# 隱藏式馬可夫模型之模糊連結法則

## A Fuzzy Tying Technique for Mixture Autoregressive HMMs

洪偉文　林義楠

Wei-Wen Hung, Yin-Nan Lin

摘要

語音辨識系統的建構中，隱藏式馬可夫模型經常被用來描述某個語音狀態之特徵向量的統計分布。然而，各個語音狀態又可以再細分成許多子狀態，而且每一個子狀態各自有自己的特徵向量平均值和變異量。為了簡化語音辨識過程中之計算量，因此，有必要將某個狀態內部之各個子狀態其特徵向量平均值(或變異量)相互連結成一個具有代表性之特徵向量平均值(或變異量)。本篇論文提出一種「模糊連結法則」，藉以連結各個子狀態之特徵向量平均值(或變異量)，為每一個語音狀態找出一個最具有代表性的特徵向量平均值(或變異量)。實驗結果證明本篇論文所提出之「模糊連結法則」能夠有效強化語音辨識系統的準確性。

關鍵詞：語音辨識系統、隱藏式馬可夫模型、特徵向量平均值和變異量、模糊連結法則。

**ABSTRACT**

In this paper, a fuzzy tying technique is developed to construct a framework for quantitatively formulating the uncertainty involved in the tying operation of Gaussian mixture autoregressive hidden Markov models (HMMs). For the proposed technique, the observation density in each Markov state is simply characterized by the convex combination of Gaussian mixture autoregressive densities that are weighted by a fuzzy membership function. By properly adjusting the fuzzy factor, we can achieve various extents of tying effect. Experimental results for recognition of continuous Mandarin telephone speech indicate that the fuzzy tying technique is useful in enhancing the robustness of HMM-based speech models.

Keywords：Fuzzy tying technique, Gaussian mixture autoregressive density, fuzzy membership function, hidden Markov model (HMM).

### I. INTRODUCTION

Hidden Markov modeling (HMM) has been extensively exploited as an effective technique that extends conventional stationary spectral analysis to the analysis of time-varying signals. It can be treated as a doubly embedded stochastic process with an underlying unobservable stochastic process that can be manifested through another stochastic process characterized by a probability distribution or a density function. This letter concentrates primarily on finite mixtures of Gaussian autoregressive densities that are the bases in the maximum likelihood formulation of hidden Markov modeling. Poritz [1] first employed a single Gaussian autoregressive density for each Markov state and welded the linear prediction analysis into the hidden Markov model methodology. Then, Juang and Rabiner [2] extended the previous work to the cases of finite mixtures of Gaussian autoregressive densities (GAM) and nearest-neighbor partitioned finite mixtures of Gaussian autoregressive densities (PGAM). In the former case, a tying technique that imposes weighted summing operation across all the Gaussian mixture autoregressive densities belonging to a state was employed to obtain the observation density for each Markov state. In contrast, the PGAM case selected only one of the Gaussian mixture autoregressive densities to characterize the observation density according to the

洪偉文　明志科技大學電子工程系教授
林義楠　明志科技大學電子工程系講師

nearest-neighbor criterion. As concluded by Bellegarda and Nahamoo [3] that a tying technique provides us a unified treatment of conventional discrete and continuous hidden Markov modeling methodologies. Especially, efficient tying allows one to design a large vocabulary speaker dependent recognition system [4]-[5] when only a relatively small amount of training data is available, which is almost always the case in practical situations.

In this paper, we propose a novel tying fuzzy technique making use of the uncertainty, that is often encountered in the formulation of the tied mixture hidden Markov models, to the benefit of the tying process. For the proposed tying technique, the observation density in each Markov state is evaluated on a frame-by-frame basis and characterized by the *convex combination* of the associated Gaussian mixture autoregressive densities. The convex combination coefficients are determined by using a fuzzy membership function. By properly adjusting the fuzzy factor, we can well approximate the embedded mixture correlation within a Markov state and achieve various extents of tying effect.

## II. FORMULATION OF THE FUZZY TYING TECHNIQUE

Assuming that we are given an $N$ -state, $M$ -mixture Gaussian-distributed hidden Markov model with parameters $\Lambda = \{N(\mu_{jk}, \Sigma_{jk}), 1 \le j \le N, 1 \le k \le M\}$ estimated from a set of training corpus. Associated with each state $s_t = j$ and mixture component $m_t = k$ of the unobservable Markov process is an observation probability density function (pdf) $b_{jk}(x_t)$ for the observation vector $x_t$ of the observation sequence $X = \{x_t, 1 \le t \le T\}$. This observation probability density $b_{jk}(x_t)$ of state $s_t = j$ and mixture component $m_t = k$ generating the observation vector $x_t$ can be characterized by a multivariate likelihood function and formulated as

$$\begin{aligned} b_{jk}(x_t) &= \Pr(x_t \mid s_t = j, m_t = k) \\ &= (2\cdot\pi)^{-D/2} \cdot \left|\Sigma_{jk}\right|^{-1/2} \times \\ &\exp\left\{-\frac{1}{2}\cdot(x_t - \mu_{jk})^T \cdot \Sigma_{jk}^{-1} \cdot (x_t - \mu_{jk})\right\}, \end{aligned} \quad (1)$$

where $D$ denotes the dimension of feature vectors and $(\mu_{jk}, \Sigma_{jk})$ are the mean vector and covariance matrix of the $k$ th mixture component in the state $s_t = j$, respectively. In the mixtures of Gaussian autoregressive densities (GAM), the observation density $b_j(x_t)$, for $j = 1,2,\cdots,N,$ has the form [2]

$$b_j(x_t)_{GAM} = \sum_{k=1}^{M} c_{jk,GAM} \cdot b_{jk}(x_t), \quad (2)$$

where $c_{jk,GAM}$ is the weight for the $k$ th mixture component and the estimate of this weighting parameter is generally given as

$$c_{jk,GAM} = \frac{\{\text{number of vectors assigned to } \textit{mixture} - j \text{ and } \textit{state} - k\}}{\{\text{number of vectors assigned to state j}\}}. \quad (3)$$

On the other hand, the nearest-neighbor partitioned mixtures of Gaussian autoregressive densities (PGAM) choose the most likely mixture component for each observation vector $x_t$ and assume the form [2]

$$b_j(x_t)_{PGAM} = \frac{1}{M} \max_{k=1,2,\cdots,M} \{b_{jk}(x_t)\}. \quad (4)$$

To further improve the performance of the GAM and PGAM approaches, we propose a fuzzy tying technique (denoted as FGAM) that offers a framework for quantitatively formulating the uncertainty involved in the tying of Gaussian mixture autoregressive densities. In the new approach, we construct the observation density $b_j(x_t)$ by linearly interpolating the set of mixture densities $b_{jk}(x_t)$, for $k = 1,2,\cdots,M,$ and express as

$$b_j(x_t)_{FGAM} = \sum_{k=1}^{M} \alpha_{tjk} \cdot b_{jk}(x_t). \qquad (5)$$

With the stochastic constraint

$$\sum_{k=1}^{M} \alpha_{tjk} = 1, \quad 1 \le t \le T, \quad 1 \le j \le N, \qquad (6)$$

this interpolation procedure is also referred as the *convex combination* [4]. The convex combination coefficient $\alpha_{tjk}$ can be interpreted as the likelihood that the observation density $b_j(x_t)$ is regarded as belonging to the mixture density $b_{jk}(x_t)$ and be defined as a conditional probability of the form

$$\alpha_{tjk} = \Pr\left(b_{jk}(x_t) \middle| b_j(x_t)\right) \quad for \quad all \quad t, \quad j \quad and \quad k. \qquad (7)$$

By employing the Bayes' rule, (7) becomes

$$\alpha_{tjk} = \frac{\Pr\left(b_j(x_t) \middle| b_{jk}(x_t)\right) \cdot \Pr\left(b_{jk}(x_t)\right)}{\Pr\left(b_j(x_t)\right)}. \qquad (8)$$

Since the probability $\Pr\left(b_j(x_t)\right)$ is independent of the mixture index $k$ and the prior probability $\Pr\left(b_{jk}(x_t)\right)$ is assumed to be identical for all mixtures, the convex combination coefficient $\alpha_{tjk}$ is then proportional to the conditional probability $\Pr\left(b_j(x_t) \middle| b_{jk}(x_t)\right)$, that is

$$\alpha_{tjk} \propto \Pr\left(b_j(x_t) \middle| b_{jk}(x_t)\right). \qquad (9)$$

Thus, in order to satisfy the stochastic constraint of (6), the convex combination coefficient $\alpha_{tjk}$ can be formulated by the following equation

$$\alpha_{tjk} = \frac{\beta_{tjk}}{\sum_{k'=1}^{M} \beta_{tjk'}}, \qquad (10)$$

where the membership function has the form [5]

$$\beta_{tjk} = \frac{1}{\sum_{k'=1}^{M} \left[ \frac{d(\mathrm{x}_t, \mu_{jk})}{d(\mathrm{x}_t, \mu_{jk'})} \right]^{1/(F-1)}} \qquad (11)$$

and $d(\mathrm{x}_t, \mu_{jk}) \cong \left\| \mathrm{x}_t - \mu_{jk} \right\|^2$. The uncertainty produced by the fuzzy tying technique is controlled by the fuzzy factor $F$ that is greater than unity. In the case of $F \to 1.0$ and $\alpha_{tjk}$ is maximum, then the other coefficient will be negligible and (5) can be reduced to the form

$$b_j(x_t)_{FGAM} = b_{jk}(x_t). \qquad (12)$$

On the other hand, increasing this parameter $F$ tends to degrade the membership toward the highest uncertainty. That is, all the convex combination coefficients become equal and

$$b_j(x_t)_{FGAM} = \frac{1}{M} \cdot \sum_{k=1}^{M} b_{jk}(x_t). \qquad (13)$$

## III. EXPERIMENTS

A continuous telephone speech database [6] was used to evaluate the schemes we discussed. Each word in the telephone speech database comprised 1~23 Mandarin syllables. From the database, we chose 8320 phonetically balanced Mandarin words (37784 syllables) spoken by 81 males and 79 females to train the right-context-dependent sub-syllable HMMs of 410 Mandarin syllables. Moreover, each syllable model contains six to seven states in which the output observation distribution is characterized by an 8-mixture Gaussian density function with diagonal covariance matrix. In the testing phase, the evaluated schemes were applied to a 500-utterance (4754 syllables) recognition task in which the testing utterances spoken by 15 males and 15 females were selected from a different set of the database. The feature vector was composed of 12-order mel frequency cepstral coefficients and their first-order time derivatives. To simulate various noisy conditions, the 500 testing utterances were corrupted by the additive white Gaussian noise (AWGN) and factory noise with SNR at 10 dB, 20 dB and 30 dB, respectively.

In the aspect of recognition for continuous Mandarin telephone speech, we evaluated the GAM, the PGAM and the FGAM schemes in terms of syllable recognition rate (S.R.R). Two kinds of noise

corruption were investigated, and to see if the FGAM scheme can achieve better syllable recognition rates than the other two schemes in various noisy conditions. In Fig. 1, we illustrated the influences of fuzzy factor on syllable recognition rates under various noisy conditions. It shows that the syllable recognition rate initially increases with the fuzzy factor $F$, attains a maximum value and then decreases with an increase in the fuzzy factor. Obviously, the optimal value of fuzzy factor is related to SNR value, i.e., the smaller the SNR value of additive white Gaussian noise, the smaller the optimal value of fuzzy factor. This phenomenon indicates the existence of uncertainty among the mixtures of Gaussian autoregressive densities.

Moreover, as shown in Table I and II, we can also observe that the FGAM scheme outperforms the widely used GAM and PGAM schemes and exhibits consistent improvements for various noisy conditions. Finally, it is worth to note that the optimal value of fuzzy factor should be heavily related to SNR value. To date, there are still no reliable criteria for the selection of the optimal fuzzy factor for a given set of training corpus. In this study, the optimal values of fuzzy factor under various noisy conditions were determined by selecting some specific values and their neighbors and comparing the corresponding syllable recognition rates.

## IV. CONCLUSIONS

This paper presented a novel fuzzy technique that provides us a framework for quantitatively formulating the uncertainty associated with the tying operation for training of Gaussian mixture autoregressive HMMs. The development of the FGAM technique was based on a more flexible strategy allowing the transition of uncertainty from lowest to highest levels as the tying process evolves. Experimental results reveal that by properly adjusting the fuzzy factor, the FGAM scheme has better capability in formulating the uncertainty and achieving higher syllable recognition rates than those of the GAM and PGAM schemes.

## REFERENCES

[1] A. B. Poritz, “Linear predictive hidden Markov models and the speech signals,” *Proc.ICASSP,* pp. 1291-1294, Paris, France, May 1982.

[2] B. H, Juang and L. R. Rabiner, “Mixture autoregressive hidden Markov models for speech signals,” *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-33, no. 6, pp. 1404-1413, Dec. 1985.

[3] J. R, Bellegarda and D. Nahamoo, “Tied mixture continuous parameter modeling for speech recognition,” *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-38, no. 12, pp. 2033-2045, Dec. 1990.

[4] G. D. Luenberger, “Optimization by vector space methods,”, pp. 14-19, Wiley, New York, 1968.

[5] N. B. Karayiannis and P. I. Pai, “Fuzzy vector quantization algorithms and their application in image compression,” *IEEE Trans. Image Processing,* vol. 4, no. 9, pp. 1193-1201, Sept. 1995.

[6] W. W. Hung and H. C. Wang, “On the use of weighted filter bank analysis for the derivation of robust MFCCs,” *IEEE Signal Processing Letters,* vol. 8, no. 3, pp. 70-73, March 2001.

Fig 1 Influences of fuzzy factor $F$ on syllable recognition rates (SRRs) [%] under various noisy conditions.

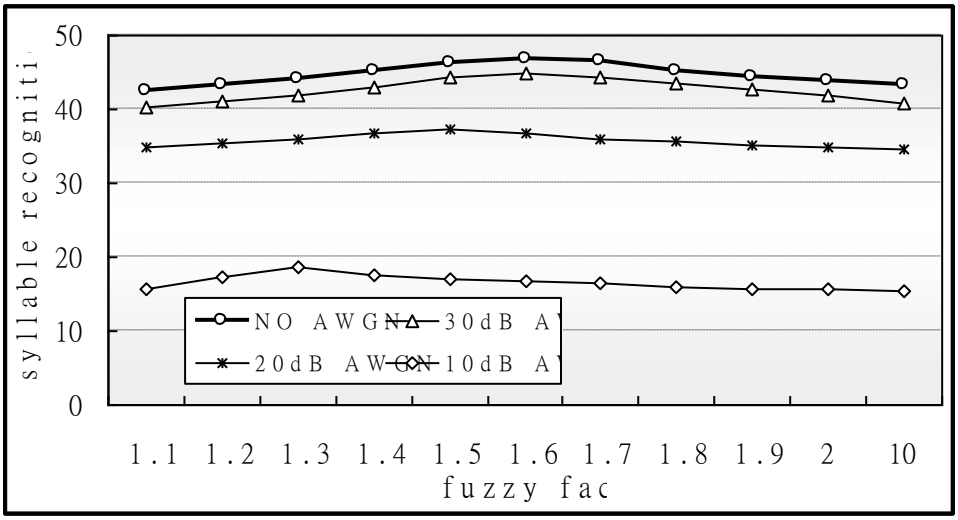


Table I. Comparisons of syllable recognition rates [%] for various schemes with white noise.

| Schemes | NO AWGN | 30 dB AWGN | 20 dB AWGN | 10 dB AWGN |
|---|---|---|---|---|
| GAM | 43.26 | 40.89 | 34.48 | 15.38 |
| PGAM | 42.51 | 40.36 | 34.96 | 15.79 |
| FGAM | 46.66 *(F = 1.6)* | 44.94 *(F = 1.6)* | 37.18 *(F = 1.5)* | 18.52 *(F = 1.3)* |

Table II. Comparisons of syllable recognition rates [%] for various schemes with factory noise.

| Schemes | 30 dB Factory Noise | 20 dB Factory Noise | 10 dB Factory Noise |
|---|---|---|---|
| GAM | 43.52 | 37.37 | 25.44 |
| PGAM | 43.21 | 37.49 | 25.67 |
| FGAM | 45.73 *(F = 1.6)* | 39.06 *(F = 1.6)* | 27.18 *(F = 1.4)* |