AN ADAPTIVE APPROACH TO ROBUST SPEECH RECOGNITION

†Jinhai Cai and ‡Zhi-Qiang Liu

Computer Vision and Machine Intelligence Lab
Department of Computer Science
The University of Melbourne, Australia
email: †cai@cs.mu.OZ.AU ‡zliu@cs.mu.OZ.AU

ABSTRACT—The performance of speech recognizers often degrades rapidly in noisy acoustic environments. The environmental noise not only disturbs speech features and affects the reliability of feature extraction, it also causes people to change their speaking manners. In this paper, we focus on the acoustic effects of noise. We propose to use the short-time modified coherence representation with a noise-adaptive approach for extraction of speech features and adaptive weighted logarithmic output probabilities of HMMs for enhancing the robustness to the errors in weak speech segments. As a result, proposed approach works well in both Gaussian white noise and computer fan noise.

INTRODUCTION

In automatic speech recognition, hidden Markov models (HMMs) have been widely accepted. This is because HMM provides a mathematically rigorous means for modeling speech signals. High performance has been achieved under laboratory conditions (Rabiner, Wilpon and Soong, 1989; Picone, 1990). However, when the background noise presents, speech recognition systems based on HMM degrade rapidly below a certain signal-to-noise ratio (SNR). Noise affects speech recognition systems in many ways. Additive noise disturbs the speech features. For example, white noise reduces the norm of cepstrum coefficient (Mansour and Juang, 1989a). Furthermore, noise causes people to change their speaking manners known as Lombard effects which also reduce the accuracy of speech recognizers. Therefore, noise is one of the most difficult problems for commercial use of speech recognition systems. In order to combat the effects of noise, many techniques have been developed, including speech enhancement (Hansen and Clements, 1991) which restores speech feature, robust distortion measure (Mansour and Juang,1989a) which immunes to noise distortion, noise adaptation (Roe,1987; Varga and Moore, 1990) which adapts the speech models to noise condition in certain domain, and better feature representation schemes such as short-time modified coherence (SMC) (Mansour and Juang, 1989b).

The SMC takes advantage of inherent coherence in adjacent segments of speech to enhance the SNR. For quasi-stationary speech signals, it was shown by Mansour and Juang (1989b) that all-pole modeling of SMC is a more robust signal representation than that of the speech itself. As a result,

SMC-based methods were able to achieve high performance if the noise is white. However, in real world, noise is usually not white. For instance, power spectra of noises generated by cruising car and computer fan show a decaying trend with the increase of frequency. Therefore, the SMC of speech, which is based on Gaussian white noise assumption, is not effective to such noise which results in poor performance.

ADAPTIVE APPROACH

In order to improve the performance of speech recognition system, we propose to use a technique based on noise-adaptive filtering and weighted probabilities of HMMs.

Noise-adaptive filtering

In order to exploit the advantages of SMC representation, we convert the colored noise to Gaussian white noise. If the noise is $d(n)$, its spectrum, $D(z)$, can be estimated by linear prediction:

$$D(z) = \frac{G}{A(z)} = \frac{G}{\sum_{i=0}^{P} a(i) z^{-i}} \tag{1}$$

where G is the gain, $a(i)$ are the linear prediction coefficients, P is the order of linear prediction. When speech is corrupted by noise, the noisy speech is

$$y(n) = s(n) + d(n) \tag{2}$$

where $s(n)$ is original speech signal. After filtered by $A(z)$, the noisy speech becomes

$$y_f(n) = s(n) * a(n) + d(n) * a(n) = s_f(n) + d(n) * a(n) \tag{3}$$

where $*$ denotes convolution. Now, the noise, $d(n)*a(n)$, becomes white noise which can be removed by the SMC representation. If the analysis window size is N, the one-side autocorrelation estimate of noisy speech is

$$\rho_{y_f}(m;l) = \sum_{n=0}^{N-1} y_f(n;l) y_f(n+m;l) \quad 0 \le m \le N \tag{4}$$

where $y_f(n;l)$ is a frame taken from $y_f(n)$, l is the index of the frame. Because the noise is assumed to be uncorrelated with speech signal, we get

$$\rho_{y_f}(m;l) = \begin{cases} \rho_{s_f}(0;l) + G^2 & m = 0 \\ \rho_{s_f}(m;l) & m > 0. \end{cases} \tag{5}$$

In order to reduce the noise, we compute the discrete Fourier transform of $\rho_{y_f}(m;l)$ by excluding $\rho_{y_f}(0;l)$

$$\Gamma_{y_f}(n;l) = \sum_{m=1}^{N} \rho_{y_f}(m;l) w(m-1) e^{-j\frac{2\pi}{N}mn} \quad 0 \le n \le N-1 \tag{6}$$

where $w(m)$ is the Hamming window, $j = \sqrt{-1}$. The speech spectrum is obtained by applying inverse filter of $A(z)$ in frequency domain

$$S(n;l) = \left| \frac{\Gamma_{y_f}(n;l)}{A^2(e^{-j\frac{2\pi}{N}n})} \right|. \tag{7}$$

The estimate of autocorrelation sequence for the original speech signal is computed using the inverse Fourier transform of $S(n;l)$

$$\hat{\rho}(m;l) = \frac{1}{N} \sum_{n=0}^{N-1} S(n;l) e^{j\frac{2\pi}{N}mn}. \tag{8}$$

In this way, our approach not only maintains the main advantages of SMC representation but also is effective to reduce colored noise.

When we apply this technique, the spectral dynamic range of SMC representation must be taken into consideration. Although the dynamic range of an autocorrelation sequence is large, the limited attenuation in sidelobes of a given analysis window results in a small dynamic range of SMC representation (Mansour and Juang, 1989b). Consequently, the highest spectral peak may mask low energy regions in spectrum. Therefore, if we use the inverse filter of $A(z)$ to restore speech features as described in equation (7) when SNR is high, some regions in speech spectrum would be over compensated. The solution to the problem of overcompensation is that different methods are introduced to reduce noise according to a given local SNR. If the local SNR is below a certain threshold, we apply the method as described above. On other hand, if the local SNR is higher than the threshold, spectral subtraction is used to reduce noise. The residual noise can be viewed as white. Therefore, standard SMC method is effective to remove such noise.

One additional problem is $A(z)$ may change the dynamic range of restored speech spectrum. For robust speech recognition, we have made all efforts to produce consistent features in adverse environments. The inconsistency of dynamic range will result in degradation in accuracy. Therefore, dynamic range normalization is essential. The normalized speech spectrum is given by

$$s(n;l) = \max\{s(n;l), \frac{S_{max}(l)}{D_y}\} \tag{9}$$

where $S_{max}(l) = \max_n\{s(n;l)|s(n;l) \geq \hat{D}(n)\}$, $\hat{D}(n)$ is spectrum of the remaining noise and D_y is the dynamic range. After normalization, the proposed approach reduces the noise effect on the restored spectrum and produces consistent spectrum as well.

Reliability-adaptive weighting

Normally, at the boundaries of a string (or substring), speech signals have low energy and their features are distorted by noise. As a result, features at boundaries will become inaccurate and unreliable in the presence of noise. It is reported (Junqua and Reaves, 1994) that, for an isolated-word recognizer, more than half of the recognition errors were due to errors at boundaries. Therefore, maintaining the robustness of the recognizer in weak signal segments is important. This can be solved as follows: the reliability of feature extraction can be defined as a function of short term speech energy in a given segment($Engl$), remaining mean noise energy(E_n) and its variance (σ_n):

$$RF_l = \begin{cases} 1 - 2\int_{Engl}^{\infty} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\frac{-(x-E_n)^2}{2\sigma_n^2} dx & Engl > E_n, \\ 0 & Engl \leq E_n. \end{cases} \tag{10}$$

In our system, we use logarithmic output probability of HMM to recognize words. An unreliable feature estimation of a given segment should have less effect on decision making. Therefore, it should have a smaller weight. We define the weights as a function of reliability

$$W_0(RF) = \min\{RF_{l-1}, RF_l, RF_{l+1}\} \tag{11}$$

and

$$W_1(RF) = RF_l. \tag{12}$$

For a standard HMM, its logarithmic output probability at state, j, for an observation vector, $(\vec{O}_l, E_l \,\&\, \Delta E_l)$, is

$$B_j(\vec{O}_l, E_l, \Delta E_l) = \log b_j(\vec{O}_l) + \log p_{1j}(E_l)^{re} + \log p_{2j}(\Delta E_l)^{re} \tag{13}$$

where re is a constant, $\vec{O}_l$ is a vector concatenating the weighted cepstral vector and the corresponding weighted delta cepstral vector (Rabiner, Wilpon and Soong, 1989), E_l and ΔE_l are the logarithmic normalized speech energy and delta logarithmic normalized speech energy for a given analysis frame repectively, $b_j(\vec{O}_l)$, $p_{1j}(E_l)$ and $p_{2j}(\Delta E_l)$ are state observation probabilities. Therefore, the weighted logarithmic output probability of the HMM is

$$\begin{aligned} WB_j(\vec{O}_l, E_l, \Delta E_l) = \; & W_0(RF) \cdot \log b_j(\vec{O}_l) + W_0(RF) \cdot \log p_{2j}(\Delta E_l)^{re} + \\ & re \cdot \log[(1 - W_1(RF)) \textstyle\sum_{E_n \le E} p_{1j}(E) + W_1(RF) \cdot p_{1j}(E_l)]. \end{aligned} \tag{14}$$

In this way, the recognition decision making is mainly based on reliable estimation.

DATABASE AND FEATURE EXTRACTION

Database

The database used for evaluating proposed approach is the Texas Instruments-Developed Studio Quality Speaker-Independent Connected-Digit Corpus (TIDIGITS). The corpus was collected at Texas Instruments in a quiet acoustic enclosure. The data was sampled at 20 KHz by a 16 bit A/D converter. In order to reduce the computation in feature extraction stage, we resampled data at 10 KHz after filtered by a 4.6 KHz antialiasing filter. In our experiments, a subset of TIDIGITS database is used which consists digit strings spoken by 16 speakers (8 men, 8 women). Half of digit strings in the subset forms training set, and the rest forms test set. There are 2048 words (digits) in the training set and 1986 words in the test set.

Feature extraction

Features are extracted from speech signals using a 25.6 ms window at 10 ms intervals by proposed approach. For each frame, coefficients of 10 order LPC are computed from the autocorrelation sequence which is obtained using equation (8). The weighted cepstrum vector and corresponding delta cepstrum vector (Rabiner, Wilpon and Soong, 1989) are

$$\begin{cases} \hat{c}(m;l) = c(m;l) \cdot [1 + \frac{Q}{2} \sin(\frac{\pi m}{Q})] \\ \Delta\hat{c}(m;l) = C \sum_{k=-K}^{K} k \cdot \hat{c}(m;l-k) \end{cases} \qquad 1 \le m \le Q \tag{15}$$

where $Q = 12$, $K = 2$, C is a constant and $c(m;l)$ is the LPC derived cepstrum vector. The vector $\vec{O}_l$ is obtained by concatenating $\hat{c}(m;l)$ with $\Delta\hat{c}(m;l)$:

$$\vec{O}_l = \{\hat{c}(m;l), \Delta\hat{c}(m;l)\}. \tag{16}$$

The restored speech energy E_l (dB) and delta energy ΔE_l are

$$\begin{cases} E_l = 10 \log Eng_l, \\ \Delta E_l = \sum_{k=-K}^{K} k \cdot E_{l-k}, \end{cases} \tag{17}$$

where $Eng_l = \hat{\rho}(0;l)$. The noise is estimated from the first and last few frames of a given digit string. The mean remaining noise energy and its variance are

$$\begin{cases} E_n = \frac{1}{2L} \sum_{l=0}^{L-1} [\hat{\rho}(0;l) + \hat{\rho}(0;T-l-1)] \\ \sigma_n = \sqrt{\frac{1}{2L} \sum_{l=0}^{L-1} \{[E_n - \hat{\rho}(0;l)]^2 + [E_n - \hat{\rho}(0;T-l-1)]^2\}} \end{cases} \tag{18}$$

where T is the frame number of the given string.

EXPERIMENTAL RESULTS

In this connected digit recognition experiments, all eleven digits, zero to nine plus 'o', are modeled using six-state discrete HMMs (DHMMs). All of testing speech signals are segmented by the method (Rabiner, Wilpon and Juang, 1986) on clean condition and known strings with DHMMs. Therefore, there are only few errors in segmentation. The noisy speech signals are generated by adding noise signals on clean speech.

In the first experiment, the speech is corrupted by Gaussian white noise which is consistent with the assumption of standard SMC method. As a result, high performance is achieved by the SMC method. This method with weighted logarithmic output probabilities of HMMs (SMC_W) can further improved the accuracy of the recognizer. The results are shown in Table 1. Because $A(z) = 1$ for Gaussian white noise, there are no difference between the standard SMC method and adaptive approach with SMC (ASMC) and between SMC_W method and corresponding adaptive method (ASMC_W).

Gaussian White Noise			
SNR	SMC	SMC_W	LPC
∞	97.8	97.8	98.9
20	96.3	97.0	87.8
15	93.4	95.3	68.4
10	89.1	93.6	38.7
5	78.7	86.6	24.2

Table 1: Word recognition accuracies (%) for connected digits corrupted by Gaussian white noise

In the second experiment, the speech is corrupted by computer fan noise whose low frequency energy is predominant. Due to the small dynamic range of the SMC representation, the spectral peak of noise may mask the low energy regions of speech spectrum and result in the loss of information when the standard SMC method is applied. Therefore, its performance is worse than that of conventional LPC method. Our approach converts colored noise into white before applying SMC method, therefore there is no noise spectral peaks. Besides, our approach excludes the noise predominant spectral regions to avoid its masking effects when the dynamic range is normalized. The experimental results shown in Table 2 confirm the improvement of our adaptive approach.

Computer Fan Noise					
	Without Weighting		With Weighting		
SNR	SMC	ASMC	SMC_W	ASMC_W	LPC
∞	97.8	97.8	97.8	97.8	98.9
20	95.7	96.8	96.4	97.1	98.1
15	91.4	95.3	93.6	96.0	93.6
10	76.5	91.2	83.4	93.5	87.2
5	58.1	82.7	72.3	87.2	69.5

Table 2: Word recognition accuracies (%) for connected digits corrupted by computer fan noise

CONCLUSIONS

This paper reports an adaptive approach which uses a noise-adaptive filtering method before applying the SMC method to reduce noise and emphasizes the reliable features by using adaptive weighted

logarithmic output probabilities of HMMs to enhance the robustness to the errors in weak speech segments. Consequently, the proposed approach has achieved better performance than that of standard the SMC method in recognizing noisy speech corrupted by Gaussian white noise and computer fan noise at all SNR conditions. Our approach also significantly improves the accuracy comparing with conventional LPC method when SNR is low. Experimental results show the proposed approach is effective to Gaussian white noise and some types of colored noises.

REFERENCES

Hansen, J.H.L. & Clements, M.A. (1991)*Constrained iterative speech enhancement with application to speech recognition*, IEEE Trans. on Signal Processing, Vol.39, No.4, pp.795-805.

Junqua, J. C., Mak, B. & Reaves, B. (1994) *A robust algorithm for word boundary detection in the presence of noise*, IEEE Trans. on Speech and Audio Processing, Vol.2, No.3, pp.406-412.

Mansour, D. & Juang, B.H. (1989a) *A family of distortion measures based upon projection operation for robust speech recognition*, IEEE Trans. on Acoust., Speech and Signal Processing, Vol.37, No.11, pp.1659-1671.

Mansour, D. & Juang, B.H. (1989b) *The short-time modified coherence representation and noisy speech recognition*, IEEE Trans. on Acoust., Speech and Signal Processing, Vol.37, No.6, pp.795-804.

Rabiner, L.R., Wilpon, J.G. & Juang, B.H. (1986) *A segmental k-means training procedure for connected word recognition*, AT&T Tech. J., pp.21-40.

Rabiner, L.R., Wilpon, J.G. & Soong, F.K. (1989) *High performance connected digit recognition using hidden Markov models*, IEEE Trans. on Acoust., Speech and Signal Processing, Vol.37, No.8, pp.1214-1225.

Picone, J. (1990) *Continuous speech recognition using hidden Markov models*, IEEE Acoust., Speech and Signal Processing Magazine, pp.26-41.

Roe, D.B. (1987) *Speech recognition with a noise-adapting codebook*, Proceedings of the IEEE International Conference on ASSP, pp.1139-1142.

Varga, A. & Moore, R.K. (1990) *Hidden Markov model decomposition of speech and noise*, Proceedings of the IEEE International Conference on ASSP, pp.845-848.