Sub-Band Based Additive Noise Removal for Robust Speech Recognition

J. Chen †, K. K. Paliwal †* and S. Nakamura**

* ATR Spoken Language Translation Research Laboratories
Kyoto, 619-0288, Japan
† School of Microelectronic Engineering, Griffith University
Brisbane, QLD 4111, Australia

E-mail: jingdong.chen@slt.atr.co.jp

Abstract

To make an automatic speech recognition system robust with respect to noise, we will probably have to solve two problems. One is the detection and identification of noise. Another is the consideration of noise effect during recognition process. In this paper, we will investigate several noise estimation approaches, such as moving average, long-term average, long-term Fourier analysis, *etc*. We will then introduce a sub-band based scheme to remove the noise effect from corrupted speech to make recognition system immune to additive noise. We will report on experiments on TI digits database and NOISEX database to justify the proposed approach.

1. Introduction

There are many sources of acoustical distortion that can degrade the performance of speech recognition systems. For many speech recognition applications, the most important source of acoustical distortion is the additive noise. How to maintain good accuracy in the presence of noise has become one of the most challenging areas of the speech recognition research currently.

There have been considerable interests in dealing with noise. The efforts include representing speech features which are robust to noise such as sub-band based features [1], and perception inspired features [2], filtering noise or estimating the parameters of the clean speech from corrupted speech signal such as Wiener filtering [3], Kalman filtering [4], spectral subtraction [5], RASTA [6], Cepstral mean removal [7], signal bias removal [8], and microphone array based front-end processing [9], and compensation of HMM model parameters to include the effects of noise or adaptation of HMM model parameters to take into account the environmental changes, such as Parallel model compensation (PMC) in log spectral or cepstral domain [10], vector Taylor series approximation based model compensation in log spectral domain [11], Jacobian approach [12] , MLLR [13], transfer vector interpolation [14], *etc*.

Although these techniques were experimented in speech recognition with certain success, there remains a great need to investigate new techniques that can accurately recognize speech in degradation environments.

To make an automatic speech recognition system robust with respect to noise, we will probably have to solve two problems. The first one is the detection and identification of noise. Many noise robust recognition approaches require noise or noise parameters. For example, spectral subtraction needs to know the power spectrum of the noise. PMC requires an accurate noise model. Most speech enhancement methods need to know the SNR or noise parameters. Currently, majority of methods estimate noise during the period of absence of speech. They operate under the assumption that noise is stationary as opposed to the time-variant speech signal and that noise has same statistics during speech or absence of speech. This noise estimation approach often needs a front-end point detector which can distinguish noise segments from speech segments.

The second problem with which we will have to face is the consideration of the effect of noise during recognition. This can be achieved through two ways which are: i) Removing noise from corrupted speech to recover clean speech parameters. ii) Compensating clean speech parameters to match the noise conditions. Spectral subtraction belongs to the first kind. It assumes that speech and noise are additive in spectral domain. Hence directly subtracting the spectrum of noise from that of the corrupted speech will recover the spectrum of clean speech signal. PMC, on the other hand, transforms the HMM model parameters trained in a noise-free speech environment to a noisy speech environment using an estimated noise model.

In this paper, we will introduce an approach called sub-band based spectral subtraction to make recognition system robust to noise. We will address two issues. One is the estimation of noise effect. To accomplish this task, we will investigate several approaches such as the moving average method, long-term average method, long-term Fourier analysis method, *etc*. Another is the removal of noise. To circumvent this problem, we will introduce a sub-band based subtraction strategy. We will report on experiments to justify our approach.

2. Estimation of Noise Effect

If $s(t)$ is the original clean speech signal, the received speech signal $y(t)$ is modeled as

$$y(t) = s(t) * h(t) + n(t) = x(t) + n(t) \quad (1)$$

where $h(t)$ is the impulse response of channel distortion and $n(t)$ the ambient noise. $*$ denotes the convolution operation, and $x(t)$ the noise-free speech.

Speech signal is time-variant and non-stationary. It is usually analyzed on the frame-by-frame basis. For one frame of speech signal, the Eq. (1) is written as

$$y(k,t) = y(t)w(t-(k-1)\tau) = x(k,t) + n(k,t) \quad (2)$$

where k denotes the frame index. $w(t)$ is a window function which only has non-zero values when t is in $[0,T]$. T is length of the window. τ is the window shift. Assume that the noise in (2) is uncorrelated with speech signal, the power spectrum of the above received speech signal is

$$Y(k,f) = X(k,f) + N(k,f) \quad (3)$$

There have been considerable efforts to estimate and to filter the noise term in the right hand side of Eq. (3). In this paper, we investigate and compare several noise estimation approaches shown below.

2.1 Estimation of noise from non-speech segments

Intuitively, noise effect can be estimated during the absence of speech. This is the way to achieve noise estimates in the ***spectral subtraction*** [15]. We call this noise estimation approach ENS (Estimation from Non-Speech segments) for short. Obviously, this approach assumes that the noise estimate achieved in the absence of speech can represent the noise in the presence of speech. In addition, a good speech signal detector which can distinguish speech segments from non-speech segments is necessary for the method.

2.2 Moving average method

To avoid a speech signal detector, a ***continuous spectral subtraction*** (CSS) was proposed [15]. This method was shown to have advantages over ***spectral subtraction***. In this approach, the average of *M* consecutive frames of short-term power spectra is used as a noise estimate, i. e.,

$$\hat{N}(k, f) = \frac{1}{M} \sum_{i=k-M+1}^{k} Y(i, f) \tag{4}$$

where $\hat{N}(k, f)$ is the noise power spectrum at the *k*th frame. We call this estimation approach moving average and abbreviate to MA.

2.3 Sequential estimation

Inspired from CSS and sequential signal bias removal [8], we introduce a sequential way to achieve noise estimates, i.e.,

$$\hat{N}(k, f) = (1-\gamma)\hat{N}(k-1, f) + \gamma Y(k, f) \tag{5}$$

where $\hat{N}(k-1, f)$ is the noise estimate at the (*k*-1)th frame and γ an updating factor. This method is called SE for short.

2.4 Long-term average

For some recognition systems, all frames of speech for a test utterance are available simultaneously. In this case, after the MA approach, we can use the average of the short-term power spectra over all frames as the noise estimate for a certain test utterance. Namely,

$$\hat{N}(k, f) = \hat{N}(f) = \frac{1}{N} \sum_{k=1}^{N} Y(k, f) \tag{6}$$

Note that in this approach, all frames in one utterance share a single noise estimate. This method is shortened for LTA.

2.5 Long-term Fourier analysis method

It was found that phonetic information of speech is encoded in the changes of the speech spectrum over time. Relatively less phonetic information is encapsulated in the long-term speech spectrum. Noise, however, can be treated as a stationary process. Long-term spectrum will provide a good estimate of noise. Based on this fact, we propose to estimate noise using long-term Fourier analysis, i.e.,

$$\hat{N}(k, f) = \hat{N}(f) = \frac{1}{\zeta} \left| F[y(l)w(l)] \right|^2 \tag{7}$$

where $F[\cdot]$ denotes Fourier Transform, $y(l)$ is the discrete version of speech signal shown in Eq. (1), $w(l)$ is a window function, ζ a normalization factor which is defined as $\zeta = \mathrm{L}\sum_{l} w(l)$, and L is the length of the Fourier Transform. Noting that the noise effect estimated from Eq. (7) has a much longer length than that of short-term power spectrum, we therefore need to warp it to have a same length as the power spectrum of each frame. We should also point out that the short-term power spectra should be normalized in a similar way as in Eq. (7) before one subtract this noise estimate from them.

3. Noise Effect Removal

Once we get the estimates of noise, the next step would be how to remove noise effect from corrupted speech signal to recover clean speech spectra or parameters. This can be done by spectral subtraction which is defined as follows [15]

$$Y_{ss}(k, f) = \begin{cases} Y(k, f) - \alpha\hat{N}(k, f), \text{if } Y(k, f) > \frac{\alpha}{1-\beta}\hat{N}(k, f) \\ \beta Y(k, f), \quad otherwise \end{cases} \tag{7}$$

where $Y_{SS}(k, f)$ is the power spectrum of the enhanced speech after spectral subtraction, α is an over-estimation factor, and β is to define the spectral flooring.

In this paper, rather than directly using Eq. (7), we introduce a sub-band based subtraction scheme. Suppose we divide the speech signal into *B* sub-bands, and two cutoff frequencies of the *i*th sub-band are denoted as f_L^i and f_U^i, the sub-band based spectral subtraction is defined as

$$\begin{aligned} Y_D^i(k, f) &= Y^i(k, f) - \alpha^i \hat{N}^i(k, f) \\ P_Y^i &= \int_{f_L^i}^{f_U^i} Y^i(k, f) df \\ P_N^i &= \int_{f_L^i}^{f_U^i} \hat{N}^i(k, f) df \\ Y_{sbs}^i(k, f) &= \begin{cases} Y_D^i(k, f), \text{if } P_Y^i > \frac{\alpha^i}{1-\beta^i} P_N^i \\ \beta^i Y^i(k, f), \quad otherwise \end{cases} \end{aligned} \tag{8}$$

where $Y^i(k, f)$, $\hat{N}^i(k, f)$ and $Y_{sbs}^i(k, f)$ are power spectrum of corrupted speech, noise estimate and power spectrum of the enhanced speech for the *i*th sub-band, and α^i and β^i are sub-band dependent over-estimation factor and spectral flooring respectively. As it will become clear soon, this sub-band based subtraction scheme has advantages over the spectrum subtraction for noisy speech recognition.

In this paper, we divide full-band into 24 sub-bands on the mel-scale. The cutoff frequencies of each sub-band are set to be same as what adopted in the mel-scale triangle filter banks. In such circumstance, Eq. (8) can be alternatively expressed as

$$E_{ss}^i(k) = \begin{cases} E_Y^i(k) - \alpha E_{\hat{N}}^i(k), \text{if } E_Y^i(k) > \frac{\alpha^i}{1-\beta^i} E_{\hat{N}}^i(k) \\ \beta^i E_Y^i(k), \quad otherwise \end{cases} \tag{9}$$

where $E_Y^i(k)$ is the output of the *i*th triangle filter when $Y(k, f)$ is passed through the triangle filter bank, and $E_{\hat{N}}^i(k)$ the output of the same filter with its input being $\hat{N}(k, f)$.

In this paper, a same α is used across all sub-bands. So is the β. The selection of sub-band dependent α and β is currently under investigation.

4. Experiments

We performed quite a few experiments for connected digit speech recognition in various noise conditions to evaluate the introduced approaches. Results from some of the experiments are reported in this paper.

4.1 Speech and noise database

The speech database used is the TI connected digits database [16]. This database contains digit strings uttered by adult speakers and children as well. However, only digit strings from 225 adult talkers are used in our experiments. These strings are originally divided into training set and test set for consistency of comparison of results among different researchers.

The vocabulary in this database consists of 11 words which include 10 digits and an "oh". Each talker uttered 77 sequences of these words, consisting of 2 tokens of each of the 11 words in isolation, and 11 strings of each of 2, 3, 4, 5, and 7 digits. The digit strings were recorded in an acoustically treated sound room with a sampling frequency being 20 kHz. For the comparison with the recognition results reported, we downsampled speech to 8 kHz using Matlab downsampling function.

To test the robustness of different approaches with respect to noise, we directly add some noise to the speech signal in the test set. The training speech is kept clean. The noise signals used are from NOISEX database [17]. The noise signal provided in this database is sampled at 16kHz. To match its bandwidth to the speech signal, we downsampled the noise signal to 8 kHz.

4.2 Recognition system

In our experiments, the HTK speech recognition system is used to perform the recognition task. This was configured as a gender-independent mixture Gaussian HMM system. The model set consists of 11 word-models, a silence model and a short pause model. Except the short pause, each model has 6 emitting states. The short pause model has only one emitting state. A mixture of 8 multivariate Gaussian distributions with diagonal covariance matrices is used for each emitting state to approximate its probability density function.

4.3 Spectral subtraction vs. sub-band based subtraction

The first experiment is performed to compare the subtraction strategy defined in the spectral subtraction (Eq.7) with the sub-band based subtraction shown in Eq. 8. The noise effect, in this experiment, is estimated from non-speech segments (ENS method). Three feature sets are investigated. They are traditional MFCCs, MFCCs computed after spectral subtraction (denoted as SSMFCC), and the MFCCs computed after sub-band based spectral subtraction (denoted as SBSMFCC). Each feature vector consists of 39 coefficients which include 12 MFCCs, normalized frame energy and their first and second order differentials. In both spectral subtraction and sub-band based subtraction, β and β^i are fixed to 0.1, and α and α^i are set to 0.6. The recognition results are shown in Table 1.

Table 1 a: Word accuracy in speech noise condition

SNR	0dB	5dB	10dB	15dB	20dB	30dB	∞
MFCC	18.8	41.3	72.6	91.6	97.2	98.7	99.0
SSMFCC	21.9	47.1	75.9	92.8	97.4	98.8	98.9
SBSMFCC	26.1	52.6	78.4	92.7	97.3	98.8	99.0

Table 1 b: Word accuracy in machine-gun noise condition

SNR	0dB	5dB	10dB	15dB	20dB	30dB	∞
MFCC	80.2	87.6	93.6	97.3	98.4	98.9	99.0
SSMFCC	81.0	87.8	93.3	96.6	98.2	98.9	98.9
SBSMFCC	83.0	90.1	95.0	97.7	98.6	99.0	99.0

Table 1 c: Word accuracy in Lynx noise condition

SNR	0dB	5dB	10dB	15dB	20dB	30dB	∞
MFCC	28.6	56.8	82.1	93.9	97.5	98.7	99.0
SSMFCC	39.4	66.6	87.0	95.4	97.9	98.8	98.9
SBSMFCC	42.3	68.8	87.8	95.7	98.1	98.9	99.0

An inspection of *Table 1* reveals that: 1) Both spectral subtraction and sub-band based noise removal approach are helpful to improve the robustness of the recognition system. 2) Sub-band based subtraction scheme consistently performs better than the spectral subtraction approach in both noisy and clean speech environments.

Since the sub-band based subtraction scheme has shown its superiority, we fix the subtraction process to sub-band based subtraction in the subsequent experiments.

4.4 Different noise estimation approaches

The second experiment is carried to test the effects of different noise estimation approaches on the recognition performance. The features used are MFCCs computed after sub-band based spectral subtraction. Each feature vector contains 39 coefficients as described before. Since we have 5 different ways to estimate noise, we have 5 different types of MFCCs after noise removal. They are denoted as ENS, MA, SE, LTA, and LTF without leading any confusion.

For noise estimation, the M parameter in the MA approach is chosen to be 30 frames and the γ factor in the SE method is set to 0.04. We should point out that, for different noise estimation approaches, the parameters used in the sub-band based subtraction should be optimized separately. This is done by performing a group of recognition experiments. We fix the β parameter to 0.1 and change α from 0.2 to 0.8 with an increment of 0.1. We observed that when $\alpha = 0.6$, the ENS method yields its best performance. The MA, SE and LTA approaches generate the highest accuracy at $\alpha = 0.5$. While LTF shows its best performance when $\alpha = 0.4$. Hence in the following comparison experiments, we fix β to 0.1, and set $\alpha = 0.6$ for the ENS, $\alpha = 0.5$ for the MA, SE and LTA methods, and $\alpha = 0.4$ for LTF approach.

The recognition accuracies for the whole test set of TI database in different noise conditions are shown in Table 2.

Table 2 a: Word accuracy in speech noise condition

SNR	0dB	5dB	10dB	15dB	20dB	30dB	∞
MFCC	18.8	41.3	72.6	91.6	97.2	98.7	99.0
ENS	26.1	52.6	78.4	92.7	97.3	98.8	99.0
MA	21.36	46.82	76.9	93.6	97.7	98.7	99.0
SE	20.8	42.9	71.3	90.6	97.2	98.5	99.0
LTA	34.8	59.8	84.3	95.2	97.9	98.8	99.1
LTF	30.3	55.3	82.3	95.1	97.9	98.9	99.2

Table 2 b: Word accuracy in machine-gun noise condition

SNR	0dB	5dB	10dB	15dB	20dB	30dB	∞
MFCC	80.2	87.6	93.6	97.3	98.4	98.9	99.0
ENS	83.0	90.1	95.0	97.7	98.6	99.0	99.0
MA	85.2	91.9	96.3	98.2	98.9	99.1	99.0
SE	82.3	89.2	94.8	97.7	98.7	99.0	99.0
LTA	86.2	92.7	96.8	98.5	98.9	99.1	99.1
LTF	84.5	92.0	96.6	98.5	98.9	99.1	99.2

Table 2 c: Word accuracy in Lynx noise condition

SNR	0dB	5dB	10dB	15dB	20dB	30dB	∞
MFCC	28.6	56.8	82.1	93.9	97.5	98.7	99.0
ENS	42.3	68.8	87.8	95.7	98.1	98.9	99.0
MA	30.2	59.3	84.5	95.5	98.2	98.7	99.0
SE	27.4	52.4	79.3	94.2	97.9	98.5	99.0
LTA	44.1	68.6	87.7	95.8	97.9	98.8	99.1
LTF	40.1	65.8	86.8	95.8	98.0	98.8	99.2

From the above experiment, we can make following observations. 1) Subtraction of noise estimates achieved from each approach is able to improve the robustness of the system. 2) LTA yields the highest robustness with respect to noise among the approaches investigated. 3) While the robustness of the LTF approach is slightly inferior to that of the LTA, it yields a little better performance in high SNR conditions.

4.5 Combination of noise removal and CMR

Cepstral mean removal (CMR) is shown to be an effective yet simple way to deal with convolution distortions. In contrast, the noise removal approach aims at coping with additive noise. This experiment is performed to combine these two methods to further improve recognition performance. The feature sets investigated include the MFCC and the MFCC computed after sub-band based noise removal followed by a CMR processing. The configuration of features and parameters used in the subtraction are same as that in section 4.4. The recognition results are shown in Figure 3, where * denotes the CMR processing.

Table 3 a: Word accuracy in speech noise condition

SNR	0dB	5dB	10dB	15dB	20dB	30dB	∞
MFCC	18.8	41.3	72.6	91.6	97.2	98.7	99.0
MFCC*	20.7	46.7	77.2	92.7	96.8	98.1	98.8
ENS*	31.7	57.5	81.0	93.4	97.4	98.8	99.1
LTA*	42.3	71.8	89.7	96.0	97.6	98.6	99.1
LTF*	38.1	68.9	89.3	96.0	97.8	98.7	99.1

Table 3 b: Word accuracy in machine-gun noise condition

SNR	0dB	5dB	10dB	15dB	20dB	30dB	∞
MFCC	80.2	87.6	93.6	97.3	98.4	98.9	99.0
MFCC*	81.9	88.8	94.2	97.4	98.5	98.8	98.8
ENS*	87.0	93.1	96.9	98.4	98.8	99.0	99.1
LTA*	90.6	95.5	97.8	98.7	99.0	99.1	99.1
LTF*	89.9	95.1	97.7	98.7	99.0	99.2	99.1

Table 3 c: Word accuracy in Lynx noise condition

SNR	0dB	5dB	10dB	15dB	20dB	30dB	∞
MFCC	28.6	56.8	82.1	93.9	97.5	98.7	99.0
MFCC*	26.9	57.0	83.9	94.6	97.1	98.2	98.8
ENS*	42.4	67.5	86.8	95.3	97.9	98.8	99.1
LTA*	53.1	79.5	92.5	96.6	97.8	98.6	99.1
LTF*	50.0	78.1	92.3	96.7	98.0	98.8	99.1

From this experiment, we observed that: 1) In most cases, the CMR can improve the speech recognition performance. 2) The combination of CMR and noise removal approach and yield further improvement. 3) Not surprisingly, the LTA, combined with a CMR process, yields the best performance.

5. Summary

Several approaches were investigated to estimate the noise effects. These included estimation of noise from non-speech segments, moving average approach, sequential estimation method, long-term average method and an approach based on long-term Fourier analysis.

A sub-band based noise removal approach was proposed and verified. Experiments showed that removing noise effect estimated using the LTA and LTF methods from corrupted speech could effectively improve recognition performance in noisy conditions, while maintaining the same or even yielding slightly better recognition accuracies in clean speech environment.

The combination of noise removal approaches and cepstral mean removal led to further improvement of recognition performance.

Acknowledgement

The research work reported in this paper is partially funded by Australian Research Council. The authors would like to thank Dr. Seiichi Yamamoto, the president of ATR Spoken Language Translation Research Laboratories, for his continuous encouragement.

References

[1] H. Bourlard, *et al*, "A mew ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands," ICSLP'96, Philadelphia, October 1996.

[2] J. R. Cohen, "Application of an Auditory Model to Speech Recognition," J. Acoustic. Soc. Amer., Vol. 85, June 1989, PP. 2623-2329.

[3] S. V. Vaseghi, *et al*, "Noise Compensation Methods for Hidden Markov Model Speech Recognition in Adverse Environments," IEEE Trans. Speech and Audio Processing, Vol. 5, No. 1, Jan. 1997. PP. 11-21.

[4] D. C. Popescu, *et al*, "Kalman Filtering of Colored Noise for Speech Enhancement," ICASSP'98, PP. 997-1000.

[5] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. 27, No. 2, April 1979. PP. 113-120.

[6] H. G. Hirsch, *et al*, "Improved speech recognition using high-pass filtering of subband envelopes," Proc. EUROSPEECH, PP. 413-416, 1991.

[7] D. Geller, *et al* ,"Improvements in speech recognition for voice dialing in the car environment," Proc. ESCA Workshop on Speech Processing in Adverse Conditions, PP. 203-206, Nov. 1992.

[8] M. Rahim, and B. -H. Juang, "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition", IEEE Trans. on Speech and Audio Processing, Vol. 4, No. 1, PP. 19-30, January 1996.

[9] T. M. Sullivan and R. M. Stern, "Multi-microphone Correlation-based Processing for Robust Speech Recognition," ICASSP'93, Vol. II, PP. 91-94.

[10] M. J. F. Gales, *et al*,, "Robust speech recognition using parallel model combination," IEEE Trans. Speech Audio Processing, Vol. 4, PP. 352-359, Sep. 1996.

[11] P. J. Moreno, *et al*, "A vector Taylor series approach for environment independent speech recognition," ICASSP'96, May 1996, PP.733-736.

[12] S. Sagayama et al, "Jacobian Approach to Fast Acoustic Model Adaptation," ICASSP'97, PP. 835-838.

[13] P. C. Woodland, *et al*, "Improving environmental robustness in large vocabulary speech recognition," ICASSP'96, May 1996, PP. 65-68.

[14] K. Ohkura, *et al*, "Speaker Adaptation Based on Transfer Vector Filed Smoothing Technique," ICSLP'1992, PP. 369-372.

[15] J. A. Nolazco Flores, *et al*, "Continuous Speech Recognition in Noise Using Spectral Subtraction and HMM Adaptation", ICASSP'94, Vol. I, PP. 409-412.

[16] http://www.ldc.upenn.edu/readme.files/tidigits.readme.html

[17] A. Varga, *et al*, "The Noise-92 Study on the Effect of Additive Noise on Automatic Speech Recognition," DRA Speech Research Unit, St. Andrew's Rd., Malvern, Worcestershire, WR14 3PS UK.