28%	87.87%
8dB	64.62%	70.47%	75.83%	77.31%	77.65%	77.33%	78.09%
4dB	45.30%	52.08%	59.33%	60.04%	60.40%	60.31%	60.97%
Average	78.67%	81.82%	83.10%	84.79%	84.71%	84.57%	85.17%

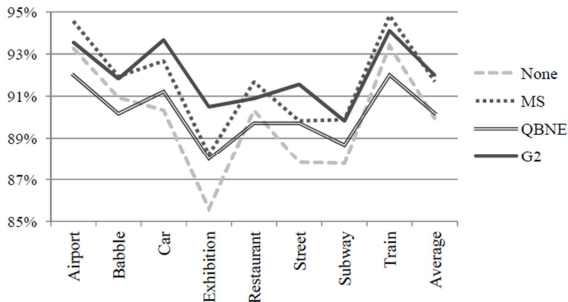
Table 5.2 Results of speech recognition experiment-1 (binary QL based methods)



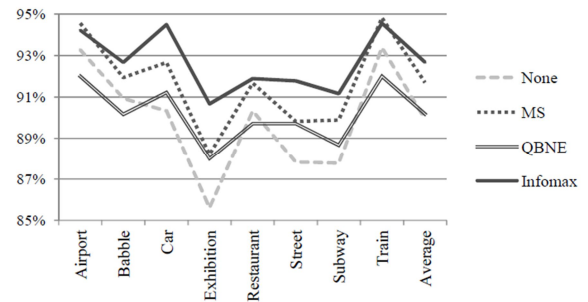
(a) Kurtosis based gaussianity measure



(b) Negentropy(G1) based gaussianity measure

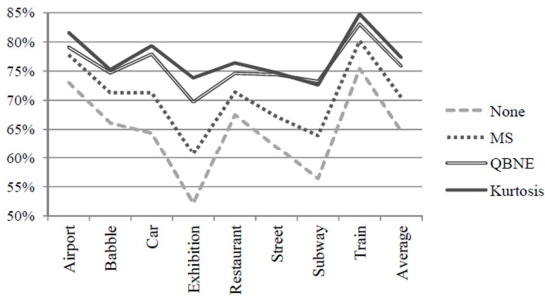


(c) Negentropy(G2) based gaussianity measure

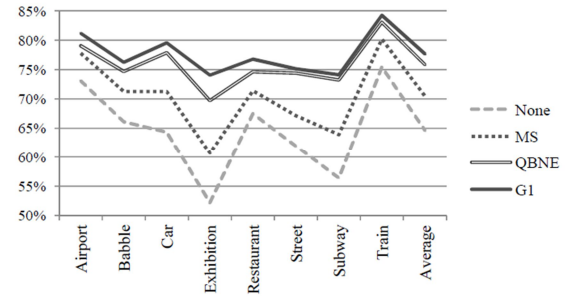


(d) Infomax based gaussianity measure

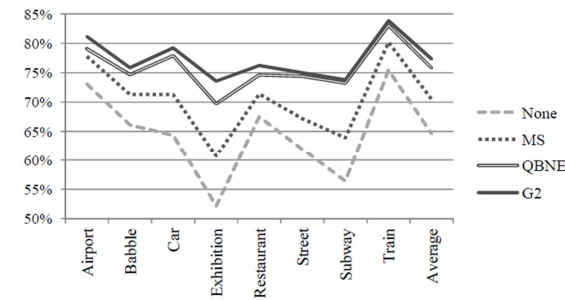
Figure 5.1 Recognition rate comparison binary QL based methods under 16 dB SNR



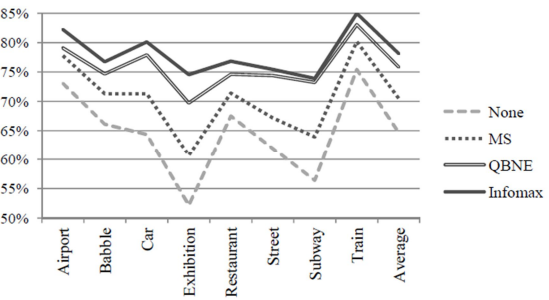
(a) Kurtosis based gaussianity measure



(b) Negentropy(G1) based gaussianity measure



(c) Negentropy(G2) based gaussianity measure



(d) Infomax based gaussianity measure

Figure 5.2 Recognition rate comparison binary QL based methods under 8 dB SNR

SNR	None	MS	QBNE	Proposed-2	
				GMM	RMM
Clean	97.56%	97.61%	95.39%	95.94%	97.11%
20dB	94.14%	94.83%	92.45%	93.25%	94.81%
16dB	89.94%	91.70%	90.19%	91.09%	92.97%
12dB	80.46%	84.26%	85.42%	86.88%	88.88%
8dB	64.62%	70.47%	75.83%	77.75%	79.60%
4dB	45.30%	52.08%	59.33%	61.17%	62.82%
Average	78.67%	81.82%	83.10%	84.35%	86.03%

Table 5.3 Results of speech recognition experiment-2 (mixture model based methods)

In accordance with type of proposed methods, we separately summarized the results of speech recognition experiments in Table 5.2 and 5.3, where “None” column lists the performance results without noise suppression. Each row is the average recognition rates over all noise types in specific SNRs.

For first proposed approach, all of the methods, based on kurtosis, negentropy and infomax, are about 6.1%, 6.0%, 5.9% and 6.5% better than “None” respectively, whereas QBNE and MS are 4.4% and 3.2% better. I found that QBNE and MS have discriminative pros and cons as shown in Figure 5.1 and 5.2. QBNE performed quite well under severe noisy conditions; on the contrary, the performance of MS becomes better as SNR increases, and best in clean condition. The binary quantile level based approaches take both advantages of QBNE and MS. In 12 dB SNR, the performances of the proposed methods are similar to MS on the average; in 8 dB SNR, they are slightly better and QBNE, and much better than the others. Among the approaches, extended infomax based noise estimation leads the best performance and kurtosis based method performs better than negentropy based methods. The major reason for better performance can be inferred that infomax algorithm and kurtosis can identify sub-Gaussian distribution in compared with negentropy.

For another type of approaches, which is founded on mixture model, the recognition rates are enhanced by approximately 5.7% and 7.4% on average compared with no processing. The performance improvement changes are analogous to binary QL noise estimation. RMM based method draws better outcome than GMM approach, thus RMM is more appropriate likelihood for power spectrum.

In summary, the speech recognition results prove that the proposed methods are quite stable and overcome the limit of conventional methods in various noise types and noise levels regardless of types of proposed noise estimation.

6. Conclusion

This thesis proposed two noise suppression approaches which do not need to use VAD. The MS based method and QBNE are conventional VAD-free methods, however they have drawbacks for various SNR environments. To overcome the limitations, I firstly proposed adjusting two different QLs based method according to estimated SNR. To apply binary QLs, the shapes of the log power spectral densities for individual frequency band are compared to Gaussian by statistical moments. The proposed methods employed three different gaussianity measurements, such as kurtosis, negentropy and extended infomax algorithm. The second noise estimation approach is based on dual mixture model. Dual GMM and RMM are applied to likelihoods of speech presence and absence, and we estimate the noise power spectrum by taking average of speech absence likelihood which has low mean and low sigma parameter, respectively. After estimating noise power spectrum, we substitute it to Wiener filtering for spectral suppression. To evaluate the proposed methods, we performed speech recognition experiments on a simple speech recognition task. Experimental results show that the proposed methods work well in various SNR conditions compared to conventional methods. Future research issues include finding new contrast functions for better approximation of noise presence.

References

- [1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113-120, Apr. 1979.
- [2] R. Martin, "Spectral subtraction based on minimum statistics," *Proc. 7th European Signal Processing Conf., EUSIPCO-94*, Edinburgh, Scotland, 13-16 September 1994, pp. 1182-1185.
- [3] Volker Stahl, Alexander Fischer, and Rolf Bippus, "Quantile based noise estimation for spectral subtraction and wiener filtering," in *Proceedings of ICASSP*, vol. 3, pp. 1875–1878, 2000.
- [4] Te-Won Lee, Mark Girolami, and Terrence J. Sejnowski, "Independent Component Analysis Using an Extended Infomax Algorithm for Mixed Subgaussian and Supergaussian Sources", *Neural Computation* 1999, 11:2, 417-441
- [5] M. Cooke, J. Hershey, and S. Rennie, "Monaural speech separation and recognition challenge," *Computer Speech & Language*, vol. 24, no. 1, pp. 1-15, 2010.
- [6] Gil-Jin Jang and Hoon-Young Cho, "Efficient spectrum estimation of noise using line spectral pairs for robust speech recognition," *Electronics Letters*, vol. 47, no. 25, pp. 1399–1401, 2011.
- [7] D. Pearce and H. Hirsch, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy condition," in *Proceedings of INTERSPEECH*, pp. 29-32, Oct. 2000.
- [8] H. G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Proceedings of ICASSP*, vol.1, pp.153-156, 9-12 May 1995.
- [9] A. Hyvärinen and E. Oja, "Independent Component Analysis: Algorithms and Applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411-430, 2000.
- [10] Intae Lee and Gil-Jin Jang, "Independent vector analysis based on overlapped cliques of variable width for frequency-domain blind signal separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 113, pp. 1–12, 2012.

Acknowledgements

Passing through past two years in Ulsan, it was continuation of the precious moments. I cannot believe I faced with last scene of master's course. Firstly, I want to express my all thanks to my family and love. Although I underwent trial and error for getting master's degree, they were always generous with support and cheers. I love you all.

Above all things, I'd like to give expression to my deepest gratitude to my advisor Prof. Gil-Jin Jang. It is great honor to be his first graduate student. He always had coached enthusiastically as well as gave lots of opportunities to widen my field of vision. Especially I cannot forget his warm heart and considerations.

My anchor Prof. Sangbae Jeong, I cannot tell him how much I appreciate what he believe me. I have learned not only expert knowledge but upright personality. I'd like to express my profound gratitude to Prof. Jae-Young Sim for giving me advice with respect to my research as a committee. I finally made it by his kind assistance.

The help of many has made today possible. I would like to thank to Machine Intelligence Labbers : Ara, Junyoung, Kibeom, Insik, Hyungju, Jiu, Doyeon, Chungho and Sungyong. I am sure they can achieve their own goal under Prof. Jang's kind guidance. I will never forget my soul mates in UNIST: Kyuyul, Taehee, Jinwoo, Younghoon, Seongsuk, Soowoong, Yongsik, Jaehwan, Hyojin, Boram and Toan. My school life was boring without these guys. I also appreciate for Gisters: Kyumin, Youngin, Unpyo, Hyunju, Sangchae, Woongbi, Hyungmin and Donghyun. Thanks so much for being with my hard time. My old and steady friendship, MS friends: Taekyun, Sinil, Sangkuk, Minki, Seunghak, Kieun, Nanhee, Yunsun, Youngmi and Bonggi. I always owe your big favors. I cannot make it without their supports, Yadang club: Minkyun, Youngjae, Soohoon, Gabhoon and Youngju. I want to be with you for all the time to come. And I am grateful for my GNU people: Soohwan, Byungil, Youngbon, Seunghoon and Minhye. I also thanks for Automatic Translate Team members of ETRI.

Lastly, I ascribe this honor to my father. He cannot share this joy with me, but I am sure he is very proud of me. I always miss you, dad.

