COMBINED SPEECH-RECOGNITION/SPEAKER-VERIFICATION SYSTEM WITH MODEST TRAINING REQUIREMENTS

Michael Wagner

School of Computing
University of Canberra

ABSTRACT - This study investigates a combined speech recognition and speaker verification system which performs well under conditions of client training being restricted to only a few repetitions per utterance. The system employs vector quantisation and discrete hidden Markov models such as to make use of VQ codeword indices for word recognition and the corresponding VQ distortions for speaker verification. Different codebook sizes are investigated on a speech corpus of 10 computer command words spoken by 8 male and 8 female speakers. At an optimum codebook size of 64, the system performs at word error rates of about 4 percent and at speaker verification equal-error rates of about 9 percent for single-word utterances. If up to 4 or 5 consecutive words are accumulated for speaker verification, the equal-error rates of the system fall to about 4 percent.

INTRODUCTION

The authentication of computer users is a problem which has come increasingly into focus as computers, networks and databases encompass ever more widening aspects of everyday life. In the vast majority of computer systems access is controlled by the typing of a password even though the level of security that is afforded by password-based user authentication is known to be modest. On the other hand, the advent of multimedia extensions to many computer systems and especially the inclusion of sound input and output facilities in most current personal computers offers the possibility of using speaker verification techniques as a means of authenticating the users of such systems.

Speaker verification systems have previously been proposed either in an entry control paradigm, e.g. by Furui (1981) or Soong et al. (1987) or in a continuous-access control paradigm (Millar et al., 1994). In entry control systems, users are required to speak the equivalent of a password before being granted access to the system. Because such systems are open to attack by recorded versions of clients' passwords, it is desirable either to train a number of possible passwords for each client or to have sufficient training data for text-independent client models. In the first case, the system prompts the user for one of the trained passwords, thus reducing the likelihood that an impostor can produce a recorded version of that particular password. In the second case the system prompts for an entirely new word which the client has never spoken before. Both these strategies, however, necessitate extensive training of the speaker verification system by the client.

The continuous access control paradigm assumes that speech-based computer user authentication is likely to be employed in systems that also make use of speech technology for the task-related aspect of the human-computer interaction, typically by using automatic speech recognition technology to enter commands, requests and data. For such systems it is an obvious option to employ a combined speech-recognition/speaker-verification system which recognises and acts upon the task-related spoken input by the user and simultaneously evaluates each spoken input with a view to verifying the identity of the user (Wagner et al., 1995).

This study proposes a paradigm which asks users at the commencement of a session to identify themselves, either by conventional login or by other means, and will then accept a number of spoken computer commands. The speech input can therefore be processed by means of a speaker-dependent speech recognition system and, following successful recognition, by an utterance-dependent speaker verification system.

In practical computer user authentication applications, it is important to minimise the need for clients to provide extensive training data to the speech recognition and speaker verification systems. It is therefore investigated how well a baseline system can perform which is constrained to having only few

repetitions of the spoken vocabulary available for the training of each user and word model. The speech corpus used in this study provides 10 training repetitions of a set of 10 computer command words, a task which a newly enrolling client could undertake relatively quickly.

THE COMBINED SPEECH RECOGNITION/SPEAKER VERIFICATION SYSTEM

The design goal for the combined speech recognition and speaker verification system is to achieve low word recognition error rates as well as low false rejection and false acceptance error rates for the clients of the system under limited-training conditions. Since a limited amount of training data necessitates a small number of model parameters, the proposed system is based on discrete hidden Markov models (DHMM) which can be represented with fewer model parameters than comparative continuous or semicontinuous hidden Markov models. Furthermore, performing VQ on the input speech offers the possibility of using the vector codes for the recognition of the utterance by means of DHMM likelihoods and using the distortions for the verification of the speaker by means of a standard VQ distortion based distance measure (Chen et al., 1994). Such a system provides an efficient combination of the tasks of speech recognition and speaker verification as the VQ process provides the input data for both tasks simultaneously. As such it is suitable for implementation on standard personal computers without special digital signal processing hardware and without compromising real-time or near real-time recognition and verification performance.

This paper addresses the question whether both satisfactory word recognition and satisfactory speaker verification can be achieved with such an algorithm. Of particular interest is the optimum size of the VQ codebook for such a system. Previous work in speaker verification indicates that a large codebook is necessary for small speaker verification error rates while the performance of the word recognition can be expected to deteriorate for codebook sizes which are too large to allow for sufficient training of the DHMM emission probabilities.

SPEECH DATA

The system has been simulated and evaluated with a set of computer commands from the commercially available TI46 speech data corpus. The TI46 corpus [ref] contains 46 utterances spoken repeatedly by 8 female and 8 male speakers, labelled f1 - f8 and m1 - m8. The vocabulary contains a set of 10 computer commands, shown in Table 1, which was used in the current study. Each speaker repeated the words 10 times in a single training session, and then again twice in each of 8 later testing sessions. The corpus is sampled at 12500 samples/s and 12 bits/sample. The data were processed in 20.48ms frames (256 samples width) at a frame rate of 125 frames/s (100 samples shift). Frames were Hamming windowed and preemphasised with μ=0.9. 46 mel-spectral bands of a width of 110mel and 20 mel-frequency cepstral coefficients (MFCC) were determined for each frame.

1. Enter	6. Rubout
2. Erase	7. Repeat
3. Go	8. Stop
4. Help	9. Start
5. No	10. Yes

Table 1. Computer Commands

SYSTEM TRAINING

In the first training step, each speaker's 100 training tokens (10 utterances $\times$ 1 training session $\times$ 10 repetitions) were used to train the speaker's VQ codebook by clustering the set of all of the speaker's MFCCs into codebooks of either 32, 64 or 128 codewords (Linde et al., 1980). In the second training step, the 10 repetitions of each utterance by each speaker were used to train 160 six-state left-to-right DHMMs, using the forward-backward and Baum-Welch algorithms.

WORD RECOGNITION RESULTS

Under the assumption that users would identify themselves at the commencement of a session, the word recognition system was tested in speaker-dependent mode. Each speaker's 160 test tokens (10 utterances × 8 testing sessions × 2 repetitions) were tested against that speaker's 10 word models. The mean word recognition error rates for the 16 speakers and for the 3 different codebook sizes are shown in Table 1.

Speaker	M=32	M=64	M=128
f1	22.50%	5.63%	2.50%
f2	18.13%	2.50%	7.50%
f3	9.38%	4.38%	3.75%
f4	14.38%	0.63%	0.00%
f5	5.63%	6.25%	1.25%
f6	5.00%	6.88%	1.88%
f7	10.00%	4.38%	2.50%
f8	25.63%	10.00%	2.50%
m1	27.63%	2.63%	6.58%
m2	34.38%	2.50%	0.63%
m3	34.39%	1.27%	19.11%
m4	14.19%	1.29%	1.29%
m5	29.30%	13.38%	12.10%
m6	52.83%	5.03%	40.88%
m7	23.13%	0.63%	1.25%
m8	23.13%	2.50%	0.63%
FEMALE	13.83%	5.08%	2.73%
MALE	29.92%	3.65%	10.32%
TOTAL	21.81%	4.37%	6.50%

Table 1. Average word recognition error rates for the 16 speakers and 3 codebook sizes.

The table shows that a VQ codebook size of 32 is clearly insufficient for the recognition of the chosen 10-word vocabulary. Error rates for this codebook size average at 13.83 percent for the female speakers and 29.92 percent for the male speakers with an overall error rate of 21.81%. For a codebook size of 64, the error rates average 5.08% for females and 3.65 for males with an overall error rate of 4.37 percent. When the codebook size is once again doubled to 128, the error rate for females falls further to 2.73 percent while that for males rises to 10.32 percent. The overall error rate is slightly higher at 6.50 percent than the error rate for a codebook size of 64.

These results show firstly that quite reasonable speaker dependent word recognition can be achieved with very limited training data. Secondly, they show that a codebook size of 64 is optimal for the conditions of this experiment, with smaller codebooks being unable to encode the phonetic richness of even a small vocabulary and with larger codebooks being unable to be trained satisfactorily with the limited amount of training data. The results of this experiment also suggest that the optimal codebook size may differ between female and male speakers.

SPEAKER VERIFICATION RESULTS

Speaker verification used the VQ distortions resulting from the above word recognition experiment. The paradigm assumed a successful, that is error-free, word recognition. Acceptance/rejection thresholds were determined a posteriori for each speaker and each of the 10 utterances. Only the 7 same-sex impostors were compared with each client model. As in all verification algorithms, it is possible to assess performance in a variety of ways which give different weightings to the false acceptance and false rejection errors. For the purpose of this study, the performance of the algorithm is measured by the equal-error rate. Table 2 shows the mean equal-error rates for the 16 speakers for the codebook sizes of 32, 64 and 128.

Speaker	M=32	M=64	M=128
f1	9.68%	9.19%	8.13%
f2	7.00%	3.88%	1.92%
f3	17.65%	16.25%	14.59%
f4	3.03%	2.43%	2.42%
f5	7.36%	7.23%	6.05%
f6	6.35%	3.63%	3.25%
f7	13.89%	11.89%	10.28%
f8	15.62%	10.97%	10.66%
m1	8.44%	9.29%	8.38%
m2	6.88%	3.77%	3.08%
m3	16.54%	13.40%	10.95%
m4	7.96%	6.31%	6.13%
m5	7.25%	6.14%	6.24%
m6	13.17%	11.80%	10.26%
m7	18.34%	15.64%	15.06%
m8	16.22%	15.65%	14.00%
FEMALE	10.07%	8.18%	7.16%
MALE	11.85%	10.25%	9.26%
TOTAL	10.96%	9.22%	8.21%

Table 2. Average speaker verification equal-error rates for the 16 speakers and 3 codebook sizes.

As expected, equal-error rates for the speaker verification experiment fall with increasing codebook size for both female and male speakers as codebooks of the sizes investigated here are barely sufficient to adequately cover the feature space which has 20 dimensions in the current experiment. It could therefore be surmised that by further enlarging the codebook one would obtain better speaker verification rates until eventually the effect of insufficient training would reverse the advantages of larger codebooks. The results of Table 2 do not suggest that there may be a different optimal codebook size for female and male speakers.

SPEAKER VERIFICATION USING LONGER UTTERANCES

Since previous research has indicated that speaker verification errors decrease proportionally to the length of the utterance over which frame distortions are averaged, speaker verification was repeated with combinations of 2, 3, 4 and 5 words as shown in Table 3.

1. Enter-erase	1. Enter-erase-go
2. Go-help	2. Help-no-rubout
3. No-rubout	3. Repeat-stop-start
4. Repeat-stop	
5. Start-yes	
1. Enter-erase-go-help	1.Enter-erase-go-help-no
2. No-rubout-repeat-stop	2. Rubout-repeat-stop-start-yes

Table 3. a) 2-word combinations, b) 3-word combinations, c) 4-word combinations, d) 5-word combinations

The resulting equal-error rates, averaged over all 16 speakers are shown in Table 4. The first row shows the results from the previous table as pertaining to an utterance length of 1 word. The following 4 rows show the average equal-error rates for the combinations of 2, 3, 4 and 5 words respectively.

The results confirm that longer utterances are advantageous for mean-distortion based speaker verification algorithms. For the smallest codebook size, verification errors decrease as the utterance length is raised from 1 to 5 words. For the codebook sizes of 64 and 128, the equal-error rates decrease up to an utterance length of 4 words and fall very slightly for an utterance length of 5 words. Clearly, speaker verification performance increases with longer utterance lengths which needs to be taken account in the design of spoken-password based entry systems. For continuous access control systems as the one proposed in this paper, speaker verification error rates of about 8% can be expected for single computer command words, but verification performance can be improved by accumulating the mean-distortion measure over several consecutive utterances.

UttLength	M=32	M=64	M=128
1 word	10.96%	9.22%	8.21%
2 words	8.73%	7.14%	6.76%
3 words	6.29%	5.30%	4.62%
4 words	5.34%	4.48%	3.77%
5 words	5.12%	4.99%	3.97%

Table 4. Speaker verification equal-error rates for different utterance lengths and codebook sizes.

It may be reasonably assumed that the use of background speakers would significantly improve the speaker verification results. For example, Chen et al. have reported experiments on a 37-speaker corpus that average equal-error rates improved from 5.5 percent to 3.6 percent when a cohort of 5 background speakers was used, and further to 1.3 percent with a hybrid cohort paradigm, thus cutting equal-error rates by a factor of about 4. Due to the limited number of speakers in the TI46 corpus, these algorithms could not be explored in the current experiment.

CONCLUSION

This study has proposed a combined speech recognition and speaker verification system, based on vector quantisation and discrete hidden Markov models, which is designed to cope with only a small number of training repetitions of the vocabulary of spoken computer commands. The system performs efficiently by using both the codeword indices and distortions of the VQ process. At an optimum codebook size of 64, the system is capable of performing at word recognition rates of about 4 percent

and at speaker verification equal-error rates of about 8 percent for single-word utterances. If up to 4 or 5 consecutive words are accumulated for speaker verification, the equal-error rates of the system fall to about 4 percent.

REFERENCES

F. Chen, J.B. Millar, M. Wagner (1994) Hybrid-threshold approach in text-independent speaker verification,. Proc Int Conf on Spoken Lang Proc, 1855-1858.

S. Furui (1981) Cepstral analysis techniques for automatic speaker verification, IEEE Trans ASSP-29, 254-272.

Y. Linde, A. Buzo, R.M. Gray (1980) An algorithm for vector quantization, IEEE Trans COM-28, 84-95.

J.B. Millar, F. Chen, I. Macleod, S. Ran, H. Tang, M. Wagner, X. Zhu (1994) Overview of speaker verification studies towards technology for robust user-conscious secure transactions. Proc. 5th Austr. Int. Conf on Speech Sci & Tech, 744-749.

F.K. Soong, A.E. Rosenberg, B.H. Juang (1987) A vector quantization approach to speaker recognition, AT&T Tech J., 66, 14-26.

M Wagner et al. (1995) Provisional Patent, Trust Project, Australian National University.