Optimizing the Implementation of MMSE Enhancement for Robust Speech Recognition

Pei Ding, Lei He, Xiang Yan, Rui Zhao and Jie Hao

Toshiba (China) Research and Development Center
5/F Tower W2, Oriental Plaza, No.1 East Chang An Avenue,
Dong Cheng District, Beijing 100738
E-mail: {dingpei, helei, yanxiang, zhaorui, haojie}@rdc.toshiba.com.cn

Abstract. In this paper several methods are proposed to optimize the implementation of minimum mean-square error (MMSE) estimation algorithm for robust automatic speech recognition (ASR). In the calculation of MMSE enhancement algorithm, the original confluent hyper-geometric function is approximated by a piece-wise linear function, which greatly reduces the computation load while keep the same performance. After enhancement, the spectrum is smoothed in both time and frequency domains with symmetric arithmetic series weights to compensate those spectrum components distorted by noise over-reduction. A tuning scheme is also proposed to control the enhancement by adjusting the a priori signal-to-noise ratio (SNR). Evaluation results confirm that the proposed methods efficiently optimize the MMSE enhancement algorithm and significantly improve the robustness of ASR system against background noises.

Keywords: Background Noise, Robust Speech Recognition, MMSE Estimation Algorithm, Speech Enhancement, Spectrum Smoothing

1 Introduction

In recent years speech recognition technologies have been greatly developed, and state-of-the-art ASR systems can achieve considerably high accuracy in clean environments. However, the recognition performance tends to be drastically degraded in the presence of background noise due to the mismatch between pre-trained acoustic models and the corrupted noisy utterances. Therefore, noise robustness is one of the most challenging problems for real ASR applications, especially in those practical situations where the ambient acoustic noise is usually inevitable, e.g. hand-free mobile devices, in-car voice-activated navigation and etc.

Many approaches have been proposed for noise robustness issues [1]. Some methods [2-5] aim at designing a robust front-end in which the interfering noise is removed from the speech or the acoustic feature is inherently less distorted by noise. Other methods [6-8] concentrate on online model adaptation technologies which reduce the mismatch between noisy speech features and the pre-trained acoustic models. Generally, the major disadvantage of model adaptation method is that they usually
cause huge computation load. Besides, the less dependency between the front-end and the recognizer can effectively reduce the complexity of ASR systems.

In this paper, a robust front-end is proposed, in which MMSE estimation algorithm [9] is used to suppress the noise in frequency domain. MMSE algorithm is more efficient that other classical enhancement method, e.g. spectral subtraction (SS) [2], because more priori knowledge is utilized, i.e. more accurate statistical model of both speech and noise. However, the direct use of MMSE algorithm has some problems and it should be optimized exactly towards robust speech recognition.

In MMSE enhancement, the calculation of the confluent hyper-geometric function results in huge computation load, which is a weakness of MMSE estimator especially when implemented in embedded ASR systems. Storing the pre-calculated function value in a lookup table is a common method to solve the problem, but the simple conversion from computation to large memory storage still restricts the application is such resource-limited situations. In this paper we propose to use a proper piece-wise linear function to substitute the confluent hyper-geometric function according to the derivative range and approximation error. Thus, the computation load can be significantly reduced, while the same noise reduction performance is kept.

In speech enhancement, some spectrum components at very low signal-to-noise ratios (SNR) tend to be floored by meaningless threshold in Mel-scaled filter binning stage because of the noise over-reduction. Even not floored, these spectrum components are prone to aggressively degrade the recognition performance. We propose to smooth the spectrum in both time and frequency indexes with arithmetic sequence weights. Thus, those unreliable spectrum components will be fed with speech energy from neighbors with high local SNRs, and the recognition rate can be efficiently improved.

Speech enhancement technologies are commonly used to benefit the human auditory perception, but normally not very suitable when directly used for robust ASR, e.g. the MMSE criterion is no longer the optimum towards recognition tasks. Noise reduction in enhancement approach is usually at the expense of speech distortion, which degrades the discriminative information of speech signals. Controlling the algorithm to make the best balance between noise reduction and speech distortion will optimize the implementation of speech enhancement for ASR. In this paper we propose to tune the enhancement by adjusting the a priori SNR in MMSE algorithm.

The remaining paper is arranged as follows. Section 2 reviews the MMSE speech enhancement algorithm. Section 3-5 describe the proposed methods, which optimize the implementation of MMSE enhancement. Section 6-7 describe the experiments in details. Finally, section 8 concludes the paper.

2  MMSE Speech Enhancement Algorithm

In [9] a short-time spectral amplitude (STSA) estimation algorithm based on a MMSE criterion is proposed to enhance the noise corrupted speech. One advantage is that MMSE estimation algorithm can efficiently suppress the background noise while at the expense of very few speech distortions. Another property of this method is that it can eliminate the residual “musical noise”.

We assume that the noise is additive and independent to the clean speech, and after fast Fourier transform (FFT) analysis of windowed speech frames each spectral
component is statistically independent and corresponds to a narrow-band Gaussian stochastic process. Let \( A(k, n), D(k, n) \) and \( R(k, n) \) denote the \( k \)th spectral component of the \( n \)th frame of speech, noise, and the observed noisy signals respectively, the estimation of \( A(k, n) \) is given as

\[
\hat{A}(k, n) = \frac{1}{2} \sqrt{\frac{\pi \xi_k}{\gamma_k (1 + \xi_k)}} M(-0.5; 1; -\frac{\gamma_k \xi_k}{1 + \xi_k}) R(k, n),
\]

where the a priori SNR \( \xi_k \) and the a posterior SNR \( \gamma_k \) are defined as:

\[
\xi_k = \frac{E(|A(k, n)|^2)}{E(|D(k, n)|^2)}; \quad \gamma_k = \frac{|R(k, n)|^2}{E(|D(k, n)|^2)}.
\]

In practice, we use a voice activity detection based noise estimation method and substitute the estimation of clean speech by enhanced spectrum of the previous frame. In Eq.(1) \( M(a; c; x) \) is the confluent hyper-geometric function that is calculated by Taylor series accumulation as follows:

\[
M(a; c; x) = 1 + \frac{a x}{c} + \frac{a(a+1)x^2}{c(c+1)2!} + \ldots = \sum_{n=c} (a)_n \frac{x^n}{n!},
\]

where \( (a)_n = a(a+1)\ldots(a+r-1) \) and \( (a)_0 = 1 \).

### 3 Approximation of the Confluent Hyper-geometric Function

From Eq.(3) we can find the calculation of the confluent hyper-geometric function leads to huge computation load, which is a major difficulty for MMSE enhancement in real application, especially for those resource-limited embedded system. To solve this problem, we propose to use a piece-wise linear function to approximate the Taylor series accumulation in the calculation of MMSE enhancement algorithm.

Let us suppose \( \frac{\gamma_k \xi_k}{1 + \xi_k} \neq 0 \) and define \( M(-0.5; 1; -\frac{\gamma_k \xi_k}{1 + \xi_k}) = M(-0.5; 1; -v) \equiv h(v) \).

A suitable piece-wise linear function \( pwl(v) \) including \( n \) segments is designed to approximate the function \( h(v) \) with \( 0 \leq v \leq 40 \) (when \( v > 40 \), \( h(v) \approx 2 \sqrt{\frac{\gamma_k \xi_k}{\pi \xi_k}} \)).

\[
h(v) \approx pwl(v) = \sum_{i=0}^{n} l_i(v), \quad 0 \leq v \leq 40,
\]

where \( l_i(v) \) is the \( i \)th linear function between the \((i-1)\)th and the \( i \)th segmentation points of \( h(v) \), denoted as \( (v^{i-1}, h(v^{i-1})), (v^i, h(v^i)) \) respectively:

\[
l_i(v) = \begin{cases} (v - v^{i-1}) \times \frac{h(v^i) - h(v^{i-1})}{v^i - v^{i-1}} + h(v^{i-1}) & v^{i-1} \leq v \leq v^i \\ 0 & \text{otherwise} \end{cases}
\]
Regarding the function $h(v)$ as the standard whose value is pre-calculated by Taylor series accumulation, we adopt the following steps to construct the piece-wise linear function, as taking into account the derivative range of $h(v)$ and the approximation errors:

**Step 1**: Add initial segmentation points, which satisfy that the difference of the derivative of $h(v)$ between two consecutive points is smaller than $T_d$;

**Step 2**: if the difference between $h(v)$ and $pwlf(v)$ is greater than $T_e$, then insert new points in the corresponding two consecutive segmentation points;

**Step 3**: repeat **Step 2** and update $pwlf(v)$.

The above procedures are illustrated in Figure 1 and in practice totally there are only 12 segments of linear function to approximate the confluent hyper-geometric function. It is obvious that the computation load is greatly reduced by using the proposed method.

### 4 Spectrum Smoothing Technology

The MMSE enhancement algorithm can be interpreted as it suppresses or emphasizes the spectrum components according to their local SNRs. The speech signals in those
components at very low SNRs are prone to be seriously distorted owing to the noise over-reduction.

Our proposed front-end is based on the framework of cepstral feature extraction, in which a threshold is usually essential to eliminate the sensitivity of logarithmic transform to very small outputs of the Mel-scaled filters. Thus, after speech enhancement, those low SNR spectrum components tend to be floored by a meaningless threshold in Mel-scaled filter binning stage, which causes the mismatch between the features and the acoustic models. Even over the thresholds, the low SNR components are also prone to aggressively degrade the recognition performance. In order to compensate the spectrum components distorted by noise over-reduction, we propose to smooth the spectrum in both time and frequency indexes with symmetrical normalized arithmetical sequence weights. The unreliable spectrum component will be filled with speech energy from neighboring bins whose local SNRs are high and avoid being floored in binning stage, consequently. Thus, the implementation of MMSE enhancement is tamed towards ASR tasks and the recognition performance is efficiently further improved.

At frame $n$ and frequency band $k$, the smoothed spectrum component $\hat{A}(k,n)$ is obtained as follows:

\[
\hat{A}(k,n) = \sum_{i=-L_f}^{L_f} \sum_{j=-L_t}^{L_t} w_f(i) \times w_t(j) \times \hat{A}(k+i, n+j),
\]

where $w_f(i)$ is the arithmetic sequence weight in the frequency index with smoothing length $F = 2 \times L_f + 1$:

\[
w_f(i) = w_f(-i) = \left(1 - w_f(0)\right) \frac{(L_f + 1 - i)}{L_f(L_f + 1)}, 1 \leq i \leq L_f,
\]

$W_f = [w_f(-L_f), \ldots, w_f(0), \ldots, w_f(L_f)]$ and $w_f(0)$ is the weight of current frequency bin. $w_t(j)$ and $W_t$ are the smoothing weights in time index and have the similar
definitions. The matrix $A_{re}$ corresponds to the spectrum block that is used for smoothing. As illustrated in Figure 2, in Eq.(6) the expression in matrix multiplication style indicates that we can firstly smooth the spectrum in frequency index and then in time index, or equivalently reverse the order.

5 Speech Enhancement Control

For speech enhancement algorithm, the noise reduction is usually at the expense of speech distortion, which impairs the nature discriminative information embedded in speech signals and tends to drop the recognition accuracy. Figure 3 visualizes the process of speech enhancement. It is indicated in the figure that MMSE algorithm is obviously better than SS in minimizing both residual noise and speech distortion. The circle 1 in Figure 3 denotes the enhanced speech when directly use MMSE algorithm. In this case the background noise is extremely suppressed but the speech distortion is also very serious, which less affects the intelligibility in human auditory perception but does considerably degrade the recognition performance. In other words the criterion in MMSE algorithm is not still optimum for robust ASR. Therefore, speech enhancement should be control to make the best balance between noise reduction and speech distortion, which leads to the “Min Distance” in Figure 3 and the recognition performance is improved, consequently. In this paper, we propose to tune the enhancement by adjusting the a priori SNR, $\xi$, in MMSE algorithm. Decreasing $\xi$ results in more aggressive noise reduction but more serious speech distortion, and increasing the factor causes the opposite effect.

Fig. 3. The relationship between noise reduction and speech distortion in speech enhancement
6  Experiment Setup

In the experiments, the speech data are sampled at 11025Hz and quantized to 16 bits. The frame length and window shift are 23.2ms and 11.6ms, respectively. In spectra processing, after MMSE speech enhancement and spectrum smoothing, 24 triangle Mel-scaled filters are applied to combine the frequency components in each bank, and the outputs are compressed by logarithmic function. Then the Discrete cosine transform (DCT) decorrelation is performed on the log-spectrum. The final acoustic feature of each frame is a 33 dimensional vector consisting of 11 Mel frequency cepstral coefficients (MFCC) and their first and second order derivatives.

Shanghai accented Mandarin database [10] is used to establish the isolated phrase recognition experiments to evaluate our proposed methods. We use 20000 utterances for training and 200 for evaluations. We adopt the model structure with moderate complexity, in which each Mandarin syllable is modeled by a right-context-dependent INITIAL (bi-phone) plus a toneless FINAL (mono-phone). Totally, there are 101 bi-phone, 38 mono-phone and one silence hidden Markov models (HMM). Each model consists of 3 emitting left-to-right states with 16 Gaussian mixtures.

To improve the robustness of ASR system we use an immunity learning scheme [11] in which the acoustic models are trained in simulated noisy environments by artificially adding car noises to clean training utterances at different SNRs. There are 12 kinds of car noises in the experiments, which are the combinations of the following three conditions:

1. Speed (km/h): 40, 60 and 100
2. Road type: “asf” (asphalt), “tun”(tunnel) and “con” (concrete)

We also use these car noises to generate the artificial noisy speech for evaluation.

7  Evaluation Results and Analysis

Twelve car noises described in section 6 are used to generate the artificial evaluation noisy speech with the SNRs from –5dB to 20dB with a 5dB step. In the experiments, the baseline denotes using the MFCC feature without special robust technologies except immunity learning scheme.

Figure 4(a) shows the word error rate (WER) averaged by 12 car noises at each SNR. We can observe that the baseline performs well in high SNR conditions, e.g. at 20dB the recognition accuracy is 99.21%. However, the baseline performance drops rapidly below 10dB, because interfering noises cause serious mismatch between the features and the models. If the MMSE speech enhancement is adopted, the noise will be efficiently suppressed and the recognition performance is efficiently improved, consequently. In experiments, directly using MMSE algorithm leads to an error reduction rate (ERR) of 61.2% versus the baseline. In Figure 4(a), Ori_MMSE denotes using the original value of the confluent hyper-geometric function, which is calculated by Taylor series accumulation. Pwlf_MMSE denotes the proposed approximation methods by using a suitable piece-wise linear function, and it is very
obvious in the figure that Pwlf_MMSE scheme keep the identical performance while with the extremely reduced computation load. If we use the proposed spectrum smoothing technology, the Pwlf_MMSE_Smooth scheme can efficiently further improve the performance of MMSE algorithm and the ERR versus baseline is 67.5%.
Figure 4(b) gives the WER averaged by the six SNRs, from which the performance difference under each car noise is analyzed. We find that the recognition performance in air-conditioner on and high speech driving conditions is obviously lower than that in the opposite conditions. The reason is that ASR performance tends to be degraded more seriously by broadband noises. In such adverse environments mentioned above the dominant noise source is the air friction from the air-conditioner and the wind outside the car, which produces the broadband white-like background noises and consequently causes dramatic performance drop for recognition. The experimental results also show that the proposed front-end can significantly improve the performance in all conditions.

![Figure 5](image_url)

**Fig. 5.** Improving the recognition performance by adjusting the *a priori* SNR

Figure 5 illustrates the perform improvement via enhancement control. In our proposed method, the *a priori* SNR is adjusted by assigning to a multiplicative amplifying factor. This factor with value 1.0 corresponds to the direct implementation of the MMSE enhancement algorithm (denoted as Direct_MMSE in Figure 5). In this figure we can observe that tuning the *a priori* SNR obviously affect the recognition performance. In the experiment, the optimal amplifying factor is about 1.15 and provides the best recognition performance, which indicates that the noise is somewhat over-suppressed in the original MMSE enhancement algorithm.
8 Conclusions

This paper proposes several methods to optimize the implementation of MMSE enhancement algorithm towards robust ASR. Substituting the confluent hypergeometric function by a piece-wise linear function significantly reduces the computation load while maintaining the same enhancement performance. Then, the noise over-reduced spectrum components are compensated by smoothing in both time and frequency indexes with arithmetic series weights. The speech enhancement is also controlled by tuning the a priori SNR in MMSE algorithm. The Experimental results confirm the efficiency of the proposed methods.

References

4. ETSl Standard, “Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms”, ETSI ES 202 050 v.1.1.1, October 2002.