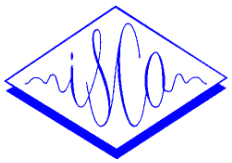

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

# FUZZY ENTROPY HIDDEN MARKOV MODELS FOR SPEECH RECOGNITION

*Dat Tran and Michael Wagner*

Human-Computer Communication Lab., School of Computing,
University of Canberra, ACT 2601, Australia
Email: {DatT, MichaelW}@ise.canberra.edu.au

## ABSTRACT

A new fuzzy technique called fuzzy entropy (FE) clustering is proposed and applied to hidden Markov models (HMMs) for speech recognition. FE-HMMs, both discrete and continuous, are proposed in this paper. Experimental results in speech recognition show good results for FE models compared with fuzzy $C$-means and conventional models.

## 1. INTRODUCTION

A successful approach in pattern recognition and cluster analysis is the fuzzy set theory-based approach [1], where fuzzy $C$-means (FCM) [2] and noise clustering (NC)[3] are widely used techniques. Applying these techniques to hidden Markov models (HMMs), Gaussian mixture models (GMMs) and vector quantisation (VQ) in speech and speaker recognition have been proposed. Speech recognition experiments using these FCM models have shown better results than conventional models [4]-[7]. However, relationships between FCM and conventional models have not been well established. As the degree of fuzziness tends to 1, FCM-HMMs, FCM-GMMs and FCM-VQ do not approach HMMs, GMMs and VQ, respectively. Instead, they approach the corresponding "hard" models. For solving this problem, a fuzzy entropy (FE) clustering technique and its NC version are proposed [8, 9]. FE clustering is as effective as, but quite different from, FCM clustering. A degree of fuzzy entropy $n > 0$ is introduced. As $n \to 0$, FE models approach their hard models. With $n = 1$, FE models are identical to conventional models. Fuzzy approaches applied to obtain FCM models are quite applicable to obtain FE models. FE-HMMs are proposed in this paper. Speech recognition experiments show lower error rates for FE-HMMs in comparison with conventional HMMs and FCM-HMMs.

## 2. FUZZY ENTROPY CLUSTERING

Let us consider the following function

$$H_n(U,\lambda;X) = \sum_{i=1}^{C}\sum_{t=1}^{T} u_{it} d_{it}^2 + n \sum_{i=1}^{C}\sum_{t=1}^{T} u_{it} \log u_{it} \qquad (1)$$

where $C > 1$ is the number of clusters, $T$ is the number of vectors in $X$, $n > 0$, $\lambda$ is the model parameter set, $d_{it}$ is the distance between vector $x_t$ and cluster $i$, and $U = [u_{it}]$, $u_{it}$ is the membership of vector $x_t$ in cluster $i$. Assuming that the matrices $U$ satisfy the following conditions

$$\sum_{i=1}^{C} u_{it} = 1 \qquad \forall t, \qquad\qquad 0 < \sum_{t=1}^{T} u_{it} < T \qquad \forall i \qquad (2)$$

which mean that each $x_t$ belongs to $C$ clusters, no cluster is empty and no cluster is all of $X$ because of $1 < C < T$.

We wish to show that minimising the function $H_n(U,\lambda;X)$ on $U$ yields solutions $u_{it} \in [0,1]$, i.e. the matrices $U$ determine fuzzy $C$-partition space for $X$ and $H_n(U,\lambda;X)$ is a fuzzy objective function. The first term on the right-hand side in (1) is the sum-of-squared-errors function $J_1(U,\lambda;X) = \sum_{i=1}^{C}\sum_{t=1}^{T} u_{it} d_{it}^2$ defined for hard $C$-means ($K$-means) clustering. Minimising $J_1(U,\lambda;X)$ over $U$ yields $u_{it} = 1$ or 0. Consider the function $E(U) = -\sum_{i=1}^{C}\sum_{t=1}^{T} u_{it} \log u_{it}$ over $U$ in the second term. This function is maximised as $u_{it} = 1/C \ \forall i$, and is minimised as $u_{it} = 1$ or 0. Therefore we can see that $u_{it}$ takes values in $[0,1]$ if the function $H_n(U,\lambda;X)$ is minimised. In fact, with the assumption in (2), minimising $H_n(U,\lambda;X)$ using Lagrange multipliers gives

$$\overline{u}_{it} = \left[\sum_{j=1}^{C} \left( e^{d_{it}^2} \Big/ e^{d_{jt}^2} \right)^{1/n}\right]^{-1} \qquad (3)$$

From (3), it can be seen that $0 \le u_{it} \le 1$ and hence the matrices $U$ determine a fuzzy $C$-partition space for $X$. The function $E(U)$ expresses the *uncertainty* of determining whether $x_t$ belongs to a given cluster or not. In other words, $E(U)$ expresses the average value of degree of nonmembership of members in a fuzzy set [10]. The function $E(U)$ was also considered [11] for mixture distributions in relating the expectation-maximisation (EM) algorithm to clustering techniques.

This proposed technique is called FE clustering to distinguish it from fuzzy $C$-means (FCM) clustering, a well-known technique in fuzzy cluster analysis. In general, the task of fuzzy entropy clustering is to minimise the fuzzy objective function $H_n(U,\lambda;X)$ on variables $U$ and $\lambda$, namely, finding a pair of $(\overline{U},\overline{\lambda})$ such that $H_n(\overline{U},\overline{\lambda};X) \le H_n(U,\lambda;X)$. This task is implemented by two alternative steps: 1) Finding $\overline{U}$ such that $H_n(\overline{U},\lambda;X) \le H_n(U,\lambda;X)$, and then 2) Finding $\overline{\lambda}$ such that $H_n(\overline{U},\overline{\lambda};X) \le H_n(\overline{U},\lambda;X)$. Finding $\overline{U}$ is obtained by using (3). To find $\overline{\lambda}$, since $E(U)$ is not dependent on $d_{it}$, determining $\overline{\lambda}$ is performed by minimising the first

term, i.e. $J_1(U,\lambda;X)$, thus the parameter estimation equations are identical to those in "hard" k-means clustering. For the Euclidean distance $d_{it}^2 = (x_t - \mu_i)^2$, we obtain

$$\overline{\mu}_i = \sum_{t=1}^{T} \overline{u}_{it} x_t \Big/ \sum_{t=1}^{T} \overline{u}_{it} \tag{4}$$

The degree of fuzzy entropy $n$ determines the partition of $X$. As $n \to \infty$, we have $u_{it} \to (1/C)$, each feature vector is equally assigned to $C$ clusters, i.e. only a single cluster. As $n \to 0$, $u_{it} \to 0$ or 1, and the function $H_n(U,\lambda;X)$ approaches $J_1(U,\lambda;X)$, FE clustering reduces to k-means clustering.

## 3. MODELLING AND CLUSTERING PROBLEMS

To apply FE clustering to statistical modelling techniques, we need to determine relationships between the modelling and clustering problems. The first task for solving these problems is to establish an optimisation criterion. For clustering purposes, considering data involves considering shapes and locations of clusters and hence the optimisation is to find optimal partitions of data. For modelling purposes, considering data structure involves considering data distributions via the use of statistical distribution functions, where optimisation means finding the right parametric form of the distributions. An advantage of statistical modelling is that it can effectively express the temporal structure of data through the use of a Markov process, a problem which is not addressed by clustering.

It would be useful if we could take advantages of both methods in a single approach. To implement this, a general distance $d_{XY}$ for clustering is defined. It denotes a dissimilarity between observable data $X$ and unobservable data (class, state) $Y$ as follows

$$d_{XY}^2 = -\log P(X,Y|\lambda) \tag{5}$$

This distance is used to relate the clustering problem to the statistical modelling problem. Indeed, since minimising this distance leads to maximising the distribution $P(X,Y|\lambda)$, grouping similar data points into a cluster becomes grouping these into a component distribution. Clusters are now represented by component distribution functions and hence characteristics of a cluster are not only its shape and location, but also the data density in the cluster and possibly the temporal structure of data if the Markov process is also applied.

## 4. MAXIMUM FUZZY LIKELIHOOD CRITERION

The distance in (5) is used to relate the minimum FE criterion in (1) to the maximum likelihood criterion. As an illustration, we consider the case that feature vectors are statistically independent and hence the log-likelihood is $L(\lambda;X) = \sum_{t=1}^{T} \log \sum_{i=1}^{C} P(x_t,i|\lambda)$. Based on (5), the distance in this case is $d_{it}^2 = -\log P(x_t,i|\lambda)$. It can be shown as $n = 1$

$$\overline{u}_{it} = P(i|x_t,\lambda) \quad \text{and} \quad L(\lambda;X) = -H_1(\overline{U},\lambda;X) \tag{6}$$

On the other hand, using the Jensen's inequality, we obtain

$$L(\lambda;X) \geq -H_1(U,\lambda;X) \tag{7}$$

(6) and (7) show that $-L(\lambda;X)$ is the minimum value of the function $H_1(U,\lambda;X)$ on $U$. The second equality in (6) shows that, if we find $\overline{\lambda}$ such that $H_1(\overline{U},\overline{\lambda};X) \leq H_1(\overline{U},\lambda;X)$ then we will have $L(\overline{\lambda};X) \geq L(\lambda;X)$. It means that, as $n = 1$, minimising the FE function in (1) on $\lambda$ using the above distance leads to maximising the likelihood function. So we define

$$L_n(U,\lambda;X) = -H_n(U,\lambda;X) \tag{8}$$

$L_n(U,\lambda;X)$ can be called the *fuzzy likelihood* function. Maximising this function (also minimising the FE function) is implemented by the *fuzzy* EM algorithm—a reestimation algorithm including an iteration of the two steps as follows

**1. Fuzzy E-step**:
Compute $\overline{U}$ such that $L_n(\overline{U},\lambda;X) \geq L_n(U,\lambda;X)$

**2. M-step**:
Compute $\overline{\lambda}$ such that $L_n(\overline{U},\overline{\lambda};X) \geq L_n(\overline{U},\lambda;X)$

## 5. FUZZY ENTROPY HMMS

### 5.1. Fuzzy Entropy Discrete HMMs

From (8), the fuzzy likelihood function for the fuzzy entropy discrete HMM (FE-DHMM) is proposed as follows

$$L_n(U,\lambda;O) = -\sum_{t=0}^{T-1}\sum_{i=1}^{N}\sum_{j=1}^{N} u_{ijt} d_{ijt}^2 - n \sum_{t=0}^{T-1}\sum_{i=1}^{N}\sum_{j=1}^{N} u_{ijt} \log u_{ijt} \tag{9}$$

where $N$ is the number of states, $d_{ijt}^2 = -\log P(O, s_t = i, s_{t+1} = j|\lambda)$, $u_{ijt} = u_{ijt}(O)$ is the fuzzy membership function denoting the degree to which the observation sequence $O$ belongs to the state sequence being state $i$ at time $t$ and state $j$ at time $t+1$ satisfying

$$0 \leq u_{ijt} \leq 1 \ \forall i,j,t, \quad \sum_{i=1}^{N}\sum_{j=1}^{N} u_{ijt} = 1, \quad 0 < \sum_{t=1}^{T} u_{iit} < T \tag{10}$$

The fuzzy EM algorithm for the FE-DHMM is as follows

**Fuzzy E-Step:** Replacing the distance $d_{ijt}$ into (3) gives

$$\overline{u}_{ijt} = \frac{\left[P(O, s_t = i, s_{t+1} = j|\lambda)\right]^{1/n}}{\sum_{k=1}^{N}\sum_{l=1}^{N}\left[P(O, s_t = k, s_{t+1} = l|\lambda)\right]^{1/n}} \tag{11}$$

**M-step:** As known in [12]-[14], $P(O, s_t = i, s_{t+1} = j|\lambda) = \alpha_t(i)\, a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$, where $\alpha_t(i)$, $\beta_{t+1}(j)$ are forward, backward variables, $A = \{a_{ij}\}$ is the state transition probability distribution, $B = \{b_j(k)\}$ is the observation probability distribution. Regrouping $L_n(U,\lambda;O)$ into separate terms for $\pi$ (the initial state distribution), $A$, $B$ and maximising this function using Lagrange multipliers and the following constraints

$$\sum_{j=1}^{N} \pi_j = 1 \qquad \sum_{j=1}^{N} a_{ij} = 1 \qquad \sum_{k=1}^{M} b_j(k) = 1 \tag{12}$$

we obtain the parameter reestimation equations

$$\overline{\pi}_j = \sum_{i=1}^{N} \overline{u}_{ij1} \qquad \overline{a}_{ij} = \sum_{t=1}^{T-1} \overline{u}_{ijt} \Big/ \sum_{t=1}^{T-1} \sum_{j=1}^{N} \overline{u}_{ijt}$$

$$\overline{b}_j(k) = \sum_{\substack{t=1 \\ s.t.\ o_t = v_k}}^{T} \sum_{i=1}^{N} \overline{u}_{ijt} \Big/ \sum_{t=1}^{T} \sum_{i=1}^{N} \overline{u}_{ijt} \tag{13}$$

As $n = 1$, the membership in (11) becomes $\overline{u}_{ijt} = P(s_t = i, s_{t+1} = j|O, \lambda)$, and the fuzzy EM algorithm is now identical to the Baum-Welch algorithm.

### 5.2. Fuzzy Entropy Continuous HMMs

Similarly, $u_{jkt} = u_{jkt}(O)$ is defined as the fuzzy membership function denoting the degree to which the observation sequence $O$ belongs to state sequence being state $s_t = i$ and to Gaussian mixture $k_t = k$ at time $t$, satisfying

$$0 \le u_{jkt} \le 1 \ \forall j, k, t, \quad \sum_{j=1}^{N} \sum_{k=1}^{K} u_{jkt} = 1, \quad 0 < \sum_{t=1}^{T} u_{jkt} < T \tag{14}$$

and the distance $d^2_{jkt} = -\log P(O, s_t = i, k_t = k|\lambda)$. The fuzzy EM algorithm for the FE continuous HMM (FE-CHMM) is as follows:

**Fuzzy E-Step:**

$$\overline{u}_{jkt} = \frac{\left[P(O, s_t = i, k_t = k|\lambda)\right]^{1/n}}{\sum_{i=1}^{N} \sum_{l=1}^{K} \left[P(O, s_t = k, k_t = l|\lambda)\right]^{1/n}} \tag{15}$$

**M-step:** the parameter estimation equations for the $\pi$ and $A$ distributions are unchanged, but the output distribution $B$ is estimated via Gaussian mixture parameters $(w, \mu, \Sigma)$

$$\overline{w}_{jk} = \frac{\sum_{t=1}^{T} \overline{u}_{jkt}}{\sum_{t=1}^{T} \sum_{k=1}^{K} \overline{u}_{jkt}} \qquad \overline{\mu}_{jk} = \frac{\sum_{t=1}^{T} \overline{u}_{jkt} x_t}{\sum_{t=1}^{T} \overline{u}_{jkt}}$$

$$\overline{\Sigma}_{jk} = \sum_{t=1}^{T} \overline{u}_{jkt} (x_t - \overline{\mu}_{jk})(x_t - \overline{\mu}_{jk})' \Big/ \sum_{t=1}^{T} \overline{u}_{jkt} \tag{16}$$

where $w_{jk}, \mu_{jk}, \Sigma_{jk}$ are the mixture weight, mean vector and covariance matrix of Gaussian mixture $k$ in state $j$. Similarly, as $n = 1$, the membership in (15) becomes $\overline{u}_{jkt} = P(s_t = j, k_t = k|O, \lambda)$ and the fuzzy EM algorithm is now identical to the Baum-Welch algorithm.

### 5.3. Noise Clustering Approach

For the fuzzy entropy HMM in the noise clustering approach (NC-FE-HMM), a separate state is used to represent outliers and is termed the *garbage state* [5]. This state has a constant distance $\delta$ from all observation sequences. The membership $u_{\bullet t}$ of an observation sequence $O$ at time $t$ in the garbage state is defined to be $u_{\bullet t} = 1 - \sum_{i=1}^{N} \sum_{j=1}^{N} u_{ijt}, \quad 1 \le t \le T$. Therefore, the membership constraint for the "good" states is effectively relaxed to $\sum_{i=1}^{N} \sum_{j=1}^{N} u_{ijt} < 1, \quad 1 \le t \le T$. This allows noisy data and outliers to have arbitrarily small membership values in good states. The fuzzy likelihood function for the FE-DHMM in the NC approach (NC-FE-DHMM)is

$$L_n(U, \lambda; O) = -\sum_{t=0}^{T-1} \sum_{i=1}^{N} \sum_{j=1}^{N} u_{ijt} d^2_{ijt} - \sum_{t=0}^{T-1} u_{\bullet t} \delta^2$$
$$-n \sum_{t=0}^{T-1} \sum_{i=1}^{N} \sum_{j=1}^{N} u_{ijt} \log u_{ijt} - n \sum_{t=0}^{T-1} u_{\bullet t} \log u_{\bullet t} \tag{17}$$

The fuzzy EM algorithm for NC-FE-DHMMs is as follows:

**Fuzzy E-Step:**

$$\overline{u}_{ijt} = \frac{\left[P(O, s_t = i, s_{t+1} = j|\lambda)\right]^{1/n}}{\sum_{k=1}^{N} \sum_{l=1}^{N} \left[P(O, s_t = k, s_{t+1} = l|\lambda)\right]^{1/n} + e^{-\delta^2/n}} \tag{18}$$

**M-Step:** identical to the M-step of the FE-DHMM in (13).

The second term in the denominator of (18) becomes quite large for outliers, resulting in small membership values in all the good states for outliers. Similarly, the FE-CHMM in the NC approach (NC-FE-CHMM) is as follows:

**Fuzzy E-Step:**

$$\overline{u}_{jkt} = \frac{\left[P(O, s_t = i, k_t = k|\lambda)\right]^{1/n}}{\sum_{i=1}^{N} \sum_{l=1}^{K} \left[P(O, s_t = i, k_t = l|\lambda)\right]^{1/n} + e^{-\delta^2/n}} \tag{19}$$

**M-Step:** identical to the M-step of the FE-CHMM in (16).

## 6. EXPERIMENTAL RESULTS

The TI46 database [7] was used for speech recognition experiments. The data were processed in 20.48 ms frames (256 samples) at a frame rate of 10 ms. Frames were Hamming windowed and preemphasised with $m = 0.9$. For each frame, 46 mel-spectral bands of a width of 110 mel and 20 mel-frequency cepstral coefficients (MFCC) were determined with a resulting 20-dimensional feature vector for each frame. Results for FCM-HMMs were reported in [4]-[7].

6-state left-to-right HMMs in speaker-dependent mode were used for all experiments. In the training phase, 10 training tokens of each word were used to train conventional DHMMs, FE-DHMMs, NC-FE-DHMMs, FCM-DHMMs and NC-FCM-DHMMs using VQ codebook sizes of 16, 32, 64, and 128. In the recognition phase, isolated word recognition was carried out by testing all 160 test tokens against conventional DHMMs, FE-DHMMs, NC-FE-DHMMs, FCM-DHMMs and NC-FCM-DHMMs of each of 16 speakers. Table 1 presents the experimental results for the recognition of the E set consisting of 9 letters *b,c,d,e,g,p,t,v,z*. Table 2 is for the 10-digit

set consisting of 10 digits from 0 to 9. Table 3 is for the 10-command set consisting of 10 commands: *enter, erase, go, help, no, rubout, repeat, stop, start, yes.* In most of the experiments, FE-HMMs show better results than conventional and FCM-HMMs.

## 7. CONCLUSION

Fuzzy entropy hidden Markov models have been presented in this paper. A parameter $n > 0$ is introduced as the degree of fuzzy entropy. As $n = 1$, fuzzy entropy HMMs reduce to conventional HMMs in the maximum likelihood scheme. An advantage obtained from this viewpoint is that we can apply fuzzy methods, such as noise clustering, to statistical models. Another advantage stems from the adjustable parameter $n$ which allows for an additional dimension of model optimisation where conventional models do not achieve adequate recognition results, such as in the case of the nine English E-set words.

Table 1: Isolated word recognition error rates (%) for the E set, $n = 2.5$ and $\delta = 1.5$

| Code book Size | HMM | FE HMM | NC FE HMM | FCM HMM | NC FCM HMM |
|---|---|---|---|---|---|
| 16 | 54.54 | 41.51 | 41.48 | 51.97 | 42.74 |
| 32 | 39.41 | 34.38 | 34.36 | 37.46 | 33.46 |
| 64 | 33.84 | 29.92 | 29.88 | 30.54 | 27.87 |
| 128 | 33.98 | 31.28 | 31.25 | 32.27 | 31.85 |

Table 2: Isolated word recognition error rates (%) for the 10-digit set, $n = 2.5$ and $\delta = 1.5$

| Code book Size | HMM | FE HMM | NC FE HMM | FCM HMM | NC FCM HMM |
|---|---|---|---|---|---|
| 16 | 6.21 | 4.84 | 4.32 | 5.83 | 4.74 |
| 32 | 2.25 | 1.81 | 1.72 | 2.16 | 2.06 |
| 64 | 0.43 | 0.37 | 0.34 | 0.39 | 0.38 |
| 128 | 0.39 | 0.35 | 0.35 | 0.38 | 0.38 |

Table 3: Isolated word recognition error rates (%) for the 10-command set, $n = 2.5$ and $\delta = 1.5$

| Code book Size | HMM | FE HMM | NC FE HMM | FCM HMM | NC FCM HMM |
|---|---|---|---|---|---|
| 16 | 15.74 | 6.45 | 6.32 | 13.78 | 13.74 |
| 32 | 4.36 | 3.64 | 3.52 | 3.88 | 3.76 |
| 64 | 2.43 | 2.27 | 2.24 | 2.28 | 2.28 |
| 128 | 1.65 | 1.45 | 1.43 | 1.60 | 1.58 |

## 8. REFERENCES


[1] L. A. Zadeh, "Fuzzy sets", *Inf. Control.*, vol. 8, no. 1, pp. 338-353, 1965.

[2] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*, Plenum Press, New York and London, 1981.

[3] R. N. Davé, "Characterization and detection of noise in clustering", *Pattern Recognition Lett.*, vol. 12, no. 11, pp. 657-664, 1991.

[4] Dat Tran and Michael Wagner, "Fuzzy Gaussian mixture models for speaker recognition", a special issue of the *Australian Journal of Intelligent Information Processing Systems*, Australia, vol. 5, no. 4, pp. 293-300, 1998.

[5] Dat Tran and Michael Wagner, "Hidden Markov models using fuzzy estimation", in *Proc. EUROSPEECH'99 Conf.*, vol. 6, pp. 2749-2752, Hungary, 1999.

[6] Dat Tran and Michael Wagner, "Fuzzy expectation-maximisation algorithm for speech and speaker recognition", in *Proc. of the North American Fuzzy Information Society* (NAFIPS'99), pp. 421-425, USA, 1999.

[7] Dat Tran and Michael Wagner, "Fuzzy hidden Markov models for speech and speaker recognition", in *Proc.* NAFIPS'99, pp. 426-430, USA, 1999.

[8] Dat Tran and Michael Wagner, "An application of fuzzy entropy clustering in speaker recognition", the Joint Conference on Information Sciences JCIS'2000, Atlantic City, NJ, USA (to appear).

[9] Dat Tran and Michael Wagner, "Fuzzy entropy clustering", the FUZZ-IEEE'2000 Conf., USA (to appear).

[10] R.-P. Li and M. Mukaidono, "Gaussian clustering method based on maximum-fuzzy-entropy interpretation", *Fuzzy Sets and Systems*, vol. 102, pp. 253-258, 1999.

[11] R. Hathaway, "Another interpretation of the EM algorithm for mixture distribution", *J. Stat. Prob. Lett.*, vol. 4, pp. 53-56, 1986.

[12] C.-H. Lee, F. K. Soong and K. K. Paliwal, *Automatic speech and speaker recognition, Advanced topics*, Kluwer Academic Publishers, USA, 1996.

[13] L. R. Rabiner and B. H. Juang, *Fundamentals of speech recognition*, Prentice Hall PTR, USA, 1993.

[14] X. D. Huang, Y. Ariki and M. A. Jack, *Hidden Markov models for speech recognition*, Edinburgh University Press, 1990.