音声言語情報処理 34－3
(2000. 12. 21)

LONG-TERM EFFECT REMOVAL FOR NOISY SPEECH RECOGNITION

J. Chen †, K. K. Paliwal †*, T. Matsui*, K. Yao*, K. P. Markov* 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: Noise speech recognition is of great interests in speech research recently. To make an automatic speech recognition system robust 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 address a new method to estimate the noise effect using a long-term Fourier analysis. We will then discuss how to remove the noise effect from corrupted speech to make recognition system immune to uncertainties.

The rationale behind our noise estimation and removal approach can be described as follows. Speech signal is a non-stationary stochastic process. Much phonetic information in 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. Hence the subtraction of long-term effect from short-term spectra will keep the discrimination information which is necessary for speech recognition, and meanwhile remove the noise effect.

We will report on experiments on DARPA speech in noise environments evaluation (SPINE) database to demonstrate the properties of the proposed approach.

Key words speech recognition, noise subtraction, long-term power spectrum, noise estimation

長時間パワースペクトル減算による雑音下音声認識

J. チェン†*、K. K. パリワル†*、松井知子*、K. ヤオ*、K. P. マルコフ*、中村哲 *

* ATR 音声言語通信研究所
〒619－0288 京都府相楽郡精華町光台 2－2－2
†グリフィス大学
ブリスベン、QLD 4111、オーストラリア
E-mail: jingdong.chen@slt.atr.co.jp

あらまし 近年、雑音下の音声認識に対する関心は高まっている。一般に音声認識システムの雑音に対する頑健性の向上には、1. 雑音区間の検出と同定、2. 認識処理中の雑音の取り扱いについて検討する必要がある。本稿では、長時間フーリエ分析を用いて雑音の影響を推定する方法を提案する。次いで、歪んだ音声から、その雑音の影響を除去する方法について述べる。この方法は、基本的には従来のスペクトル・サブトラクションと同様の処理である。雑音が混入した観測信号の短時間パワースペクトルから、その長時間パワースペクトルを差し引いたものを特徴量とすることで、雑音の影響を受けない認識を行う。従来法の連続スペクトル・サブトラクションでは、短時間パワースペクトルの長時間平均が用いられるが、本方法ではその代わりに長時間パワースペクトルを用いることで、音声認識で重要な音韻の識別的な情報を保ち、かつ雑音の影響を除去したスペクトルのより良い推定を得る。本稿では DARPA の雑音環境における評価用のデータベース（SPINE）を用いて認識実験を行い、本方法の効果を示す。

キーワード 音声認識、雑音除去、長時間パワースペクトル、雑音推定

1. INTRODUCTION

Significant advances have been made in recent years in the area of automatic speech recognition. It is now possible to use speech recognition successfully in a controlled environment. However, the performance of a speech recognizer suffers dramatic degradation when there is a mismatch between training and testing conditions [1-3]. There are many factors that contribute to the mismatch. The main factor, that causes the mismatch, is the presence of noise. Maintaining good accuracy in noisy conditions 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 may be roughly classified into three categories.

1. Filtering noise from noisy speech signal in the front-end processing stage. This approach aims at removing noise components or estimating the parameters of the clean speech from corrupted speech signal. The representative methods in this category include spectral subtraction [4], RASTA [5], Cepstral bias removal [6], iterative signal bias removal [7], etc. Recently, microphone array in which several to tens of microphone elements are arranged in a specific configuration has been intensively investigated for speech analysis and speech recognition [8]. The fact that a microphone array can get different realizations of noise signal and noise corrupted speech signal provides considerable potential in filtering noise and improving the effective SNR of speech as it is input to the recognition system. The widely application of microphone array, however will depend upon many factors such as the size of an array, the cost of an array, to what degree can we deal with the reverberation when an array is used somewhere other than an anechoic chamber, etc.
2. Compensation of HMM model parameters to include the effects of noise or adaptation of HMM model to take into account the environmental changes. Parallel model compensation (PMC) in log spectral or cepstral domain [9], vector Taylor series approximation based model compensation in log spectral domain [10], MLLR [11] based model compensation are representatives of this type. Amongst them, the PMC like methods are intensively studied for robust speech recognition. Though reported results showed their significant advantages for noisy speech recognition, the effectiveness of these methods depends on an accurate estimation of a noise model. Besides, the heavy computational load casts another barrier for the application of this type of methods in real the world.
3. Representing speech features which are more robust to noise. This has been being of great interests for decades. Achievements in this area [12-13] have led tremendous improvements of speech recognition. Recently, a sub-band based feature representation drew much attention because some evidences showed that human beings process speech on a sub-band basis. In this technique[14-16], the full-band speech is divided into several sub-bands and each sub-band is represented individually. Various experiments have demonstrated the substantial advantages of this technique for robust speech recognition in colored noise conditions. The performance of this method, however will hinge upon our knowing which sub-band is more or less corrupted by noise.

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 compared to 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: 1. Removing noise from corrupted speech to recover clean speech parameters. 2. 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 which trained in a noise-free speech environment to noisy speech environment using an estimated noise model.

In this paper, we will report a new method to make recognition system robust to noise. We will address how to estimate noise spectrum using a long-term Fourier analysis. We will then discuss how to take into account the noise effect during speech recognition. The method used is long-term subtraction which remove noise spectrum from the spectrum of corrupted. We shall report experiments to justify our approach.

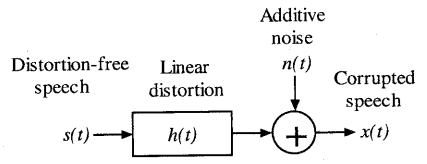


Fig. 1. A model for distortions

2. NOISE EFFECT ESTIMATION

Many signal processing algorithms use an implicit model shown in Fig. 1 for environment degradation. Assuming that the distortion-free speech signal $s(t)$ is first passed through a linear filter whose output is then corrupted by uncorrelated additive noise $n(t)$, we write the corrupted speech signal $y(t)$ as

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

where $x(t)$ denotes noise-free speech.

Many robust speech recognition algorithms involve the estimation of the linear distortion and the noise effect from the corrupted speech. Since channel distortions and the noise are additive to speech signal in different domains, they cannot be simultaneously estimated and removed in one domain. The linear distortion, which is additive in log-spectral or cepstral domain, is often estimated by an expectation (or averaging) operation in these two domains. This is well addressed in many papers [5-7][17-18]. Noise effect, however, is found to be additive in spectral domain, and is often estimated and suppressed in linear or power spectral domain.

Due to the fact that speech signal is a non-stationary stochastic process and that much of the phonetic information in speech is encoded in the changes of the speech spectrum over time, speech signal is often analyzed on a frame basis in which signal is divided into segments of tens of milliseconds and each segment is represented separately. Rewrite Eq. (1) as

$$\begin{aligned} y(t) &= y(t)w(t) + y(t)w(t-\tau) + \cdots + y(t)w(t-(L-1)\tau) \\ &= y_1(t) + y_2(t) + \cdots + y_k(t) + \cdots + y_L(t) \end{aligned} \quad (2)$$

where $w(t)$ 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. L is the number of frames. $y_k(t)$ is the k^{th} frame of speech signal which is

$$\begin{aligned} y_k(t) &= y(t)w(t-(k-1)\tau \\ &= [x(t)+n(t)]w(t-(k-1)\tau) = x_k(t) + n_k(t) \end{aligned} \quad (3)$$

In short-term analysis, a set of transformations is applied to the k^{th} frame of speech signal, $y_k(t)$, to convert it to several parameters. Attempts to remove noise effect from the short-term parameters are often done in power spectral domain in which the corrupted speech can be written as

$$Y_k(f) = X_k(f) + N_k(f) \quad (4)$$

Assuming noise to be a stationary process, the above equation can be expressed as

$$Y_k(f) = X_k(f) + N(f) \quad (5)$$

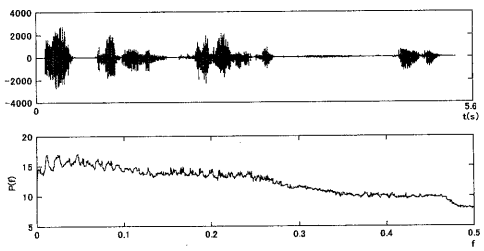
The noise effect estimation in short-term analysis becomes to estimate the noise term in the right hand side of Eq. (5). Much work has addressed how to estimate the noise effect. One typical method is to use a speech activity detector to distinguish speech segments from noise segments. Then the noise effect is estimated from the noise segments [4]. This method has two drawbacks. One is that it depends on the effectiveness of the speech activity detector. Another, as it will become clear soon, is that in some applications, non-speech segments do not contain noise as much as that in speech segments (such as push-to-talk case). In this case, the effects estimated from non-speech segments do not represent the noise during the presence of speech. To avoid a speech activity detector, a continuous spectral subtraction (CSS) was proposed [20-21]. In this approach, the average of short-term power spectra over both speech and noise frames is used as the estimate of noise. The estimated power spectrum of noise, however, is found far from being satisfaction. In this paper, we propose to use long-term analysis to estimate noise effect. This approach is shown to be able to yield a good estimate of noise.

From Eq. (1), the long-term power spectrum of corrupted speech signal can be written as

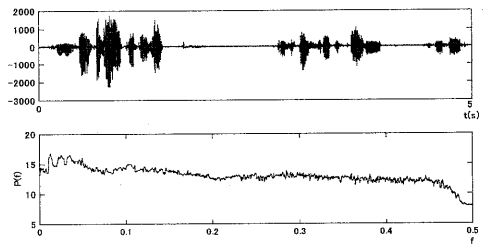
$$\overline{Y}(f) = \overline{X}(f) + \overline{N}(f) \quad (6)$$

The long-term power spectrum is estimated by taking Fourier transform for the whole utterance, and followed by a magnitude square operation. Since noise is assumed to be stationary, $\overline{N}(f)$ in (6) is approximately equal to $N(f)$ in (5). Thus equation (6) can be further written as

$$\overline{Y}(f) = \overline{X}(f) + N(f) \quad (7)$$



S1: It won't I 'm on aim and on arm and it won't do anything



S2: See I'm trying to figure out if I did what I was supposed to before that I think I did

Fig. 2. Long-term effects for two utterances

Fig. 2 plots the long-term power spectra for two utterances. For power spectra, the y-axis is plotted in a log-scale. The speech signals are from SPINE evaluation database. (For detailed description of this database, refer to Section 4.)

An inspection of Fig. 2 reveals that the long-term power spectra for two utterances resemble each other even though the phonetic compositions for the two utterances are different. This may indicate that the long-term power spectra do not contain much phonetic information which is necessary for speech recognition. On the contrary, the long-term power spectrum of noise provides a good estimate of noise due to its stationary property. Based on this observation, we treat (7) as an estimated noise effect, i.e.,

$$\hat{N}(f) = \overline{Y}(f) = \overline{X}(f) + N(f) \tag{8}$$

In the following section, we will discuss how to remove noise effect for speech recognition.

3. NOISE EFFECT REMOVAL

Previous section discussed the estimation of noise. This section will address the problem in suppressing the noise effect in spectral domain. This can be done by subtraction of noise effect from the short-term power spectrum as follows:

$$\hat{Y}_k(f) = Y_k(f) - \hat{N}(f) \tag{9}$$

where $\hat{Y}_k(f)$ is the short-term power-like spectrum after noise suppression. $Y_k(f)$ is the short-term power spectrum of k^{th} frame of speech signal. $\hat{N}(f)$ is the noise estimate.

Substituting (4) and (8) to (9) gives:

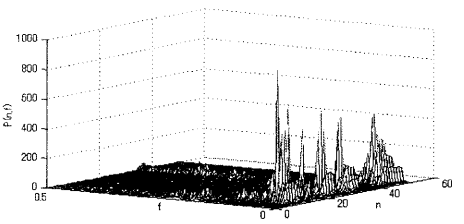
$$\begin{aligned}\hat{Y}_k(f) &= [X_k(f) + N_k(f)] - [\overline{X}_k(f) + \hat{N}(f)] \\ &\approx X_k(f) - \overline{X}_k(f)\end{aligned} \tag{10}$$

One can see from this equation that the noise effect is removed. We call $\hat{Y}_k(f)$ the noise suppressed power spectrum (NSP).

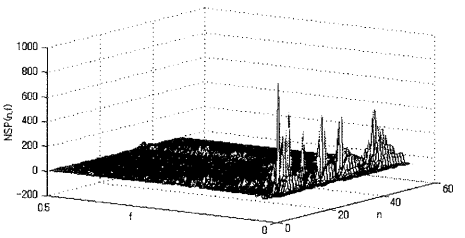
Fig. 3 shows the power spectra and noise suppressed power spectra of one utterance in both noise-free and noise conditions. The clean speech signal is from RM database. (Sr0009.wav from ADG0_4.) In short-term analysis, the window length is 24ms. To show both power spectrum and noise suppressed spectrum in noise condition, we add some noise to the clean speech signal. The noise samples is from NOISEX database (06.ns). The signal-to-noise ratio is controlled to be 10dB.

An inspection of a) and b) shows that power spectrum and noise suppressed power spectrum are similar, excepted that NSP may have some negative values. This indicates that the removal of long-term effect will not affect the phonetic information which is encapsulated in the power spectrum.

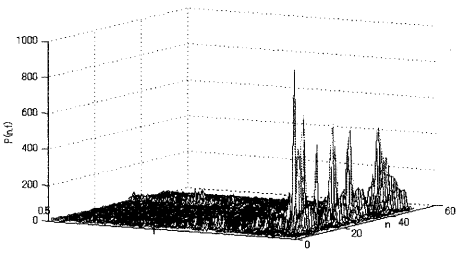
Comparing a) and c), and b) and d), one may find that the noise suppressed power spectrum is less affected by noise. This suggests that the NSP may be more robust to noise. This point will be verified through experiments reported in the next section.



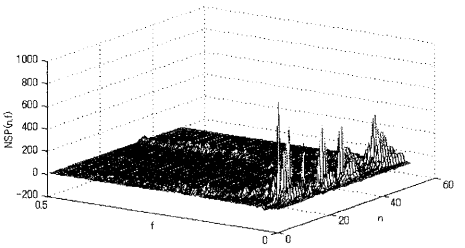
a). Short-term power spectrum of noise-free speech



b). Short-term NSP of noise-free speech



c) Short-term power spectrum of noise speech



d) Short-term NSP of noise speech

Fig. 3. Power spectrum and noise suppressed power spectrum of noise-free and noisy speech

Most of the current speech recognition systems do pattern matching in cepstral domain in which the convoluative distortions are additive. To convert $\hat{Y}_k(f)$ to cepstral parameters, we pass it through a set of mel-scale triangle filter banks whose outputs are further transformed by a log operation and a DCT. However, since $\hat{Y}_k(f)$ may have negative values, the outputs of the filter banks are not guaranteed to be positive. Hence the

log operation used in the estimation of the MFCCs is not applicable in such case. Two ways may be used to deal with this issue. One is to set a positive floor. If a certain output falls below this threshold, it is compulsively set to this floor so that all outputs are guaranteed to be positive. In this paper, however, we introduce another way—redefine a new log operation. Denote the outputs of the filter bands as $E[k,n]$, where $n=1,2,\cdots,N$, N is the total number of filters. The log-operation to the outputs is defined as

$$\log E[k,n] = \log|E[k,n]| + i \arg E[k,n] \qquad (11)$$

where

$$\arg E[k,n] = \begin{cases} 0, & if\ E[k,n] \geq 0 \\ \pi, & if\ E[k,n] < 0 \end{cases} \qquad (12)$$

This log operation yields complex values. In this paper, we take the magnitude of the output of log operation as the inputs to a DCT.

After these transformations, we get an MFCC-like feature set. We call these features noise suppressed MFCCs (NSMFCCs).

4. EXPERIMENTS

We carried out various speech recognition experiments to test the efficiency of the proposed approach. We here select one experiment based on a DARPA noise speech database issued recently for speech in noisy environment evaluation to show the validity of the method.

4.1 SPINE database

This noisy speech database was developed by DARPA to provide a first step forum for assessing the state of the art and practice in speech recognition technology for noisy military environments. The speech data was generated by ARCON corp. for the DoD digital Voicing Processing Consortium (DDVPC) under controlled conditions. The data contained two parts, i. e. training set and test set. However, since only training part is released to public, we will use this part as a test bed to evaluate algorithms and will report results on it. We use SPINE database to refer to the training part of the original database.

This speech data consists of conversations between two communicators working on a collaborative task in which they seek and shoot at targets (ARCON Communicability Exercise, ACE). The speakers can talk freely, but the total vocabulary used is fairly limited. Each person is seated in a sound chamber in which a previously recorded military background noise environment is accurately reproduced. The participants use the microphone that is resident to the particular environment (e. g., ship or helicopter). The whole data include 10 talker pairs. Each talker pair contains twelve 5-minute conversations, and hence about one hour speech. The whole database has about 600-minute speech in total, which include 4 military noise environments: Quiet, Navy Aircraft Carrier CIC, Army HMMWV and Air Force E3A AWACs. We should point out that Push-to-Talk has been used during recording some conversations. This cause the non-speech segments of these conversation do not contain background noise. Hence some noise robust algorithms such as spectral subtraction are not applicable.

The male/female distribution in the database is show below:

P02: male, male
p03: female, female
p04: female, male
p06: female, male
p07: male, male
p08: female, male
p09: female, male
p10: male, male
p11: female, male
p22: female, female

where pxx denotes the index for certain talker pair. We see from above that the whole database have 11 male speakers and 9 female speakers.

Speech is digitized at a sampling rate of 16kHz with 16-bit quantization value for each sample. More detailed decription of this database please refer to [19].

4.2 SPEECH RECOGNITION SYSTEM

In the experiment, the HTK large vocabulary speech recognition system is used to perform the recognition task. This is configured as a gender-independent word-internal triphone mixture Gaussian HMM system. The base phone set contains 42 phonemes, a silence model and a short pause model. A set of word-internal triphone is formed from a dictionary which contains 5250 words (5000 words are from switch board corpus). A left-to-right tee HMM model with 3 emitting states is constructed for each phone. A 4-component mixture Gaussian distribution is used for each emitting state to approximate the probability density function. Word pair language model provided by CMU is included in the decoding processing.

4.3 RECOGNITION RESULT

In the first experiment, we divide the whole database into two sets. P22 which contains two female talkers is used for evaluation. All the other nine pair is used for training. The feature sets we investigated include:

MFCC: 12 MFCCs along with the energy plus the first and second differentials of these features.

NSMFCC: 12 NSMFCCs along with the energy plus the first and second differentials of these features.

Table 1 shows the recognition results for both training and test sets.

Feature set	Training set word accuracy (%)	Testing set word accuracy (%)
MFCC	68.57	58.38
NSMFCC	69.21	60.44

Table 1. Recognition accuracy for P22

In the second experiment, we use P11 as the test bed, while all the other nine pairs are used to train HMM model parameters. Table 2 presents the results for this experiment.

Feature set	Training set word accuracy (%)	Testing set word accuracy (%)
MFCC	68.63	49.49
NSMFCC	68.69	52.21

Table 2. Recognition accuracy for P11

One can see from Table 1 and Table 2 that the noise suppressed MFCCs yield a little bit bitter performance than MFCCs for the training set. For test set, about 2 percent improvement of the word accuracy is obtained. This justify the validity of the NSMFCCs.

SUMMARY

This paper addressed two problems involved in the noise robust speech recognition. One was the estimation of noise effect. To achieve an estimate of noise, we presented a long-term Fourier analysis method. Experiment showed that this approach could provide a good estimate of noise as long as the noise to be dealt with was a stationary process. Another problem addressed was the consideration noise effect during speech recognition. To accomplish this task, we proposed a long-term effect removal to estimate the noise suppressed power spectra from the corrupted speech and a new approach to convert the noise suppressed power spectra to a MFCC-like feature sets. Experiments performed on the SPINE database released by DAPRA recently justified the validity of the proposed approach.

ACKNOWLEDGEMENT

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

REFERENCE

[1] David S. Pallett, *et al*, "1996 Preliminary Broadcast News Benchmark", Proceedings of the 1997 DARPA Speech Recognition Workshop, International Conference Center Chantilly, Virginia, February 2-5, 1997

[2] David S. Pallett, *et al*, "1997 Broadcast News Benchmark Test Results: English and Non-English", Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne Conference Resort, Lansdowne, Virginia, February 8-11, 1998.

[3] David S. Pallett, *et al* "1998 Broadcast News Benchmark Test Results: English and Non-English Word Error Rate Performance Measures", Proceedings of the DARPA Broadcast News Workshop, Hilton at Washington Dulles Airport Herndon, Virginia, February 28-March 3, 1999.

[4] 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.

[5] H. G. Hirsch, P. Meyer, and H. W. Ruehl, "Improved speech recognition using high-pass filtering of subband envelopes," Proc. EUROSPEECH, PP. 413-416, 1991.

[6] D. Geller, R. Haeb-Umbach and H, Ney, "Improvements in speech recognition for voice dialing in the car environment," Proc. ESCA Workshop on Speech Processing in Adverse Condiions, PP. 203-206, Nov. 1992.

[7] 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.

[8] C. H. Lee, F. K. Soong and K. K. Paliwal, "Automatic Speech and Speaker Recognition," Kluwar Academic Publishers, 1996.

[9] M. J. F. Gales and S. J. Young, "Robust speech recognition using parallel model combination," IEEE Trans. Speech Audio Processing, Vol. 4, PP. 352-359, Sep. 1996.

[10] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment independent speech recognition," ICASSP'96, May 1996, PP.733-736.

[11] P. C. Woodland, M. J. F. Gales and D. Pye, "Improving environmental robustness in large vocabulary speech recognition," ICASSP'96, May 1996, PP. 65-68.

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

[13] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," J. Acoustic. Soc. Amer., Vol.87, April 1990, PP.1738-1752.

[14] H. Bourlard and S. Dupont, "A mew ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands," ICSLP'96, Philadelphia, October 1996.

[15] S. Okawa, T. Nakajima and K. Shiria, "A Recombination Strategy for Multi-band Speech Recognition Based on Mutual Information Criterion," EUROSPEECH'99, PP. 603-606.

[16] K. K. Paliwal, "Decorrelated and Liftered Filter-Bank Energies for Robust Speech Recogntion," EUROSPEECH'99, PP. 85-88.

[17] A. Acero, Acoustical and Environmental Robustness in Automatic Speech Recognition, Kluwer Academic Publishers, Boston, MA, 1993.

[18] F. H. Liu, R. M. Stern, A. Acero and P. Moreno, "Environment Normalization for Robustness Speech Recognition Using Direct Cepstral Comparison," ICASSP'94, Vol. II, April 1994, PP. 61-64.

[19] http://www.aic.nrl.navy.mil/

[20] J. A. Nolazco Flores and S. J. Young, "Continuous Speech Recognition in Noise Using Spectral Subtraction and HMM Adaptation," ICASSP'94, PP. 409-412.

[21] D. V. Compernolle, "Noise Adaptation in a Hidden Markov Model Speech Recognition System," Computer Speech and Language, (1989) 3, PP. 151-167.