Improved Robust Speech Recognition Considering Signal Correlation Approximated by Taylor Series

Jia-lin Shen, Jeih-weih Hung, Lin-shan Lee
Institute of Information Science, Academia Sinica
Taipei, Taiwan, Republic of China
jlshen@iis.sinica.edu.tw

ABSTRACT

In this paper, an improved mismatch function by considering signal correlation between speech and noise is proposed to better estimate the noisy speech HMM's. A linearized model based on Taylor series expansion approach is used to approximate the proposed mismatch function. The parameters of the noisy speech HMM's can be estimated more precisely by combining the parameters of the clean speech and noise HMM's in the log-spectral domain or cepstral domain. Experimental results show that improved robustness for speech recognition in the presence of white noise as well as colored noise can be obtained.

1. INTRODUCTION

The mismatch between training and testing environments is the major cause for the performance degradation of many recognition techniques. Quite several approaches were proposed to adapt the HMM's trained by clean speech to a new environment [1-5]. Apparently, the mismatch function for the additive noise and the clean speech is non-linear in the log-spectral domain or cepstral domain, and was very often expressed as [1]

$$x = \log(\exp(s)+\exp(n)), \qquad (1)$$

where x, s and n denote the log-spectral representation of the corrupted noisy speech, clean speech and noise, respectively. The cepstral representation can be easily obtained by applying the discrete cosine transform (DCT). The noisy speech HMM's can thus be derived based on the above non-linear mismatch function. In Parallel Model Combination (PMC) method [1], the clean speech HMM's are transformed from the cepstral domain to the linear-spectral domain, combined with the parameters of the noise HMM's in that domain, and then inversely transformed back to the cepstral domain for recognition. As an alternative, the above non-linear mismatch function can be approximated by a linearized model using the Taylor series expansion (TSE) approach [2-4], so that the clean speech and noise HMM's can be directly combined in the log-spectral domain or cepstral domain. In this way, the recognition can be performed in matched training and testing conditions using the estimated noisy speech HMM's.

In our previous study [5], we found that a correlation term between speech and noise will appear in the power spectrum of the noisy speech signal, which is ignored in the widely used mismatch function as shown in equation (1). This correlation term can be ignored in the estimation of the mean of noisy speech based on the assumption of zero-mean of noise. However, this assumption is not necessarily suitable especially for the short-term spectral analysis for speech recognition [5]. In this paper, it is proposed to model the correlation term between speech signals and noise, at least to some extent, and tries to obtain a tractable solution with a linearized model based on Taylor series expansion

approach [6]. The above equation (1) can be modified to

$$x = g(s,n) = \log(s^l + n^l + r\sqrt{s^l}\sqrt{n^l}), \quad (2)$$

where $s^l = \exp(s)$ and $n^l = \exp(n)$ mean the power spectrum of the clean speech and noise, respectively. The first two terms in equation (2) are in fact the same as in equation (1), while the last term represents the correlation, with r being a correlation factor. Based on the modified mismatch function proposed here, the parameters of the noisy speech HMM's can be estimated more precisely in terms of the parameters of the clean speech and noise HMM's.

2. ESTIMATION OF MODEL PARAMETERS IN NOISY ENVIRONMENTS

As mentioned above, the corrupted noisy speech can be represented by a mismatch function composed of clean speech and noise. Based on the mismatch function, the mean and variance of the corrupted noisy speech can be estimated if the Gaussian distributions are assumed to model the speech.

2.1 Modified Mismatch Function

In the short-term spectral analysis for speech recognition, the power spectrum of the noisy speech, $||X(w)||^2$, can be derived from the power spectrum of clean speech and noise as shown in the following :

$$\begin{aligned} \|X(w)\|^2 &= \|S(w)+N(w)\|^2 \\ &= \|S(w)\|^2 + \|N(w)\|^2 + 2\,\mathrm{Real}(S(w)\cdot\overline{N}(w)) \\ &\le \|S(w)\|^2 + \|N(w)\|^2 + 2\|S(w)\|\|N(w)\|. \end{aligned}$$

As a result, the power spectrum of the noisy speech can then be expressed as shown in equation (2), where $s^l = ||S(w)||^2$ and $n^l = ||N(w)||^2$, respectively, and it can be noted that the absolute value of the correlation factor r shown in equation (2) is located between 0 and 2.

2.2 Estimating Parameters by Taylor Series Expansion (TSE)

From [6], if the mismatch function $g(s,n)$ in equation (2) is sufficiently smooth near the point (μ_s, μ_n), with μ_s and μ_n representing the mean of clean speech and noise in the log-spectral domain, respectively, then the mean and variance of noisy speech can be estimated in terms of the mean and variance of s and n :

$$\mu_x = g(\mu_s,\mu_n) + \frac{1}{2}\left(\frac{\partial^2 g(\mu_s,\mu_n)}{\partial s^2}\sigma_s^2 + \frac{\partial^2 g(\mu_s,\mu_n)}{\partial n^2}\sigma_n^2\right), \quad (3)$$

$$\sigma_x^2 = \left(\frac{\partial g(\mu_s,\mu_n)}{\partial s}\right)^2\sigma_s^2 + \left(\frac{\partial g(\mu_s,\mu_n)}{\partial n}\right)\sigma_n^2. \quad (4)$$

Here the clean speech and noise are assumed uncorrelated and the Taylor series expansion of order 2 is used [5]. The formulations for Taylor series expansion with higher order can refer to [2]. Therefore the estimated mean and variance of the noisy speech signals can be obtained according to equations (3)(4) :

$$\begin{aligned} \mu_x = {} & \log(e^{\mu_s} + e^{\mu_n} + re^{\mu_s/2}e^{\mu_n/2}) + \\ & \frac{1}{2}\frac{\frac{r}{4}e^{\frac{3}{2}\mu_s}e^{\frac{1}{2}\mu_n} + e^{\mu_s}e^{\mu_n} + \frac{r}{4}e^{\frac{1}{2}\mu_s}e^{\frac{3}{2}\mu_n}}{(e^{\mu_s} + e^{\mu_n} + re^{\frac{1}{2}\mu_s}e^{\frac{1}{2}\mu_n})^2}(\sigma_s^2 + \sigma_n^2), \end{aligned} \quad (5)$$

$$\begin{aligned} \sigma_x^2 = {} & \left(\frac{e^{\mu_s} + \frac{r}{2}e^{\frac{1}{2}\mu_s}e^{\frac{1}{2}\mu_n}}{e^{\mu_s} + e^{\mu_n} + re^{\frac{1}{2}\mu_s}e^{\frac{1}{2}\mu_n}}\right)^2\sigma_s^2 + \\ & \left(\frac{e^{\mu_n} + \frac{r}{2}e^{\frac{1}{2}\mu_s}e^{\frac{1}{2}\mu_n}}{e^{\mu_s} + e^{\mu_n} + re^{\frac{1}{2}\mu_s}e^{\frac{1}{2}\mu_n}}\right)^2\sigma_n^2. \end{aligned} \quad (6)$$

As a comparison, the estimated parameters of noisy speech HMM's using Parallel Model Combination (PMC) approach based on the proposed mismatch function are expressed as [5] :

$$\begin{aligned} \mu_{x^l} &= \mu_{s^l} + \mu_{n^l} + r\sqrt{\mu_{s^l}\cdot\mu_{n^l}}, \\ \sigma_{x^l}^2 &= \sigma_{s^l}^2 + \sigma_{n^l}^2 + \rho_1\mu_{s^l}\cdot\mu_{n^l} + \\ &\quad \rho_2\sigma_{s^l}\sqrt{\mu_{s^l}\cdot\mu_{n^l}} + \rho_3\sigma_{n^l}\sqrt{\mu_{s^l}\cdot\mu_{n^l}}. \end{aligned} \quad (7)$$

where ρ_1, ρ_2 and ρ_3 are weighting parameters. Note

that the composition process is performed in the linear-spectral domain and then the corresponding mean and variance in the log-spectral domain can be obtained accordingly [1] :

$$\mu_x = \log(\mu_{x^l}) - \frac{1}{2}\log\left(\frac{\sigma_{x^l}^2}{\mu_{x^l}} + 1\right)$$
$$\sigma_x^2 = \log\left(\frac{\sigma_{x^l}^2}{\mu_{x^l}} + 1\right) \quad (8)$$

Based on the mismatch function, the parameters of the noisy speech HMM's can be derived either from the *direct* estimation by TSE in equations (5)(6) or from the *indirect* estimation using PMC approach as shown in equations (7)(8). However, no matter what approaches are adopted, the more accurate the mismatch function, the more precise the estimated parameters.

2.3 Modified Mismatch Function for Dynamic Features

The proposed mismatch function can be also applied to dynamic features and the corresponding estimated mean and variance for noisy speech HMM's can be therefore obtained :

$$\Delta x = x(t+k) - x(t-k)$$
$$= \log\left(\frac{e^{\Delta s}}{1+\frac{1}{C}+r\sqrt{\frac{1}{C}}} + \frac{e^{\Delta n}}{1+C+r\sqrt{C}} + \frac{re^{\frac{\Delta s}{2}}e^{\frac{\Delta n}{2}}}{r+\sqrt{C}+\sqrt{\frac{1}{C}}}\right) \quad (9)$$

where Δ denotes the delta operation and a simple difference over a window width, *k*, is used. In addition, *C* is a function of signal-to-noise ratio (SNR) with $C = 10^{SNR/10}$.

3. EXPERIMENTAL RESULTS

3.1 Speech database

The speech database included 4 sets of 1345 isolated syllables in Mandarin Chinese produced by two male speakers. 3 sets were used for training and 1 for testing. The recognition rates quoted are the average of the rates for each of the speakers. All the speech data were obtained in an office-like laboratory environment. They were low-pass filtered, digitized by an Ariel S-32C DSP board with sampling frequency 16kHz. After end-point detection was performed, 20 ms Hamming window was applied every 10 ms with a pre-emphasis factor of 0.95. 14-order mel-frequency cepstral coefficients derived from the power spectrum filtered by a set of 30 triangular band-pass filters were used for each frame. In order to include additive noise in the speech database, the noise from NOISEX92 database was added to clean speech for different level of SNR's. In addition, noise HMM's for different levels of noise were individually trained, composed of one state and one mixture per state.

3.2 Experiments

In the first experiment as shown in Table 1 in the presence of white noise, the recognition rates were 48.33%, 14.42% and 2.01% for 30dB, 20dB and 10dB of SNR, respectively, which were immediately increased to 69.81%, 44.98% and 24.16%, respectively, using the PMC method. In addition, these rates were improved to 78.66%, 65.65% and 42.60% for 30dB, 20dB and 10dB of SNR, respectively, using the previously proposed TSE methods [2-3]. The PMC method based on the proposed correlated mismatch function led to the recognition rates of 78.36%, 59.26% and 36.80% for 30dB, 20dB and 10dB of SNR's, respectively. Here the correlation factor *r* was set to 1. Furthermore, the TSE approach based on the proposed correlated model provided the recognition rates up to 78.29%, 66.17% and 44.16% for 30dB, 20dB and 10dB of SNR's, respectively. It can be found that direct estimation using TSE approach outperforms indirect estimation using PMC approach. Moreover, the recognition rates can be improved especially under low SNR conditions when the signal correlation between speech and noise is considered. Also, in the case of 30 dB the

recognition rates were very near to those using matched HMM as shown in the last row of Table 1. This is probable the reason why the correlated model becomes not helpful.

Type	30dB	20dB	10dB
Clean HMM	48.33	14.42	2.01
PMC	69.81	44.98	24.16
Correlated PMC	78.36	59.26	36.80
TSE	78.66	65.65	42.60
Correlated TSE	78.29	66.17	44.16
Matched HMM	80.67	71.15	49.74

Table 1. Recognition rates for different versions of models in the presence of white noise.

As shown in Table 2, the recognition rates under F16 noisy environments were degraded to 66.25%, 29.81% and 8.55% for 30dB, 20dB and 10dB of SNR, respectively. Similarly, the PMC and TSE approaches can provide significant improvements and TSE outperforms PMC in any cases. Moreover, the compensation based on the proposed modified mismatch function can be further improved. The results listed in Table 2 indicated that TSE compensation based on the correlated model gave the recognition rates up to 83.72%, 75.17% and 52.08% for 30dB, 20dB and 10dB of SNR, respectively. Improvements under low SNR conditions were observed as compared to original TSE approach. We believe that the estimation of noisy speech HMM's can be further improved if more precise estimation of correlation between speech and noise can be applied.

4. CONCLUSION

In this paper, we proposed a modified mismatch function by considering signal correlation for speech signals in the presence of additive noise. Based on the mismatch function, the parameters of the noisy speech HMM's can be estimated more precisely and improved robustness can be therefore obtained.

REFERENCES

1. M.J.F. Gales and S.J. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination", *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 5, Sep. 1996, pp. 352-359.
2. N.S. Kim, "Statistical Linear Approximation for Environment Compensation", *IEEE Signal Processing Letters*, Vol. 5, No. 1, Jan. 1998, pp. 8-10.
3. D.Y. Kim, N.S. Kim, C.K. Un, "Model-based Approach for Robust Speech Recognition in Noisy Environments with Multiple Noise Sources", *Eurospeech*, Vol. 3, 1997, pp. 1123-1126.
4. P.J. Moreno, B. Raj and R.M. Stern, "A Vector Series Approach for Environment Independent speech recognition", *Int. Conf. Acoustics, Speech, Signal Processing(ICASSP)*, May, 1996, pp. 733-736.
5. Jeih-weih Hung, Jia-lin Shen, Lin-shan Lee, "Improved Robustness for Speech Recognition Under Noisy Conditions Using Correlated Parallel Model Combination", *Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Vol. 1, 1998, pp. 553-556.
6. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, New York, McGraw-Hill, 1984.

Type	30dB	20dB	10dB
Clean HMM	66.25	29.81	8.55
PMC	78.74	58.22	25.87
TSE	83.64	74.57	50.93
Correlated TSE	83.72	75.17	52.08
Matched HMM	84.54	81.04	69.22

Table 2. Recognition rates for different versions of models in the presence of F16 noise (colored noise).