ROBUST FEATURES DERIVED FROM TEMPORAL TRAJECTORY FILTERING FOR SPEECH RECOGNITION UNDER THE CORRUPTION OF ADDITIVE AND CONVOLUTIONAL NOISES

Kuo-Hwei Yuo and Hsiao-Chuan Wang

Department of Electrical Engineering, National Tsing Hua University,
Hsinchu, Taiwan 30043 ROC
E-mail: hcwang@ee.nthu.edu.tw

ABSTRACT

This paper presents a novel method using robust features for speech recognition when the speech signal is corrupted by additive and convolutional noises. This method is conceptually simple and easy to be implemented. The additive noise and the convolutional noise are removed by temporal trajectory filtering in autocorrelation domain and cepstral domain, respectively. No prior information of noise corruption is necessary. A task of multi-speaker isolated digit recognition is conducted to demonstrate the effectiveness of using these robust features. The cases of channel filtered speech signal corrupted by additive white noise and color noise are tested. Experimental results show that significant improvements can be achieved as comparing with some traditional features.

1. INTRODUCTION

It is well known that the performance of automatic speech recognition systems may drastically degrade in case of mismatch between training and test environments. This problem is inevitably encountered when the speech recognizers are deployed in real environment where noise or channel effect always exists.

The techniques for robust speech recognition may be classified into four categories: speech enhancement methods [1], robust feature representation [2-5], model compensation methods [6-7], and robust distance measures [8-9] .They may be designed to deal with additive noise, convolutional noise, or the combination of both noises. Some of these methods had applied the technique of temporal trajectory filtering. The method of RASTA uses temporal filtering in logarithmic spectrum domain[3]. Hirsch *et al*. [10] use temporal filtering in linear subband domain. Avendano *et al*. [11] extend this temporal filtering method to DFT magnitude trajectory for dereverberation of speech. If we look at the case of corruption by additive and convolutional noises simultaneously, we find that many methods need the prior information of noise or spend much computation effort. These disadvantages result in the restriction for practical applications.

This research has been sponsored by the National Science Council, Taiwan, ROC, under contract number NSC-86-2745-E-007-010.

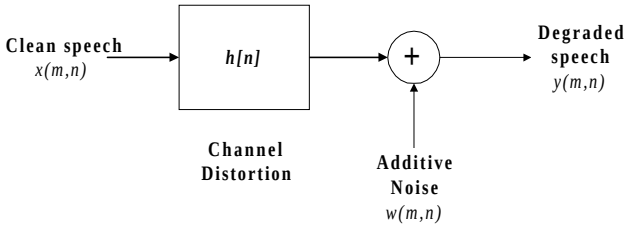

Figure 1. Block diagram of the degradation

This paper proposes a method using new features that are robust to stationary noise. In short-term analysis, a set of feature parameters is derived for each frame so that a speech is represented by a sequence of feature vectors. If the effect of stationary noise is a constant added to the features for all the frames, this constant can be removed by filtering the temporal trajectory of feature vectors along the frames. Here we remove the additive noise by filtering the trajectory of autocorrelation coefficients and the channel effect by filtering the trajectory of cepstral coefficients. These two noises can be simultaneously removed in our proposed method without too much effort in computation.

2. ROBUST FEATURES

In this paper, the focus is on the effect of the additive and convolutional noises to the speech recognition. We assume that the channel and the additive noise are stationary. The degradation of speech signal is shown in Figure 1. This speech is modeled as

$$y(m,n) = x(m,n) \otimes h(n) + w(m,n), \qquad (1)$$

$$0 \le m \le M-1,\ 0 \le n \le N-1$$

where m is frame index, n is discrete time index within a frame, $x(m,n)$ is clean speech, $y(m,n)$ is degraded speech, $h(n)$ is impulse response of channel, $w(m,n)$, is additive noise, and "$\otimes$" denotes the convolution operation. If $x(m,n)$, $w(m,n)$, and $h(n)$ are uncorrelated, the autocorrelation of the noisy speech can be expressed as

$$r_{yy}(m,k) = r_{xx}(m,k) \otimes h(k) \otimes h(k) + r_{ww}(m,k) \qquad (2)$$

$$0 \le m \le M-1,\ 0 \le k \le N-1$$

where $r_{yy}(m,k)$, $r_{xx}(m,k)$ and $r_{ww}(k)$ are the short-term autocorrelation sequences of noisy speech, clean speech, and additive noise, respectively, and k is the autocorrelation sequence

index within a frame. Because the additive noise is assumed to be stationary, $r_{ww}(m,k)$ does not change for all frame index *m* and thus the frame index can be dropped off. Eq. (2) becomes

$$r_{yy}(m,k) = r_{xx}(m,k) \otimes h(k) \otimes h(-k) + r_{ww}(k) \qquad (3)$$

$$0 \le m \le M-1,\ 0 \le k \le N-1 \ .$$

In this paper, only one-sided autocorrelation sequence of each frame is computed,

$$r_{yy}(m,k) = \frac{1}{N-k} \sum_{j=0}^{N-1-k} y(m,j)y(m,j+k) \qquad (4)$$

$$0 \le m \le M-1,\ 0 \le k \le N-1 \ .$$

2.1 Removal of Additive Noise in Autocorrelation Domain

Differentiating both sides of (3) with respect to frame index *m* for all *k* yields

$$\frac{\partial}{\partial m} r_{yy}(m,k) = \frac{\partial}{\partial m}\left\{r_{xx}(m,k)\right\} \otimes h(k) \otimes h(-k) \qquad (5)$$

$$0 \le m \le M-1,\ 0 \le k \le N-1 \ .$$

The sequence, $\{\partial r_{yy}(m,k)/\partial m\}_{k=0}^{N-1}$, is named the relative autocorrelation sequence (RAS) of degraded speech at *m*th frame. Eq. (5) demonstrates that the effect of the additive noise is removed from the RAS of degraded speech but the convolutional noise still exists. The RAS's can be obtained by polynomial approximation in a manner similar to the derivative of delta cepstral coefficient from cepstral coefficients [12]. Therefore, RAS's are approximated by

$$\frac{\partial r_{yy}(m,k)}{\partial m} \cong \frac{1}{T_L} \sum_{t=-L}^{L} t \cdot r_{yy}(m+t,k) \qquad (6)$$

$$0 \le m \le M-1,\ 0 \le k \le N-1$$

where

$$T_L = \sum_{t=-L}^{L} t^2 \ . \qquad (7)$$

Eq. (6) can be interpreted as a filtering process of the temporal autocorrelation trajectory using an FIR filter,

$$H(z) = \frac{1}{T_L} \sum_{t=-L}^{L} t \cdot z^t \ . \qquad (8)$$

2.2 Removal of Convolutional Noise in Cepstral Domain

In Eq. (5), we can see that the effect of convolutional noise still exists. This convolutional noise becomes a bias when RAS is transformed into cepstral domain. If we consider the RAS as another short-term time-domain representation of speech, we can compute its mel-frequency cepstral coefficient (MFCC). This new feature is denoted RAS-MFCC.

$$C_{RAS,y}(m,p) = C_{RAS,x}(m,p) + 2C_h(p) \qquad (9)$$

$$0 \le m \le M-1,\ 0 \le p \le P-1$$

were $C_{RAS,y}(m,p)$ and $C_{RAS,x}(m,p)$ denote cepstra of RAS of degraded speech and clean speech, respectively. $C_h(p)$ denotes cepstrum of convolutional noise and *p* denotes cepstral index. It is well known that cepstral mean normalization method [2] and delta cepstral coefficients [12] are two effective methods for removing bias in traditional MFCC feature. Here we also apply these two methods to RAS-MFCC for removing bias and then obtain two new features named CMN-RAS-MFCC and delta RAS-MFCC. Both features are robust to additive noise and convolutional noise. The operation of these two method can be interpreted as temporal filtering in cepstral domain. Figure 2 illustrates the process for computing CMN-RAS-MFCC and delta RAS-MFCC.

2.2.1 Cepstral mean normalization of RAS-MFCC (CMN-RAS-MFCC)

The CMN-RAS-MFCC is computed by subtracting the mean of RAS-MFCC's in a speech utterance. For a set of *M* RAS-MFCC's in an utterance, the mean vector is

$$V(p) = \frac{1}{M} \sum_{m=0}^{M-1} C_{RAS,y}(m,p),\ 0 \le p \le P-1 \qquad (10)$$

and the CMN-RAS-MFCC's are given by

$$C_{RAS,y}^{cmn}(m,p) = C_{RAS,y}(m,p) - V(p) \qquad (11)$$

$$0 \le m \le M-1,\ 0 \le p \le P-1 \ .$$

2.2.2 Delta coefficients of RAS-MFCC (Delta RAS-MFCC)

Because the bias is invariant for frame index *m*, the partial differential of RAS-MFCC with respect to frame index *m* can remove the bias and yields the delta RAS-MFCC

$$\Delta C_{RAS,y}(m,p) \qquad (12)$$

$$= \frac{\partial}{\partial m} C_{RAS,y}(m,p) \cong \frac{1}{T_s} \sum_{t=-S}^{S} t \times C_{RAS,y}(m+t,p)$$

$$0 \le m \le M-1,\ 0 \le p \le P-1 \ ,$$

where

$$T_S = \sum_{t=-S}^{S} t^2 \ . \tag{13}$$

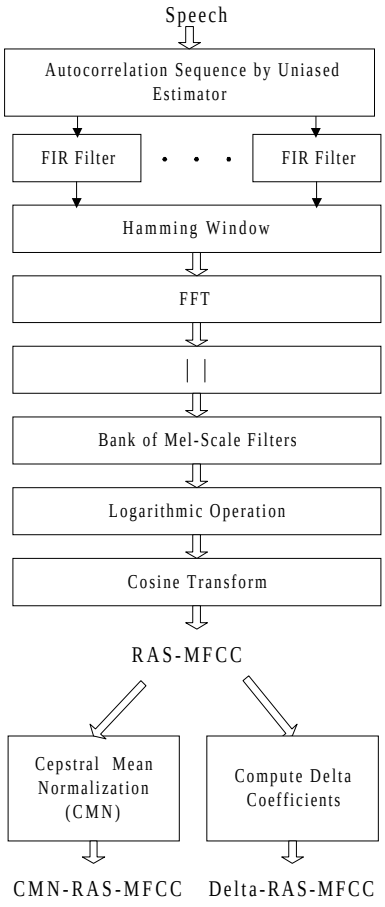

Figure 2. Block diagram of computing CMN-RAS-MFCC and Delta-RAS-MFCC

3. EXPERIMENTS

Some experiments were conducted to evaluate the proposed features for speech recognition in additive noise at different signal to noise ratios and different channel distortions.

Utterances from a Mandarin isolated-digit database, collected from 100 speakers (50 males and 50 females) by 8 kHz sampling rate in noise-free environment, were used as clean speech. Three sessions had been recorded. In a session, each speaker uttered one repetition of ten isolated-digits. The first two sessions were used for training. The third session artificially corrupted by additive noise and convolutional noise was for testing. The white Gaussian noise was artificially generated by computer, while the color noise was obtained from NOISEX-92. A total of 41 channel filter models collected from real telephone networks in Taiwan were randomly chosen for producing the channel effect to speech signal. Each digit was modeled by a left-to-right HMM without skips. Each HMM had seven to nine states depending on the duration of digit. Each state was represented by a mixture of four Gaussian components with diagonal covariance matrix. The first and last states of each HMM were tied together as silence state. Note that all the HMM models are trained from clean speech.

The experimental results for various types of features in three testing conditions are shown in Table 1. A total of eight types of features were evaluated. The first three features are traditional MFCC, delta MFCC, and their concatenation. The fourth feature, CMN-MFCC, represents traditional MFCC compensated by CMN method. The fifth feature is the concatenation of CMN-MFCC and delta-MFCC. The final three features are CMN-RAS-MFCC, delta-RAS-MFCC, and their concatenation.

Three testing conditions are labeled as “I”, “II”, and “III" in Table 1. The first condition is that the tested speech is not corrupted by any noise. The second condition is that the tested speech is only corrupted by convolutional noise. The third condition is that the speech is corrupted by both convolutional noise and additive noise. Two types of additive noise, white noise and factory noise, were used.

3.1 No additive and convolutional noise corruption

For the case that test speech signals are not corrupted by any noise, the concatenation of MFCC and delta-MFCC achieves the best performance. The use of features based on RAS-MFCC can not gain any advantages. The reason is simple. Since the test speech is also clean speech, any noise removal operation is meaningless and even makes worse. However, the concatenation of CMN-RAS-MFCC and delta-RAS-MFCC is comparative to the concatenation of traditional MFCC and delta-MFCC.

3.2 Only convolutional noise corruption

When the test speech is only corrupted by convolutional noise, the recognition accuracy of traditional MFCC is reduced from 95.7 % to 88.7 %. CMN-MFCC and the concatenation of CMN-MFCC and delta MFCC still keep their performance well because they do have removed the channel effect. However, both CMN-RAS-MFCC and delta-RAS-MFCC lose about 5 % in recognition rate as comparing to their clean speech test. The concatenation of CMN-RAS-MFCC and delta-RAS-MFCC shows some resistance to this reduction and get only 3 % loss in recognition rate.

3.3 Simultaneous convolutional and additive noise corruption

When the test speech is corrupted by convolutional noise and additive noise simultaneously, the recognition rate of MFCC is seriously decreased. Delta-MFCC, CMN-MFCC, and their concatenation also lose their recognition rates even they are better than MFCC. Both CMN-RAS-MFCC and delta-RAS-MFCC show remarkable performance over those traditional MFCC features. The concatenation of CMN-RAS-MFCC and delta-RAS-MFCC can get further improvement in most cases.

It is worth to note that the proposed features are good in cases of white noise and color noise.

Table 1 The recognition rates for using various types of features.

I. No corruption by any noise. II. Only corruption by convolutional noise. III. Corruption by additive and convolutional noise

Feature		Dim	I	Convolutional noise									
				Additive noise									
				II	III White					III Factory			
			clean	∞	20dB	15dB	10dB	5dB	0dB	20dB	15dB	10dB	5dB	0dB
1	***MFCC)***	12	0.957	0.887	0.583	0.439	0.291	0.165	0.107	0.548	0.394	0.311	0.256	0.2
2	***ΔMFCC***	12	0.971	0.956	0.878	0.778	0.551	0.326	0.184	0.898	0.798	0.633	0.372	0.191
3	***MFCC+ΔMFCC***	24	0.985	0.953	0.738	0.586	0.348	0.128	0.101	0.719	0.516	0.311	0.172	0.117
4	***CMN-MFCC***	12	0.979	0.973	0.822	0.65	0.384	0.214	0.143	0.812	0.629	0.429	0.237	0.142
5	***CMN-MFCC + ΔMFCC***	24	0.983	0.979	0.906	0.775	0.441	0.194	0.148	0.883	0.681	0.454	0.241	0.152
6	***CMN-RAS-MFCC***	12	0.966	0.912	0.918	0.895	0.855	0.71	0.435	0.89	0.848	0.771	0.593	0.392
7	***Δ-RAS-MFCC***	12	0.935	0.883	0.886	0.848	0.785	0.657	0.457	0.866	0.831	0.781	0.612	0.415
8	***CMN-RAS-MFCC + Δ-RAS-MFCC(24)***	24	0.978	0.947	0.948	0.932	0.871	0.714	0.411	0.926	0.899	0.795	0.582	0.348

4. CONCLUSION

In this paper, we have derived new robust features, CMN-RAS-MFCC and delta RAS-MFCC, for speech recognition under the corruption of both additive noise and convolutional noise. In deriving these features, no prior information about noise is necessary. The proposed method is applicable to white noise or color noise corruption. The approach is conceptually simple and easy to be implemented for practical applications.

REFERENCES

[1] S. F. Boll. "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech. Signal Processing*, vol. 27. pp. 113-120, Apr. 1979.

[2] A. Acero and R. M. Stern, "Robust speech recognition by normalization of the acoustic space," in *Proc. IEEE Int. Conf Acoust.. Speech, Signal Processing*, 1991, pp. 893-896.

[3] H. Hermansky and N. Morgan, "RASTA Processing of Speech," *IEEE Trans. Speech Audio processing*, vol. 2, pp. 578-589, October, 1994

[4] D. Mansour and B. H. Juang, "The short-time modified coherence representation and noisy speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 795-804, June 1989.

[5] J. Hernando and C. Nadeu, "Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition," *IEEE Trans. Speech Audio processing*, vol. 5, no. 5, pp. 80-84, Jan. 1997.

[6] M. J. F. Gales and S. J. Young, "Robust Speech Recognition in Additive and Convolutional Noise Using Parallel Model Combination", *Computer Speech and Language*, pp. 289-307, Sep. 1995.

[7] M. G. Rahim and B. H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 4, no. 1, pp. 19-30, Jan. 1996

[8] D. Mansour and B. H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition", *IEEE Trans. Acoust., Speech. Signal Processing*, vol. 37, no. 11, pp. 1659-1671, Nov. 1989.

[9] B. A. Carlson and M. A. Clements, "A projection-based likelihood measure for speech recognition in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 2, no. 1, part 1, pp. 97-102, Jan. 1994.

[10] H. G. Hirsch, P. Meyer, and H. Ruchl, "Improved speech recognition using high-pass filtering of subband envelopes," *Proc. EUROSPEECH*'91, (Genova), 1991, pp. 413-416.

[11] C. Avendano and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," in *Proc. ICSLP'96*, vol. 2, pp.

[12] S. Furui, "Speaker-independent isolated word recognition based on emphasized spectral dynamics," in *Proc. IEEE Int. Conf Acoust.. Speech, Signal Processing* (Tokyo), 1986, pp. 1991-1994.