SUBTRACTION OF ADDITIVE NOISE FROM CORRUPTED SPEECH 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

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. Much research effort in robust speech recognition has been devoted to compensation for the effects of additive noise. In this paper, we will present an approach to remove the additive noise from corrupted speech signal to make speech front-ends immune to additive noise. We address two problems, i.e., noise estimation and noise removal. For noise estimation, we introduce a new method called long-term Fourier estimation. We also discuss how to convert the spectra after subtraction to MFCC-like feature coefficients. We report on experiments on DARPA speech in noisy environments evaluation (SPINE) database to justify the proposed approach.

1. NOISE EFFECT ESTIMATION

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

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

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

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

where index k denotes the k^{th} frame. $w(t)$ 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

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

Compared with speech signal, the effect of noise varies much slowly and is often treated as a stationary process. Hence Eq. (3) can be rewritten as

$$Y(k,f) = X(k,f) + N(f) \qquad (4)$$

There has been considerable interest to compensate the noise component in the right hand side of Eq. (4). One technique called ***spectral subtraction*** [1] has proved to be an important strategy to cope with additive noise. In general, spectral subtraction approach attempts to achieve an estimate of noise power spectrum during the absence of speech, and then subtracts the estimate from the power spectra of corrupted speech. If the noise estimate is $\hat{N}(f)$, the k^{th} frame of spectrum, after spectral subtraction, is written as

$$Y_{SS}(k,f) = Y(k,f) - \hat{N}(f) \qquad (5)$$

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.

To avoid a speech signal detector, a ***continuous spectral subtraction*** (CSS) was proposed [2-4]. In this approach, the average of N consecutive frames of short-term power spectra is used as a noise estimate, i. e.,

$$\hat{N}(t,f) = \frac{1}{N}\sum_{i=t-T}^{t} Y(i,f) \qquad (6)$$

The subtraction scheme in the CSS method is often defined as follows [3]

$$Y_{CSS}(k,f) = Y(k,f) - \alpha\hat{N}(k,f) \qquad (7)$$

where α is an over-estimation factor.

The key issues to a successful implementation of the CSS to an application are the selection of suitable α and N in Eq. (6) and Eq. (7).

In this paper, instead of estimating noise power spectrum in absence of speech or using average of short-term power spectra as noise estimate, we will use the long-term power spectrum of speech signal as noise estimate, i.e.,

$$\hat{N}_L(f) = Y(f) \qquad (8)$$

where $Y(f)$ is the power spectrum of the corrupted speech signal in Eq. (1), which is estimated by taking a Fourier transform of the whole signal and followed by a square magnitude operation.

The subtraction strategy is same as what adopted in spectral subtraction, i.e.,

$$Y_{LTR}(k,f) = Y(k,f) - \hat{N}_L(f) \qquad (9)$$

We call this approach long-term effect removal (LTR). The reason beneath the LTR can be described as below. It was found that 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.

2. FROND-ENDS REPRESENTATION

Another issue with the subtraction-based approaches is the conversion of the spectra after subtraction to some feature coefficients. Since most of the current speech recognition systems do pattern matching in cepstral domain in which the convoluative distortions are additive, we will convert $Y_{LTR}(k,f)$ to cepstral parameters, This can be done by passing $Y_{LTR}(k,f)$ through a set of mel-scale triangle filter banks whose outputs are further transformed by a log operation and a DCT. However, since $Y_{LTR}(k,f)$ 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 an over-estimation flooring factor as adopted in [3]. In this paper, however, we introduce another method — 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

$$\log E[k,n] = \log|E[k,n]| + i \arg E[k,n] \tag{9}$$

where

$$\arg E[k,n] = \begin{cases} 0, & if \quad E[k,n] \geq 0 \\ \pi, & if \quad E[k,n] < 0 \end{cases} \tag{10}$$

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).

3. 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.

3.1 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.

3.2 SPINE DATABASE

Speech in noise environment (SPINE) database was developed by DARPA to assess the state of the art and practice in speech recognition technology for noisy military environments. The database contained conversational speech talker pairs in military noise environments for limited vocabulary. 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.

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

Pair Indices	P02	P03	P04	P06	P07
M/F distribution	MM	FF	FM	FM	MM
Pair Indices	P08	P09	P10	P11	P22
M/F distribution	FM	FM	MM	FM	FF

where Pxx denotes the index for certain talker pair. 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 [5].

3.3 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 + energy) + Δ.

NSMFCC: (12 NSMFCCs + energy) + Δ.

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 justifies 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 of 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. Work is in progress to compare the LTR with CSS, SS and some other noise reduction methods for speech recognition.

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 Research Laboratories, for continuous encouragement.

REFERENCE

[1] 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.

[2] J. A. Nolazco Flores and S. J. Young, “Continuous Speech Recognition in Noise Using Spectral Subtraction and HMM Adaptation”, ICASSP’94, Vol. I, PP. 409-412.

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

[4] M. Shozakai, S. Nakamura and K. Shikano, “Robust Speech Recognition in Car Environments,” ICASSP’98, PP. 269-272.

[5] http://www.ldc.upenn.edu