

Text-Independent Speaker Recognition Using VQ, Mixture Gaussian VQ and Ergodic HMMs

Xiaoyuan Zhu, Yuqing Gao, Shuping Ran, Fangxin Chen,
Iain Macleod, Bruce Millar and Michael Wagner

Abstract—Alternative techniques are evaluated for text independent speaker recognition in a speech activated menu navigation task, typical of windows-based interactive computing. Even though the vocabulary employed may be relatively small, ease of management in the target application makes text independence highly desirable. The main techniques studied were weighted and unweighted vector quantisation, mixture Gaussian VQ and ergodic continuous hidden Markov models (CHMM). Data from 25 speakers was acquired in several sessions, with five repetitions of each utterance in each session and an inter-session interval of one or more weeks. The overall results with between session training/test data showed that unweighted conventional VQ was inferior to variance weighted VQ, mixture Gaussian VQ and CHMM. The latter three techniques gave similar performances, achieving a recognition accuracy of about 97 to 98% with utterances from the training vocabulary. Short utterances from outside the training vocabulary gave a recognition accuracy of approximately 93%.

Keywords—Speaker recognition, vector quantisation, hidden Markov models.

1. INTRODUCTION

This study investigates the performance of a range of vector quantisation (VQ) based methods for text independent speaker identification during a menu navigation task typical of interactive computing. The length of typical spoken commands in this context ranges from about half to one second.

The target application for the technology we are developing is that of improved security of access to sensitive data. Although text dependent speaker recognition techniques have potential for greater recognition accuracy [1], there are several management and design advantages in the application area which make text independent speaker recognition quite attractive. The information used to characterise each of a large number of users (speakers) is likely to be stored on a remote secure data base. Text independent information can be relatively concise (of no more than a few thousand bytes per user) and predictable in size. Enrolment of users can also be a “once off” operation which does not need to be repeated as the command vocabulary for various computing tasks changes.

Two different modes of usage are envisaged — implicit and explicit assertion of identity. In the former mode, which is the main focus of the current study, users continually assert their identity as they interact with the computer system in a normal task-directed manner. In the

The authors are with the TRUST (Technology for Robust User-conscious Secure Transactions) Project at the Australian National University, Canberra, ACT 0200, with the exception of Dr Yuqing Gao who is now with Apple-ISS Research Center, ISS, National University of Singapore.

latter, which may for example be invoked at log-on time or prior to completion of an operation with high security requirements, users would read aloud a phrase displayed on the monitor screen (chosen at random from a set of phrases with good ability to discriminate between speakers).

The present study examines the performance of various methods within a speaker identification (recognition) paradigm. The envisaged target application involves a closed (but possibly large) set of registered legitimate users and an effectively open set of potential impostors who are to be prevented from accessing sensitive data. While a full evaluation of the level of performance achievable in such an application requires a speaker verification paradigm, which in turn depends on a determination of the characteristics of the test speakers in relation to the total population, the recognition paradigm employed here provides a valid basis for comparing the relative performance of different methods.

2. METHOD

A training vocabulary for our experiments was chosen according to the following criteria: the individual utterances should be representative of the type of spoken command which might be employed in interactive menu navigation (and other simple computing tasks) within a graphical user environment; the set of utterances should include a substantially complete phonemic inventory of Australian English (taking into account relative frequencies of occurrence); and speakers should give the training utterances (as displayed on a monitor screen) consistent pronunciation and inflection. A series of pilot experiments was conducted to refine the data gathering procedure and to identify any problem utterances.

The training vocabulary finally adopted consisted of 30 utterances — mainly one or two word commands (eg. “open file,” “help index” and “search”), plus several longer utterances such as might be used for explicit identity checking. Speech data from 25 speakers was acquired in three sessions, with five repetitions of each utterance in each session. The intervals from the first to second and second to third sessions were about one and three weeks respectively. The second and third sessions included two repetitions each of 10 additional utterances to be used for evaluating text independence.

The set of 25 test speakers comprised 12 males and 13 females, of varying ages and predominantly speakers of general or cultivated Australian English. Training data sets were recorded via an interactive presentation and collec-

tion procedure controlled by the speakers. Subsequent automatic processing located start and end points for each utterance; the resulting speech segments were checked visually and auditorily and modified where necessary to maintain consistent and accurate end-pointing.

Four different methods were evaluated for text independent speaker recognition: conventional VQ, which used Euclidean distances between input frames and codebook vectors for both training and testing; variance-weighted VQ, in which distances were weighted on a per-codebook basis according to observed variances in the training data; mixture Gaussian VQ (equivalent to a single-state CHMM); and multi-state CHMM. In accordance with earlier research [2], all methods used cepstral coefficients as input data vectors based on mel-frequency analysis (order 20, 10kHz sampling rate, with frames of 25.6ms and 61% overlap). Various codebook sizes and numbers of mixtures were evaluated, to ascertain appropriate parameters for each method, with training and test data generally being taken from different sessions. The models were trained on the complete data for each speaker in the training session (excluding the 10 utterances designed to assess the extent of text independence). Each identification test was based on a single short utterance (as short as 0.3 second, the average duration of the 170 test stimuli from each speaker was about 1.1 seconds).

3. RESULTS

Initial studies with training and test data from different sessions showed that unweighted conventional VQ was consistently poorer than the other methods. This lack of robustness is in part likely to be a consequence of the conventional VQ method itself with its non-parametric set of prototype vectors, which have limited ability to represent speaker-specific patterns. Further, the partition operation in the acoustic space of conventional VQ is a strict division into cells, which will aggravate quantisation errors. In contrast to conventional VQ's nonparametric representations, variance weighted and mixture Gaussian VQ (together with its multi-state extension to CHMM) use parameterised representations which take account of distributions in the data, and thus have theoretical advantages compared with conventional VQ. These representations can be regarded as acoustic prototypes which help reduce the impact of partition errors. Our tests suggested that these theoretical advantages are reflected in practice: in one case, a recognition accuracy of 94.4% with conventional VQ increased to 98.1% with mixture Gaussian VQ. As a result of its theoretical limitations and poorer performance, we omitted conventional VQ from subsequent experiments.

In determining appropriate codebook sizes for variance weighted VQ and numbers of mixtures for mixture Gaussian VQ, we need to give the models sufficient complexity to represent most of a given speaker's speech with reasonable accuracy. At the same time, we need to keep the number of states in the model small enough that other speakers are not modelled well. Figure 1 shows the recognition accuracy obtained as a function of codebook size for variance weighted VQ and number of mixtures for mixture Gaussian

VQ, training and testing with data from sessions one and two respectively. The complexity of the relevant model is best measured in terms of the number of free parameters rather than codebook size or the number of mixtures as such. In our case, mixture Gaussian VQ has a little more than twice as many free parameters as variance weighted VQ when the number of mixtures is equal to the size of the codebook.

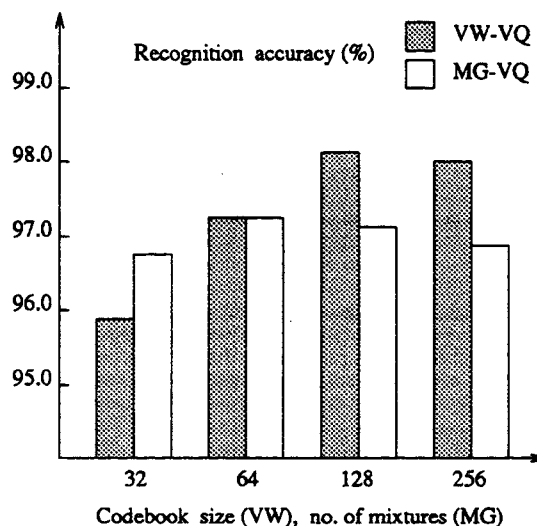


Fig. 1. Recognition performance versus model complexity

Allowing for this difference in the number of free parameters, we found that the performance of the variance weighted and mixture Gaussian VQ methods was comparable. Figure 1 shows the results for training on session one and testing on session two. We conducted a further experiment to see if less complex models gave better results when the time between acquisition of the training and test data was longer, given the possibility that more complex models may encode details which are less stable over time. Figure 2 compares the performance of variance weighted VQ models of differing complexities, training/testing with data from sessions one/two and one/three respectively. The previously observed decrease in recognition accuracy with time [3] is evident. Contrary to our expectations, the recognition accuracy of the session one/three case improved as the model complexity increased. We are conducting further experiments in an attempt to account for this behaviour.

The influence of test utterance length on recognition performance was the next question investigated. Previous research [4] suggests that best results are obtained with utterances of 5 seconds or longer. Such a length is impractical in our application, where a maximum command length is dictated by considerations of user acceptability and efficiency. Figure 3 shows recognition accuracy versus utterance length for variance weighted VQ with a codebook size of 128; the training and test data were from sessions two and three respectively. Fortunately, our experimental results suggest that the majority of the gain in recognition accuracy with length has already occurred with utterances of one to one and a half seconds in length — a more accept-

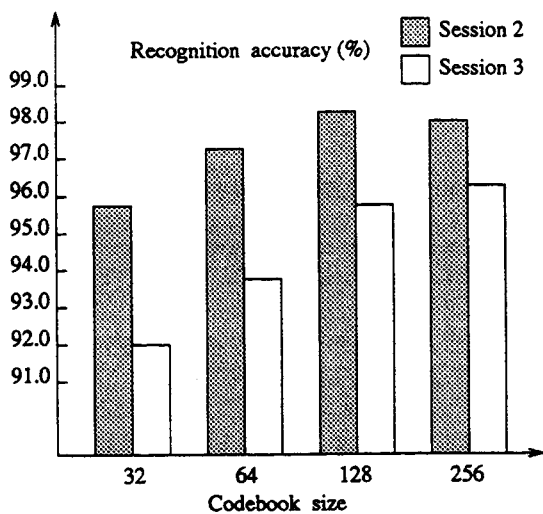


Fig. 2. Evaluation with longer time interval

able value for spoken commands. It appears that the task constraints involved with spoken commands of the type we investigated act to improve recognition accuracy with short utterances (the overall stress contour with such commands, for example, is relatively well defined).

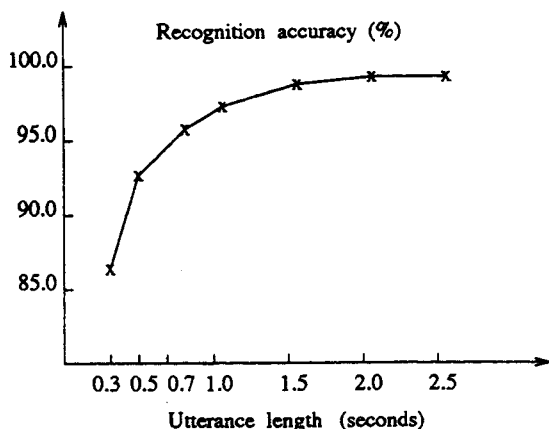


Fig. 3. Recognition accuracy versus utterance length

A fourth series of experiments attempted to assess the degree to which our speaker identification methods were text independent, by using different utterances for training and testing. The VQ model used and training and test data sessions were as for the previous experiment. The 10 additional utterances recorded for assessing text independence were of the type used in implicit assertion of identity (ie. typical of spoken commands). They were thus on average shorter than the main set of 30 utterances, which included several longer examples of a type suitable for explicit assertion of identity. To remove the confounding effect of improving accuracy with utterance length shown in Figure 3, we tested our model on utterances of similar length. On the basis of rather limited data, Figure 4 contrasts the recognition accuracies obtained for test utterances inside and outside the training set. While a distinct difference is

observable, the overall performance on utterances outside the training set is still quite good, suggesting that the VQ models are capturing rather general aspects of speakers' vocal tract characteristics and speech behaviour in addition to certain details pertaining to the training set. While our training data sets included a complete inventory of Australian English phonemes, they did not include examples of each phoneme in a representative range of phonetic contexts and stress levels. Some dependence on the training set of utterances is thus expected.

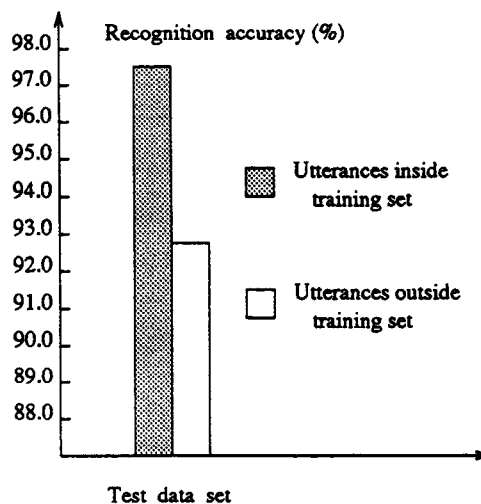


Fig. 4. Assessment of text independence

A final set of experiments investigated the question of whether better recognition accuracies would be obtained by using a single state model with a larger number of mixtures or a multi-state model in which each state had a smaller number of mixtures, keeping the overall model complexity constant (as measured by the product of the number of states and the number of mixtures). The models used here were ergodic CHMMs (ie. multi-state versions of mixture Gaussian VQ). Ergodic models have been shown to be effective in this area [5].

Research at NTT [6] has shown that (given sufficient training data) the speaker identification rate increases with the total number of mixtures (the number of states times the number of mixtures assigned to each state) and is independent of the number of states. Since complex CHMMs can only be trained well with substantial amounts of training data, the number of states and mixtures usable in practice is task dependent.

In our experiments we fixed the total number of mixtures at 64 (according to the good results shown in Figure 1 for mixture Gaussian VQ with this number of mixtures) and varied the number of states and number of mixtures from 1 by 64 through to 64 by 1. The ergodic CHMMs were trained via a simplified segmental k-means training procedure [7]. Our expectation was that some tradeoff of number of mixtures for number of states would yield a model which captured simple aspects of speech dynamics while remaining within a text independent framework. As shown in Figure 5, however, it was not clear that multi-state mod-

els had any overall advantage compared with a single state mixture Gaussian VQ model of equal complexity; any gains through modelling of dynamics appear to have been offset by the reduced resolution for each state.

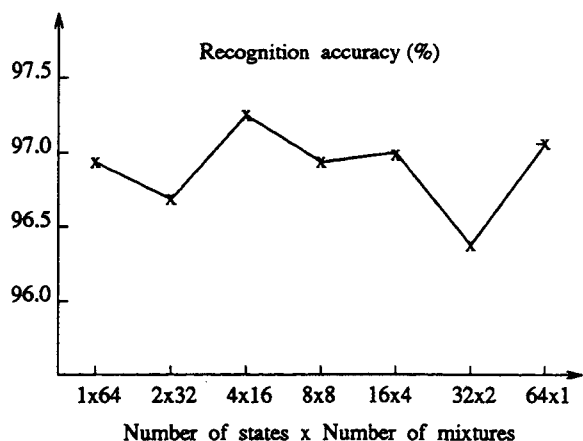


Fig. 5. Performance of multi-state models

4. DISCUSSION

Our experiments suggest that useful information regarding a computer user's identity can be gained from spoken commands of about one second in length, such as might be used for certain common operations performed within a graphical user environment. Half a second is too short, but the gain in recognition accuracy with utterance length is relatively small after one second. The results obtained in these experiments indicate that the task constraints improve the recognition accuracy with short utterances.

Variance weighted and mixture Gaussian VQ were both clearly superior to conventional unweighted VQ in our application. We have yet to establish whether one or other of

these enhanced forms of VQ gives consistently better performance in relation to speech acquired under operational rather than experimental conditions.

The experiments with multi-state CHMM models depicted in Figure 5 tend to support Matsui and Furui's finding [6] that the product of the number of states and number of mixtures is an important parameter in determining overall performance. In the next phase of our work (where we will have more extensive training data available), we will study multi-state models with greater complexity to see if they can capture elementary aspects of speech dynamics, leading to better speaker discrimination.

ACKNOWLEDGEMENT

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

REFERENCES

- [1] J.M. Naik, "Speaker verification: A tutorial," *IEEE Communications Magazine*, pp.42-48, Jan. 1990.
- [2] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on ASSP*, Vol.29, pp.254-272, Apr. 1981.
- [3] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features," *IEEE Transactions on ASSP*, Vol.29, pp.342-350, Jun. 1981.
- [4] D. O'Shaughnessy, *Speech Communication: Human and Machine*, Addison-Wesley, 1987.
- [5] M. Savic and S. K. Gupta, "Variable parameter speaker verification system based on hidden Markov modeling," *Proceedings of ICASSP90*, pp.281-284, New Mexico, Apr. 1990.
- [6] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs," *Proceedings of ICASSP92*, pp.II-157-II-160, San Francisco, Mar. 1992.
- [7] X. Zhu, "A combined neural network and hidden Markov model approach for speaker recognition," *Proceedings of The 1993 IEEE Region 10 International Conference on Computers, Communications, Control and Power Engineering*, Vol.2, pp.1074-1077, Beijing, Oct. 1993.