A COMPARISON OF PBDHMM AND CHMM FOR ISOLATED WORD RECOGNITION

Yaxin Zhang, Chee Wee Loke, Roberto Togneri, Mike Alder
Center for Intelligent Information Processing Systems
Department of Electrical and Electronic Engineering
The University of Western Australia

Abstract

Using phoneme-based Gaussian mixture as a VQ codebook in DHMM speech recognition system (PBDHMM) is an efficient way to improve the system performance. This paper compares the performances of PBDHMM system with that of the well known continuous HMM system for isolated word recognition task. The results shown that PBDHMM system obtained better results than CHMM system, especially for phoneme-distinct data.

INTRODUCTION

A discrete hidden Markov model (DHMM) speech recognition system is a combination of vector quantization and hidden Markov modeling. The DHMM speech recognizer was first successfully used for a speaker-independent isolated digit recognition task by Rabiner and his colleagues in 1982 [5]. After that it was widely employed for various speech recognition tasks, not only for the small vocabulary isolated word recognition, but also the large vocabulary continuous speech recognition such as the SPHINX system [3]. The main advantages of DHMM are its concise concepts and low computational costs. The computational simplicity arises from the fact that the calculation of the output probability for each observation vector of a test utterance is displaced by consulting the probability from the existed models in which store all possible output probabilities for the finite observations, which are from the vector quantizer.

There is an important disadvantage of an DHMM-based speech recognition system. It is that the use of a VQ codebook to represent the speech spectral vectors creates an inherent spectral distortion in representing the actual analysis vector. Since there is only a finite number of codebook vectors, the process of choosing the "best" representation of a given spectral vector is equivalent to quantizing the vector and leads to quantization error. This is further worsened by an inappropriate quantization or clustering of the data. As the size of the codebook increases, the size of the quantization error decreases. However, with any finite codebook there will always be some nonzero level of quantization error. Many researchers have realized this problem and tried to use various techniques to reduce the effects of quantization error or inappropriate clustering. For example, Lee *et al* used a method called *multiple codebook integration* in the SPHINX system for continuous speech recognition [3]. In this system a set of three VQ codebooks were generated from three different speech features, bilinear-transformed LPC cepstrum coefficients, differenced bilinear-transformed LPC cepstrum coefficients, and a weighted combination of power and differenced power. Nishimura and Toshioka reported a 50% error reduction comparing with the conventional VQ in a small vocabulary isolated word recognition task by using a so called *multi-dimensional multi-labeling* technique [4]. The similarity of this technique to Lee's method is that it also uses multiple speech features to generate the multi-codebook. However, instead of the best match vector (the nearest codeword) in the codebook as the output, Nishimura's vector quantizer outputs few best matched codewords for each input vector from each codebook and the outputs are decided by a predetermined threshold. Another technique is based on the probability density function (pdf) smoothing in the HMM parameter estimation stage [6]. It can be seen as a compensative technique for the information loss caused by conventional VQ.

Some researchers assume that the inappropriate clustering problem is caused by serious quantization distortion [2, 3, 5]. However, this is not necessarily the best explanation. From the distortion point of view, one could reduce the quantization distortion by increasing the quantization level, *i.e.*, codebook size of VQ. When the quantization level or codebook size $\longrightarrow \infty$, the quantization distortion $\longrightarrow 0$. This means that the larger the codebook size is, the better performance the recognizer has. However, in a DHMM speech recognition system with a conventional vector quantization strategy, this does not always work. For example, in a small vocabulary isolated word recognition task, say the 10 digits, a codebook with 64 or 128 codewords may give the best recognition result. When the codebook size is increased to 256 or more, the quantization distortion must be smaller than before, but the recognition result may be worse. The reason is that the speech signal space is inappropriately classified. In fact, the speech space consists of finite clusters which certainly reveal some statistical characteristics.

Investigations have shown that the speech signal space consists of finite clusters which represent phoneme data sets from male or female speakers and reveal Gaussian distributions [1]. In our speaker-independent isolated word recognition system, we used a substitute vector quantization strategy, phoneme-based Gaussian mixture VQ codebook, instead of the conventional VQ algorithm, and obtained significant improvements of system performance.

In a phoneme-based Gaussian mixture codebook, each phoneme of each sex of speakers in the training sequence is represented by a Gaussian codeword. This means that we treated the phonemes as data classes in the clustering produced by VQ. The codebook size, however, is normally larger than the number of phonemes of training sequence. The reason is that the phoneme space is only a subspace of the speech signal. In other words, the speech signal space includes not only the phonemes, but also some detailed sound between the phonemes and silence or background noise. Then we generated a Gaussian mixture codebook in which some of codewords were determined by the individual phonemes and others by the rest of training sequence.

In this paper we compare the performance of the continuous hidden Markov modeling (CHMM) system and the phoneme-based VQ discrete hidden Markov modeling (PBDHMM) system for isolated word recognition. Two databases were used for the comparison with better results obtained by the PBDHMM system.

PHONEME-BASED VECTOR QUANTIZATION

Investigations have shown that the distribution of phonemes reveal a Gaussian-like distribution [1]. Normally a vowel or a consonant can be described by one Gaussian while a diphthong is represented by two partially overlapped Gaussians. In fact, according to our observations, most phonemes occupy separate or partially separate sub-spaces, and the consonants show more separation than the vowels. Another interesting observation is that the same phoneme data from different sex speakers occupy partially separate sub-spaces. Based on these observations, we know that each phoneme data class could be treated as an affine sub-space in the first step modeling, vector quantization. In our previous work, we proposed an improved training procedure for the speech recognition system in which the phonemes were manually extracted from the training sequence and used to form the Gaussian models individually. We generated a Gaussian model for each phoneme and used these Gaussians to initialize the training procedure of the EM algorithm which was employed to produce a Gaussian mixture model for the training speech data.

Because the phoneme space is not the whole speech space but only a subspace, we should generate a Gaussian mixture model which not only represents the the individual phonemes, but also the detailed sound between and out of the phonemes, and also the background noise. For this VQ training procedure, we run the EM algorithm with only the E step for the fixed Gaussians. This means that the clustering procedure only changes the mixture weighting parameters, but does not change the means or the covariance matrices of the fixed Gaussians. The remaining Gaussians in the codebook are generated depending on the probability distribution of speech data from the whole training sequence. After the EM training procedure, the training sequence is classified as N clusters which are represented by the N Gaussians in which M Gaussians ($M \leq N$) were generated individually from the M phonemes extracted before. These Gaussians more accurately describe the distribution characteristic of the phonemes in the speech signal space. Therefore better classification of speech space is achieved and the performance of whole recognition system is improved. The system was evaluated for isolated word recognition task.

DATABASES AND SPEECH FEATURE PREPARATION

Two English databases were used for the training and testing. The first database is the Studio Quality Speaker-Independent Connected-Digits Corpus (TIDIGITS) from the National Institute of Standards and Technology (NIST) in the USA. The data is sampled at 20 KHZ sampling rate and digitized to 16 bit resolution. A subset of the database used in our experiments comprised a small vocabulary of eleven isolated digits (from zero to nine and oh) spoken by 112 speakers (55 males and 57 females) and test data spoken by 113 different speakers (56 males and 57 females). There are 1232 utterances in the training sequence and 1243 in the testing sequence. The second database is the TI46-word Speaker-dependent Isolated Word Corpus from the same organization as above. The database comprises 46 isolated words, 10 digits, 10 computer command words, and 26 English alphabets. The data is sampled at 12500 HZ and digitized to 14 bit resolution. Here we used only a subset of the 26 alphabets. There are 16 speakers (8 male and 8 female) in the database and each word was repeated 26 times by the speakers. The first 10 repetitions were used as the training set and the remaining 16 as the testing set. There are 4160 utterances in the training sequence and 6656 in the testing sequence.

The feature extraction procedure for both databases is the same. The FFT of the speech data was computed every 10 ms using a 25 ms Hamming window. The FFT coefficients were binned into 12 Mel-spaced values to produce 12-dimensional feature vectors corresponding to the frequency range from 60 to 5000 HZ. The adjacent frequency bands have 25% overlapping.

EXPERIMENTS AND RESULTS

According to our observation that the phoneme data from male and female speakers occupy different regions, the training sequence was divided into two parts according to the speaker's sex. The phonemes were manually extracted from the speech waveforms for the training sequence. We extracted 19 phonemes from 11 isolated digits of TIDIGITS for each sex and 24 from the 26 alphabets of TI46(alphabet) for each gender. In the training procedure we estimated the parameters of two Gaussian models for each vowel or consonant, one for the data from the male speaker and one for female's. For each diphthong data two mixture models of two Gaussians were produced by the EM algorithm. This is based on the observation that the diphthong data look like two overlapped Gaussian clusters. For the all phonemes processed, we had 44 Gaussian models for TIDIGITS and 58 Gaussian models for TI46(alphabet).

Suppose we have M Gaussian models from the individual phonemes. In this case, it is reasonable that the code book should have more elements than M because of the presence of background noise, silence between the speech utterances, as well as some detailed sound not included in the phoneme extraction. In other words, the phoneme space is not the whole speech space, it is a subset of the observation sequence. Thus the EM algorithm was used to generate a Gaussian mixture model which has N ($N > M$) Gaussians for the whole training sequence. The previously obtained M Gaussian models were used to initialize the EM training procedure. The means and covariance matrices of the M Gaussians were fixed in the estimation iterations but the probabilities (or weights) of the M Gaussians were recalculated depending on the probability distribution of the observation sequence. The rest of the N-M Gaussians were also produced in the EM training procedure. A set of phoneme-based Gaussian mixture codebooks with different codebook size were generated and an optimization procedure was applied to

Database	TIDIGITS	TI46(alphabet)
CHMM	97.87%	95.62%
PBDHMM	98.3%	96.56%

Table 1: The accuracy rates of English alphabet recognition.

the codebooks [10]. Finally, we had an optimized codebook with 81 codebook elements, *i.e.*, N=81.

A hidden Markov Modeling system with continuous Gaussian mixture density was employed for the comparison. Both the PBDHMM and CHMM speech recognizers used a 5 states left-to-right constraint model structure described by Rabiner at [5]. A Gaussian model with full covariance matrix was used for each state each recognition model in the CHMM system. This means that there are total 55 full covariance matrices for TIDIGITS and 130 for TI46(alphabet). Table 1 shows the experimental results.

CONCLUSIONS

In both the cases, PBDHMM system obtained better results than CHMM system. In the TIDIGITS case, the recognition accuracy gain is 0.43%. In the TI46(alphabet) case, the recognition accuracy was improved by 0.94%.

In PBDHMM system we generated the mixture of 81 Gaussian models as the codebook for the two databases. While in the CHMM system 55 and 130 Gaussian models were employed for the two databases respectively. Because the main calculation cost in the recognition step is that for the log-likelihood of each input utterance with the Gaussians, we assume that the calculation complexity of the recognition system is directly proportional to the number of Gaussian models it used. As such, the calculation cost of PBDHMM system is higher in the TIDIGITS case but is lower in the TI46(alphabet) case than the CHMM system.

Comparing the results not only with the CHMM system, but also with current results for this type vocabulary, (about 99.9% for 10 digits recognition and 95.5% for 36 alphabetdigits recognition) [11], we can say that the phoneme-based Gaussian mixture codebook for a DHMM speech recognition system is an efficient method for the recognition of phoneme-distinct data such as alphabets.

ACKNOWLEDGEMENT

This work was supported in part by The University Fee-Waver Scholarship and The University Research Studentship of the University of Western Australia.

REFERENCES

[1] Zhang, Y., Alder, M., & Togneri, R., "Using Gaussian Mixture Modeling in Speech Recognition", Proceedings of ICASSP 1994, April 1994, Adelaide, Australia. pp. I613-616.

[2] Huang, X. D. & Jack, M. A., (1988), "Hidden Markov Modeling of Speech Based on a Semicontinuous Model," Electronic Letters, Vol. 24(1), Jan. 1988, pp 6-7.

[3] Lee, K. F., Hon, H. W., and Reddy, R., (1990), "An Overview of the SPHINX Speech Recognition System," IEEE Trans. on ASSP, Vol. ASSP-38(1), Jan. 1990, pp. 35-45.

[4] Nishimura, M., and Toshioka, K., (1987), "HMM-Based Speech Recognition Using Multi-Dimensional Multi-Labeling," Proceedings of ICASSP-87, pp. 1163-1166.

[5] Rabiner, L. R., Levinson, S. E., & Sondhi, M. M., (1983), "On the Application of Vector Quantization and Hidden Markov Model to Speaker-Independent, Isolated Word recognition," AT&T Tech. J., Vol. 62(4), April 1983, pp. 1075-1105.

[6] Schwartz, R., *et al*, (1989), "Robust Smoothing Methods for Discrete Hidden Markov Models," Proceedings of ICASSP-89, pp. 548-551.

[7] 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, Dec. 1992, pp. 334-340.

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

[9] Zhang, Y., Alder, M., and Togneri, R., (1993), "Using Gaussian Mixture Modeling for Phoneme Classification", Proceedings of ANZIIS-93 (Australian and New Zealand Conference on Intelligent Information Systems), Perth, Australia, Dec. 1-3, 1993, pp. 649-652.

[10] Zhang, Y., Togneri, R., deSilva, C., and Alder, M., (1994), "Optimization of Phoneme-Based Codebook in a DHMM System", Proceedings of The Fifth Australian International Conference on Speech Science and Technology, Dec. 1994.

[11] Rabiner, L. R., (1994), "Applications of Voice Processing to Telecommunications", Distinguished Lecturer Seminar, Perth, Australia, 4 July, 1994.