SPEAKER RECOGNITION USING CONTINUOUS ERGODIC HMMs

Shuping Ran, Bruce Millar, William Laverty, Iain Macleod, Michael Wagner and Xiaoyuan Zhu

TRUST Project
Research School of Information Sciences and Engineering
Australian National University

ABSTRACT

This paper aims to investigate the effect of training the transition probabilities of Continuous Ergodic Hidden Markov Models (CEHMMs) and the effect of the choice of number of states and number of mixtures for CEHMMs when using them in a speaker recognition task. Speaker recognition experiments with and without training the transition probabilities using different combination of number of states and mixtures were carried out. The length of training and testing utterances was from 0.3 to 2.5 seconds with an average length of about one second.

Using a different data set, the results confirm Matsui and Furui's finding that the total number of mixtures (i.e. the product of the number of states times the number of mixtures per state) is an important parameter in determining speaker recognition performance. Training of the transition probabilities did not improve the overall recognition rate. We suggest a possible explanation for our failure to derive any benefits here.

INTRODUCTION

Continuous Ergodic Hidden Markov Models (CEHMMs) (Figure 1) have become popular for speaker recognition. They were originally used to model time series, and have been applied to speech signals as one type of time series (Poritz, 1982). Although several researchers have reported good CEHMM performance when used for text-independent speaker recognition (Matsui and Furui, 1992; Savic and Gupta, 1990), many issues have not been fully understood or studied. For example:

- how useful are the transition probabilities between states for speaker recognition?
- in what ways is speaker recognition performance affected by choice of the number of states and the number of mixtures?

In a comparative study of Vector Quantisation and Hidden Markov Modelling for speaker recognition, Matsui and Furui (1992) set all transition probabilities for each state to be equal (summing to unity). They reported that speaker recognition performance was correlated with the total number of mixtures (i.e. the product of the number of states and the number of mixtures per state). From this result they concluded that transition probabilities were ineffective for text-independent speaker recognition. They used acoustic data from 36 speakers (23 male and 13 female). The length of their sentences was 4 seconds for both training and testing. Linear Predictive Cepstral Coefficients (LPCC) were used as speech features.

Intuitively, one would think that use of CEHMMs for speaker recognition essentially models the signals produced by each speaker as a sequence of intervals with differing internal properties. Transition probabilities should be thus form part of this modelling. This paper reports our experimental results using CEHMMs for speaker recognition, and addresses the two questions raised above. We set up different experimental conditions where, for each selection of the number of states and the number of mixtures per state, experiments both with and without training of transition probabilities were carried out. We used a different data set from Matsui and Furui's with much shorter utterances (see below).

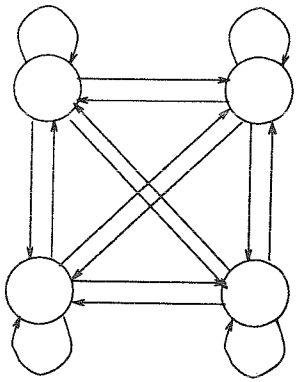

Figure 1: A Continuous Ergodic Hidden Markov Model of 4 states.

DATA CORPUS AND ANALYSIS

We used acoustic data from 24 Australian English speakers (12 male, 12 female). The utterances, typical of speaker commands to a computer, varied in length from 0.3 to 2.5 seconds with an average of about one second.

Data from two recording sessions (recorded one week apart) were used. There were 5 repetitions per speaker of each of the 30 utterances in each session.

These data were digitised at 20 KHz and then down-sampled to 10 KHz, after bandlimiting to 60 – 4800 Hz. Fourier analysis of order 8 was performed on each 25.6 ms frame with a 10 ms frame advance, providing 128 spectral magnitude coefficients. 20 mel-frequency cepstral coefficients (MFCCs) were computed from the 128 coefficients (Millar et al., 1994).

EXPERIMENTS

Two sets of experiments were conducted, setting the total number of mixtures to 8, 16, 32, and 64, varying the number of states and the number of mixtures per state. In one set of experiments, the transition probabilities were trained; in a second set, the transition probabilities were not trained and for each state all transition probabilities leaving that state (including the self probability) were set to be equal. Both sets of experiments were repeated by exchanging the training and testing data sets.

RESULTS

Tables 1 and 2 summarise the results, giving averaged recognition rates across all 24 speakers for each experimental condition. Table 1 reports the results of training the CEHMMs using the first data set, and testing with the second data set. Table 2 reports the results of training the CEHMMs using the second data set, and testing with the first data set.

Experimental conditions of 32 states and 64 states were not included because the utterances were too short to be usefully divided into such a large number of states.

Figure 2 reports the results of tests of statistical significance of the recognition rates for all speakers. Labels such as "Xs/Ym" mean X states with Y mixtures per state.

Total No of Mixtures	*No of States*	*No of Mixtures per state*	*Average Recognition Rate*	
			with Trained TP	*with Equal TP*
8	4	2	90.1	90.2
	8	1	89.2	88.7
16	4	4	92.9	93.9
	8	2	93.2	93.3
	16	1	92.7	92.3
32	4	8	95.7	95.4
	8	4	95.7	95.8
	16	2	96.0	95.6
64	4	16	96.7	96.5
	8	8	96.9	96.6
	16	4	96.4	96.4

Table 1: Results of training CEHMMs using the first data set and testing with the second data set. (TP means Transition Probabilities)

Total No of Mixtures	*No of States*	*No of Mixtures per state*	*Average Recognition Rate*	
			with Trained TP	*with Equal TP*
8	4	2	88.5	91.9
	8	1	89.7	89.7
16	4	4	94.7	94.6
	8	2	95.1	94.6
	16	1	94.3	94.4
32	4	8	96.6	96.7
	8	4	96.7	96.7
	16	2	96.4	96.1
64	4	16	97.3	97.6
	8	8	97.7	97.8
	16	4	97.9	97.6

Table 2: Results of training CEHMMs using the second data set and testing with the first data set.

STATISTICAL ANALYSIS OF THE RESULTS

Speaker recognition rates for all speakers were collected for all 11 experiments (ranging from 4 states with 2 mixtures per state to 16 states with 4 mixtures per state), both with and without training of the transition probabilities. The results were analysed statistically as summarised below.

Analysis was performed using a repeated measures design with three repeated measures factors:

- the number of states and mixtures (11 levels);
- the presence of transition probabilities (2 levels – with or without transition probabilities present);
- the session used as training data (2 levels – session 1 for training and session 2 for testing; session 2 for training and session 1 for testing).

Prior to analysis the data were transformed using the angular transformation ($\arcsin\sqrt{p}$) in order to stabilise variance (Scheffé, 1959).

Only the main effect of the States/Mixtures factor was significant. No other main effects or interactions were significant. The presence of transition probabilities thus had no effect on the recognition rate. As expected, there was no significant difference in recognition rate dependent on which session was used for training and which session was used for testing.

A post hoc comparison of the effect of combinations of states/mixtures on recognition rates was performed using Tukey's method of multiple comparison (Tukey, 1953). The following groups {4s/2m, 8s/1m}, {4s/4m, 8s/2m, 16s/1m} and {4s/8m, 8s/4m, 16s/2m, 4s/16m, 8s/8m, 16s/4m} were found to have no significant differences within groups; all comparisons between groups exhibited significant differences.

We examined the trained transition probabilities to see if any clear patterns emerged. The results showed that for each state, the probability of staying in that state (ranging from 65% to 90% or so) was typically much greater than any of the transition probabilities to other states. These latter transition probabilities varied from effectively zero up to about 25%. These trained probabilities show the expected pattern – once in a given state the most likely subsequent state is the current state, and certain state sequences will be much more likely than others – but intriguingly, they did not lead to any significant benefit (or even penalty) with the testing data.

DISCUSSION AND CONCLUSION

We have confirmed Matsui and Furui's finding that the total number of mixtures is an important system parameter when using HMMs for speaker recognition, using a speech data set with quite different characteristics from theirs. They found (as we also have) that for a given total number of mixtures, variations in the number of states and number of mixtures per state had little effect on the overall recognition performance, implying that variations in transition probabilities between states did not form an important component of the overall speech models for speaker recognition. They thus did not train these probabilities and set them to be equal.

In contrast to the above implication, we found that having trained the transition probabilities they varied widely, with the expected result that the self-transition probabilities were typically quite high. We did not expect, however, that having set varying transition probabilities during the training procedure, which presumably gave benefits with the training data set, these benefits would not carry over to the test data set.

Using our text-independent paradigm, the trained CEHMM speaker models with equal transition probabilities will allocate their states and mixtures in a manner somewhat akin to a mixture-Gaussian VQ codebook. In computing distortion measurements with VQ, the "transition probabilities" between codewords are equal – the best matching codeword is used for each successive frame independent of the currently selected codeword.

The process of determining the CEHMM state sequence which maximises the likelihood of observing a given input sequence is based on maximising the sum of the products of state output observation probabilities and state transition probabilities over all the possible state sequences. With a CEHMM with trained transition probabilities, the maximum likelihood will depend on both the output observation probabilities and the transition probabilities. Because the self-transition probabilities were much larger than those for transitions to different states, this will lead to the process staying in the current state for extended periods when successive frames have similar speech features. Therefore the obtained maximum likelihood will be dominated by self-transition probabilities and the output observation probabilities. With a CEHMM in which the transition probabilities are set to be equal, the maximum likelihood which best matches a given input will be determined purely on the basis of the state observation probabilities. Because our training and testing data sets used the same 30 utterances they should therefore have had similar overall statistics. This will again lead to the process staying in the current state for extended periods when the speech features are changing slowly. The calculated maximum likelihood will then be dominated by self-transition probabilities and the output observation probabilities. Therefore, if the encoded observation probabilities within corresponding states are similar, the state sequence which maximises the likelihood of observing the test utterance in both cases (equal and trained transition probabilities) should be similar, and hence the speaker recognition results should be similar too. In the next

phase of our research, we hope to be able to evaluate this suggested explanation.

ACKNOWLEDGEMENT

This research has been carried out on behalf of the Harry Triguboff AM Research Syndicate.

REFERENCES

Matsui, T. and Furui, S. (1992), "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs", Proc. of ICASSP, pp.II-157–II-160.

Millar, B., Chen, F. Macleod, I. Ran, S. Tang, H. Wagner, M., and Zhu, X. (1994), "Overview of speaker verification studies towards technology for robust user-conscious secure transactions", Proc. of SST-94.

Poritz, A. B. (1982), "Linear predictive hidden Markov Models and the speech signal", Proc. of ICASSP, pp.281–284.

Savic, M. and Gupta, S. K. (1990), "Variable parameter speaker verification system based on hidden Markov modeling", Proc. of ICASSP, pp.281–284.

Scheffé, H. (1959), The analysis of variance, John Wiley, New York, pp.364–368.

Tukey, J. W. (1953), "The problem of multiple comparison", roneod MS of 395pp, Princeton University.

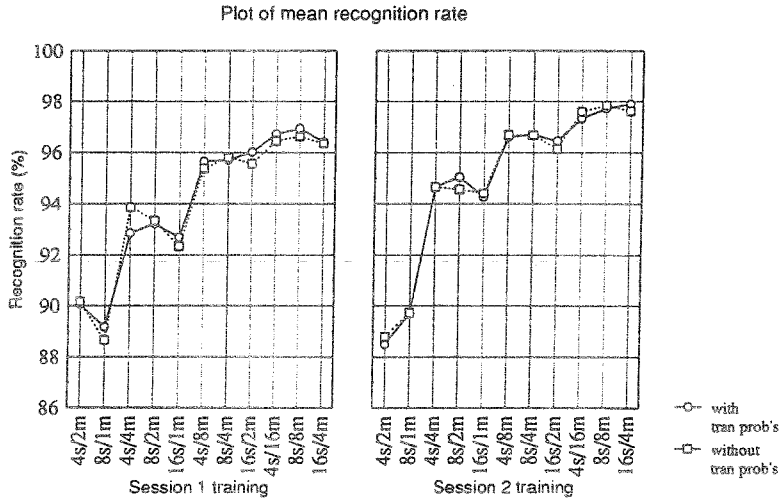

Figure 2: Results of training CEHMMs using the first data set and testing with the second data set; and training CEHMMs using the second set and testing with the first data set.