OPTIMAL COHORT DESIGN IN VQ-DISTORTION BASED TEXT-INDEPENDENT SPEAKER VERIFICATION

Iain Macleod, Fangxin Chen, Bruce Millar and William Laverty

TRUST[1] Project
Research School of Information Sciences and Engineering
Australian National University

ABSTRACT – The "cohort normalised" method of speaker verification computes for each input utterance its relative distance from models of the client and a cohort of speakers drawn from the same population. It is assumed that variations which reduce the utterance fit to the client model will tend to have similar effects with respect to the cohort speaker models. The use of "relative distance" can lead to improved client/impostor discrimination.

This paper explores several issues related to the design of suitable cohorts. Using VQ codebooks in multidimensional cepstral space as the basic speaker models, we show that pairs of codebooks can be related geometrically in terms of vector differences between their centroids in cepstral space. In a well-designed cohort, the cohort members give adequate "coverage" of the client's codebook in multidimensional space.

Cohort members are usually chosen on the basis of their similarity to the client. Experiments in which cohort members were instead chosen according to their position relative to the client led to a slight improvement in verification performance, suggesting that joint consideration of similarity and position would give even better results. However, with a limited set of speakers, it will often be difficult to find cohort members who meet these simultaneous requirements. Preliminary tests indicate that in certain cases it may well prove possible to synthesise suitable "phantom" codebooks based on those of real speakers.

INTRODUCTION

In the classic procedure for speaker verification, an input utterance is accepted or rejected according to a threshold on its goodness of fit with a model of the client's speech. While such a measure truly reflects absolute deviations between the client's model and input utterances, it is sensitive to overlapping client and impostor distributions which arise because of the effects of intra-speaker variation, recording environment change and phonetic variation. This in turn leads to a high Equal Error Rate (EER).

An alternative approach (Rosenberg *et al.*, 1992) uses a "cohort" of speakers, with speech models similar to that of the client, allowing relative measures of similarity or difference to be computed and reducing problems due to the above-mentioned variations. Similarity is judged on the basis of the mean distortion of a potential cohort speaker's utterances with respect to the client speaker's VQ model.

Tests of the cohort method show that it is subject to problems with false acceptance of impostor utterances which are quite dissimilar to those of the client (eg. from a speaker of opposite sex to the client) but which still give a better fit to the client model than to any of the cohort models. We gave a tentative geometrical explanation of this problem in Chen, Millar and Wagner (1994), suggesting that the problem arises from inadequate "coverage" of the client by cohort members. Choice of suitable cohort members needs to be based on an understanding of the relationship between pairs of codebooks — our aim in the following is to explicate and quantify at least one component of this relationship, using VQ codebooks in multidimensional cepstral space as the basic speaker models.

APPROACHES TO THE MEASUREMENT OF CODEBOOK DIFFERENCES

In our experiments, we characterise speakers by codebooks of 128 codewords (vectors in 15-D mel-frequency cepstral space) chosen such as to minimise the encoding error (distortion) with the training data sets. The number of muscle groups used in articulating speech sounds is much less than 15. We also know (Davis & Mermelstein, 1980) that most of the relevant information for phonetic discrimination in the speech of two males could be represented with about six cepstral coefficients. Hawkins, Macleod and Millar (1994) have further shown that high-dimensional cepstral data relating to vocalic speech tend to fall on low-dimensional quadratic surfaces (predominantly parabolic) which can be characterised in terms of only four parameters. Important components of the codeword distributions will thus have lower intrinsic dimensionality than that of their space of representation; the overall distributions can thus be expected to show significant clustering in cepstral space.

One way of assessing the similarity of a pair of codebooks is to measure the distortion when one codebook is used to encode the speech data on which it was trained, and then to compare this to the distortion obtained with this data using the other codebook. Given that the codebooks for all speakers have been trained on the same set of utterances, the regions most densely occupied by codewords should be similar for pairs of similar codebooks.

In terms of our verification paradigm, the similarity of codebooks measured in such a way represents the similarity of speakers. The ratio of distortions is a scalar magnitude; as a directionless quantity it thus gives no indication as to which of two given codebooks would yield the smaller distortion when encoding the training data for a third, for example. As a similarity measure it provides an estimate of how "close" the regions of cepstral space occupied by the two codebooks are, but it does not indicate their relative positions. Scalar measures are thus of only limited use in diagnosing problems with a given cohort or in choosing cohort members for a given client.

A simple vector measure considers the relative differences between codebook centroids (formed from the average of all vectors in a codebook). While we can be confident that pairs of speaker models which give relatively small errors in encoding each other's training data will have similar distributions of codewords in cepstral space, with pairs of similar codebooks there may still be considerable interspersion of codewords. The question arises as to whether any differences (in magnitude and direction) between the centroids of such codebooks are meaningful in the statistical sense. Given the inhomogeneity and complexity of the codeword distributions in cepstral space, simple statistical characterisations (based on variances of these distributions) are not appropriate for answering this question. An alternative method associates the codewords in one book with neighbours in the other (eg. on the basis of closest Euclidean distance, as used in the following) and then analyses the distributional properties of the resulting set of 128 difference vectors, to see if they cluster in particular directions. We now consider a method for analysing these properties.

STATISTICAL STRUCTURE OF CODEWORD-PAIR DIFFERENCES

This section develops a method to test the statistical significance of vectorial relationships between codebooks. In analysing the difference vectors, we (i) determine a mean directional component, (ii) test the statistical significance of this component, and (iii) subtract the mean from all vectors before analysing the residuals with Principal Components Analysis (PCA). The mean vector between codeword pairs can simply be shown to be equal to the vector between the corresponding codebook centroids. PCA is used to check to what extent directional variability between codeword pairs is concentrated in a few directions.

We have analysed distributions of difference vectors with relatively similar and dissimilar pairs of codebooks (with similarity being assessed in terms of average distortion), using Hotelling's T^2 statistic to test the hypothesis that the mean vector of the difference vectors was non-zero. For all pairs of codebooks examined, the hypothesis was confirmed ($p < 0.0001$) showing that the difference vectors tend to point in a consistent direction. As a result of the low intrinsic dimensionality of the cepstral distributions of vocalic speech (Hawkins, Macleod and Millar, 1994), a significant proportion of the codewords will tend to cluster on hypersurfaces of lower dimensionality. If, however, there was substantial interspersion of the codeword distributions in the codebooks being compared, the difference vectors would have a less consistent orientation. The results of our analysis here show that the degree of interspersion is limited, thus indicating that distributions of codewords from similar codebooks have similar shapes and that the concept of relative displacements between codebook pairs has statistical validity. After subtracting the mean vector from each difference vector, analysis of the residuals with PCA revealed one distinct non-noise directional component with dissimilar pairs of codebooks and two orthogonal components with similar codebook pairs (one component being somewhat larger than the other) — see Figure 1.

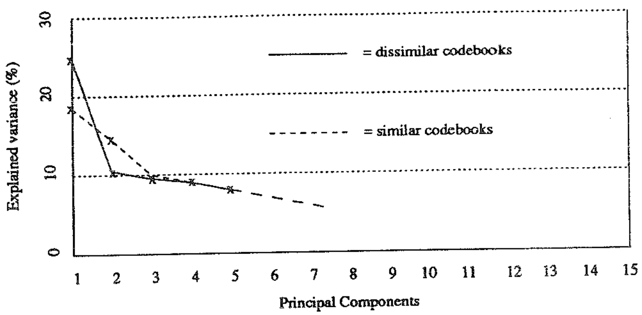

Figure 1: Analysis of residuals with similar and dissimilar codebooks.

The presence of non-noise Principal Components in the residuals, after the mean vector is subtracted, means that there are further systematic variations in the relationships between pairs of codeword distributions in addition to the mean displacement. Two codebooks with similar centroids may thus give large distortions when encoding each other's training data (eg. if one codebook had a greater span in certain directions than the other).

We can estimate our progress towards explaining the total relationship between two codebooks by computing the length of the (vector) sum of the difference vectors and comparing this to the sum of the scalar lengths of the individual vectors. If all difference vectors point in the same direction, these two lengths will be the same. If the difference vectors are randomly oriented, the summed vector length will be only a small fraction of the scalar sum of lengths. On examining the codebooks of potential cohort members in relation to given client codebooks, we found that this length ratio varied from about 25% to 40%, a much larger than expected length for the sum of random vectors. In addition to supporting the statistical finding that the difference between codebook centroids is real, this result means that a large enough component of the total relationship is captured that clear benefits should follow from taking relative codebook positions into account when constructing cohorts.

CHOOSING COHORT MEMBERS

We now have statistical justification in using relative centroid positions to consider the extent to which the members of a cohort "enclose" a client or leave "gaps" in the coverage, given possible interspersion of the codeword distributions of similar speakers. The minimum distortion among the cohort models and the mean distortion across cohort models have both been proposed for use in the client/cohort comparison. The following tests are based on use of the *min* statistic.

An optimal cohort is one in which for each potential impostor there is a cohort member whose codebook encodes impostor utterances with lower distortion than that achieved with the client's codebook. We have to be careful not to falsely reject the client's speech, so such cohort codebooks need to encode the client's training data with a significantly (but not dramatically) larger distortion than that obtained with the client's codebook: we want the cohort members to be similar, but not too similar, to the client. A percentage of impostors with speech very similar to the client will thus be falsely accepted, but this is unavoidable. Referring to Figure 2, imagine a hyperellipsoid (concentric with the client), which contains the centroids of codebooks for speakers similar to the client. The members of one potential cohort could then be distributed on the surface of a second larger hyperellipsoid with roughly twice the diameters of the first, so that (on average) utterances made by speakers whose codebook centroids lay outside the first hyperellipsoid would be attributed to a cohort member, and utterances made by speakers whose codebook centroids lay inside would be attributed to the client. By varying the size of the smaller hyperellipsoid, we could then achieve the desired balance between Type I and Type II errors. (Hyperellipsoids are assumed instead of hyperspheres, because of the fact that other codebooks are unlikely to be evenly distributed about the client's.)

In the usual case, only a limited set of speakers (and their trained codebooks) will be available for cohort

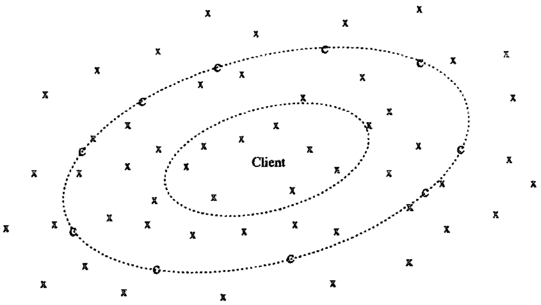

Figure 2: Hypothetical 2-D geometrical depiction of codebook centroids in cepstral space.

construction. The most similar speakers in this set to a given client may well be less (or sometimes more) similar than desired. Nevertheless, a functional cohort of size N can be formed by choosing the N most similar codebooks. Just as the codeword distributions themselves will be of lower intrinsic dimensionality than that of the representation space, so we might expect that the relative positions of codebook centroids (and thus of cohort members) will also be unevenly distributed. For example, cepstral features will tend to vary in a systematic manner with changes in parameters such as vocal tract length and shape.

In terms of our geometric analogy, anomalous acceptance of dissimilar impostors with a cohort chosen from the speakers most similar to the client arises because the client is "covered" too sparsely or too unevenly. An alternative procedure for assembling a cohort is as follows. Choose a speaker who is similar (but not too similar) to the client as the first cohort member. Test the remaining speaker population to see which speaker (of about the desired similarity to the client) gives the highest percentage of false acceptances with this cohort of size one. This speaker will lie in a direction which is not well covered by the first cohort member and is chosen as the second cohort member. The procedure is repeated until a cohort of the required size has been formed.

The speech data used in our studies is described in Millar *et al.* (1994). We randomly divided our population of 45 speakers into two — a cohort formation population of 25 speakers and a client/test population of 20 speakers (10 male and 10 female). Using the method of assessment outlined in Chen, Millar and Wagner (1994), we then tested the verification performance of cohorts assembled (i) from the speakers most similar to the client, and (ii) by starting with the most similar speaker to the client, adding the speaker who gave the greatest number of false acceptances with this cohort of size one, and so on as each new member was added. Because of the limited speaker population available, the similarity of cohort members to the client was not considered in building up the cohort using the "optimum direction" method (which was intended to identify and then fill gaps in the cohort coverage of the client). The results given in Table 1 showed a slight advantage for the direction method, even though (apart from the first cohort member) similarity to the client was not considered. For several clients, the EER with the "optimum direction" procedure increased slightly as the cohort size increased from three to five; in this case the final one or two cohort members chosen must have led to false rejections of the client (ie. these members were too similar to the client).

Analysis of the EERs achieved with Min5 and Sel5 showed that the observed improvement with Sel5 was not statistically significant. Our limited experiments thus indicate that the direction and similarity methods produce cohorts of similar quality. Given the different basis of these two methods of assembling cohorts, simultaneous consideration of both coverage and similarity promises to improve overall performance.

CONSIDERATIONS IN SYNTHESISING "PHANTOM" CODEBOOKS

Given the difficulties we encountered with locating suitable cohort members (because of our limited population of speakers), the question arises as to whether we can form synthetic codebooks with the desired properties. An obvious starting point here would be to modify the client's codebook to get a new

Client	Abs_VQ	Min1	Sel1	Min2	Sel2	Min3	Sel3	Min4	Sel4	Min5	Sel5	Sel5′
1	2.46	4.99	4.99	3.39	1.29	0.70	0.71	0.60	0.62	0.16	0.56	0.56
2	11.05	30.34	30.34	15.39	23.72	12.01	9.39	10.62	8.23	9.39	8.24	8.23*
3	11.38	23.63	23.63	18.21	17.12	14.70	14.90	11.19	9.32	9.39	8.79	8.79
4	4.74	3.56	3.56	3.56	2.48	1.10	2.38	0.71	1.77	0.62	1.72	1.72
5	1.18	3.37	3.37	3.52	2.36	2.38	3.54	2.33	3.52	2.93	3.49	3.49
6	3.54	3.70	3.70	3.80	3.65	2.89	3.52	2.44	3.42	2.31	2.86	2.86
7	5.29	4.69	4.69	0.59	2.55	0.50	2.28	0.45	1.69	0.11	0.57	0.57
8	1.77	3.49	3.49	3.52	3.49	5.43	3.43	4.78	3.03	4.27	3.09	3.03*
9	1.30	0.64	0.64	0.64	0.64	0.62	0.62	0.75	0.62	0.62	0.64	0.62*
10	3.98	1.17	1.17	1.76	1.77	0.58	1.26	0.59	1.13	0.51	0.53	0.53
11	10.13	3.64	3.64	2.47	1.16	1.62	1.19	1.31	0.70	1.31	0.66	0.66
12	2.93	10.60	10.60	1.88	4.11	1.72	2.34	1.18	1.66	1.75	1.18	1.18
13	1.79	8.24	8.24	2.92	6.41	2.78	5.17	1.77	3.54	1.77	2.33	2.33
14	6.47	1.18	1.18	1.18	1.18	1.18	1.18	1.18	1.18	1.16	1.16	1.16
15	7.03	9.83	9.83	9.41	5.28	7.11	5.17	7.69	5.15	8.03	5.57	5.57
16	3.51	3.51	3.51	2.34	2.31	2.33	2.31	2.28	2.36	2.25	1.77	1.77
17	5.88	11.51	11.51	7.11	8.46	6.46	5.82	4.72	4.47	4.19	4.11	4.11
18	10.03	5.34	5.34	0.47	0.64	0.11	0.62	0.06	0.05	0.06	0.05	0.05
19	11.78	11.10	11.10	4.74	8.84	6.38	5.31	5.60	3.63	4.15	4.10	3.52*
20	4.83	11.51	11.51	8.99	8.24	8.71	6.91	8.54	5.95	7.92	6.02	5.87*
Mean	5.55	7.80	7.80	4.79	5.29	3.97	3.90	3.44	3.09	3.15	2.87	2.83

Table I: Achieved Equal Error Rate percentages with an absolute threshold (ABS_VQ) and with selected cohorts of size *n* chosen conventionally (Min*n*) and according to false acceptances (Sel*n*). The final column (Sel5′) shows the improved results obtained with several clients (marked with *) through use of a final synthetic cohort member.

codebook which is just sufficiently dissimilar (ie. gives the desired amount of distortion when encoding the client's training data with respect to the balance of Type I and Type II errors). We could, for example, disturb 1 or more of the 15 coefficients in each codeword at a time to yield synthetic cohorts displaced a desired distance from the client in the direction of the altered coefficients. Experiments showed that codebooks synthesised in this manner had little practical utility — they usually did not encode impostor utterances as efficiently as the client's codebook and thus did not lead to improvements in speaker verification performance. The source of the problem here is that we are working with codeword distributions which are most likely densely clustered in only a small region of the 15-D cepstral space. In synthesising "phantom" codebooks we need to ensure that the synthetic codewords are representative of those of typical speakers similar to the client. Working in a space which is known to be inhomogeneously occupied, we can minimise errors arising from inhomogeneities by using codeword pairs from similar real speakers and interpolating synthesised values, thereby staying "close" to known real values.

Experiments in synthesising codebooks by either adding or subtracting a fixed vector displacement to or from all codewords in a real speaker's codebook, either the client's or a (potential) cohort member's, were quite instructive. The fixed displacement was usually 50% of the difference vector between the client's and cohort's codebook centroids. In a typical example, the client's codebook encoded a set of test client utterances with a distortion of 2783, the cohort's codebook gave a distortion of 3323, the client's codebook displaced by either + or – 50% of the difference vector between the centroids gave distortions of 2811 and 2799 respectively, and the cohort's codebook displaced by + or – 50% of this difference vector gave distortions of 3422 and 3255 respectively. Two points to be noted here are that (i) the observed increases and decreases in distortion are consistent with our geometric interpretation, and (ii) when the client codebook is displaced halfway towards the cohort, the distortion increases but is still substantially smaller than the (reduced) distortion obtained when the cohort codebook is displaced halfway towards the client.

The second point above provides further evidence that the distributions of codewords vary in ways other than overall position — speakers are characterised by the shapes of their codeword distributions as well. We thus tried a second method of interpolation which aimed to indirectly capture something of these other dimensions of variation. Instead of adding a fixed vector displacement to all codewords, we interpolated (or extrapolated) on the basis of individual difference vectors between codeword pairs. As we add an increasing percentage of these difference vectors to the codewords in the client codebook, so the synthesised codebook will gradually change from one that is similar to the client codebook into one that is similar to the cohort codebook. For the example client and cohort codebooks considered above, a synthetic codebook interpolated using 50% of the individual difference vectors for codeword pairs gave a distortion of 3078, which was close to halfway (3053) between the respective client and cohort distortions of 2784 and 3323.

Table I shows that with some clients the EER increased from Sel4 to Sel5. In these cases, we used the chosen fifth cohort member to construct an extrapolated synthetic codebook (moving the chosen cohort codebook further away from the client) and recalculated the EER (shown as Sel5′). In all cases this procedure prevented the EER from increasing between Sel4 and Sel5′; in two cases (clients 19 and 20) the synthetic cohort member reduced the EER between Sel4 and Sel5′. The reduction in the overall error rate to 2.83% was not, however, sufficient to make the difference between Min5 and Sel5′ statistically significant.

DISCUSSION

The overall results of our experiments provide evidence that the distributions of codewords in 15-D MFCC space are rather complex. Although we can show statistically that the observed mean displacements between similar codebooks are real and do not occur just by chance, the distributions of codewords in given codebooks will vary in shape and extent as well as position. Our concept of relative codebook positions captures an important part, but only a part, of the total relationship between similar codebooks.

In the next phase of our work, we aim to conduct more-extensive tests of cohorts formed from a mixture of real and synthetic codebooks. In particular, we want to explore better methods of forming codeword pairs than simple nearest neighbour and to evaluate the utility of synthetic cohort members formed via fixed displacements of real codebooks relative to those formed by adding some chosen percentage of individual difference vectors between pairs of codewords.

ACKNOWLEDGEMENT

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

NOTE

(1) Technology for Robust User-conscious Secure Transactions.

REFERENCES

Chen, F., Millar, B. and Wagner, M. (1994), "Hybrid threshold approach in text-independent speaker verification," *Proc. Int. Conf. on Spoken Language Processing*, Yokohama, 1855–1858.

Davis, S. B. and Mermelstein, P. (1980), "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. ASSP-28, 357–366.

Hawkins, S., Macleod, I. and Millar, B. (1994), "Modelling individual speaker characteristics by describing a speaker's vowel distribution in articulatory, cepstral and formant space," *Proc. Int. Conf. on Speech Science and Technology*, Perth, to appear.

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. Int. Conf. on Speech Science and Technology*, Perth, to appear.

Millar, B., Chen, F. and Wagner, M. (1994), "The efficacy of cohort normalisation in a speaker verification task under different types of speech signal variance," *Proc. Int. Conf. on Speech Science and Technology*, Perth, to appear.