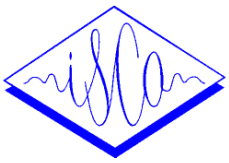

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

SPEAKER IDENTIFICATION USING VECTOR QUANTISATION WITH CODEWORD-SPECIFIC DERIVATIVE CODING

Michael Wagner, John S. Mason[1], J. Bruce Millar
e-mail: miw@trust.anu.edu.au
Trust Project[2], Research School of Information Sciences and Engineering,
Australian National University
Canberra ACT 0200
AUSTRALIA

ABSTRACT

This paper investigates improvements to the vector quantisation (VQ) distortion method of text-independent speaker identification, using a conventional codebook of instantaneous cepstral vectors from each speaker's training data, and one second-level codebook of transitional cepstral vectors for each codeword of the instantaneous codebook. Results on a 20-speaker database of 30 phonetically rich utterances show a reduction of the error rate from 6.5% for a conventional codebook of size 128 to 5.5% for a codebook which contains 16 transitional codewords for each of the 128 instantaneous codewords (128×16). Results on a 20-speaker database of spoken digits show a reduction of error rate from 3.1% for a conventional (128×0)-codebook to 0.9% for a (128×4)-codebook. Alternatively, a constant error rate can be maintained at a reduced number of codeword comparisons using codeword-specific transitional codebooks. Results also show that, given a sufficient size of transitional codebook, transitional distortion scores after instantaneous preclassification can be superior to purely instantaneous distortion scores.

1. INTRODUCTION

Previous research has shown that average VQ distortion scores can be successfully employed in text-independent speaker identification [1], [2], [3]. Both instantaneous and transitional cepstral information have been found useful for speaker discrimination [4] and both contribute relatively independently to the performance of text-independent speaker identification systems [5]. The question arises whether the transitional cepstral information can be represented by a codebook of instantaneous cepstral vectors with a second-level codebook of transitional vectors for each codeword of the instantaneous codebook. In this case, the two-level codebook represents the likelihood of specific directions of movement being associated with particular regions of the cepstral space.

The objective of the study is twofold: Firstly, it is hypothesised that speaker identification results will improve when using a second-level codebook of transitional vectors as compared to a codebook of instantaneous vectors. Secondly, it is hypothesised that equivalent speaker identification performance is achieved by using a smaller two-level codebook than would be possible with a single-level codebook, thus achieving a saving in the number of codeword comparisons necessary for each cepstral vector tested.

2. FIRST EXPERIMENT

2.1. Data

For Experiment 1, the first 20 speakers of the Trust corpus [6] were used. The 30 utterances recorded in that corpus are listed in Table 1. The corpus contains three sessions recorded at approximately weekly intervals with each session comprising five repetitions of each of the utterances. The VQ codebooks were trained using the ten repetitions of the commands recorded in sessions 1 and 2. The test data comprised the five repetitions of the commands recorded in session 3. The data were digitised at 10,000 samples/s and 16 bits/sample, and 20 mel-frequency cepstral coefficients were computed for 25.6ms frames with a 10ms frame advance. Transitional cepstra were determined by taking the cepstral differences between frames which were ±50ms removed from the target frame, except for the first five and last five frames of the utterance where transitional cepstra were extrapolated from the available frames. Each speech frame was therefore represented by one instantaneous cepstral vector and one transitional cepstral vector.

2.2. Method

Conventional instantaneous codebooks of sizes C=16, 32, 64 and 128 were constructed from the instantaneous cepstral vectors of the training utterances using the LBG algorithm [7]. In a second scan of the training data, the transitional cepstral vectors of the training utterances were divided into C subsets, each

[1] Dr Mason was on leave from the University of Wales, Swansea, U.K.

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

10.21437/Eurospeech.1995-101

transitional vector being assigned to that subset whose index c ($1 \leq c \leq C$) is the code of the corresponding instantaneous vector. Each of these C sets of transitional cepstral vectors was then clustered into transitional codebooks of sizes D=2, 4, 8 and 16.

1. My name is ...	
2. My office phone number is 249-6898	
3. The system access number is 175-093	
4. New window	
5. Line number	6. Grammar check
7. Send mail	8. Help index
9. Print preview	10. Clear screen
11. Edit file	12. Show ruler
13. Change font type	14. Read only
15. Open footer	16. Measure width
17. Open data manager	18. Page set up
19. View preferences	20. Save file
21. Resume task	22. Print merge
23. Fill down	24. Move cursor
25. Text format	26. Undo typing
27. Append text	28. Row height
29. Insert table	30. Compare with

Table 1. List of 30-word Trust Speech Corpus.

During the testing of new speech frames against the trained codebooks, the VQ distortion of each frame against the combined codebook was determined by first finding the closest codeword of the instantaneous codebook with respect to the instantaneous cepstral vector, yielding an index c ($1 \leq c \leq C$) and distance γ. The closest codeword of the c-th transitional codebook with respect to the transitional cepstral vector was then found, yielding an index d ($1 \leq d \leq D$) and distance δ. The instantaneous and transitional distortions γ and δ were then combined according to a weighting factor α which was varied between 0 and 1.

2.3. Results

The speaker identification results for the 30 phonetically rich utterances are shown in Tables 2 and 3.

The rightmost column (α=1.0) of Table 2 shows the baseline error rate for speaker identification based on instantaneous information only. Increasing the instantaneous codebook size from 16 to 128, reduces the error rate from 13.6% to 6.5%.

The leftmost column of Table 2 (α=0.0) shows that the value of the transitional VQ distortion measure increases strongly with increasing transitional codebook size. It is noteworthy that in all cases, after preclassification with a conventional instantaneous codebook, the sole use of the transitional distortion (α=0.0) is superior to the sole use of the instantaneous distortion (α=1.0) as long as the transitional codebooks are at least of size 16 (compare [5]).

16	**0.00**	**0.20**	**0.40**	**0.60**	**0.80**	**1.00**
2	42.96	16.03	14.83	13.94	13.66	13.59
4	23.90	13.87	13.45	13.45	13.52	13.59
8	15.18	11.97	12.32	12.88	13.31	13.59
16	10.06	10.73	11.75	12.42	13.13	13.59
32	**0.00**	**0.20**	**0.40**	**0.60**	**0.80**	**1.00**
2	34.91	13.41	11.93	11.40	11.05	10.73
4	18.78	10.62	10.80	10.91	10.94	10.73
8	12.46	9.46	10.13	10.45	10.70	10.73
16	8.93	8.01	9.32	10.10	10.55	10.73
64	**0.00**	**0.20**	**0.40**	**0.60**	**0.80**	**1.00**
2	26.19	9.85	8.51	8.08	7.91	7.80
4	13.80	8.19	7.70	7.73	7.70	7.80
8	9.07	6.71	7.17	7.45	7.62	7.80
16	6.95	6.25	6.85	7.45	7.59	7.80
128	**0.00**	**0.20**	**0.40**	**0.60**	**0.80**	**1.00**
2	17.68	7.66	7.09	6.78	6.64	6.46
4	9.99	6.67	6.64	6.42	6.57	6.46
8	7.41	6.07	6.11	6.35	6.32	6.46
16	5.58	5.54	5.97	6.25	6.25	6.46

Table 2. Percentage error rates for the phonetically rich data for instantaneous codebook sizes C=16, 32, 64 and 128, transitional codebook sizes D=2, 4, 8 and 16, and with combination weights α ranging from 0.0 to 1.0.

Table 2 also shows that with instantaneous and transitional codebook sizes of at least 32 and 8 respectively, the optimum use of instantaneous and transitional distortions is a combination of the two with α=0.2.

	0	**2**	**4**	**8**	**16**
16	13.59	13.59	13.45	11.97	10.06
32	10.73	10.73	10.62	9.46	8.01
64	7.80	7.80	7.70	6.71	6.25
128	6.46	6.46	6.42	6.07	5.54

Table 3. Percentage error rates for the phonetically rich data for instantaneous codebook sizes C=16, 32, 64 and 128, and transitional codebook sizes D=0, 2, 4, 8 and 16 with optimum combination weights.

Table 3 shows that consistently for any instantaneous codebook size, the speaker identification performance increases by combining the instantaneous codebook with a transitional codebook and that performance

also increases with the size of that instantaneous codebook.

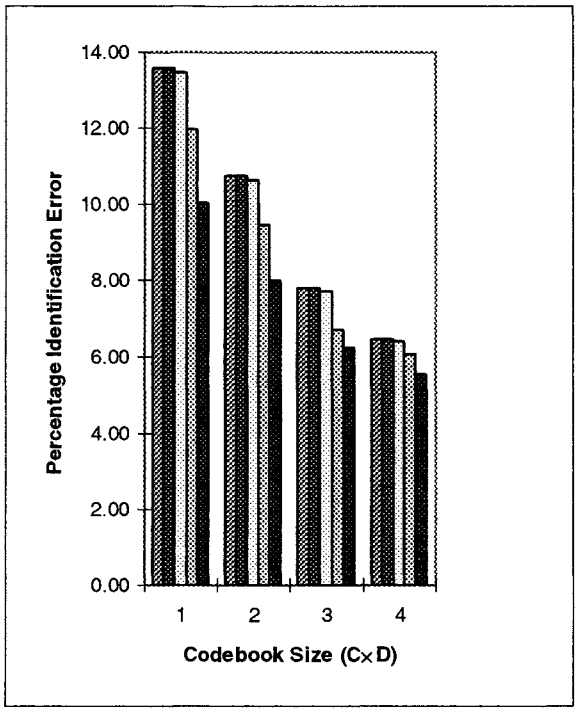

Figure 1. Percentage error rates for the phonetically rich data. The 4 groups of bars show the instantaneous codebook sizes C=16, 32, 64 and 128. Each group comprises 5 bars for the transitional codebook sizes D=0, 2, 4, 8 and 16 with optimum combination weights.

It is also interesting to note that combined instantaneous and transitional codebooks allow the reduction of the number of codeword comparisons in order to achieve a given speaker identification performance. Comparing the performance of the (32×16)-codebook with that of the (64×0)-codebook in Table 3 shows similar error rates of 8.0% and 7.8% with only 48 codeword comparisons required in the first case as against 64 in the second. In the case of the (64×16)-codebook with 6.25% error compared with the (128×0)-codebook with 6.5% error, the performance of the former is superior to that of the latter despite a reduced number of 80 codeword comparisons as against 128.

3. SECOND EXPERIMENT

3.1. Data

Experiment 2 used a database of the ten digits of British English spoken by 20 speakers. The VQ codebooks were trained using 20 repetitions of the digits recorded over several weeks. The test data comprised 5 repetitions of the digits recorded about 2 weeks later. The data were digitised at 8,000 samples/s and 14 mel-frequency cepstral coefficients were computed for 32ms frames with a 16ms frame advance. Transitional cepstra were determined by taking the difference cepstra of frames which were 96ms apart.

3.2. Method

The method of training VQ codebooks and testing utterances against those codebooks was essentially identical to that of Experiment 1. Instantaneous codebooks of sizes C=16, 32, 64 and 128 were built from the training data before the transitional cepstra for each codeword were clustered into second-level codebooks of sizes D=2, 4, 8 and 16, except in the case of C=128 where insufficient training data prevented the construction of meaningful transitional codebooks of sizes D=8 and 16.

3.3. Results

The speaker identification results for the 10 digits are shown in Tables 4 and 5. As expected, the overall speaker identification error rates are lower than in Experiment 1 because, for the ten-digit vocabulary, the codebook has to cover a smaller phonetic variance than for the vocabulary of Experiment 1.

16	**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
2	25.8	15.0	9.7	8.2	9.7	12.4
4	12.1	8.2	5.8	6.0	8.5	12.4
8	8.7	5.9	5.1	5.5	7.3	12.4
16	5.8	4.3	4.0	4.9	6.7	12.4
32	**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
2	17.1	8.7	5.3	4.4	4.2	5.7
4	8.7	5.3	3.5	3.0	3.7	5.7
8	5.6	2.9	2.0	2.7	3.1	5.7
16	5.3	3.0	2.4	2.0	3.4	5.7
64	**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
2	9.7	4.7	2.2	1.9	2.4	4.1
4	6.7	3.9	2.8	1.7	2.2	4.1
8	4.8	2.8	1.6	1.5	2.1	4.1
16	3.7	2.9	1.3	1.3	2.1	4.1
128	**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
2	6.8	3.4	2.0	1.1	1.5	3.1
4	4.9	2.0	1.4	1.0	1.5	3.1

Table 4. Percentage error rates for the digit data for instantaneous codebook sizes C=16, 32, 64 and 128, transitional codebook sizes D=2, 4, 8 and 16, and with combination weights α ranging from 0.0 to 1.0.

The results of Table 4 confirm that for a transitional codebook size of D=16, after preclassification with a conventional instantaneous codebook, the sole use of the transitional distortion (α=0.0) is superior to the

sole use of the instantaneous distortion (α=1.0). In contrast to the richer vocabulary of Experiment 1, the benefit of the transitional distortion can already be seen at the smaller transitional codebook sizes of D=2, 4 and 8 where the optimum α values are in the range of 0.4-0.8 (compare [5]).

As in the first experiment, the addition of small transitional codebooks for each instantaneous codeword yields significant reductions of the speaker identification error rate.

	0	2	4	8	16
16	12.4	8.2	5.8	5.1	4.0
32	5.7	4.2	3.0	2.0	2.0
64	4.1	1.9	1.7	1.5	1.3
128	3.1	1.1	1.0		

Table 5. Percentage error rates for the digit data for instantaneous codebook sizes C=16, 32, 64 and 128, and transitional codebook sizes D=0, 2, 4, 8 and 16 with optimum combination weights.

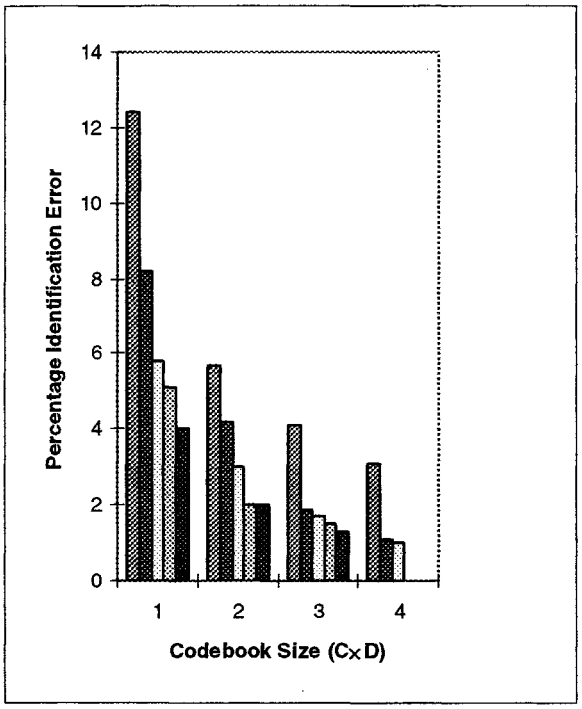

Figure 2. Percentage error rates for the digit data. The 4 groups of bars show the instantaneous codebook sizes C=16, 32, 64 and 128. Each group comprises 5 bars for the transitional codebook sizes D=0, 2, 4, 8 and 16 (for C=128 only D=0, 2 and 4) with optimum combination weights.

From the perspective of minimising the number of codeword comparisons, the results of the second experiment are even more promising. Table 5 shows that the speaker identification performance of a (128×0)-codebook can be bettered by a (64×2)-codebook or a (32×4)-codebook, a (64×0)-codebook is bettered by a (32×4) or a (16×16)-codebook, and a (32×0)-codebook is surpassed by a (16×8)-codebook, with the corresponding savings in codeword comparisons.

4. CONCLUSION

Two experiments have shown that moderate improvements of text-independent speaker identification rates can be achieved by employing a two-level codebook comprising a conventional instantaneous cepstral codebook together with one transitional cepstral codebook for each of the instantaneous codewords.

5. REFERENCES

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

[2] J.S. Mason, J. Oglesby, L. Xu, Codebooks to optimise speaker recognition. Proc. Eurospeech, 1989, pp 267-270.

[3] X. Zhu, B. Millar, I. Macleod, M. Wagner, F. Chen, S. Ran, A comparative study of mixture-Gaussian VQ, ergodic HMM and left-to-right HMMs for speaker recognition. Proc IEEE Int Symp on Speech, Image Proc. and Neural Networks, 1994, pp 618-621.

[4] S. Furui, Cepstral analysis techniques for automatic speaker verification, IEEE Trans ASSP-29, 1981, pp 254-272.

[5] F.K. Soong, A.E. Rosenberg, On the use of instantaneous and transitional spectral information in speaker recognition, IEEE Trans ASSP-36, 1988, pp 871-879.

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

[7] Y. Linde, A. Buzo, R.M. Gray, An algorithm for vector quantisation, IEEE Trans Com-28, 1980, pp. 84-95.