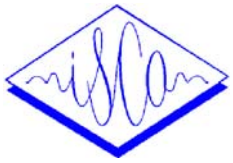

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

HIDDEN MARKOV MODELS USING FUZZY ESTIMATION

Dat Tran and *Michael Wagner*
Human-Computer Communication Lab, School of Computing
University of Canberra Belconnen, ACT 2601, Australia
(dat, miw)@hcc1.canberra.edu.au

ABSTRACT

In the conventional hidden Markov model, the model parameters are reestimated by an iterative procedure known as the Baum-Welch method. This paper proposes an alternative procedure using fuzzy estimation, which is generalised from the fuzzy c-means and the Baum-Welch methods. An extension of this approach, which uses a garbage state to deal with outlier data is also proposed. Experiments using the TI46 speech data corpus show this approach can be applicable to speech and speaker recognition.

10.21437/Eurospeech.1999-605

1. INTRODUCTION

The hidden Markov model (HMM) approach is the well-known and widely used statistical method of characterising the spectral properties of the time frames of a speech pattern. The HMM method provides a natural and highly reliable way of recognizing speech for a wide range of applications [1]-[7]. In HMM theory, the algorithm used to estimate the HMM parameters is the well-known Baum-Welch (B-W) algorithm. It can be regarded as an expectation-maximisation (EM) [8]-[10] algorithm.

The fuzzy *c*-means (FCM) clustering is one of the most widely used approaches for cluster analysis [12]. It is an extension of the *k*-means algorithm (the hard *c*-means algorithm) generalised by Bezdek [13][14]. Gustafson and Kessel [15] proposed a modification of the FCM algorithm in terms of geometric shapes of clusters.

Most of the previous fuzzy approaches to speech and speaker recognition concentrated on applying the FCM algorithm, known as fuzzy vector quantisation (FVQ), instead of conventional vector quantisation (hard c-means). In speaker recognition, FVQ is used to generate speaker models called fuzzy codebooks. In HMM-based speech recognition, FVQ makes a soft decision about which codeword (mean vector) is closest to the input vector, and generates an output vector whose components indicate the relative closeness of each codeword to the input [16]-[19]. However, the above approaches are only applicable to discrete HMMs (observations are discrete symbols). There has not been a fuzzy approach to continuous HMMs in which observations are continuous and modelled by probability density functions. Therefore, finding a fuzzy approach that can be applied to discrete and continuous HMMs should be studied. This paper proposes in Section 2 an alternative procedure using fuzzy estimation, which is generalised from the FCM and the B-W methods. An extension of this approach to deal with outlier data is also proposed in Section 3. Section 4 reports speech and speaker recognition experiments using the Texas Instruments (TI46) speech data.

2. HMMS USING FUZZY ESTIMATION

2.1 Discrete HMMs

Consider a generalised *Q*-function as follows [20][21]

$$Q_m(U,\bar{\Lambda}) = \sum_{all\ S} u_S^m(O)\log P(O,S\mid\bar{\Lambda}) \qquad (1)$$

where

- $\bar{\Lambda}=\{\bar{\pi},\bar{A},\bar{B}\}$ denotes the complete parameter set of the HMM reestimated from $\Lambda=\{\pi,A,B\}$, with
1. $\pi=\{\pi_i\}$, $1\le i\le N$: the initial state distribution
2. $A=\{a_{ij}\}$, $1\le i,j\le N$: the state transition probability distribution
3. $B=\{b_j(k)\}$, $1\le j\le N$, $1\le k\le M$: the observation symbol probability distribution.

- $P(O,S\mid\bar{\Lambda})$ is the joint probability of the observation sequence $O=o_1,o_2,...,o_T$ and the state sequence $S=s_1,s_2,...,s_T$, given the model $\bar{\Lambda}$
- $u_S(O)$ is the membership function, denoting the degree to which the observation sequence *O* belongs to the state sequence *S*, and satisfies

$$0\le u_S(O)\le 1,\qquad \sum_{all\ S} u_S(O)=1 \qquad (2)$$

- $m\ge 1$ is a weighting exponent on each fuzzy membership $u_S(O)$ and is called the degree of fuzziness. As shown in [21], the generalised *Q*-function reduces to the *Q*-function for the conventional HMM for $m = 1$

$$u_S(O)=P(S\mid O,\Lambda) \qquad \text{for } m=1 \qquad (3)$$

The task of this approach is to maximise the generalised *Q*-function in (1) on variables *U* and Λ, e.g., finding a

pair of $(\bar{U}, \bar{\Lambda})$ such that $Q_m(\bar{U}, \bar{\Lambda}) \geq Q_m(U, \Lambda)$. For maximising on variable *U*, (1) is rewritten as

$$Q_m(U, \bar{\Lambda}) = -\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N} u_{ijt}^m d_{ijt}^2 \tag{4}$$

where $u_{ijt} = u_{ijt}(O)$ is the membership function, denoting the degree to which, the observation sequence *O* belongs to state *i* at time *t* and to state *j* at time *t* + 1, satisfying

$$0 \leq u_{ijt} \leq 1 \qquad \sum_{i=1}^{N}\sum_{j=1}^{N} u_{ijt} = 1 \tag{5}$$

and

$$\begin{aligned} d_{ijt}^2 &= d_{ijt}^2(O) = -\log P(O, s_t = i, s_{t+1} = j \mid \Lambda) \\ &= -\log\{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)\} \end{aligned} \tag{6}$$

d_{ijt}^2 denotes the dissimilarity between the observation sequence *O* and the model Λ at time *t*, $\alpha_t(i), \beta_t(j)$ are the forward and backward variables, respectively [8]. The expression on the right-hand side of (4) is of the form of the fuzzy objective function in the FCM method [14], and hence $\bar{U}$ can be determined as follows

$$\bar{u}_{ijt} = \left\{ \sum_{k=1}^{N}\sum_{l=1}^{N} [d_{ijt} / d_{klt}]^{\frac{2}{m-1}} \right\}^{-1} \qquad m > 1 \tag{7}$$

As mentioned in (3), using computations in the conventional HMM [7], we obtain

$$\bar{u}_{ijt} = \xi_t(i, j) = P(s_t = i, s_{t+1} = j \mid O, \Lambda) \qquad m = 1 \tag{8}$$

where

$$\xi_t(i, j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)} \tag{9}$$

For maximising on variable $\bar{\Lambda}$, (1) is regrouped into three terms as follows

$$\begin{aligned} Q_m(U, \bar{\Lambda}) = &\sum_{j=1}^{N}\left(\sum_{i=1}^{N} u_{ij1}^m\right)\log \bar{\pi}_j + \\ &\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\sum_{t=1}^{T-1} u_{ijt}^m\right)\log \bar{a}_{ij} + \\ &\sum_{j=1}^{N}\sum_{k=1}^{M}\left(\sum_{t\in o_t = v_k}\sum_{i=1}^{N} u_{ijt}^m\right)\log \bar{b}_j(k) \end{aligned} \tag{10}$$

Using the Lagrange multiplier method, the parameters of HMMs are reestimated as

$$\bar{\pi}_j = \sum_{i=1}^{N} \bar{u}_{ij1}^m, \quad \bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \bar{u}_{ijt}^m}{\sum_{t=1}^{T-1}\sum_{j=1}^{N} \bar{u}_{ijt}^m}, \quad \bar{b}_j(k) = \frac{\sum_{t\in o_t = v_k}\sum_{i=1}^{N} \bar{u}_{ijt}^m}{\sum_{t=1}^{T}\sum_{i=1}^{N} \bar{u}_{ijt}^m} \tag{11}$$

It can be seen that, as *m* = 1, replacing (8) and (9) into (11), we obtain the B-W reestimation formulas in the conventional HMM.

2.2 Continuous HMMs

Observations are often continuous vectors, and the most general representation of the model output probabilities is a mixture of Gaussians [8]

$$b_j(o_t) = \sum_{k=1}^{M} w_{jk} N(o_t, \mu_{jk}, \Sigma_{jk}) \tag{12}$$

where o_t, *t* = 1,…, *T*, are the *n*-dimensional observation vectors being modelled, w_{jk}, *j* = 1,…, *N*, *k* = 1,…, *M* are mixture coefficients, and $N(o_t, \mu_{jk}, \Sigma_{jk})$ is a Gaussian with mean vector μ_{jk} and covariance matrix Σ_{jk} for the *k*th mixture component in state *j*. The following constraints are satisfied

$$w_{jk} > 0, \ \sum_{k=1}^{M} w_{jk} = 1, \text{ and } \int_{\Re^n} b_j(o_t) do_t = 1 \tag{13}$$

Similarly, let $u_{jkt} = u_{jkt}(O)$ be the membership function, denoting the degree to which the observation sequence *O* belongs to state *j* and mixture *k* at time *t*, satisfying

$$0 \leq u_{jkt} \leq 1 \qquad \sum_{j=1}^{N}\sum_{k=1}^{M} u_{jkt} = 1 \tag{14}$$

It can be shown that the fuzzy reestimation formulas for the coefficients of the mixture density are of the form

$$\begin{aligned} \bar{w}_{jk} &= \frac{\sum_{t=1}^{T} \bar{u}_{jkt}^m}{\sum_{t=1}^{T}\sum_{k=1}^{M} \bar{u}_{jkt}^m}, \qquad \bar{\mu}_{jk} = \frac{\sum_{t=1}^{T} \bar{u}_{jkt}^m o_t}{\sum_{t=1}^{T} \bar{u}_{jkt}^m}, \\ \bar{\Sigma}_{jk} &= \frac{\sum_{t=1}^{T} \bar{u}_{jkt}^m (o_t - \bar{\mu}_{jk})(o_t - \bar{\mu}_{jk})'}{\sum_{t=1}^{T} \bar{u}_{jkt}^m} \end{aligned} \tag{15}$$

where the prime denotes vector transposition,

$$\bar{u}_{jkt} = \left\{ \sum_{i=1}^{N}\sum_{l=1}^{M} [d_{jkt} / d_{ilt}]^{\frac{2}{m-1}} \right\}^{-1} \qquad m > 1 \tag{16}$$

$$\bar{u}_{jkt} = \eta_t(i,j) = P(s_t = j, k_t = k \mid O, \Lambda) \qquad m = 1 \quad (17)$$

$$d^2_{jkt} = d^2_{jkt}(O) = -\log P(O, s_t = j, k_t = k \mid \Lambda)$$
$$= -\log\left\{\sum_{i=1}^{N} \alpha_t(i) a_{ij} w_{jk} N(o_t, \mu_{jk}, \Sigma_{jk}) \beta_{t+1}(j)\right\} \quad (18)$$

and

$$\eta_t(j,k) = \frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \times \frac{w_{jk} N(o_t, \mu_{jk}, \Sigma_{jk})}{\sum_{k=1}^{M} w_{jk} N(o_t, \mu_{jk}, \Sigma_{jk})} \quad (19)$$

3. FUZZY ESTIMATION WITH GARBAGE STATE

Both fuzzy and B-W estimation methods presented in the above section have a common disadvantage in the problem of sensitivity to outlier vectors. We can see this by considering the sum in (5). It means that for both clean data and noisy data, the degrees of belonging of the observation sequence *O* at time *t* across states at time *t* and time *t* + 1 always sum to one. However, it would be more reasonable that, if the observation at time *t* in the sequence *O* comes from noisy data or outliers, the degrees of belonging at that time should be as small as possible for all states, namely, the sum in (5) should be smaller than one. This property is important since all HMM parameters are computed from these degrees of belonging. In [22], Dave proposed the idea of a noise cluster to deal with noisy data for fuzzy clustering methods. This idea can be applied to the HMM. In this approach, the noise is considered to be a *separate state* that has a constant dissimilarity δ from all observations. The membership $u_{\bullet t}$ of the sequence *O* at time *t* in this garbage state is defined to be

$$u_{\bullet t} = 1 - \sum_{i=1}^{N}\sum_{j=1}^{N} u_{ijt} \quad \text{then} \quad \sum_{i=1}^{N}\sum_{j=1}^{N} u_{ijt} < 1 \quad (20)$$

This allows outlier data to have arbitrarily small membership values in "good" states. The generalised *Q*-function in (4) is modified as follows

$$Q_m(U, \bar{\Lambda}) = -\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N} u_{ijt}^m d_{ijt}^2 - \sum_{t=1}^{T} \delta^2 \left(1 - \sum_{i=1}^{N}\sum_{j=1}^{N} u_{ijt}^m\right)^m \quad (21)$$

and the following membership can be derived by differentiating (21) with respect to u_{ijt}

$$\bar{u}_{ijt} = \left\{\sum_{k=1}^{N}\sum_{l=1}^{N} [d_{ijt} / d_{klt}]^{\frac{2}{m-1}} + [d_{ijt} / \delta]^{\frac{2}{m-1}}\right\}^{-1} \quad m > 1 \quad (22)$$

From the second term on the right-hand side of (22), we can see that with a suitable value for δ, u_{ijt} has a smaller value than that in (7) as the dissimilarity d_{ijt} is large. In the case of continuous HMM, a similar modification can be obtained for the membership function in (14)

4. EXPERIMENTAL RESULTS

A comparison between the B-W estimation (m = 1) and fuzzy estimation (m > 1) for isolated word-recognition in speaker-dependent mode was performed. Three data sets of the TI46 corpus were used for speech recognition using discrete models: the *E* set including the 9 letters: {*b, c, d, e, g, p, t, v, z*}; the 10-*digit* set: {0, 1,…, 9}; and the 10-*command* set: {*enter, erase, go, help, no, rubout, repeat, stop, start, yes*}. The 10-*command* set was also used for text-dependent speaker identification of 16 speakers using continuous models. These data sets were derived from the TI46 speech data corpus, uttered by 16 speakers, 8 female and 8 male, labelled f1-f8 and m1-m8, respectively. Each speaker repeated the words 10 times in a single training session, and then again twice in each of 8 later testing sessions. The corpus was sampled at 12500 samples per second and 12 bits per sample. The data were processed in 20.48 ms frames (256 samples) at a frame rate of 125 frames per second (100 sample shift). Frames were Hamming windowed and preemphasised with μ = 0.9. For each frame, 46 mel-spectral bands of a width of 110 mel and 20 mel-frequency cepstral coefficients (MFCC) were determined [23]. Table 1 presents the experimental results for the recognition of the E set using 6-state left-to-right HMMs in speaker-dependent mode. Here m = 1, m = 1.2, and m = 1.2, δ = 1.5 stand for the B-W estimation, the fuzzy estimation, and the fuzzy estimation with garbage state, respectively. The codebook size was varied from 16 to 128 and the results show that fuzzy estimation with garbage state is consistently better than fuzzy estimation, which is in turn consistently better than B-W estimation.

Codebook Size	Recognition Error Rate (%) for		
	m = 1	m = 1.2	m = 1.2, δ = 1.5
16	54.54	51.97	43.74
32	39.41	37.46	33.46
64	33.84	30.54	27.87
128	33.98	32.27	31.85

Table 1: Speech recognition for the *E* set

Codebook Size	Recognition Error Rate (%) for	
	HMMs	Fuzzy HMMs
16	6.21	5.83
32	2.25	2.16
64	0.43	0.39
128	0.39	0.38

Table 2: Speech recognition for the 10-*digit* set

Table 2 presents the experimental results for the recognition of the 10-*digit* set using 6-state left-to-right

models, in speaker-dependent mode, and Table 3 presents the equivalent results for the 10-command set. Speaker identification results on 16 speakers using 3-state 2-mixture and 3-state 4-mixture models, in text-dependent mode are shown in Table 4.

Codebook Size	Recognition Error Rate (%) for	
	HMMs	Fuzzy HMMs
16	15.74	13.78
32	4.36	3.88
64	2.43	2.28
128	1.65	1.60

Table 3: Speech recognition for the 10-*command* set

Models	Identification Error Rate (%) for	
	HMMs	Fuzzy HMMs
3 states 2 mixtures	2.52	2.04
3 states 4 mixtures	1.39	1.36

Table 4: Speaker identification results for 16 speakers

In these 3 experiments we show only the results for fuzzy estimation and B-W estimation because the fuzzy estimation with garbage state produces very similar results to the fuzzy estimation. This is probably due to the fact that clusters in the 10-*digit* and 10-*command* sets are better separated than those in the *E* set.

5. CONCLUSIONS

Fuzzy estimation and a modification for noisy data have been proposed to the HMM in this paper. Both states and mixtures are regarded as fuzzy subsets. Time-dependent fuzzy membership functions are also an alternative approach to the FCM method. Experimental results have shown the effectiveness of this fuzzy approach to HMMs.

REFERENCES

[1] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi. *An introduction to the application of the theory of Probabilistic functions of a Markov process to automatic speech recognition.* The Bell System Technical Journal, vol. 62, no. 4, 1983, pp 1035-1074, 1983

[2] L. R. Rabiner, S. E. Levinson, and M. M. Sondhi. *On the application of vector quantisation and hidden Markov models to speaker-independent, isolated word recognition.* The Bell System Technical Journal, vol. 62, no. 4, 1983, pp 1075-1105, 1983

[3] B. H. Juang. *Maximum likelihood estimation for multivariate observations of Markov sources.* AT&T Technical Journal, 64, pp 1235-1239, 1985

[4] L. R. Rabiner, and B. H. Juang. *An introduction to hidden Markov models.* IEEE Acoustics, Speech, and Signal Processing Society Magazine, vol. 3, no. 1, pp. 4-16, Jan. 1986.

[5] L. R. Rabiner. *A tutorial on hidden Markov models and selected applications speech recognition.* In Proc. IEEE, vol. 77, no. 2, pp. 257-286, Feb. 1989

[6] Vidyadhar G. Kulkarni. *Modeling and analysis of stochastic systems.* Chapman & Hall, UK, 1995

[7] X.D. Huang, Y. Ariki, and M.A. Jack. *Hidden Markov models For Speech Recognition.* Edinburgh University Press, 1990

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

[9] 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, Ser. B, vol. 39, pp. 1-38, 1977.

[10] C. F. J. Wu. *On the convergence properties of the EM algorithm.* Ann. Statist. vol. 11, pp. 95-103, 1983.

[11] L. Zadeh. *Fuzzy Sets*. Inform. Contr., vol. 8, pp. 338-353, 1965.

[12] James C. Bezdek and Sankar K. Pal. *Fuzzy models for pattern recognition.* IEEE Press, 1992.

[13] J. Dunn. *A fuzzy relative of the ISODATA process and its use in detecting compact well-separated cluster*. J. Cybernetics, vol. 3, pp. 32-57, 1974.

[14] James C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York and London, 1987.

[15] D. E. Gustafson, and W. Kessel. *Fuzzy clustering with a fuzzy covariance matrix.* In Proc. IEEE-CDC, vol.2 (K. S. Fu, ed.), pp. 761-766, IEEE Press, Piscataway, New Jersey, 1979.

[16] H-P. Tseng, M. J. Sabin and E. A. Lee. *Fuzzy vector quantisation applied to hidden Markov modelling.* In Proc. Int. Conf. Acoustics, Speech & Signal Processing (ICASSP), pp. 641-644, 1987.

[17] M. Koo and C. K. Un. *Fuzzy smoothing of HMM parameters in speech recognition.* Electronic Letters, vol. 26, pp. 7443-7447, 1990.

[18] E. Tsuboka and J. Nakahashi. *On the fuzzy vector quantisation based hidden Markov model*. In Proc. Int. Conf. Acoustics, Speech & Signal Processing (ICASSP), vol. 1, pp. 637-640, 1994.

[19] H. J. Choi, Y. H. Oh. *Speech recognition using an enhanced FVQ based on a codeword dependent distribution normalisation and codeword weighting by fuzzy objective function.* In Proc. Int. Conf. Spoken Language Processing (ICSLP), vol. 1, pp. 354-357, USA, 1996.

[20] Dat Tran and Michael Wagner. *Fuzzy Expectation-Maximisation Algorithm for speech and speaker Recognition*. In Proc. the 18th Int. Conf. of the North American Fuzzy Information Society (NAFIPS'99), 1999, USA (to appear).

[21] Dat Tran and Michael Wagner. *Fuzzy HMMs for speech and speaker Recognition*. In Proc. the 18th Int. Conf. of the North American Fuzzy Information Society (NAFIPS'99), 1999, USA (to appear).

[22] R. N. Dave. *Characterization and detection of noise in clustering.* Pattern Recognition Letters, vol. 12, no. 11, pp. 657-664, 1991.

[23] M. Wagner. *Combined speech-recognition/ speaker-verification system with modest training requirements*. Proceedings of the Sixth Australian International Conference on Speech Science and Technology, Adelaide, Australia, pp. 139-143, 1996.