MODELING VARIANCE VARIATION IN A VARIABLE PARAMETER HMM FRAMEWORK FOR NOISE ROBUST SPEECH RECOGNITION

Xiaodong Cui[1] *and Yifan Gong*[2]

Department of Electrical Engineering
University of California, Los Angeles, CA 90095[1]
Microsoft Corporation
One Microsoft Way, Redmond, WA 98052[2]
Emails: xdcui@icsl.ucla.edu, ygong@microsoft.com

ABSTRACT

Variance variation with respect to a continuous environment-dependent variable is investigated in this paper in a variable parameter Gaussian mixture HMM (VP-GMHMM) for noisy speech recognition. The variation is modeled by a scaling polynomial applied to the variances in the conventional hidden Markov acoustic models. The maximum likelihood estimation of the scaling polynomial is performed under an SNR quantization approximation. Experiments on the Aurora 2 database show significant improvements by incorporating the variance scaling scheme into the previous VP-GMHMM where only mean variation is considered.

1. INTRODUCTION

In noisy speech recognition, it is well known that the mean and variance of a Gaussian mixture HMM (GMHMM) vary with the environment. This can be clearly observed in Fig.1 where the mean (upper panel) and variance (lower panel) of the Gaussian distribution trained with MFCC features vary over signal-to-noise ratio (SNR). Hence, the acoustic models have to be adapted to a particular environment in order to achieve good performance [1][2].

In [3], a variable parameter GMHMM (VP-GMHMM) is investigated where the variation of the mean of Gaussian is modeled by a polynomial over a continuous environment-dependent variable and the variance remains constant. Significant improvements have been reported in literature such as [4] and [5] when variance variation is considered in noise robust speech recognition. In this paper, the variation of the variance of Gaussian is modeled in the previous VP-GMHMM framework. Therefore, instead of a constant Gaussian variance trained and applied, both mean and variance change with the environment in the VP-GMHMM discussed in this paper.

The remaining of the paper is organized as follows. In Section 2, a polynomial variance scaling approach is introduced and the maximum likelihood (ML) estimation of the polynomial by an SNR quantization scheme is described in Section 3. Experimental results are shown in Section 4 and finally a summary is given in Section 5.

This work was started when Y. Gong was with Texas Instruments.

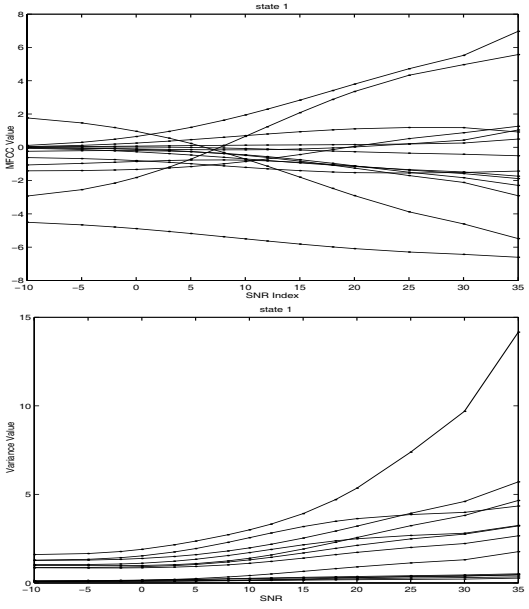

Fig. 1. Variation of the mean (upper panel) and variance (lower panel) ($C_1, \cdots, C_7; \triangle C_1, \cdots, \triangle C_7$) against SNR for the first state of the /ah/ sound of female speakers with MFCC features.

2. POLYNOMIAL VARIANCE SCALING IN VP-GMHMM

In a conventional Gaussian mixture HMM (CV-GMHMM) each state has a multivariate Gaussian mixture distribution:

$$p(\mathbf{o}_t|s_t = i) = \sum_k \alpha_{ik} b_{ik}(\mathbf{o}_t) \quad (1)$$

where $b_{ik}(\mathbf{o}_t) \sim \mathcal{N}(\mathbf{o}_t; \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik})$ is the kth multivariate Gaussian mixture in state i with weight α_{ik}, and $\boldsymbol{\mu}_{ik}$ and $\boldsymbol{\Sigma}_{ik}$ are the mean and covariance associated with it. Typically, $\boldsymbol{\Sigma}_{ik}$ is assumed diagonal.

In the VP-GMHMM discussed in [3], the mean vector $\boldsymbol{\mu}_{ik}$ in Eq. 1 is described as a P-th order polynomial function of a environment-dependent scalar variable v:

$$\boldsymbol{\mu}_{ik}(v) \quad = \quad \sum_{j=0}^{P} \mathbf{c}_{ikj}^{m} v^j$$

where $\mathbf{c}_{ikj}^{m} = [c_{ik0j}^{m}, \cdots, c_{ikDj}^{m}]^T$ are the coefficients of the mean polynomial and D is the feature dimension.

In this paper, the variance variation is taken into account by a polynomial scaling factor applied to the original variances which can be written as:

$$\boldsymbol{\Sigma}_{ik}(v) \quad = \quad \boldsymbol{\Lambda}(v)\boldsymbol{\Sigma}_{ik}^{0} \tag{2}$$

where $\boldsymbol{\Sigma}_{ik}^{0}$ is the original diagonal covariance matrix and the scaling matrix $\boldsymbol{\Lambda}$ is a diagonal matrix with scaling factors along the principal diagonal:

$$\begin{bmatrix} e^{\sum_{j=0}^{P} c_{ik0j}^{v} v^j} & & \\ & \ddots & \\ & & e^{\sum_{j=0}^{P} c_{ikDj}^{v} v^j} \end{bmatrix}$$

In Eq. 2, the scaling factor is a polynomial of the environment v and the exponential form guarantees the positiveness of the variance. One of the advantages of employing a variance scaling polynomial rather than modeling the variance itself by a polynomial is that the variance scaling polynomial gives flexibility of tying at various level of granularity.

3. ESTIMATION OF VARIANCE SCALING POLYNOMIAL

If no variance variation is considered, as in [3], a constant variance is trained in the VP-GMHMM. In this case, the covariance matrix, instead of the scaling matrix, is estimated and we can readily have the re-estimation formula as:

$$\boldsymbol{\Sigma}_{ik} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} \gamma_t^r(i,k)(\mathbf{o}_t^r - \sum_{j=0}^{P} \mathbf{c}_{ikj}^{m} v_r^j)(\mathbf{o}_t^r - \sum_{j=0}^{P} \mathbf{c}_{ikj}^{m} v_r^j)^T}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} \gamma_t^r(i,k)} \tag{3}$$

where $\gamma_t^r(i,k) = p(s_t^r = i, \xi_t^r = k | \mathbf{O}^r, \overline{\lambda})$ is the probability of being in state i mixture component k at time t given the r-th utterance $\mathbf{o}_t^r$ and previous model parameters $\overline{\lambda}$. R is the number of utterances and T_r is the number of frames in the r-th utterance.

If the scaling matrix $\boldsymbol{\Lambda}$ is applied to the original covariance matrix, the ML estimation is performed over the diagonal elements of $\boldsymbol{\Lambda}$. Since diagonal covariance matrices are employed in this paper, the multivariate Gaussian PDF is simply the product of the univariate Gaussian PDF of individual dimensions. For the simplicity of mathematical derivation, the estimation of the scaling matrix is carried out accordingly dimension by dimension.

Rewrite the Gaussian mixture PDF as

$$b_{ik}(\mathbf{o}_t^r) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\sigma_{ikd}^2 e^{\sum_{j=0}^{P} c_{ikdj}^{v} v_r^j}}} e^{-\frac{(o_{td}^r - \mu_{ikd}(v_r))^2}{2\cdot\sigma_{ikd}^2 e^{\sum_{j=0}^{P} c_{ikdj}^{v} v_r^j}}}$$

The ML estimation of the coefficients of the variance scaling polynomial can be performed using the EM algorithm [6] by defining the auxiliary Q-function as

$$\begin{aligned} Q_b(\lambda;\overline{\lambda}) \quad &= \quad \sum_{r=1}^{R}\sum_{i\in\boldsymbol{\Omega}_s}\sum_{k\in\boldsymbol{\Omega}_m}\sum_{t=1}^{T_r} \gamma_t^r(i,k)\cdot\log b_{ik}(\mathbf{o}_t^r) \\ &= \quad \sum_{r=1}^{R}\sum_{i\in\boldsymbol{\Omega}_s}\sum_{k\in\boldsymbol{\Omega}_m}\sum_{t=1}^{T_r} \gamma_t^r(i,k)\cdot \\ &\quad \sum_{d=1}^{D}[\log\frac{1}{\sqrt{2\pi}} + \log\frac{1}{\sigma_{ikd}^2} + \log\frac{1}{\sqrt{e^{\sum_{j=0}^{P} c_{ikdj}^{v} v_r^j}}} \\ &\quad -\frac{(o_{td}^r - \mu_{ikd}(v_r))^2}{2\cdot\sigma_{ikd}^2\cdot e^{\sum_{j=0}^{P} c_{ikdj}^{v} v_r^j}}] \end{aligned}$$

Taking the derivative of $Q_b(\lambda;\overline{\lambda})$ over c_{ikdj}^{v} and setting to zero, we get:

$$\begin{aligned} &\sum_{r=1}^{R} e^{-\sum_{p=0}^{P} c_{ikdp}^{v} v_r^p}\cdot v_r^j \sum_{t=1}^{T_r}\gamma_t^r(i,k)\cdot\frac{(o_{td}^r - \mu_{ikd}(v_r))^2}{\sigma_{ikd}^2} \\ = \quad &\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_t^r(i,k)\cdot v_r^j, \qquad j = 0,1,\cdots,P \end{aligned}$$

which can be written as the following $P+1$ simultaneous nonlinear equations:

$$\begin{cases} \mathbf{F}_0(c_{ikd0}^{v},\cdots,c_{ikdP}^{v}) = 0 \\ \quad\vdots \\ \mathbf{F}_P(c_{ikd0}^{v},\cdots,c_{ikdP}^{v}) = 0 \end{cases} \tag{4}$$

Theoretically, root finding algorithms can be applied to Eq. 4 to solve the $P+1$ simultaneous nonlinear equations for the $P+1$ unknowns $\{c_{ikd0}^{v},\cdots,c_{ikdP}^{v}\}$, and obtain the re-estimation of the P-th order variance scaling polynomial for the d-th dimension. Practically, it is difficult to find a good solution since the number of summation terms in $\mathbf{F}_j$ depends on the number of utterances participated in the training and the algorithms are very easy to get stuck at local minima. Therefore, certain approximation has to be considered to make the solution feasible.

To reduce the number of summation terms, the SNR range is quantized into $P+1$ values,

$$(\upsilon_j, \upsilon_{j+1}] \mapsto \upsilon_j^*, \qquad j = 0, \cdots, P.$$

where $\upsilon_0 = -\infty$ and $\upsilon_{P+1} = +\infty$. This quantization is based on the assumption that there is no significant difference among the variances within a reasonable SNR interval.

The quantization converts the $P+1$ simultaneous nonlinear equations in Eq. 4 into $P+1$ simultaneous linear equations:

$$\boldsymbol{\Psi}_{ikd} \cdot \boldsymbol{\kappa}_{ikd} = \boldsymbol{\Phi}_{ikd} \tag{5}$$

where

$$\boldsymbol{\Psi}_{ikd} = \begin{bmatrix} \boldsymbol{\Psi}_{ikd}(0,0) & \cdots & \boldsymbol{\Psi}_{ikd}(0,P) \\ \vdots & \boldsymbol{\Psi}_{ikd}(j,p) & \vdots \\ \boldsymbol{\Psi}_{ikd}(P,0) & \cdots & \boldsymbol{\Psi}_{ikd}(P,P) \end{bmatrix}$$

with

$$\boldsymbol{\Psi}_{ikd}(j,q) \triangleq \sum_{\upsilon_r \in [\upsilon_q, \upsilon_{q+1})} \sum_{t=1}^{T_r} \gamma_t^r(i,k) \cdot \frac{(o_{td}^r - \mu_{ikd}(\upsilon_q^*))^2}{\sigma_{ikd}^2} \cdot \upsilon_q^{*j};$$

and

$$\boldsymbol{\Phi}_{ikd} = [\boldsymbol{\Phi}_{ikd}(0), \cdots, \boldsymbol{\Phi}_{ikd}(P)]^T$$

with

$$\boldsymbol{\Phi}_{ikd}(q) \triangleq \sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_t^r(i,k) \cdot \upsilon_q^{*j};$$

and

$$\boldsymbol{\kappa}_{ikd} = [\boldsymbol{\kappa}_{ikd}(0), \cdots, \boldsymbol{\kappa}_{ikd}(P)]^T$$

with

$$\boldsymbol{\kappa}_{ikd}(q) \triangleq e^{-\sum_{p=0}^{P} c_{ikdp}^{v} \upsilon_q^{*p}}$$

From Eq. 5, $\boldsymbol{\kappa}_{ikd}$ can be obtained as:

$$\boldsymbol{\kappa}_{ikd} = \Psi_{ikd}^{-1} \cdot \Phi_{idk}$$

After $\boldsymbol{\kappa}_{ikd}$ are available, c_{ikdp} can be readily resolved from another linear system equation by taking logarithm.

4. EXPERIMENTAL RESULTS

Experiments are performed on the Set A of the Aurora 2 database. There are four types of noise in the training and test data of Set A which include subway, babble, car and exhibition noise. For each type of noise, training data are recorded under five SNR conditions: clean, 20 dB, 15 dB, 10 dB and 5 dB while test data consist of six SNR conditions: clean, 20 dB, 15 dB, 10 dB, 5 dB and 0 dB. There are 8440 utterances in total for the four types of noise contributed by 55 male speaker and 55 female speakers. For the test set, each SNR condition of each noise type consists of 1001 utterances and 24024 in total from 52 male speakers and 52 female speakers.

Mel-Frequency Cepstral Coefficients (MFCC) features are used to train the acoustic models. The frame length is 25 ms and the frame shift is 10 ms. The speech feature for each frame contains 12 static MFCCs (excluding C0) plus log energy (E) and their first and second order derivatives. Therefore, there are 39 components in each feature vector.

The HMMs adopt a left-to-right topology and are word-based models with 16 emission states for each digit, 3 states for the silence model and 1 state for the short pause model. We investigate cases in which each state has one or two Gaussian mixtures. All the Gaussian mixtures have diagonal covariance matrices.

Utterance SNR is chosen as the environmental variable υ and it is estimated by the minimum statistics tracking algorithm proposed in [7]. Both mean and variance scaling polynomials are chosen to be 2nd-order polynomials and the SNR range is quantized into three levels:

$$\begin{aligned} (-\infty,\ 10] &\mapsto 5 \\ (10,\ 25] &\mapsto 15 \\ (25, +\infty) &\mapsto 35 \end{aligned}$$

In the training stage, training data are categorized into several subsets in terms of their SNRs (e.g. clean, 20dB, 10 dB, etc.) and CV-GMHMMs are trained for each subset. The Gaussian means of the same mixture from different subset CV-GMHMMs are regressed with respect to SNR to obtain the initial mean polynomials. The mean polynomials are first estimated without variance scaling and after 25 EM iterations mean and variance scaling polynomials are joint estimated for another 2-3 iterations. In the recognition stage, utterance SNR is estimated for the speech signal and one set of environment-dependent HMM parameters (Gaussian mixture means and variances) is instantiated based on the SNR estimate to decode the speech signal.

Fig. 2 shows the estimated mean (upper panel) and variance scaling (lower panel) polynomials of the energy component from the first state of digit "oh" for the four types of background noise from Set A. The HMMs used in the plot have one mixture per state. It can be observed from the figure that the polynomials capture the trend of mean and variance with respect to SNR.

To compare VP-GMHMM and CV-GMHMM, acoustic HMMs with one and two Gaussian mixtures in each state are used for the experiments for the four types of background noise. Table. 1 shows the recognition accuracy averaged across the four types of noise. The variance scaling polynomials are state-tied. In the table, VP-GMHMM1 stands for the models that only mean polynomials are applied (Eq. 3 is used for variance estimation) while VP-GMHMM2 for the models in which both mean and variance scaling polynomials are used. It can be observed from the table that, both VP-GMHMM1 and VP-GMHMM2 outperform CV-GMHMM in all the test conditions. By modeling the variance variation,

	Experimental Conditions						
	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	Ave.
1M (CV-GMHMM)	96.46	96.21	95.38	92.36	83.28	53.18	86.14
1M (VP-GMHMM1)	97.91	97.23	96.29	93.23	85.95	61.77	88.73
1M (VP-GMHMM2)	98.12	97.55	96.86	94.48	88.45	70.38	90.97
2M (CV-GMHMM)	97.89	97.25	96.68	94.81	88.51	64.23	89.89
2M (VP-GMHMM1)	98.65	97.94	97.49	95.63	89.87	70.97	91.76
2M (VP-GMHMM2)	98.65	98.12	97.78	95.97	90.81	73.31	92.44

Table 1. Average recognition accuracy on four types of noise in set A of the Aurora 2 database with one Gaussian mixture (1M) and two Gaussian mixtures (2M) in each state.

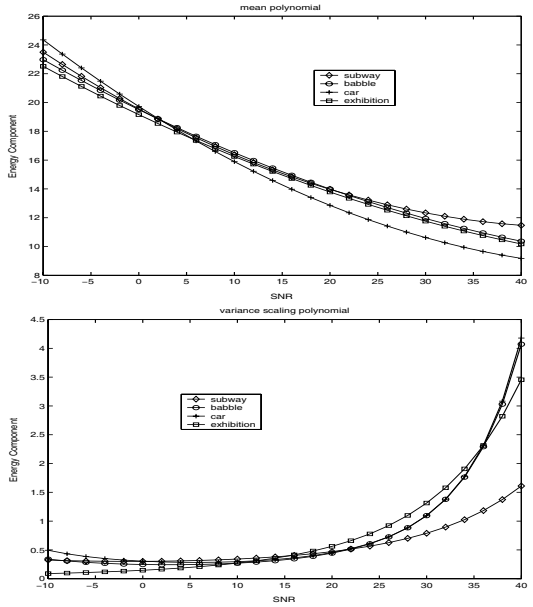

Fig. 2. Estimated mean (upper panel) and variance scaling (lower panel) polynomials of the energy component from the first state of digit "oh" for the four types of background noise.

VP-GMHMM2 yields better performance than VP-GMHMM1, especially in the low SNR conditions.

5. SUMMARY AND CONCLUSIONS

In this paper, variance variation with respect to an continuous environment-dependent variable is investigated in a VP-GMHMM framework. The variation is described by a scaling polynomial applied to the variances in the original acoustic models. The maximum likelihood estimation of the scaling polynomial is given under an SNR quantization approximation scheme. By considering both mean and variance variations, VP-GMHMM yields significant improvements in the experiments performed in the Aurora 2 database.

6. REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: a survey," *Speech Communication*, vol. 16, pp. 261–291, 1995.

[2] M. Gales and S. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 5, pp. 352–359, 1996.

[3] X. Cui and Y. Gong, "Variable parameter Gaussian mixture hidden Markov modeling for speech recognition," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, pp. 12–15, 2003.

[4] M. Gales, D. Pye, and P.C. Woodland, "Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation," *Proc. of Int Conf. on Spoken Language Processing*, 1996.

[5] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of hmm variances using the feature enhancement uncertainty computed from a parametric model of speech distorti," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, pp. 412–421, 2005.

[6] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[7] R. Martin, "An efficient algorithm to estimate instantanous SNR of speech signals," *Proc. of European Conf. on Speech Communication and Technology*, pp. 1093–1096, 1993.