Using Gaussian Mixture Modeling for Phoneme Classification

Yaxin Zhang, Mike Alder, Roberto Togneri
Centre for Intelligent Information Processing System
Department of Electrical and Electronic Engineering
The University of Western Australia
Nedlands, WA 6009, AUSTRALIA

ABSTRACT

Phoneme recognition is a key characteristic in large-vocabulary speech recognition system. The recent reports have shown that the accuracies of speaker-independent English phoneme recognition are around 60% while in the speaker-dependent case the recognition accuracies are under 70%. This paper describe a new method in which the Gaussian mixture modeling was employed for speaker-independent phoneme recognition. The Expactation-Maximization (EM) algorithm was used to generated the Gaissian mixture models (GMMs). Two English databases were used for both the system training and testing. The phonemes were manually extracted from 11 isolated digits (from zero to nine and oh). The testing results are higher than that of recent reports. Some related observations are also reported.

**This paper was published in the proceedings of ANZII, Dec. 1993 Perth, Australia.*

1.INTRODUCTION

Phoneme recognition has been commonly used in speech recognition systems for large-vocabulary continuous or isolated speech recognition [8]. Using phonemes as the basic unit of recognition template makes the system more flexible for adapting to new words and expanding the vocabulary. It is obvious that the accuracy of phoneme recognition is a key issue in such systems. The most popular methods used for phoneme recognition are hidden Markov modeling, dynamic time warping, and the use of various artificial neural networks. The recent reports have shown that the accuracies of speaker-independent English phoneme recognition are around 60%. Even in the speaker-dependent case, the recognition accuracies are under 70% [6,7]. It is desirable to look for effective methods for phoneme recognition to improve the performance of speech recognition system.

Previous investigations tell us that the phonemes, especially vowels , reveal Gaussian-like distributions [9]. This allows us to classify the phoneme space depending on its statistical distribution. In this paper, we report a new method for speaker-independent phoneme recognition. The new method is composed of two phases, training and testing. In the training phase, we extract the phonemes from a speech training sequence and form a sub-sequence of training for each phoneme. A Gaussian mixture model (GMM) is generated for each phoneme data sequence by using the Expectation-Maximization (EM) algorithm [1]. It means that each phoneme is describe by a GMM in which there are two Gaussians for each diphthong and one for each vowel. In the testing phase, each input data point is tested using the all GMMs. The input points are decided to belong to a certain phoneme when its GMM gives the highest likelihood. Compared with other methods, our new method gets much higher recognition accuracies. Also, it has less computational complexity and looks simple in concept.

2.GAUSSIAN MIXTURE MODEL AND EM ALGORITHM

Gaussian mixture models are an effective description of data sets comprising clusters of vectors which are more complex than simple Gaussian distribution. A Gaussian mixture model is defined as

$$f(x) = \sum_{i=1}^{N} p_i g(x, \mu_i, \Sigma_i)$$

where $g(x, \mu_i, \Sigma_i)$ is the Gaussian probability density function with mean μ_i and covariance matrix $\Sigma_i = (\sigma_i^{j\,k})$, x is a random D-dimensional vector, $x = (x^1, x^2, ..., x^D)$, and the p_i are weights which describe the relative likelihood of classes being generated from each of the clusters and must satisfy $\sum_{i=1}^{N} p_i = 1$, where N is the number of classes.

In cases where the distances between means are large in comparison to the square roots of the variances, this model describes a set of isolated clusters of ellipsoidal shape. It is corresponding to the observations that the most phoneme clusters are separated or partly separated in the speech space and have ellipsoid-like shape.

In order to generate the GMMs from the phoneme training sequence, we employed the Expectation-Maximization (EM) algorithm. The EM algorithm for maximum-likelihood estimation of the parameters of a GMM is an iterative procedure in which each iteration consists of two steps: an estimation step (E-step), followed by a maximization step (M-step). In the E-step, the likelihoods, means, and covariance matrices of GNNs are estimated depending on the observation sequence. In the M-step, the new values of the estimation of the parameters of the GMMs are computed.

Suppose we have a sample of S points $x_j = (x_j^1, x_j^2, ..., x_j^D)$, $j = 1, 2, ..., S$, drawn from a set of points which are assumed to lie in N clusters. We initialize N Gaussians with probabilities $p_1 = p_2 = ... = p_N = 1/N$, means $\mu_1, \mu_2, ..., \mu_N$, which can either be random or set equal to N of the data points with a small perturbation, and covariance matrices $\Sigma_1, \Sigma_2, ..., \Sigma_N$, set equal to the identity matrix or a multiple thereof.

In the E-step we compute:

the total likelihood

$$t_j = \sum_{i=1}^{N} p_i g(x_j, \mu_i, \Sigma_i),\ j = 1, 2, ..., S; \tag{1}$$

where g is the Gaussian probability density function;

the normalized likelihoods

$$n_{ij} = p_i g(x_j, \mu_i, \Sigma_i)/t_j,\ i = 1, 2, ..., N;\ j = 1, 2, ..., S; \tag{2}$$

the notional counts

$$C_i = \sum_{j=1}^{S} n_{ij},\ i = 1, 2, ..., N; \tag{3}$$

the notional means

$$\bar{x}_i = \sum_{j=1}^{S} x_j n_{ij}/C_i,\ i = 1, 2, ..., N; \tag{4}$$

and the notional sums of squares

$$SS_i^{p\,q} = \sum_{j=1}^{S} x_j^p x_j^q n_{ij}/C_i,\ i = 1, 2, ..., N,\ and\ p, q = 1, 2, ..., D. \tag{5}$$

In the M-step we compute new values of the parameters of the Gaussian model as follows:

$$p_i = C_i/S;$$

$$\mu_i = \bar{x}_i;$$

$$\Sigma_i^{p\,q} = SS_i^{p\,q} - \bar{x}_i^p \bar{x}_i^q;$$

where $i = 1, 2, ..., N$.

Dempster *et al* [1] have shown that the EM algorithm is convergent and that the convergence rate is quadratic. According to our observations, in most cases, about ten iterations are sufficient to yield useful estimates of the parameters of HMMs for speech data.

3. DATA BASE AND EXPERIMENTAL DESIGN

Two English databases were investigated for the new method. The first database is the Studio Quality Speaker-Independent Connected-Digits Corpus (TIDIGITS) from the National Institute of Standard and Technology in the USA. The training data comprised a small vocabulary of eleven isolated digits (from zero to nine and oh) spoken by 112 speakers (57 males and 55 females) and test data spoken by 113 different speakers (57 males and 56 females). The second database is CDIGITS corpus which we collected ourselves. The CDIGITS was recorded in a normal laboratory environment which means the data has quite high background noise level. The average signal-to-noise ration (SNR) is about 35 dB. It includes 22,000 utterances from 1108 speakers and was sampled at 16 KHZ sampling rate and digitized to 14 bit resolution. In our experiments 2,200 utterances from 100 speakers (50 males and 50 females) were used as the training sequence and 1,100 utterances from 50 different speakers (25 males and 25 females) formed the testing sequence.

The speech data was windowed and 512-point FFT's were computed with a 200-point (12.5ms) advance between frames. The FFT coefficients were binned into 12 Mel-spaced values to produce 12-dimensional feature vectors.

The both training sequence was divided into two parts according to the speaker's sex. For the TIDIGITS database, we extracted 11 vowels from 11 isolated digits for each part. For the CDIGITS database, we extracted 28 phonemes, which included vowels, diphthongs, and consonants, from 11 isolated digits for each part. All the phoneme are manually segmanted from the two databases.

The training phase consists of two steps. First, The EM algorithm was used to generated the GMMs individually. One Gaussian was produced for each vowel or consonant. Two Gaussians were generated for each diphthong. It is based on the observation that the probability distribution of a typical diphthong in the multi-dimension space looks like two Gaussian-like clusters. The observation was performed by projecting from 12 dimension of the data to the computer screen. When all the phonemes were done, we combined all the GMMs into a codebook in which each codeword is a GMM. Second, we input the whole training sequence and the codebook into the EM training procedure and ran it with only the E-step. It means the means μ_i and the covariance matrices Σ_i of the GMMs would not be changed. But the probabilities p_i of the GMMs will be re-computed according to the probability distribution of whole training sequence. It is clear that the phonemes and the GMMs share a one-to-one relationship. In the testing phase, the likelihoods of each input data point of phonemes with the GMMs were computed. The highest likelihood given by a certain GMM indicates that the input point belong to this cluster. A set of input data points is classified as a particular phoneme only when a majority of them are classified as belonged to the appropriate cluster.

4.RESULTS AND CONCLUSIONS

Table 1 shows the Accuracy rates of phoneme recognition which were obtained using the new method of Gaussian mixture modeling for the two database.

The results of CDIGITS are worse than that of TIDIGITS because the database of CDIGITS has quite high background noise while the TIDIGITS is a clean studio quality database. Comparing with the accuracies of isolated digits recognition for the two corpora using the conventional VQ/HMM method, 98.3% for TIDIGITS and 78% for CDIGITS, we are not be surprised at the big differences between the accurate rates of phoneme recognition of the two database. These results are much better than that of other recent reports. A matter of disputation about these results is that the number of phonemes we used is limited (11 vowels from TIDTGITS and 28 phonemes from CDIGITS). We intend to do more experiments using the common databases such as TIMIT corpus.

Table 2 shows a part of the confusion matrix of phoneme recognition for the CDIGITS database. The testing sequence is same as the training one .It tells us that some vowels are more confused with some certain vowels than others. For instance, /i/ and /e/, /Λ/ and /a/, /u/ and /o/, are the confused pairs. While /i/, /a/, /u/ are hardly confused. This observation is corresponding to the result of Person and Barney [10]. In fact, we observed the same situation among the consonants. For example, /s/ and /z/, /θ/ and /f/ are confused while /f/ is hardly confused with /s/. Consonants are never confused with vowels or diphthongs. This suggests that vowels and consonants occupy separate regions in the speech space.

The results illustrate that it is reasonable that we generated two Gaussians for each diphthong. We observed that in the recognition phase the output of diphthongs always transferred from one Gaussian to another certain one and gave almost same recognition accuracies as the vowels.

database	training	testing
TIDIGITS	89.73	84.85
CDIGITS	72.32	59.49

Table 1: The accuracy rates of phoneme recognition

	/a/	/Λ/	/e/	/i/	/o/	/u/	/f/	/s/	/θ/	/z/
/a/	176	11	5	0	2	1	0	0	0	0
/Λ/	21	163	1	0	1	0	0	0	0	0
/e/	7	4	149	16	2	1	0	0	0	0
/i/	0	1	12	163	2	3	0	0	0	0
/o/	3	1	0	1	185	13	0	0	0	0
/u/	2	2	3	0	23	161	0	0	0	0
/f/	0	0	0	0	0	0	149	1	16	2
/s/	0	0	0	0	0	0	0	186	2	14
/θ/	0	0	0	0	0	0	29	3	116	4
/z/	0	0	0	0	0	0	2	23	3	161
accuracy	87.2	58.3	70.2	78.1	85.0	61.6	65.5	75.1	55.3	60.8

Table 2: Part of the confusion matrix for phoneme recognition

5. ACKNOWLEDGEMENT

This work was supported in part by The University Fee-Waiver Scholarship and The University Research Studentship of the University of Western Australia. The author would like to thank Dr. Chrise deSilva for his help in the programming of the EM algorithm.

6.REFERENCE

[1] Dempster, A. P., Laird, N. M., and Rubin, D. B.(1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm", Proc. R. Stat. Soc. B, 39(1), pp.1-38.

[3] Wolf, J. H.(1970), "Pattern Clustering by Multivariate Mixture Analysis", Multivariate Behavioural Research, Vol. 5, pp. 329-350.

[4] Zhang, Y., and deSilva, C. (1991), "An Isolated Word Recognizer using the EM Algorithm for Vector Quantization", Proceeding of IREECON 91,(Sydney, Australia), Sep. 1991, pp. 289-292.

[5] Zhang, Y., deSilva, C., Attikiouzel Y., and Alder, M. (1992), "A HMM/EM Speaker-Independent Isolated Word Recognizer", The Journal of Electrical and Electronic Engineering, Australia, Vol. 12, No.4, pp. 334-240.

[6] Huang, X. D.(1992), "Phoneme Classification Using Semicontinuous Hidden Markov Models", IEEE Trans. on Signal Processing, Vol.40, pp.1062-

[7] Nobuo Hataoka, and Alex H. Waibel (1990), "Speaker-Independent Phoneme Recognition on TIMIT Database Using Integrated Time-Deley Neural Network (TDNNs)", Proceedings of IJCNN International Joint Conference on Neural Networks, June, 1990, Vol.1, pp.57-

[8] Wayne A. Lea, "Trends in Speech Recognition", Speech Science Publications, 1986, pp.3-18.

[9] Pijpers, M., and Alder, M.D. (1992), "Finding Structure in the Speech Space ", Proceedings of First Australian and New Zealand Conference on Intelligent Information System ", Dec., 1993.

[10] Rabiner, L.R., and Schater, R.W., "Digital Processing of Speech Signals ", Printice-Hall Inc., 1978, pp.42-45.