A MULTI-HMM ISOLATED WORD RECOGNIZER

Yaxin Zhang, Christopher J. S. deSilva,
Roberto Togneri, Mike Alder, and Yianni Attikiouzel
Centre for Intelligent Information Processing System
The University of Western Australia

Abstract

A multi-HMM speaker-independent isolated word recognition system is described. In this system,three vector quantization methods are used for the classification of speech space. This multi-HMM system results in an improvement of about 50 per cent in the error rate in comparison to the single model system.

INTRODUCTION

Currently, the most popular approach to speech recognition is the combination of Vector Quantization (VQ) with a Hidden Markov Modelling (HMM) We can consider HMM/VQ to be a two step modelling technique. The first step, vector quantization, is used to divide the signal space into a number of cells or sub-spaces and produce a codebook of vectors. The second step, Hidden Markov Modelling, is used to produce a set of models which represent possible sequences of codebook vectors which arise from words that the system is to recognize.

The commonest VQ algorithm is LBG algorithm. It has the advantages of being simple and not requiring exessive computation. The LBG algorithm does not guarantee that the classification of speech space is globally optimal. This means that some of the codebook vectors may not be typical of the vectors in the cells they represent. The shortcomings of the LBG algorithm lead to an inappropriate classification of the speech space and inadequate matching with the Hidden Markov Modelling, and consequently a limited recognition accuracy for the whole system.

Observations have shown that different utterences of the same speech sound form a cluster around some centre, which represents some average or fiducial production of the sound. The variations about the mean will occur at random when a large population of speakers is considered, so the points in the cluster may be distributed according to a multidimensional Gaussian probability density function. This view of the speech

production process suggests that classification of the speech space is better done on the basis of a Gaussian Mixture Model (GMM), in which the points are clustered around the means according to Gaussian distributions and each cluster is assigned a weight representing the frequency with which points in the cluster occur. A method now known as Expectation-Maximization (EM) algorithm for estimating the parameters of GMMs was described by Wolfe [2]. The EM algorithm can be used as a substitute for the LBG algorithm for quantization of the speech space. Experiments have shown that the EM algorithm matches the HMM quite well and leads to a better recognition accuracy. However, the EM algorithm is more computation intensive than the LBG algorithm and is sensitive to background noise. Another classification method that we have devised is the MGC algorithm which is similar to the EM algorithm but requires less computation in training and produces slightly rougher classification than the EM algorithm.

In this paper we describe a multi-HMM (MHHM) speaker-independent isolated word recognizer in which the three VQ algorithms mentioned above are used independently of each other. These quantizations of the speech space are then used to produce three HMMs for each word in the vocabulary using the Baum-Welch algorithm. In the recognition step, the Viterbi algorithm is used in three sub-recognizers. The probabilities of the observation sequences matching the models are multiplied by the weights and summed to give the probability that the utterance is of a particular word in the vocabulary. We report the results of comparing this method with the use of a single vector quantization algorithm. This results is an reduction of about 50 per cent in the error rate in comparison to the best single VQ/HMM system.

SYSTEM CONSTRUCTION

We constructed a multi-Hidden Markov Model (MHHM) speaker-independent isolated word recognizer. The system is composed of three sub-recognizers, each of which uses one vector quantization method for the first step modelling.

In the training step, the three classification algorithms mentioned above are employed for a parallel vector quantization and a codebook is generated for every sub-recognizer. Thus, three models for each word of the vocabulary are produced by the Hidden Markov Modelling (Baum-Welch) algorithm. In the recognition step, the Viterbi algorithm is used in parallel with the three sub-recognizers. The probabilities of the observation sequences matching the models are multiplied by weights determined by the individual recognition accuracies. The weighted probabilities for each word to be recognized are summed. Then the model which produces the highest probability is the output.

Table 1 demonstrates the recognition improvement of the MHMM/VQ system. The numbers in the table are the output scores of individual HMM/VQ and MHMM/VQ recognizers when the input word is "ONE". The highest score is the system output. The whole system gave a correct output even when two of three single systems, LBG and MKF, gave incorrect recognitions.

	HMM/LBG	HMM/MKF	HMM/EM	MHMM/VQ
ONE	-134	-122	**-111**	**-539**
TWO	-191	-188	-184	-841
THREE	-173	-165	-169	-759
FOUR	-156	-148	-151	-680
FIVE	-166	-169	-163	-746
SIX	-189	-197	-200	-885
SEVEN	-178	-202	-191	-803
EIGHT	-195	-186	-183	-840
NINE	**-129**	-131	-140	-606
ZERO	-177	-173	-177	-791
OH	-141	**-119**	-144	-608
OUTPUT	NINE	OH	ONE	ONE

Table 1: The scores of hidden Markov models matching the input word "ONE" with three vector quantization algorithms

EXPERIMENT AND RESULTS

A set of evaluation tests was performed on the MHMM/VQ system. Our data base was the *Studio Quality Speaker-Independent Connected-Digits Corpus* (TIDIGITS) published by the National Institute of Standards and Technology in the U.S.A.. The training data comprised a small vocabulary of eleven isolated digits (from zero to nine and oh) spoken by 112 speakers (55 male and 57 female) and testing data spoken by 113 different speakers (56 male and 57 female).

For preparing the input for the VQ system, the speech data was windowed and feature vectors were constructed for each window. The first pre-processing step was the computation of the power spectrum of the windowed signal using a FFT routine, followed by summation of the components of the power spectrum to simulate a bank of twelve band-passs filters.

A set of preliminary investigations was performed using the individual HMM/VQ systems. Different codebook sizes were used and Table 2 shows the results.

Table 2 shows that among the individual recognizers the HMM/EM gave the best recognition accuracies, and the HMM/LBG the worst. This is consistent with our exspectation that the Gaussian mixture model is a good description of speech features which matches Hidden Markov Modelling very well.

Table 3 shows the results achieved by the multi-model system. Compared with the best single recognizer, HMM/EM, the MHMM/VQ system obtained 38.8%, 54.9%, and 53.1% reduction in the recognition error rates for the codebook sizes 32, 64, and 128 respectively.

Codebook size		32	64	128
Accuracy of recognition	LBG	86.18%	93.09%	94.33%
	MKF	87.78%	93.13%	95.34%
	EM	91.05%	93.59%	97.23%

Table 2: The results of individual HMM/VQ recognizers

Codebook size	32	64	128
Accuracy of recognition	94.53%	97.11%	98.70%

Table 3: The results of Multi-HMM/VQ recognizer

CONCLUSION

Consideration of the speech production process suggests that Gaussian Mixture Models offer a good description and the EM algorithm is an effective classification method for the first step modelling in a HMM/VQ system.

The multi-Hidden Markov Model speech recognizer gave much better recognition results. This was proved by the performances of combinations of any two of three sub-recognizers and combination of three of them together. The best results achieved by MHMM/VQ recognizer represent a reduction in the even rate of about 50 per cent in comparison to the HMM/EM recognizer, the best single recognizer.

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] 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.

[2] Linde, Y., Buzo, A., and Gray, R. M.(1980), "An Algorithm for Vector Quantizer Design", IEEE Trans. on Comm., COM-28(1), pp. 84-95.

[3] Rabiner, L. R.(1989), "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proc. IEEE, Vol. 77(20), pp. 257-286.

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