The Effect of Membership Concentration in FVQ/HMM for Speaker-Independent Speech Recognition

Chang-Young Lee* · Ho-Soo Nam* · Hyun-Seok Jung* · Chai-Bong Lee*

ABSTRACT

We investigate the effect of membership concentration on the performance of the speaker-independent recognition system by FVQ/HMM. For the membership function, we adopt the result obtained from the objective function approach by Bezdek. Membership concentration is done by varying the exponent in the membership function. The number of selected clusters is constrained to two for the sake of cheap computational cost. Experimental results showed that the recognition rate has its maximum value when the membership function was taken to be inversely proportional to the distance of the input vector from the cluster centroid. When the membership concentration was two weak or too strong, the performance was found to be relatively poor as expected. Except these extreme cases, the membership concentration was not shown to affect the recognition rate significantly. This is in accordance with the general observation that the fuzzy system is not much sensitive to the detailed shape of the membership function as long as it is overlapped over multiple classes.

Keywords: Speech Recognition, Fuzzy Vector Quantization, Membership, HMM

I. Introduction

Up to date, the hidden Markov model (HMM) has been shown to be one of the most successful techniques for automatic speech recognition. It has been proved to yield good performance in wide ranges including speaker-independent large-vocabulary continuous speech recognition [1].

The front-end observation events for HMM are frequently obtained by vector quantization (VQ) process. As the most primitive form of VQ, a single cluster that best matches the input vector is selected and its codebook index is transferred to HMM. The quantization errors incurred during this 'crisp' VQ is quite high due to the lost information of all the other clusters than the selected one's. To circumvent this problem, continuous observation density HMM [2] and semicontinuous HMM [3] have been studied. Since more information is passed to these HMM versions, the recognition rate is significantly enhanced. However, the increase in

* Div. of Information System Engineering, Dongseo University

computational cost and structural complexity is unavoidable. Considering the practical applications where real time response is mandatory [4], the problems regarding computational load should not be underestimated.

In order to cope with the low recognition rate caused by large quantization errors while keeping the computational cost as cheap as in the conventional discrete VQ/HMM, fuzzy vector quantization (FVQ) and its application to HMM have been developed and studied by many authors [5][6][7]. In this scheme, multiple cluster indices are chosen according to some criterion and each selected index is assigned appropriate membership value that represents the degree to which the cluster centroid matches the input vector. Instead of 'crisp' (hard) decision for participating or not in the subsequent HMM process, candidates have some 'membership' (eligibility) between 0 and 1 for training of HMM parameters.

There are numerous ways to get the membership values. Included are intuition, inference, inductive reasoning, neural networks and genetic algorithms [8]. In a classical work [9], Bezdek used an objective function approach in obtaining the membership function.

As an application of this result to speaker-dependent English alphabet recognition, Tseng et. al. [5] found that the recognition rate is improved compared to the conventional discrete VQ/HMM. Their experiments were performed on the two cases of fuzziness. A peculiar result of their work is that the recognition rate did not increase monotonically with the number of training data, which is in contradiction with the common expectation that 'the more the training data the better the results.' Another problem in their work is that all the clusters from the codebook were used for training of HMM. For large codebook size, this requires heavy computational load and thus deteriorates the merit of the discrete VQ/HMM.

Tsukoba et. al. [6] improved the theory as given in [5] and achieved the recognition rate superior to the result acquired by continuous HMM with unimodal Gaussian distribution at each state. There have also been attempts to enhance the recognition rate by incorporating FVQ with neural networks [10][11]. However, the effect of the membership concentration on the recognition rate has not yet been studied.

In this paper, we reinvestigate the FVQ/HMM for isolated word recognition. To alleviate the computational cost, we restrict the number of clusters for FVQ be two. The degree of fuzziness will be varied by concentration of membership functions and the corresponding effects on the recognition rate will be examined.

II. Theory of FVQ/HMM

In a classical work on the fuzzy partitioning [9], Bezdek derived that the objective membership function for class i to yield the minimization of the objective function

$$O = \sum_{i=1}^{M} [\mu_O(i)]^m d_i$$

is given by

$$\mu_O(i) = \frac{1}{d_i^{1/(m-1)} \left(\sum_{k=1}^{M} \frac{1}{d_k^{1/(m-1)}} \right)} \quad , \quad i = 1 \sim M. \tag{1}$$

M is the size of the partitions (classes) and d_i is the distance between the input vector to be soft-classified and the centroid of the i-th class. Usually, the distance is given by a Euclidean norm in multi-dimensional hyperspace. In speech processing, the feature vectors are usually extracted from the speech signals by either linear predictive cepstral coefficients (LPCC) or mel-frequency cepstral coefficients (MFCC) [12]. An extensive survey on feature extraction can be found in [13]. The adjustable parameter m, called the 'exponential weight', affects the degree of distinction between the close and the far class centroids from the input vector. The larger $m > 1$, the classification becomes fuzzier or less 'concentrated' in the language of Zadeh [14].

Mathematically, the Eq. (1) can be obtained simply (intuitively) by assuming the power law

$$\mu_O(i) \propto d^{-\alpha} \tag{2}$$

with $\alpha = 1/(m-1)$ and applying the normalization condition

$$\sum_{i=1}^{M} \mu_O(i) = 1 \tag{3}$$

In view of this observation, the approach of Bezdek might also be interpreted as intuitive.

Along with the objective function, a constraint function is in general necessary in fuzzy decision making [15]. Since the codebook size needs to be somewhat large for good performance in HMM [16], it is inevitable to have any sort of constraints to cut out the clusters whose distances are relatively large from the input vector. That is, membership values should fall to zero for clusters with distances above a certain specified value. Otherwise, all the cluster indices would participate in HMM with nonzero membership values as in [5].

As a straightforward candidate for the constraint membership function, one may use

$$\mu_C(i) = 1 - \theta(d_i - d_C) \tag{4}$$

where $\theta(\cdot)$ is the unit step function and d_C is the 'cut-off' distance. Zero membership values

are given to those cluster centroids whose distances with the input vector are larger than this value. The final 'decided' membership function is then obtained by the operation [15]:

$$\mu(i) = \mu_O(i) \wedge \mu_C(i) = \min\{\mu_O(i),\ \mu_C(i)\} \tag{5}$$

By the constraint membership function, the clusters outside the hypersphere of radius d_C centered at the input vector are discarded.

In our present work, however, fixed number of two nearest clusters from the input vector were selected. The choice of this number is motivated by the work of [16] and it was also found in our separate study that such a choice yields the best results. The limitation of the maximum number of VQ candidates to two helps also in alleviating the computational load and expediting the calculation, at the cost of losing mathematical rigor and consistency in that we were applying different cutoff distances from frame to frame. The rigorous use of (4) and (5) and study of its effect on the recognition rate awaits further study.

The HMM is specified by the 3-tuple parameter set $\lambda=(\pi, A, B)$ for each word. π is the initial state probability, $A=[a_{ij}]\ (1\le i, j\le N)$ is the state transition probabilities, and $B=[b_i(j)]\ (1\le i\le N,\ 1\le j\le M)$ is the probability of observing the event j at state i. N is the number of states which is taken to be 20 in our experiments. The training of the HMM parameters needs to be modified in order to accommodate the FVQ results.

In FVQ, the events are not exclusive. An observation event is distributed fuzzily over several clusters, and thus the observation at time t is changed as follows:

$$o_t \Rightarrow \vee_k\{o_t(k) \text{ with degree } \mu(k)\}\ ,\ 1\le k\le K$$

where K is the number of selected clusters, which is chosen to be 2 in our study. The accommodation of the multiple observation probability at state i at time t might then be achieved by

$$b_i(o_t) \Rightarrow b_i[\vee_k\{o_t(k)\}] = \frac{\sum_{k=1}^{K} w_k\, b_i(o_t(k))}{\sum_{k=1}^{K} w_k} \tag{6}$$

with some weighting factors ω_k. It should be admitted that this 'weighted-mean' prescription has no a priori theoretical background. It is just a heuristic choice and needs further study.

If we adopt for the weighting factor as

$$w_k = \mu(k)$$

and apply the normalization condition (3) with M replaced by K, then Eq. (6) is reduced to

$$b_i(o_t) \Rightarrow \sum_{k=1}^{K} \mu(k)\ b_i(o_t(k)) \tag{7}$$

By this prescription, the probability of an event observation is replaced by the 'likelihood' of multiple events' observation. Though there's no constraint to the membership values other than $\mu(k) \in [0, 1]$, we normalized that the sum of the membership values for a given input vector be unity. The detailed reestimation formulas are referred to [17] with the modification as given by Eq. (7).

III. Experimental Setup

Our study was performed on a set of phone-balanced 200 isolated Korean words. 40 persons, 20 male and 20 female, participated in speech production. Each utterance was sampled at 16 kHz and quantized by 16 bits. 512 data points corresponding to 32ms of time duration were taken to be a frame. The next frame was obtained by shifting 170 data points, thereby overlapping the adjacent frames by 2/3. For each frame, Hanning windowing was applied after pre-emphasis for spectral flattening. As the feature vector, MFCC of order 13 was obtained with the mel-scale adopted from [13].

A codebook of size 2,048 was generated by LBG clustering algorithm [18]. The distances between the input vector and the codebook cluster centroids were calculated with Euclidean norm and quick-sorted in order of distance. Two nearest centroids were selected and assigned membership values according to Eq. (2) and then normalized.

In spite of insufficient training data, speech utterances of 40 persons were divided into three disjoint groups. The first group consisting of 30 persons' speech was used for training of HMM parameters. After each training iteration, the recognition rate was examined on the second group consisting of 4 persons' speech. The HMM model parameters $\lambda = (\pi, A, B)$ for each word that yields the best recognition rate for this second group were used for the final test of the speaker-independent speech recognition on the third group of 6 persons.

For the HMM, a non-ergodic left-right (or Bakis) model of 20 states was adopted. Initial estimation of the parameters was obtained by K-means segmental clustering [19] after the first training. By this procedure, the convergence of the iteration was so fast that enough

convergence was reached only after several iterations. Backward state transitions were prohibited by suppressing a_{ij} with $i > j$ to a very small value (10-20) but skipping of states was allowed.

Parameter reestimation was performed by Baum-Welch reestimation formula with "scaled multiple observation sequences" [20] to avoid machine-errors caused by repetitive multiplication of small numbers. After each iteration, the parameters $b_i(j)$ were boosted above 10-6 in view of the results from [20]. Three features were monitored during the iterations: (1) the recognition rate for the second group, (2) the total probability likelihood of events for all the words of the training set according to the trained model, and (3) the event observation probabilities for one word. Iteration was terminated when the convergences for these three features were thought to be sufficient.

In order to examine the effect of membership concentration, we measured the recognition rate for 8 values of the parameter α in Eq. (2), i.e.,

$$\alpha = 0, 1, 2, 4, 8, 16, 32, \infty$$

If we interpret the membership value as the 'eligibility for participation in HMM', repetitive doubling of α correspond to

'Eligible', 'Very Eligible', 'Very Very Eligible', ...

since the linguistic hedge 'very' is accomplished by 'arithmetic squaring of membership value' according to Zadeh [14] (Doubling of α is equivalent to squaring of μ. See Eq. (2)).

IV. Results

Table 1 and Figure 1 shows our experimental results, the effect of the membership concentration on the recognition rate. As α in Eq. (2) is varied from 0 to ∞, the difference in membership values of the two clusters becomes larger, i.e., the membership concentration gets stronger.

The first row in the table is for the case of no discrimination of membership values between the two clusters. The last row corresponds to the discrete VQ/HMM. These extreme cases did not show good performance compared to the other ones as expected.

Table 1. The details of our experimental results

Parameter α in Eq. (2)	Membership versus Distance	Average of the First Membership	Average of the Second Membership	Membership Concentration	Recognition rate
0	-	0.500	0.500	Extremely Weak	93.67%
1	$\mu \propto \frac{1}{d}$	0.539	0.461	Very Very Weak	95.50%
2	$\mu \propto \frac{1}{d^2}$	0.575	0.425	Very Weak	95.08%
4	$\mu \propto \frac{1}{d^4}$	0.638	0.362	Weak	94.83%
8	$\mu \propto \frac{1}{d^8}$	0.726	0.274	Strong	94.92%
16	$\mu \propto \frac{1}{d^{16}}$	0.818	0.182	Very Strong	94.50%
32	$\mu \propto \frac{1}{d^{32}}$	0.892	0.108	Very Very Strong	93.92%
∞	-	1.000	0.000	Extremely Strong	87.50%

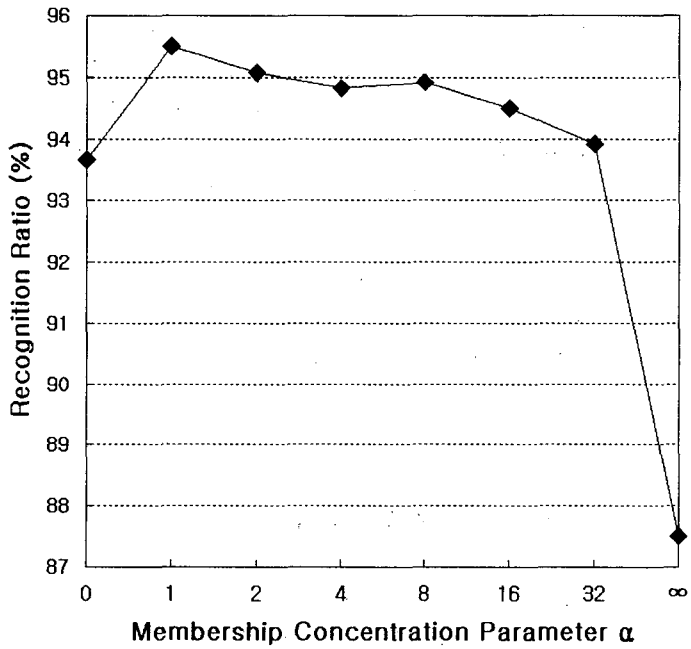

Figure 1. Recognition rate versus membership concentration. As α is increased, the distinction between the two membership values becomes larger. The right-most data corresponds to the case of discrete VQ/HMM.

We see form the Figure 1 that the recognition rate reaches its maximum when $\alpha = 1$. This means that the best performance in FVQ/HMM is obtained by assuming that the membership value is inversely proportional to the distance of the input vector to the centroid. There's a region of membership concentration (α=1~8) within which the recognition rate falls within some acceptable bounds from the maximum. In that region, that is, unless the membership concentration is neither too weak nor too strong, the recognition rate is not so sensitive to the membership values. This result is in accordance with the general observation that the detailed shape of the membership function is not of much importance in the overall performance of a fuzzy system as long as it is reasonably overlapped over multiple class candidates.

V. Conclusion

In this paper, we investigated the effect of membership concentration on the recognition rate in speaker-independent speech recognition by fuzzy VQ/HMM. Given an input feature vector, the membership function of a codebook cluster for HMM training was assumed to be proportional to some inverse powers of the distance between the input vector and the cluster centroid. This is as obtained from the objective function approach by Bezdek. Membership concentration was done by varying the exponent in geometric series of factor 2. This corresponds to the realization of the linguistic hedge 'very' as suggested by Zadeh. In an effort to acquire good performance while keeping the computational complexity and cost at low level, we applied a constraint that the number of clusters selected be two. As a result, it has been observed that the recognition rate has its maximum value when the membership function was taken to be inversely proportional to the distance between the input vector and the cluster centroid. When the membership concentration was too weak (soft, fuzzy) or too strong (hard, crisp), the performance was relatively not good as expected. Except these extreme cases, concentration of the membership values was shown to yield minor changes in the recognition rate.

References

[1] Lee, K., Hon H., & Reddy, R. 1990. "An Overview of the SPHINX Speech Recognition System." *IEEE Trans. ASSP*, Vol. 6, No. 1, pp. 35-45.

[2] Rabiner, L. & Juang, B. 1993. *Fundamentals of Speech Recognition.* Prentice-Hall International, Inc., pp. 350-352.

[3] Huang, X. D. 1992. "Phoneme Classification Using Semicontinuous Hidden Markov Models."

IEEE Trans. on Signal Processing, Vol. 40, No. 5, pp. 1062-1067.

[4] Rabiner, L. & Juang, B. 1993. *op. cit.*, p. 486.

[5] Tseng, H-P., Sabin, M. J., & Lee, E. A. 1987. "Fuzzy Vector Quantization Applied To Hidden Markov Modeling." *Proc. ICASSP*, pp. 641-644.

[6] Tsukoba, E. & Nakahashi, J. 1994. "On the Fuzzy Vector Quantization Based Hidden Markov Model." *Proc. ICASSP*, pp. I637-I640.

[7] Tran, D. & Wagner, M. 1999. "Fuzzy Hidden Markov Models for Speech and Speaker Recognition." *NAFIPS: 18th International Conference of the North American*, pp. 426-430.

[8] Ross, T. J. 1995. *Fuzzy Logic with Engineering Applications.* McGraw-Hill, pp. 87-125.

[9] Bezdek, J. C. 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press.

[10] Jang, C. S. & Un, C. K. 1992. "Integrating Time-Delay Neural Networks and Fuzzy HMM for Phoneme-Based Word Recognition." *Electronics Letters*, Vol. 28, No. 25, pp. 2319-2320.

[11] Cong, L., Xydeas, C., & Erwood, A. F. 1994. "Combining Fuzzy Vector Quantization and Neural Network Classification for Robust Isolated Word Speech Recognition." *Singapore ICCS '94. Conference Proc.*, Vol. 3, pp. 884-887.

[12] Wang, Jia-Ching., Wang, Jhing-Fa., & Weng, Yu-Sheng. 2002. "Chip Design of MFCC Extraction for Speech Recognition." *The VLSI Journal*, Vol. 32, pp. 111-131.

[13] Picone, J. W. 1993. "Signal Modeling Techniques in Speech Recognition." *Proc. IEEE*, Vol. 81, No. 9, pp. 1215-1247.

[14] Zadeh, L. 1972. "A Fuzzy-Set Theoretic Interpretation of Linguistic Hedges." *J. Cybern.*, Vol. 2, No. 2, pp. 4-34.

[15] Zimmermann, H. J. 1991. *Fuzzy Set Theory and Its Applications.* Kluwer Academic Publishers, pp. 241-248.

[16] Peinado, A. M., Segura, J. C., Rubio, A. J., Sanchez, V. E., & Garcia, P. 1994. "Use of Multiple Vector Quantization for Semicontinuous-HMM Speech Recognition." *IEEE Proc. Vis. Image Signal Process*, Vol. 141, No. 6, pp. 391-396.

[17] Rabiner, L. & Juang, B. 1993. *op. cit.*, pp. 321-389.

[18] Linde, Y., Buzo, A., & Gray, R. M. 1980. "An Algorithm for Vector Quantizer Design." *IEEE Trans. on Communications*, Vol. 28, pp. 84-95.

[19] Rabiner, L. R., Juang, B. H., Levinson, S. E., & Sondhi, M. M. 1985. "Recognition of Isolated Digits Using Hidden Markov Models With Continuous Mixture Densities." *AT&T Tech. J.*, Vol. 64, No. 6, pp. 1211-1234.

[20] Levinson, S. E., Rabiner, L. R., & Sondhi, M. M. 1983. "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition." *Bell System Tech. J.*, Vol. 62, No. 4, pp. 1035-1074.

received: November 10, 2005
accepted: December 10, 2005

▲ Chang-Young Lee
Div. Of Information System Engineering
Jurye San 69-1, Sasang, Pusan 617-716, Korea
Tel: +82-51-320-1719 Fax: +82-51-320-2389
E-mail: seewhy@dongseo.ac.kr

▲ Ho-Soo Nam
Div. Of Information System Engineering
Jurye San 69-1, Sasang, Pusan 617-716, Korea
Tel: +82-51-320-1716 Fax: +82-51-320-2389
E-mail: hsnam@dongseo.ac.kr

▲ Hyun-Seok Jung
Div. Of Information System Engineering
Jurye San 69-1, Sasang, Pusan 617-716, Korea
Tel: +82-51-320-1714 Fax: +82-51-320-2389
E-mail: hsjung@dongseo.ac.kr

▲ Chai-Bong Lee
Div. Of Information System Engineering
Jurye San 69-1, Sasang Pusan 617-716, Korea
Tel: +82-51-320-1755 Fax: +82-51-320-2389
E-mail: lcb@dongseo.ac.kr