ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

INTEGRATION OF MODEL ADAPTATION AND MISSING FEATURE THEORY FOR ROBUST SPEECH RECOGNITION

An-Tze YU

Department of Electrical Engineering
National Tsing Hua University, Hsinchu
yuat@ms25.hinet.net

Hsiao-Chuan WANG

Department of Electrical Engineering
National Tsing Hua University, Hsinchu
hcwang@ee.nthu.edu.tw

ABSTRACT

This paper presents a method to integrate the model adaptation technique and the missing feature theory for robust speech recognition. The recognition performance can be further improved as comparing with that utilizing only a single approach. The basic idea comes from that the integration of two techniques will compensate two improper assumptions made separately by these two techniques. The evaluation of the proposed method on the task of isolated Mandarin digits recognition shows that the system robustness is upgraded substantially.

1. INTRODUCTION

A set of line spectrum frequencies (LSFs) is an alternative representation of linear prediction coefficients (LPCs) [1]. A vocal tract model represented by LSFs can tolerate bigger quantization errors. Because of this property, LSF representation is more attractive and extensively used in the area of speech coding and synthesis [2]. It is appealing to use LSFs in the task of speech or speaker recognition within a digital communication based system [3].

Additive noises distort the LSFs and hence substantially degrade the recognition performance [4]. Many methods have been proposed for the robustness issue [5][6]. Though the model adaptation technique and the missing feature theory are proved effective in improving performance separately. The integration of the two methods should further enhance the recognition robustness. The model adaptation adjusts a clean speech model to match a testing environment. This allows a more accurate pattern matching during speech recognition phase. The missing feature theory categorizes noisy features into reliable and unreliable ones according to noise masking criteria. Using this information, different weights can be applied to reliable features and unreliable features to reduce the impact of the additive noise on the recognition performance.

The basic idea comes from that the integration of these two techniques will compensate two improper assumptions made separately by them. First, model adaptation technique assumes that all features are reliable under the influence of additive noise. This is not true because the speech signal is unreliable when the level of noise is higher than the level of speech signal. Second, the missing feature theory assumes that the reliable features do not change in the presence of additive noise. However, the fact is that all features will be influenced, though having different extents. For this reason, we expect the achievement of improved robustness as these two assumptions are compensated.

In the proposed method, the energy distribution of noise in spectral domain is first estimated by a noise estimator. The energy distribution is then fed into a model adaptation module to adapt the clean speech model to match the noisy condition. And a missing feature module also uses this energy distribution to category the input features of each frame into reliable ones and unreliable ones. The marginalization method is taken here. Hence, those unreliable features are ignored. The evaluation of the proposed method on the task of isolated Mandarin digits recognition shows that the system robustness is upgraded substantially.

2. A BRIEF REVIEW OF LSFS

This section briefly reviews the preliminary of LSFs for completeness to the explanations in next sections.

2.1. Linear Prediction

In a LP analysis, the model of speech production is expressed by a discrete-time equation as

$$s(n) = \sum_{i=1}^{M} a(i)s(n-i) + Ge(n) , \tag{1}$$

where $a(1), a(2),.., a(M)$ are the LP coefficients, M is the system order, $e(n)$ is the excitation, and G is the gain of the excitation. The expression of Eq.(1) in the z-domain is

$$S(z) = \frac{GE(z)}{A(z)} , \tag{2}$$

where

$$A(z) = 1 - \sum_{i=1}^{M} a(i) z^{-i} \tag{3}$$

is the inverse filter, $S(z)$ and $E(z)$ are the signal and the excitation, respectively. $\frac{G}{A(z)}$ is called the LPC model, and often used to characterize the spectral envelope of the speech signal.

2.2. Line Spectrum Frequencies

LSFs can generally be obtained from LPC model by defining a symmetrical polynomial, *P(z),* and an anti-symmetrical polynomial, *Q(z),* from the inverse filter, *A(z)*,

$$\begin{aligned} P(z) &= A(z) + z^{-(M+1)} A(z^{-1}) \\ Q(z) &= A(z) - z^{-(M+1)} A(z^{-1}) \end{aligned} . \tag{4}$$

The zeros of $P(z)$ and $Q(z)$ are on the unit circle and interlaced. These zeros are complex conjugates and their angles are the LSFs.

LSF can also be computed by defining the Ratio Filter as,

$$R(z) = z^{(M+1)} \frac{A(z)}{A(z^{-1})} , \tag{5}$$

Then the phase of the Ratio Filter is given by

$$\phi(\omega) = (M+1)\omega + 2\theta(\omega) , \tag{6}$$

where $\phi(\omega)$ and $\theta(\omega)$ are the phases of $R(\omega)$ and $A(\omega)$ respectively. The LSFs are frequencies at which the phase of the ratio filter is a multiple of π -radians, i.e.

$$\phi(\omega_k) = k\pi , \quad k = 1,2,...,M . \tag{7}$$

3. NOISE COMPENSATION METHODS

When speech contaminated additive noise, the features of speech are distorted according to the extent of the additive noise. The distortion will cause severe degradation on the recognition performance. Many methods have been proposed for tackling this problem. This section reviews two compensation methods applied here.

3.1. The Model Adaptation for LSFs

The LSFs of a speech signal will be distorted by an additive noise. According to [5], the LSFs of the noisy signal can be expressed as,

$$\omega_y^k = \lambda(\omega_y^k)\omega_x^k + (1 - \lambda(\omega_y^k))\omega_w^k , \tag{8}$$

where ω is a LSF, and *y, x* and *w* are the short notation for noisy speech, clean speech and noise, respectively. Hence ω_y^k, ω_x^k and ω_w^k in Eq.(8) are the *k*-th LSFs respectively for noisy speech, clean speech and noise. The ω_y^k must be computed iteratively, because the desired parameter ω_y^k is needed in computing $\lambda(\omega_y^k)$. The algorithm is summarized as follows:

Initialization

1. *n=0*

2. $\omega_y^k[0] = \frac{\omega_x^k + \omega_w^k}{2}$

3. $r = \frac{G_x}{G_w}$

4. $\alpha = \frac{r}{r+1}$

Recursion

5. $\theta_d(\omega_y^k[n]) = \theta_w(\omega_y^k[n]) - \theta_x(\omega_y^k[n])$

6. $r_A(\omega_k^y[n]) = \frac{\left|A_x(\omega_k^y[n])\right|}{\left|A_w(\omega_k^y[n])\right|}$

7. $\beta(\omega_y^k[n]) = \frac{1}{\theta_d(\omega_y^k[n])} \tan^{-1} \frac{\alpha \sin\theta_d(\omega_y^k[n])}{(1-\alpha) r_A(\omega_y^k[n]) + \alpha\cos\theta_d(\omega_y^k[n])}$

8. $\phi_x(\omega_y^k[n]) = (M+1)\omega_y^k[n] + 2\theta_x(\omega_y^k[n])$

9. $\phi_w(\omega_y^k[n]) = (M+1)\omega_y^k[n] + 2\theta_w(\omega_y^k[n])$

10. $m_x(\omega_y^k[n]) = \frac{\phi_x(\omega_x^k) - \phi_x(\omega_y^k[n])}{\omega_x^k - \omega_y^k[n]}$

11. $m_w(\omega_y^k[n]) = \frac{\phi_w(\omega_y^k[n]) - \phi_w(\omega_w^k)}{\omega_y^k[n] - \omega_w^k}$

12. $\lambda(\omega_y^k[n]) = \frac{\beta(\omega_y^k[n]) m_x(\omega_y^k[n])}{\beta(\omega_y^k[n]) m_x(\omega_y^k[n]) + (1 - \beta(\omega_y^k[n])) m_w(\omega_y^k[n])}$

13. $\omega_y^k[n+1] = \lambda(\omega_y^k[n])\omega_x^k + (1 - \lambda(\omega_y^k[n]))\omega_w^k$

14. *if* $\left|\omega_y^k[n+1] - \omega_y^k[n]\right| > \delta$

then $n \Leftarrow n + 1$, and go to 5,

otherwise go to 15.

Termination

15. $\omega_y^k = \omega_y^k[n]$

3.2. The Missing Feature Theory [6]

Recent works have shown that the application of the missing feature theory in speech and speaker recognition under noisy conditions improves the recognition performance. In this approach the time-frequency representation of the noisy speech signal is partitioned into reliable and unreliable regions according to noise masking criteria. In the framework of the approach two main techniques in the classification process have been implemented: marginalization - when unreliable data are ignored, data imputation - when unreliable data are estimated. Missing features theory often directly applied for spectral features, such as filter bank energy, and indirectly applied to non-spectral features such as Cepstral features.

In HMMs, each state is defined by emission and transition probabilities. For a single state model λ, the probability to emit vector $X = [\omega_x^1, ..., \omega_x^k, ..., \omega_x^M]^T$ is expressed as

$$\Pr ob(X \mid \lambda) = \sum_{i=1}^{M} p_i \prod_{k=1}^{M} N(\omega_x^k, \mu_{i,k}, \sigma_{i,k}^2) \quad (9)$$

where p_i is the weight for i-th Gaussian pdf, X is a vector containing the LFSs and the delta-LSFs, and $\mu_{i,k}$ and $\sigma_{i,k}^2$ are the k-th components of the vectors of mean and variance for i-th Gaussian pdf. The components of X can be divided into reliable and unreliable features according to noise masking criteria. In this paper, the energy distributions of a testing speech and the estimated noise in spectral domain are compared. A component will be classified as reliable if its corresponding spectral energy is larger than that of the estimated noise. In Eq. (9) the contribution of the reliable and unreliable components can be expressed as follows

$$\Pr ob(X \mid \lambda) = \sum_{i=1}^{M} p_i \prod_{k:reliable} N(\omega_x^k, \mu_{i,k}, \sigma_{i,k}^2) \prod_{k:unreliable} N(\omega_x^k, \mu_{i,k}, \sigma_{i,k}^2) \quad (10)$$

In the marginilization method, only the reliable components were used for the recognition task. In this case, the marginal pdfs are used to compute the emission probabilities.

$$\Pr ob(X \mid \lambda) = \sum_{i=1}^{M} p_i \prod_{k:reliable} N(\omega_x^k, \mu_{i,k}, \sigma_{i,k}^2) \quad (11)$$

4. INTEGRATION OF MODEL ADAPTATION AND MISSING FEATURE THEORY

The model adaptation technique and the missing feature theory are integrated in this study. Fig. 1a describes the general architecture for a speech recognition system. First, the feature extraction module extracts a features sequence from the input speech. And then, given a trained speech model, the features sequence is recognized by the recognizer. Fig. 1b describes the proposed architecture for the integration. First, The noise estimator estimates the energy distribution of noise in spectral domain from a testing utterance. Given the energy distribution, the clean speech model is adapted to match the noise environments using the algorithm described in the previous section. And, the noise information is used in each frame to classify each component of the features into reliable or unreliable. The component will be classified as reliable if its corresponding spectral energy is larger than that of the estimated noise signal. When the speech model has been adapted and missing features have been defined. The recognition is then performed.

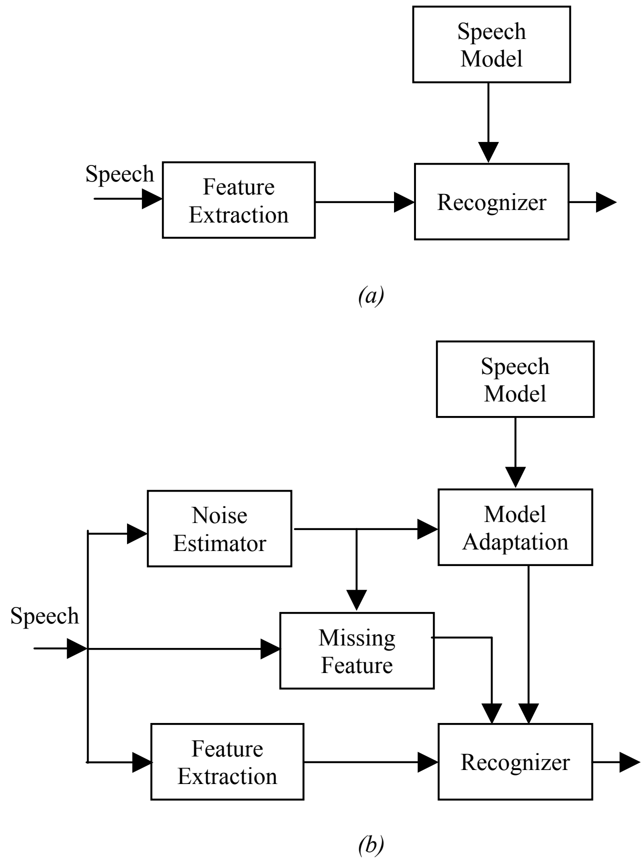

Figure 1. (a) Architecture of general speech recognition system (b) Architecture of the proposed method

5. EXPERIMENTS

A series of experiments were conducted to evaluate the performance of the proposed methods. The task is to recognize isolated Mandarin digits in noisy environment. The speech data provided by 50 male and 50 female speakers were collected in a sound treated environment. There are three sessions of data collection. Each speaker uttered a set of ten Mandarin digits in a recording session. Two sessions are used for training the speech models and the other session is for testing.

The additive noise is added to the clean speech with specific SNR values for generating artificial noisy speech. Three noises are tested. They are white noise, factory noise, and the cabin noise of an F16 fighter. White noise is artificially generated by using random number generating program. The other two noises are picked up from NOISEX-92 database.

Each Mandarin digit is modeled by a HMM. The HMM contains seven states beginning with a pre-silence state and ending with a post-silence state. The silence states for all digits are tied together. Each state is modeled by a mixture of 4 Gaussian densities. All covariance matrices are diagonal. The recognition feature is composed of 10 LSFs and 10 delta-LSFs. Without performing any model adaptation or noise compensation, the recognition rates are shown in table 1. Then model adaptation and missing feature techniques are applied separately to test their compensation ability. The results are displayed in table 2 and 3. Finally, the proposed integration method is performed. The results are shown in table 4. Table 5 summarizes the results. This table indicates that the model adaptation gives better performance than the missing feature. And the proposed method produces the best performance in all cases.

6. CONCLUSIONS

This paper presents a method to integrate the model adaptation technique and the missing feature theory for robust speech recognition. The basic idea comes from that the integration of two techniques will compensate the two improper assumptions made separately by these two techniques. The evaluation of the proposed method on the task of isolated Mandarin digits recognition shows that the system robustness is upgraded substantially.

Contaminating noise	SNR				
	20 dB	15 dB	10 dB	5 dB	0 dB
White noise	75.4	62.6	46.4	36.9	27.4
Cabin noise	95.4	92.0	81.1	70.3	50.5
Factor noise	95.8	95.5	91.6	81.3	66.7

Table 1. Recognition rates without noise compensation

Contaminating noise	SNR				
	20 dB	15 dB	10 dB	5 dB	0 dB
White noise	85.3	78.3	68.6	54.4	43.8
Cabin noise	97.2	96.6	94.1	89.2	77.7
Factor noise	98.1	97.3	95.8	91.9	84.3

Table 2. Recognition rates with model adaptation

Contaminating noise	SNR				
	20 dB	15 dB	10 dB	5 dB	0 dB
White noise	80.2	74.3	61.4	46.7	37.9
F16 noise	97.1	96.3	93.6	89.4	77.0
Factor noise	97.8	97.0	94.9	90.2	82.7

Table 3. Recognition rates with missing Feature

Contaminating noise	SNR				
	20 dB	15 dB	10 dB	5 dB	0 dB
White noise	87.1	79.6	70.2	56.8	46.5
F16 noise	98.4	97.2	95.7	91.3	80.7
Factor noise	98.3	97.8	96.0	93.4	86.1

Table 4. Recognition rates with the proposed method

	white	F16	Factory	Average	Improve-ment
Baseline	49.7%	77.9%	86.2%	71.3%	---
Model Adaptation	66.1%	90.8%	93.4%	83.4%	17.1%
Missing Feature	60.1%	90.5%	92.5%	81.1%	13.8%
Integration method	68.0%	92.7%	94.3%	85.0%	19.3%

Table 5. Summary of recognition performance for various compensation methods

7. ACKNOWLEDGEMENT

This research was sponsored by the National Science Council, Taiwan, under contract numbers NSC-89-2213-E- 007-129 and NSC-90-2213-E-007-028.

8. REFERENCES

[1] Itakura F. . "Line spectrum representation of linear predictive coefficients of speech signals," J. Acoust. Soc. Amer., vol. 57, Suppl., no. 1, S35, 1975.

[2] Paliwal, K. K. , "A study of LSF representation for speaker-depedent and speaker-independent HMM based speech recognition systems," *Proc. ICASSP'90*, pp. 804-807, 1990

[3] Yu, A. T. and Wang, H. C., “Effect of noise on line spectrum frequency and a robust speech recognition method for the low bit-rate encoded speech,” *Proc. Int. Conf. on Phonetic Science’99*, 1999

[4] Yu, A. T. and Wang, H. C., "Model adaptation in line spectrum domain", *Proc. Int. Conf. Acoustics, Speech, Signal Processing'00,* pp.1547-1550 ,2000

[5] Yu, A. T. and Wang, H. C., "An algorithm for finding line spectrum frequencies of added speech signals and its application to robust speech recognition" *Proc. European Conference on Speech Communication and Technology,* 2001.

[6] Renevey, P. and Drygajlo, A., "Estimation of unreliable data for robust speech recognition", Proc. ICASSP, Istanbul, Turkey, 2000