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

STRUCTURE-BASED COMPENSATION USING AN IMPROVED STATISTICAL LINEAR APPROXIMATION FOR MANDARIN SPEECH RECOGNITION OVER TELEPHONE

Zhao-Bing HAN, Hua-Yun ZHANG, Bo XU

National Lab of Pattern Recognition, Institute of Automation

Chinese Academy of Sciences, Beijing, 100080
{zbhan, hyzhang , xubo }@nlpr.ia.ac.cn

ABSTRACT

In this paper, a Vector Piecewise Polynomial (VPP) approximation algorithm is proposed for robust speech recognition in telecommunication environments. The method is formulated in a statistical framework in order to perform the optimal compensation of noise effect given the observed noisy speech, a model describing the statistics of the speech recorded in clean reference environment and the estimation of the noisy recognition environment.

The VPP algorithm is an extension of P.J.Moreno's Vector Taylor Series (VTS) approximations for dealing with the distortion due to channel effects and background noise. We use a piecewise polynomial, namely two linear polynomials and a quadratic polynomial, to approximate the environment function (f(v)). Moreno replaced f(v) by its vector Taylor series approximation. It is well known that VTS is not precise if variables (v) are not close to the Taylor expansion points (v0). The VPP algorithm can overcome this defect. In addition, VPP estimates the parameters of the environment by the expectation-maximization (EM) algorithm.

Experimental results are presented in the paper on the application of this approach in improving the performance of Mandarin large vocabulary continuous speech recognition (LVCSR) due to different transmission channels (Such as fixed telephone line and GSM) and the background noise. The proposed VPP algorithm is found to converge fast. The method can reduce the average character error rate (CER) by about 12 %.

1. INTRODUCTION

In real-world applications, the mismatch between training and testing conditions often results in significant degradation in performance of automatic speech recognition (ASR) system. Some sources of mismatch include additive noise, channel and transducer mismatches that contribute to a spectral tilt and a spectral shaping, speaker mismatch, different accents, stress, and different speaking styles [1]. Our work has been concentrated on the problems of channel effects and additive noise.

Over the past years, many compensation approaches have been proposed to deal with the problem of speech recognition in noisy environment. These methods fall into three broad categories [2]. (1) Training-based compensation. It assumes the availability of a pair of stereo training sets, one corresponding to speech utterances recorded in clean and undistorted conditions and the other corresponding to data recorded simultaneously through a second set of distorted channel and microphone conditions. Such examples are empirical cepstral compensation (Stern et al., 1996), variance adapted likelihood measure (VALM) algorithms (Chien et al, 1996) and so on. (2) Blind compensation. This approach assumes that there is uncertainty in parameter estimation of the train model or the observed feature representation of the test utterance. It is known as the minimax classification approach (Merhav and Lee, 1993), HMM Inversion (Moon and Hwang, 1995) and so forth. (3) Structure-based compensation. Compensation is accomplished by modifying the parameters of the assumed structures. So that we can apply the structural constrains to guide the compensation process. In this category, the methods are the richest, for instance, stochastic matching(SM) framework (Sankar and Lee, 1996), codeword-dependent cepstral normalization (CDCN; Acero 1990), parallel model combination (PMC) (Gales and Young, 1995), maximum likelihood linear regression (MLLR; Leggetter and Woodland, 1995), vector Taylor series (VTS; P. J. Moreno, 1996), etc.

In this paper, we present a structure-based noise compensation method to remove the effect of channel and background noise over telephone line. As mentioned earlier, this approach (VPP) assumes a model of the environment and uses the piecewise polynomial to approximate f(v). VTS approximates the function by its Taylor series expansion around the mean of the clean speech. But it is known that Taylor series is not precise approximation if the distribution of the variable (v) isn't around the point of Taylor's expansion. The experimental result [4] has shown that there is much residual mismatch. In addition, the Vector Polynomial approximationS (VPS) algorithm (Bhiksha Raj, 1996) is analogous to VPP. But VPS uses the triangular function to approximate the second derivative of f(v). Although it may be very exact to approximate the derivative, its approximation of the function f(v) which is derived from integrating the derivative twice may be not precise.

2. VPP APPROXIAMTION ALGORITHM

2.1 A Model of the Environment

In the previous papers, we assume a model of the environment in which speech is corrupted by unknown additive stationary noise and linearly filtered by a telephone channel. The Model is shown in Fig.1 [3].

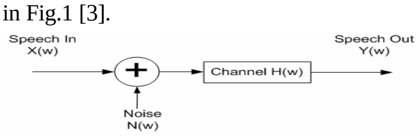

Fig.1. Environment model of a distorted telephone network.

$$Y(w) = X(w)\,|H(w)|^2 + N(w) \qquad (1)$$

Where Y(w) represents the power spectrum of the degraded speech, X(w) is the power spectrum of the clean speech, H(w) is the transfer function of the linear filter, and N(w) is the power spectrum of the additive noise.

In the log-spectral domain this relation can be expressed as:

$$y = x + h + \log(1 + e^{n-x-h}) \qquad (2)$$

or in more general terms:

$$y = x + f(x, n, h)$$

Where h is an unknown parameter that represents the effect of linear filter and n represents the effect of background noise. f(x,n,h) denotes the environment function.

We also assume that the PDF of the log-spectra of the speech signal can be well represented by a summation of multivariate Gaussian distributions:

$$p(x) = \sum_{k=0}^{M-1} p[k] N_x(m_x \Sigma_{x,k})$$

Furthermore, we assume that the statistics of the noise can be well represented by a single Gaussian $N_n(m_n, \Sigma_n)$.

2.2 VTS algorithm

To get an analytical solution, VTS replaces the environmental vector function f(x,n,h) by its vector Taylor series approximation. So a vector Taylor series is defined around a set of points m_x, n_0 and h_0.

$$y = x + f(x,n,h) \cong x + f(m_x, n_0, h_0) + \nabla_x f(m_x, n_0, h_0)$$
$$(x - m_x) + \nabla_n f(m_x, n_0, h_0)(n - n_0) + \nabla_h f(m_x, n_0, h_0)$$
$$(h - h_0) + \ldots$$

Therefore its mean and variance are shown as:

$$m_{y,k} = (I + \nabla_x f)' m_{x,k} + (\nabla_n f)' m_n + (\nabla_h f)' h +$$
$$f(x_0, n_0, h_0)$$

and

$$\Sigma_{y,k} = (I + \nabla_x f)' \Sigma_{x,k} (I + \nabla_x f) + (\nabla_n f)' \Sigma_n (\nabla_n f)$$

As we all know, the approximation of VTS is not exact if x, n and h are not close to the expansion points m_x, n_0 and h_0.

The VPP algorithm can deal with this problem.

2.3 VPP algorithm

We define $g(v) = \log(1 + e^v)$ for convenience. g(v) is the nonlinear part of the environment function f(v). If g(v) can be approximated precisely, the approximation of f(v) will be exact. We firstly investigate the property of g(v), then use the piecewise polynomial to replace it.

2.3.1 The property of g(v)

From Fig.2 , g(v) is monotonically increasing, with asymptotes at g(v)=0 for $v \to -\infty$ and g(v)=v for $v \to +\infty$. In the ranges $(-\infty$,a) and (b,$+\infty$) the plot of g(v) is very close to a line; In the range (a , b) , the plot is a curve. In our experiment it proves that a quadratic polynomial is enough to approximate g(v) in the rang (a, b), and doesn't need a cubic or higher order polynomial. Therefore, we adopt a quadratic polynomial to approximate this curve.

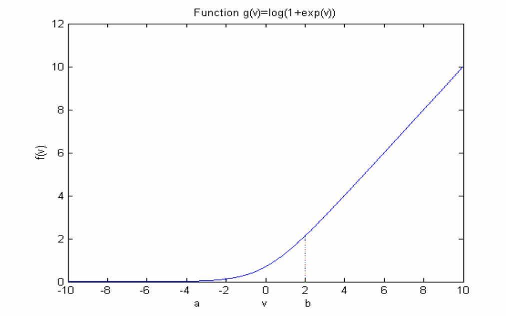

Fig.2 the function g(v)=log(1+exp(v))

2.3.2 The distribution of the range of v

We can get the distribution of v through the experiment of VTS. The statistical property of v is shown in Fig.3. The distribution function of v is like a single Gaussian very much.

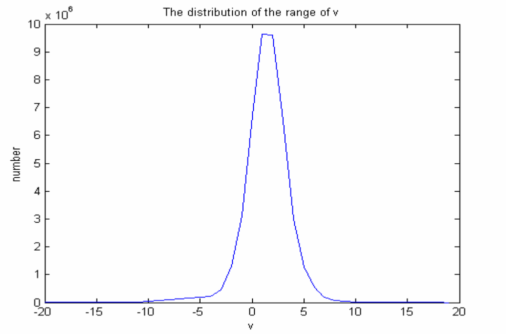

Fig.3. The distribution of the range of v

It is important to approximate the variable around the peak value. VPP takes this case into account.

2.3.3 The piecewise polynomial approximation

In our approach, linear approximation in the range of $(-\infty$,a) and (b,$+\infty$) is obtained by minimizing the mean squared error. The range of (a, b) is replaced by a quadratic Chebyshev Polynomial.

The Chebyshev approximation formula is defined below:

$$g(x) \approx \sum_{k=0}^{N-1} C_k T_k(x) - \frac{1}{2} C_0 \qquad (3)$$

$$C_j = \frac{2}{N} \sum_{k=1}^{N} g(x_k) T_j(x_k) \qquad (4)$$

A comparative plot of the actual function and our approximation is shown in Fig.4. As can be seen from this figure, the approximation can not be distinguished from the actual function.

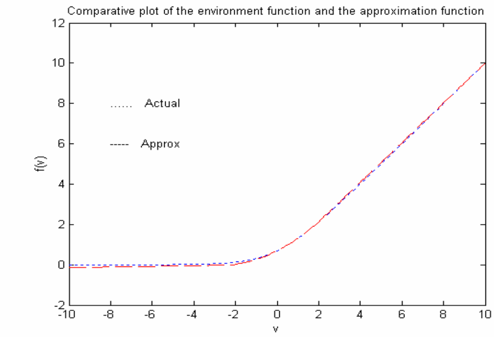

Fig.4 Comparative plot of the environment function and the approximation function.

3. VPP COMPENSATION FRAMEWORK

In [5], it was shown that it is reasonable for Gaussian densities for clean speech to be assumed to transform into Gaussian densities for noisy speech. This assumption implies a linear transformation between the log spectra of clean and noisy speech for each Gaussian density component of the PDF of clean speech. Therefore, to estimate the noise and channel parameters, we approximate the environment function (Eq.(2))by the statistical linear transformation[6].

$$\tilde{y} = (1 - A_k)x + (1 - A_k)h + A_k n + B_k \tag{5}$$

So the estimate of mean and variance of the K[th] Gaussian of the noisy speech can be got as functions of A_k and B_k:

$$m_{y,k} = (1 - A_k)(m_{x,k} + h) + A_k m_n + B_k \tag{6}$$

$$\Sigma^2{}_{y,k} = (1 - A_k)^2 \Sigma^2{}_{x,k} + A^2{}_k \Sigma^2{}_n \tag{7}$$

Since the estimate of $m_{y,k}$ and $\Sigma^2{}_{y,k}$ can be got from the VPP algorithm. From (8) and (9), we can compute A_k and B_k. Therefore, m_n, Σ_n and h can be derived by EM algorithm.

The optimal parameter is l = {m_n, Σ_n, h }. The auxiliary function, Q, is defined by

$$Q(l, \bar{l}) = E[\log p(X, N, K \mid \bar{l} = \{m_n, \Sigma_n, h\}) \mid y, l = \{m_n, \Sigma_n, h\}]$$

Where **N**={$\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_T$} denotes the noise vector sequence which is statistically independent of the clean feature vector sequence X, and K={$k_1, k_2, \ldots, k_T$} is a hidden sequence of mixture components. By following the way similar to that used in [7], we can obtain the formulas of the estimation of the environment parameters:

$$\hat{m}_n = \frac{\sum_{t=1}^{T}\sum_{k=1}^{K} P(k \mid y_t, l) m_n(y_t, k, l)}{\sum_{t=1}^{T}\sum_{k=1}^{K} P(k \mid y_t, l)}$$

$$\hat{\Sigma}_n^2 = \frac{\sum_{t=1}^{T}\sum_{k=1}^{K} P(k \mid y_t, l)[\Sigma_n(y_t, k, l) + m_n(y_t, k, l) m_n^T(y_t, k, l)}{\sum_{t=1}^{T}\sum_{k=1}^{K} P(k \mid y_t, l)}$$
$$- \hat{m}_n \hat{m}_n^T$$

$$\hat{h} = \sum_{t=1}^{T}\sum_{k=1}^{K} P(k \mid y_t, l)(I - A)^{-1} \Sigma_{x,k}^{-1} (y_t - (I - A) m_{x,k} - An - B)$$

Where

$$m_n(y_t, k, l) = \tilde{\Sigma}_n (\Sigma_n + \tilde{\Sigma}_n)^{-1} m_n + \Sigma_n (\Sigma_n + \tilde{\Sigma}_n)^{-1} \bar{m}_n$$

$$\Sigma_n(y_t, k, l) = [\Sigma_n^{-1} + \tilde{\Sigma}_n^{-1}]^{-1}$$

Where

$$\bar{m}_n = (I - A)^{-1}(y_t - (I - A) m_x - An - B)$$

$$\tilde{\Sigma}_n = (A^T)^{-1} (I - A)^T \Sigma_{x,k} (I - A) A^T$$

The environment compensation algorithm for data compensation proceeds as follows (B. Raj, 1996):

- (1) Get initial estimates for {m_n, Σ_n, h}.
- (2) Compute $m_{y,k}$ and $\Sigma^2{}_{y,k}$ using the estimates of { m_n, Σ_n, h} based on VPP approximation.
- (3) Obtain the values of A_k and B_k using the estimates of m_n, Σ_n, h, $m_{y,k}$ and $\Sigma^2{}_{y,k}$.
- (4) Perform a single iteration of the EM algorithm to re-estimate the values of {m_n, Σ_n, h}.
- (5) If the likelihood of the observed noisy data has not converged, return to Step 2.
- (6) Estimate the clean feature vector sequence based on the minimum mean squared error MMSE) criterion.

4. EXPERIMENTAL RESULTS

To evaluate the effectiveness of the proposed algorithm, we performed a series of experiments on telephone quality speaker-independent (SI) Mandarin speech recognition.

In order to obtain telephone quality speech materials for the acoustical model training, we utilize the Mandarin 863 speech database. This database was developed by Chinese national 863 program for LVCSR. It contains about 70 hours' speech. The speech data are 16000Hz sampled and 16bit linear quantized. We disposed this database with three methods: 1) resample the database to 8000Hz and u-law quantization; 2) pass the database through the real PSTN network by Dialogic telephone cards plugged in PCs; 3) pass the database through GSM full rate (GSM FR 06.10) coder and decoder. All these three transcoded databases are used as training data for the acoustical model.

The acoustic features consist of energy, pitch, 12 mel-cepstral with delta and delta-delta features. The vocabulary of this task consists of more than 40K words. Tri-gram statistics are used for language modeling.

There are three test sets in our experiments named as TELTEST, GSMTEST, SIMUTEST respectively. TELTEST is gathered through the PSTN network, GSMTEST is gathered through GSM-FR codec and SIMUTEST is derived from artificially contamination. Every contains about 240 continuous Mandarin sentences from 4 different speakers.

We compare our VPP algorithm with other four widely used compensation algorithms.

1. Long term CMS(Cepstral Mean Subtraction over the duration of a whole Mandarin sentence).
2. RASTA filtering [8].
3. SBR (Signal Bias Removal) [9].
4. VTS [3].

The performance is measured by character error rate (CER) using NIST's SCLITE.

In Tab.1, we present the results from the experiment using speaker-independent (SI) Mandarin speech recognition in telecommunication environment.

	CMS	SBR	RASTA	VTS	VPP
TELTEST	76.1	79.4	75.1	78.3	80.7
GSMTEST	75.6	76.3	73.7	79.8	81.3

Tab.1. The results on telephone quality speaker-independent (SI) Mandarin speech recognition

Fig.6 shows performance of VPP algorithm at several SNR's on telephone quality speaker-independent (SI) Mandarin speech recognition.

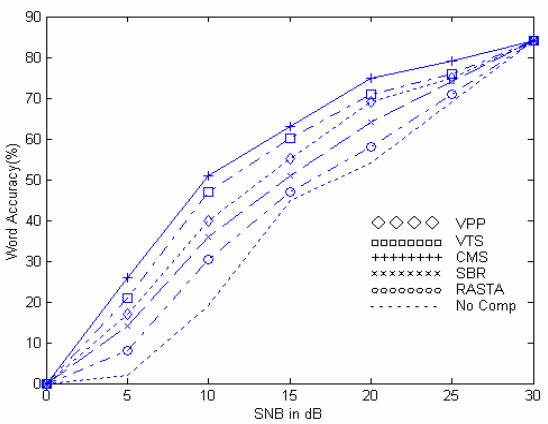

Fig.6 Performance of VPP algorithm at several SNR's

5. CONCLUSION

In this paper, we have presented a novel approximation –based approach to compensate for the effects of linear filtering and background noise given the parameters of the PDF of clean speech. The proposed experimental results shown that the approximation of environment function using the VPP algorithm is more precise than VTS. In the real telephone network, the algorithm presented here provides significant improvement over the previous work in data compensation. Compared to other channel compensation approaches, VPP shows more adaptability to the variations introduced by channel and noise, then obtains about 12% character error rate decreasing in Mandarin telephone speech recognition.

6. ACKNOWLEDGEMENTS

The research work described in this paper is supported by the National Nature Science Foundation of China under grant number 69835003 and the National Science Foundation of China under grand number 60175012.

7. REFERENCES

[1] A.Sankar and C.-H..Lee. " A maximum likelihood approach to stochastic matching for robust speech recognition", IEEE Trans. Speech Audio Processing, vol.4, pp.190-202, May 1996.

[2] C.-H.Lee."On stochastic feature and model compensation approaches to robust speech recognition," Speech Communication, vol.25, pp. 29-47, Aug, 1998.

[3] P.J.Moreno. "Speech Recognition in Noisy Environments", PhD thesis, Carnegie Mellon University, Pittsburgh, Pensilvania, April 1996.

[4] J.C.Segura,M.C.Benitez,A.de la Torre, S.Dupont, A.J. Rubio."VTS residual noise compensation" *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol.1, pp. 409-412,May, 2002.*

[5] P.J.Moreno, B.Raj, E.B.Gouvea,, and R.M.Stern."Multivariate Gaussian Based Cepstral Normalization for Robust Speech Recognition". *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1995, pp. 137-140*.

[6] B. Raj, E. Gouvea, and R. M. Stern."Cepstral Compensation using Statistical Linearization," Proc. of the ESCA Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-au-Mousson, France, April, 1997.

[7] D. Y. Kim, C. K. Un, N. S. Kim, "Speech recognition in noisy environments using first-order vector Taylor series," *Speech Communication*, Vol. 24, No. 1, pp. 39-49, 1998

[8] H.Hermansky, N. Morgan, RASTA Processing of Speech, IEEE Trans on Speech and Audio Processing, vol.2, No..4 pp 578-589, Oct, 1994.

[9] Mazin G.Rahim and Biing-Hwang Juang, "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition", IEEE Trans. On Speech and Audio Processing, Vol.4, No.1, pp.19-30, Jan, 1996.