VARIABLE PARAMETER GAUSSIAN MIXTURE HIDDEN MARKOV MODELING FOR SPEECH RECOGNITION

*Xiaodong Cui**

Department of Electrical Engineering
University of California, Los Angeles, CA
xdcui@ucla.edu

Yifan Gong

Speech Technologies Laboratory
Texas Instruments Inc., Dallas, TX
Yifan.Gong@ti.com

ABSTRACT

To improve recognition, speech signals corrupted by a variety of noises can be used in speech model training. Published hidden Markov modeling of speech uses multiple Gaussian distributions to cover the spread of the speech distribution caused by the noises, which distracts the modeling of speech event itself and and possibly sacrifices the performance on clean speech. We extend GMHMM by allowing state emission parameters to change as function of an environment-dependent continuous variable. At the recognition time, a set of HMMS specific to the given the environment is instantiated and used for recognition. Variable parameter (VP) HMM with parameters modeled as a polynomial function of the environment variable is developed. Parameter estimation based on EM-algorithm is given. With the same number of mixtures, VPHMM reduces WER by 40% compared to conventional multi-condition training.

1. INTRODUCTION

Speech recognition in a noisy environment using hidden Markov models requires modeling speech distributions in the given environment, otherwise severe performance degradation may occur [1]. Approaches of such a modeling include using noisy speech during the training phase [2, 3, 4, 5, 6, 7] which can be generalized to multi-condition training [8, 9] in which available speech data collected in a variety of environments are used in model training.

Published Gaussian mixture hidden Markov modeling of speech uses multiple Gaussian distributions to cover the spread of the speech distribution caused by the noises. Three problems with this approach can be mentioned:
/1/ Since no noise model is incorporated and since the recognition accuracy is only optimized to the intensity and characteristics of the training noises, recognition performance could be sensitive to noise levels [10].
/2/ At the recognition time, a speech signal can only be produced in a particular environment. However, for a given noisy environment, the distributions of all conditions in the training phase are all open to the search space. The variety of noisy speech distributions decreases model discriminability. Therefore, the improvement of recognition robustness to noisy speech is obtained at a cost: sacrificing the recognition rate of any given environment including the clean one.
/3/ Since is difficult to collect all varieties of noisy data to cover all possible types of noises at all SNRs, the performance on unseen noises remains unpredictable.

We propose an extension to the conventional HMM, referred to as variable parameter Gaussian mixture HMM (VPHMM). VPHMM allows HMM parameters to change as function of a continuous variable that depends on the environment. At the recognition time, a set of HMMs specific to the given the environment is instantiated. Speech recognition is therefore based on the environment-specific models, instead of on the models with distributions averaged over all training environments.

Modeling multi-dimensional Gaussians distributions with noise-dependent mean and covariance has been used in the past in speech recognition, for instance in utterance rejection in noisy environments [11, 12].

In this paper, we develop variable parameter Gaussian mixture HMM where parameters are modeled as a polynomial function of the environment variable. ML parameter estimation based on EM-algorithm is given. We give the solution for the mean vectors of state emission PDF.

2. VARIABLE PARAMETER HMM

Fig-1 plots the variation of mean vector components of an HMM state as function of SNR change. The plot is

*The work was performed while the author was a summer student intern at Texas Instruments.

obtained by computing the mean of uniformly segmented states for a given phoneme under specific SNRs. The plot

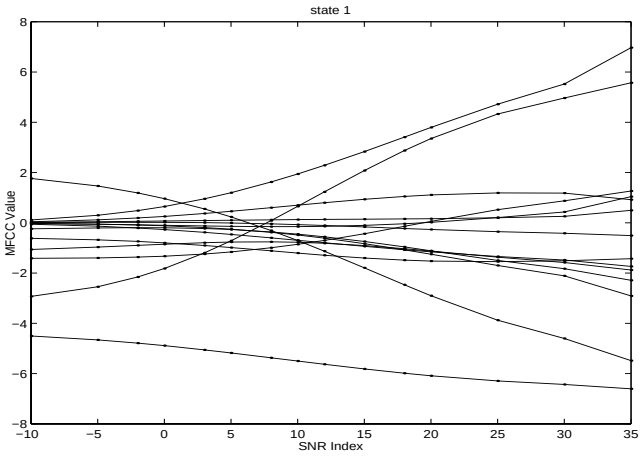

Fig. 1. Variation of MFCC mean components $(C_1, \cdots, C_7; \triangle C_1, \cdots, \triangle C_7)$ vs. SNR for the first state in /ah-1/ of female speakers

clearly shows that, for each dimension, the distribution of the observed feature value is a function of SNR. Pooling such distributions together and train SNR independent models, as in multi-condition training, inevitably yields flat distributions. Our task is to model such variation under the GMHMM framework. We will focus on the state emission PDFs. Other HMM parameters, including transition probabilities, can be solved by following the same procedure.

Let N be the number of HMM states and M be the mixture number. Let $\Omega_s \triangleq \{1, 2, \ldots N\}$ be the set of states, and $\Omega_m \triangleq \{1, 2, \ldots M\}$ be the set of mixture indicators. For an observed speech sequence of T vectors: $\mathbf{O} \triangleq \mathbf{o}_1^T \triangleq (\mathbf{o}_1, \mathbf{o}_2, \ldots \mathbf{o}_T)$, we introduce state sequence

$$\Theta \triangleq (\theta_0, \ldots \theta_T) \text{ where } \theta_t \in \Omega_s$$

and mixture indicator sequence

$$\Xi \triangleq (\xi_1, \ldots \xi_T) \text{ where } \xi_t \in \Omega_m.$$

We assume that, at an HMM state i, the emission probability density function is a multi-variate Gaussian mixture distribution with parameters depending on ν, a scalar representing the environment [13].

$$\begin{aligned} p(\mathbf{o}_t|\theta_t = i; \nu) &= \sum_k \alpha_{i,k}(\nu) b_{i,k}(\mathbf{o}_t) \qquad (1) \\ &= \sum_k \alpha_{i,k}(\nu) N(\mathbf{o}_t; \mu_{i,k}(\nu), \mathbf{\Sigma}_{i,k}(\nu)) \end{aligned}$$

where $\mu_{i,k}(\nu)$ is the mean vector of the mixing component k at the state j, $\mathbf{\Sigma}_{i,k}(\nu)$ is the covariance matrix of the mixing component k at the state i, $\alpha_{i,k}(\nu) \triangleq Pr(\xi_t = k|\theta_t = i; \nu)$ is the a priori probability of component k at the state i, and $\sum_k \alpha_{i,k}(\nu) = 1$ Notice that $\mu_{i,k}(\nu)$, $\mathbf{\Sigma}_{i,k}(\nu)$ and $\alpha_{i,k}(\nu)$ are all function of environment ν. However, in the next only $\mu_{i,k}(\nu)$ will be elaborated.

3. ML PARAMETER ESTIMATION

3.1. Criterion

We use maximum likelihood criterion to find model parameters, which can be solved by EM algorithm [14]. Two kinds of data are involved in EM: observable $\mathbf{X}$ and non-observable $\mathbf{Y}$. The EM algorithm maximizes, w.r.t. the new parameter set λ, the mathematical expectation of the log-likelihood of $\{\mathbf{X}, \mathbf{Y}\}$, conditioned on the observed data $\mathbf{X}$, and for a value $\bar{\lambda} \in \Lambda$ of the parameter. The expectation is taken over the sample space of the unobservable data Y:

$$\mathcal{Q}(\lambda|\bar{\lambda}) \triangleq \mathbf{E}_{\mathbf{Y}} \left\{ \log p(\mathbf{X}, \mathbf{Y}|\lambda) | \mathbf{X}, \bar{\lambda} \right\} \qquad (2)$$

3.2. Formulation

We will use x^r to denote any variables or functions instantiated with the r-th observation sequence. Suppose we have R observation sequences: $\mathbf{X} = (\mathbf{O}^1, \mathbf{O}^2, \ldots \mathbf{O}^R)$ with corresponding unobservable variables: $\mathbf{Y} = (\Theta^1, \Xi^1, \Theta^2, \Xi^2, \ldots \Theta^R, \Xi^R)$. We have:

$$p(\mathbf{X}, \mathbf{Y}|\lambda) = \prod_{r=1}^{R} p(\mathbf{O}^r, \Theta^r, \Xi^r|\lambda) \qquad (3)$$

Using Eq-3, Eq-2 can be instantiated as

$$\begin{aligned} \mathcal{Q}(\lambda|\bar{\lambda}) =& \sum_{r=1}^{R} \sum_{\Theta^r} \sum_{\Xi^r} p(\Theta^r, \Xi^r|\mathbf{O}^r, \bar{\lambda}) \log p(\mathbf{O}^r, \Theta^r, \Xi^r|\lambda) \\ =& \mathcal{Q}_u(\lambda|\bar{\lambda}) + \mathcal{Q}_a(\lambda|\bar{\lambda}) + \mathcal{Q}_c(\lambda|\bar{\lambda}) + \mathcal{Q}_b(\lambda|\bar{\lambda}) \end{aligned}$$

$\mathcal{Q}_u$, $\mathcal{Q}_a$ and $\mathcal{Q}_c$, which share the same mathematical structure, are respectively function of HMM initial state occupancy probabilities, HMM transition probabilities, and mixing probabilities. The maximization of the related equations are similar and can be solved by following [16]. $\mathcal{Q}_c$ is a function of state observation parameters:

$$\begin{aligned} \mathcal{Q}_b(\lambda|\bar{\lambda}) \triangleq& \sum_{r=1}^{R} \sum_{j \in \Omega_s} \sum_{k \in \Omega_m} \sum_{t=1}^{T^r} p(\theta_t^r = j, \xi_t^r = k|\mathbf{O}^r, \bar{\lambda}) \cdot \qquad (4) \\ & \log b_{jk}(\mathbf{o}_t^r) \end{aligned}$$

The maximization of $\mathcal{Q}_b(\lambda|\bar{\lambda})$ for VPHMM requires solution of linear systems and will be presented in section 3.3.

3.3. Solution to observation parameters

We derive the parameters involved in maximizing the quantity $Q_b(\lambda|\bar{\lambda})$ as defined by Eq-4. As introduced in Eq-1, $b_{jk}(\mathbf{o}_t)$ is the observation pdf of $\mathbf{o}_t$ at state i, assuming mixture component k. We now specify the form of dependence of mean vector on the environment ν. Many continuous function of ν can be used to model such dependences. Polynomial function of ν is exploited here, for three reasons: /1/ With higher enough degree, polynomial approximation can be arbitrarily close to any continuous function. /2/ The derivatives of polynomial function are easy to obtain, which makes it attractive in parameter estimation. /3/ As shown by Fig-1, the dependence is smooth and can be modeled by low order polynomials.

We assume that the observation mean vector is a polynomial function of environment ν:

$$\mu_{i,k}(\nu) \triangleq \sum_j \mathbf{c}_{i,k,j}\nu^j \tag{5}$$

With Eq-5, equating the partial derivative of Eq-4 w.r.t. $\mathbf{c}_{i,k,j}$ to zero, we have, after some arrangement [13]:

$$\sum_{p=0}^{P_k^i}\sum_{r=1}^{R}\sum_{t=1}^{T^r} p(\theta_t^r = i, \xi_t^r = k|\mathbf{O}^r, \bar{\lambda})(\mathbf{\Sigma}_{i,k}(\nu_r))^{-1}\nu_r^{p+j}\mathbf{c}_{i,k,p}$$
$$= \sum_{r=1}^{R}\sum_{t=1}^{T^r} p(\theta_t^r = i, \xi_t^r = k|\mathbf{O}^r, \bar{\lambda})(\mathbf{\Sigma}_{i,k}(\nu_r))^{-1}\nu_r^j \cdot \mathbf{o}_t^r \tag{6}$$

Let, for each state i and each mixing component k,

$$\mathbf{l}_{i,k}(\zeta, \eta, \alpha, \beta) \triangleq \tag{7}$$
$$\sum_{r=1}^{R}\sum_{t=1}^{T^r} p(\theta_t^r = i, \xi_t^r = k|\mathbf{O}^r, \bar{\lambda})(\mathbf{\Sigma}_{i,k}(\nu_r))^{-1}\zeta^\alpha\eta^\beta$$

Eq-6 can be written as

$$\sum_{p=0}^{P_k^i}\mathbf{l}_{i,k}(\nu_r, \nu_r, p, j)\cdot\mathbf{c}_{i,k,p} = \mathbf{l}_{i,k}(\nu_r, \mathbf{o}_t^r, j, 1) \tag{8}$$

Eq-8 describes, for each state i and mixture component k, a linear equation system with $P_{i,k}+1$ variables defined in the vector space R^D, which can be written in a compact form:

$$\mathbf{A}_{i,k}\mathbf{c}_{i,k} = \mathbf{b}_{i,k} \tag{9}$$

Where $\mathbf{A}_{i,k}$ is a $(P_{i,k}+1)\times(P_{i,k}+1)$ dimension matrix:

$$\mathbf{A}_{i,k} \triangleq \begin{bmatrix} \mathbf{u}_{i,k}(0,0) & \dots & \mathbf{u}_{i,k}(0,P_{i,k}) \\ \vdots & \mathbf{u}_{i,k}(p,j) & \vdots \\ \mathbf{u}_{i,k}(P_{i,k},0) & \dots & \mathbf{u}_{i,k}(P_{i,k},P_{i,k}) \end{bmatrix}$$

where $\mathbf{u}_{i,k}(p,j)$ is itself a D by D matrix:

$$\mathbf{u}_{i,k}(p,j) \triangleq \mathbf{l}_{i,k}(\nu_r, \nu_r, p, j) \tag{10}$$

$\mathbf{b}_{i,k}$ is a $(P_{i,k}+1)$ dimensional vector:

$$\mathbf{b}_{i,k} \triangleq [\mathbf{v}_{i,k}(0), \dots \mathbf{v}_{i,k}(j), \dots \mathbf{v}_{i,k}(P_{i,k})]^T$$

where $\mathbf{v}_{i,k}(j)$ itself is a D-dimensional vector:

$$\mathbf{v}_{i,k}(j) \triangleq \mathbf{l}_{i,k}(\nu_r, \mathbf{o}_t^r, j, 1) \tag{11}$$

and $\mathbf{c}_{i,k}$ a $(P_{i,k}+1)$ dimensional vector:

$$\mathbf{c}_{i,k} \triangleq [\mathbf{c}_{i,k}(0), \dots \mathbf{c}_{i,k}(j), \dots \mathbf{c}_{i,k}(P_{i,k})]^T$$

$\mathbf{c}_{i,k,j}$ can be obtained by any adequate linear system solution method. In solving 9, the value of $(\mathbf{\Sigma}_{i,k}(\nu_r))^{-1}$ can be substituted by the one in the $\bar{\lambda}$ set, since such a substitution will still guarantee the increase of Eq-4 as required by EM procedure. Finally, for diagonal covariance matrix case, the computation is substantially simpler [13, 11].

4. EXPERIMENTAL RESULTS

The goal of the experimentation reported in this paper is to verify, on the same noisy training and testing data sets, if VPHMM can achieve lower recognition error rates than conventional multi-condition training.

Speech database used in the experiments is TIDIGITS database, corrupted digitally with highway car noise recorded with a hand free microphone. Training data contains 8603 utterances. The SNR ranges from 40 dB to -5 dB, in an uniform distribution. Testing database contains 8507 utterances, not used as training data. The testing SNR is either 30, 25, 20, 15, 10, or 5 dB, with equal probabilities.

In the experiments, the environment variable ν (Eq-5) measures the SNR of an utterance. Second order of polynomials are used. At the recognition time, a set of HMMS specific to the given the environment is instantiated according to Eq-5 and used for recognition.

Fig-2 compares conventional HMM (CVHMM) with VP-HMM, both with a single mixture per HMM state. The two schemes have thus the same degree of freedom in distribution mode representation. It can be seen that over SNR ranging from 30 to 5 dB, VPHMM gives 40% lower word error rates.

5. CONCLUSION

Conventional GMHMM uses a constant set of parameters to cover all noisy environments, resulting in HMM distri-

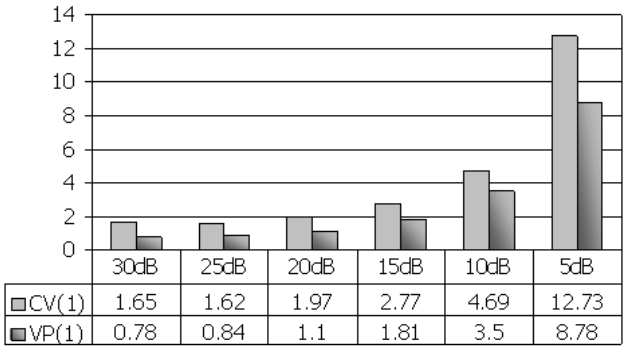

Fig. 2. WER vs. SNR with HMM trained using conventional multi-condition training and using VPHMM training

butions that do not match any of the environments. variable parameter incorporates environment variables (e.g. SNR) into conventional HMMs to make the HMM environment dependent. By explicitly modeling the Gaussian mean vectors as polynomials of SNR, VPHMM adjusts its model parameters based on the environmental SNR to obtain accurate speech distribution under any given SNR level. With the same number of mixtures, VPHMM reduces by 40% the word error rates given by CVHMM for SNR ranging from 5 to 30 dB. Interesting directions for further study includes polynomial function tying among states, the use of vectorial environment variable rather than scalar variable, and modeling variance as environment dependent quantity.

6. REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, April 1995.

[2] S. Morii, T. Morii, and M. Hoshimi, "Noise robustness in speaker independent speech recognition," in *Internat. Conf. on Spoken Language Processing*, Nov 1990, pp. 1145–1148.

[3] S. Furui, "Toward robust speech recognition under adverse conditions," in *ESCA Workshop Proceedings of Speech Processing in Adverse Conditions*, Cannes, France, 1992, pp. 31–41.

[4] S. V. Vaseghi, B. P. Milner, and J. J. Humphries, "Noisy speech recognition using cepstral-time features and spectral-time filters," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Adelaide, Australia, April 1994, vol. II, pp. 65–68.

[5] C. Mokbel and G. Chollet, "Speech recognition in adverse environments: speech enhancement and spectral transformations," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 1991, pp. 925–928.

[6] S. Das, R. Bakis, A. Nadas, D. Nahamoo, and M. Picheny, "Influence of background noise and microphone on the performance of the IBM TANGORA speech recognition system," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 1993, vol. II, pp. 71–74.

[7] S. Das, A. Nadas, D. Nahamoo, and M. Picheny, "Adaptation techniques for ambience and microphone compensation in the IBM Tangora speech recognition system," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Adelaide, Australia, April 1994, vol. I, pp. 21–23.

[8] R. P. Lippmann, E. A. Martin, and D. B. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Dallas, TX, April 1987, pp. 705–708.

[9] M. Blanchet, J. Boudy, and P. Lockwood, "Environment adaptation for speech recognition in noise," in *EUSIPCO*, 1992, vol. VI, pp. 391–394.

[10] T. Kitamura, S. Ando, and E. Hayahara, "Speaker-independent spoken digit recognition in noisy environments using dynamic spectral features and neural networks," in *Internat. Conf. on Spoken Language Processing*, Banff, Alberta, Canada, October 1992, vol. I, pp. 699–702.

[11] Y. Gong, "Noise-dependent Gaussian mixture classifiers for robust rejection decision," *IEEE Trans. on Speech and Audio Processing*, Feb. 2002.

[12] Y. Gong, "Noise-robust open-set speaker recognition using noise-dependent gaussian mixture classifier," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Orlando, FL., 2002.

[13] Y. Gong, "Variable parameter gaussian mixture hidden markov modeling for speech recognition," Tech. Rep., Technical Activity Report, Texas Instruments, June 2002.

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

[15] C. F. J. Wu, "On the convergence properties of the EM algorithm," *The Annals of Statistics*, vol. 11, no. 1, pp. 95–103, 1983.

[16] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.