THE EFFICACY OF COHORT NORMALISATION IN A SPEAKER VERIFICATION TASK UNDER DIFFERENT TYPES OF SPEECH SIGNAL VARIANCE

J.Bruce Millar, Fangxin Chen and Michael Wagner.

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

ABSTRACT

This paper examines the influence of three types of speech signal variance on the performance of a text-independent speaker verification system and the efficacy of cohort normalisation as a technique to compensate for this influence. The three forms of variance comprise that generated by repetition of utterances over time, the addition of extraneous noise to test utterances, and the inclusion of test utterances which use phonemic sequences that do not occur in the training data. A statistical analysis of the results for a gender-balanced set of 20 client/impostor speakers and an independent population of 25 cohort speakers is presented. The results indicate that conventional cohort normalisation is moderately successful in combatting repetition variance and phonetic variance but gives no significant improvement in the presence of moderate levels of noise. A hybrid form of cohort normalisation is shown to combat the former two variance types very effectively and to provide a weakly significant improvement for moderate levels of noise induced variance.

INTRODUCTION

One of the major problems of speaker verification is the setting of thresholds against which to test measures of similarity or difference. The use of a fixed threshold is not robust under normal variance experienced in the capture of incoming speech in practical systems. Cohort normalisation has been shown to produce improved performance for speaker verification when applied to speech which has been captured under different conditions to those existing when the speaker models were created.

The basic idea is attributed to Higgins et al.(1991), and has been applied in other research on speaker verification (Rosenberg et al., 1992; Matsui and Furui, 1994). The TRUST project is exploring the use of this technique and has already highlighted one of its limitations (Chen et al., 1994a). A further paper in these proceedings (Chen et al., 1994b) extends our exploration of this technique.

The present paper aims to differentiate between the benefits of cohort normalisation when confronted with different types of speech signal variance. We examine the impact of three distinct forms of variance. The first form of variance is that produced by simple repetition of utterances by a speaker where the repetitions are spread across many days. The second form of variance is that produced when test speech has an ambient acoustic background which is different to that under which the speaker models were trained. Additive noise representing the presence of many speakers talking in the background is used in order to simulate this situation. The third form of variance is that produced by the use of phonetic sequences which were not a part of the training data from which the speaker models were created. Data fitting this specification were included in the initial TRUST data corpus (Millar et al., 1994).

While the first form of variance has been explicitly tested by others, the second and third forms of variance, although acknowledged have not been specifically explored. The aim of this work is to generate an effectiveness profile for cohort normalisation when challenged by explicitly controlled forms of speech variance.

The three forms of variance included in this paper are indicative of major problems facing speaker verification. Change in the client over time requires frequent updating of models unless a robust normalisation

technique can be used. The presence of extraneous noise which is different to that during the most recent enrolment is a constant hazard which cannot always be controlled for. The use of phonetic sequences which did not occur in training will occur due to speaking errors, changes in speaker physiology, and maybe minor task variations.

THE BASIC METHOD

The basic concept behind cohort normalisation is to provide for each client speaker a 'cohort' of similar speakers whose speaker models provide a local environment within which distance measurements to incoming test speech can be normalised. The cohort of similar speakers is selected by measuring the distance of prospective cohort members from the speaker model of the client speaker. The group of N speakers from a 'cohort set' who are closest to the client speaker are chosen to be members of that speaker cohort. This criterion may not be optimal (Chen et al., 1994b) but was the standard criterion in use at the time this work was performed.

The normalisation process involves the measurement of the distances between an incoming speech sample and the speaker models of the client and the members of the client's cohort. The unnormalised 'client distance' is normalised by subtracting some statistic of the set of cohort distances from the client distance. Our simple cohort normalised score is given by:-

$$D(norm) = D(client) - Minimum\ [D(cohort)]$$

This technique makes the assumption that the introduction of variance in the testing signal will cause a shift in the analytic space in which the speakers are modeled for the test utterance concerned. Thus the boundary between utterances which come from the client and those which come from other speakers will shift also. Verification that is based on a simple 'distance threshold' will be disturbed by such a shift. The 'cohort normalisation' formulation assumes that the variance-induced shift will vary the distance to the client model and to the models of similar speakers by a similar amount. A hybrid cohort normalised score is obtained by setting 'D(norm)' to a very large value when 'D(client)' exceeds a preset but large threshold lying well outside the clients normal speech range (Chen et al., 1994a).

THE DATA CORPUS FOR THESE EXPERIMENTS

In these experiments the initial TRUST data corpus (Millar et al., 1994) was used in the following way. Firstly 10 male and 10 female speakers were randomly selected to act as clients. A cohort of 5 similar speakers was chosen for each speaker from among an additional set of 25 speakers. The remaining 19 client speakers in the data corpus were used as impostors.

REPETITION VARIANCE

Repetition variance occurs when the same utterance is produced on successive occasions. Zhu et al. (1994) have shown that for the data from the initial TRUST corpus (Millar et al., 1994) the speaker identification performance of a vector quantisation (VQ) model deteriorates in a way that is dependent on the time interval between the production of the material on which the model was based and the production against which the model was tested. This is entirely consistent with earlier reports on the temporal variability of speaker characteristics (e.g. Furui, 1974; Soong et al., 1987)

In the current experiment VQ models were produced from the 5 repetitions of the corpus recorded in the first recording session (session A) for the 20 client speakers. Testing data comprised client data from the second (B) and third (C) sessions which were recorded at approximately one week intervals following the first session. Each session comprised 5 repetitions of the corpus. Impostor data was drawn from the utterances of the 19 other speakers in the client-set during sessions B and C.

The performance of each client model is expressed as an equal error rate (EER) percentage for each test session. The experiment was then repeated using simple and hybrid cohort normalisation providing results for both methods in the form of EER percentage for each client. These results are reported in table 1.

SPEAKER number	EER % Session B absolute	EER % Session B cohort	EER % Session B hybrid	EER % Session C absolute	EER % Session C cohort	EER % Session C hybrid
1	6.59	1.76	1.76	2.48	0.62	0.60
2	4.01	20.68	5.88	11.15	18.80	9.53
3	4.60	1.17	1.17	9.66	7.81	5.92
4	4.69	0.50	0.47	6.40	1.61	2.33
5	0.59	0.03	0.03	2.33	2.36	2.33
6	4.05	5.34	3.55	4.77	2.39	2.93
7	1.65	0.00	0.00	4.16	0.59	4.13
8	0.61	1.80	0.61	2.25	6.33	2.25
9	1.67	1.68	0.03	1.86	1.16	1.86
10	6.45	0.59	0.53	5.28	1.44	1.10
11	7.00	3.67	2.38	12.41	2.33	8.95
12	1.31	0.50	0.13	4.11	1.74	1.77
13	2.01	0.66	0.66	2.36	3.01	2.36
14	8.77	0.48	0.10	7.65	1.10	1.10
15	5.20	4.13	2.96	5.88	8.77	6.44
16	2.94	1.18	1.18	3.52	1.91	1.77
17	13.20	7.18	7.06	7.67	5.45	5.29
18	16.47	8.23	7.48	12.36	2.28	1.86
19	14.69	2.96	2.96	22.93	5.45	5.26
20	7.07	6.78	5.27	6.46	8.12	6.30
AVG	5.68	3.47	2.21	6.78	4.16	3.70
STD	4.59	4.78	2.44	5.06	4.33	2.63

Table 1: Performance of speaker verification for 20 speakers using absolute, cohort normalised and hybrid cohort normalised scores when test data is displaced in time from the training data

A two-way analysis of variance on these results showed that the increase in EER across subsequent sessions was in fact only weakly significant ($p < 0.07$), but that there was a significant difference between the three different thresholding methods applied ($p < 0.003$). There was no significant interaction between these factors. A post hoc Tukey HSD test revealed that the observed average benefit of simple cohort normalisation over absolute thresholding was moderately significant ($p \simeq 0.03$) and that of hybrid thresholding was highly significant ($p < 0.003$).

NOISE VARIANCE

In this experiment the mean energy of the test data from each speaker was computed on a token by token basis and two levels of multi-speaker babble noise from the NOISEX-92 data corpus (Varga, 1993) were added to each speaker's test data to give signal-to-noise ratios of 10dB and 20dB.

The performance of each client model is expressed as an EER for a particular signal-to-added-noise ratio in the test data. The experiment was performed for absolute thresholds, simple cohort normalisation and hybrid cohort normalisation. The results are reported in table 2.

A two-way analysis of variance of these results showed that noise level caused a highly significant ($p << 0.001$) difference in the speaker verification performance, and that the form of thresholding used caused differences in performance which were only weakly significant ($p < 0.07$). The opposite average effects of simple cohort normalisation between the 10dB and 20dB single-to-noise ratio conditions fell short of creating a significant interaction between the factors. A post hoc analysis using the Tukey HSD test indicated that the simple cohort normalisation gave no significant improvement and that the weakly significant improvement was entirely due to the use of the hybrid cohort normalisation method.

SPEAKER number	EER % SNR 10dB Absolute	EER % SNR 10dB Cohort	EER % SNR 10dB Hybrid	EER % SNR 20dB Absolute	EER % SNR 20dB Cohort	EER % SNR 20dB Hybrid
1	29.09	17.53	14.66	15.09	4.67	4.72
2	19.50	22.25	21.05	10.31	20.65	19.74
3	11.90	19.18	8.61	6.99	2.96	2.93
4	25.21	47.11	26.53	9.35	8.27	6.47
5	3.52	7.74	3.54	0.69	2.96	1.85
6	22.39	35.92	23.53	13.41	10.02	10.00
7	12.90	3.54	12.74	3.60	0.69	0.65
8	14.29	44.54	15.58	4.87	12.04	6.00
9	11.22	9.94	11.93	2.90	3.55	1.73
10	41.08	36.91	24.15	16.55	9.99	6.43
11	25.02	24.99	23.41	13.72	13.57	10.18
12	21.45	52.59	21.86	7.06	8.96	7.79
13	26.06	26.69	26.85	7.27	1.46	1.35
14	41.11	10.02	9.89	21.19	3.50	3.50
15	31.35	44.15	43.34	8.93	12.76	11.81
16	10.06	7.06	6.02	4.61	4.02	3.10
17	21.06	31.39	31.43	9.84	8.77	8.77
18	38.70	8.27	7.68	25.16	11.15	11.15
19	36.16	23.96	22.68	16.59	9.55	9.30
20	23.54	28.29	20.60	8.24	16.47	13.52
AVG	23.28	25.10	18.80	10.32	8.30	7.05
STD	10.77	14.91	9.75	6.30	5.31	4.87

Table 2: Performance of speaker verification using absolute, cohort normalised and hybrid cohort normalised scores when fixed amounts of noise which were not present in the training data are added to the test data

INTER-PHONETIC VARIANCE

In two of the recording sessions additional material was recorded comprising two repetitions of 10 tokens. These data are used to evaluate speaker verification against models constructed from the basic 30 tokens which were recorded in all three sessions. Although the basic 30 tokens were designed to cover all the phonemes of the language, particular combinations of phonemes used in the additional data did not occur in the training data. The 10 additional tokens comprised 38 syllables, of which 27 did not occur in the training material. This form of variance will vary from test token to test token within a session rather than being associated with a session as such. The results are reported in table 3.

A two-way analysis of variance showed that there was a highly significant difference ($p < 0.001$) in performance between situations in which the test data comprised just those phonemic sequences that existed in the training material and those in which test data included additional sequences, and also when different thresholding methods were applied ($p < 0.001$). There was no significant interaction. A post hoc Tukey HSD test showed that the use of simple and hybrid cohort normalisation gave highly significant benefits ($p < 0.001$) over absolute thresholding for the situations whether the phonemic sequences were, or were not, included in the training material. Further the benefit of the hybrid method over the simple method was significant for the situation where the phonemic sequences were not included. The apparent removal of the performance deficit caused by such additional sequences by use of the hybrid method was at best only weakly significant ($p < 0.08$).

DISCUSSION

These three experiments illustrate the way in which cohort normalisation is effective against various forms of variance. The previous literature has focussed on the variance caused by the change of a telephone handset microphone and has shown benefit for this case where there is a global change in

SPEAKER number	EER % within absolute	EER % within cohort	EER % within hybrid	EER % outside absolute	EER % outside cohort	EER % outside hybrid
1	3.73	0.83	0.83	9.82	5.08	5.15
2	8.58	19.31	7.27	13.53	22.19	13.53
3	8.56	6.92	5.42	11.49	7.22	7.49
4	5.60	1.00	1.65	7.56	4.61	2.61
5	1.66	0.66	1.67	4.47	7.05	2.75
6	3.33	3.02	2.58	7.29	7.15	3.09
7	2.68	0.38	0.99	6.88	0.00	6.88
8	1.70	4.03	2.08	5.02	10.44	5.02
9	1.72	1.11	0.34	2.10	2.10	0.07
10	5.99	1.05	0.99	5.02	2.88	2.34
11	10.54	1.68	3.73	10.58	7.49	4.47
12	3.37	1.38	1.00	0.41	2.20	0.54
13	2.06	1.36	1.00	4.02	4.42	0.73
14	6.33	0.04	0.03	15.05	2.47	2.54
15	5.10	5.67	4.07	10.37	7.76	7.49
16	2.91	1.59	1.34	5.43	2.47	2.47
17	10.24	5.95	5.41	12.92	7.70	7.56
18	14.62	6.69	6.41	15.32	0.41	0.41
19	17.16	3.59	4.07	27.49	9.97	10.03
20	5.75	6.95	5.08	10.58	9.97	10.03
AVG	6.08	3.66	2.08	9.27	6.18	4.76
STD	4.37	4.39	2.20	6.01	4.93	3.72

Table 3: Performance of speaker verification using absolute, cohort normalised and hybrid cohort normalised scores when test data contains phonetic sequences that are 'within' the training data or contains phonetic sequences that are 'outside' the training data.

the frequency response of the input channel (Rosenberg et al., 1992).

The first pattern that can be noted in the results of our first experiment is that the deterioration (increase) of average verification EERs, when increasing the time between training and testing, can be observed for all three analysis methods. This can be seen by comparing columns 2 & 5, 3 & 6, and 4 & 7 in table 1. It can also be observed that on average simple cohort normalisation improves the verification performance over that obtained with the absolute threshold method, and that the hybrid cohort normalisation exceeds this performance for test data from session B and from session C. This can be observed by comparing columns 2,3 & 4, and columns 5, 6 & 7.

The statistical analysis of the results indicated a very strong effectiveness of both cohort and hybrid cohort normalisation in coping with this kind of variance. The fact that only weakly significant difference was found in the results for sessions B and C emphasises the need to explore time differences greater than the order of one or two weeks between model training and model testing.

Our second experiment indicated that the simple cohort normalisation method was not effective in significantly reducing the impact of additive noise for signal-to-noise levels up to 20dB and in fact created a deficit with respect to the use of an absolute threshold if the signal-to-noise ration was as low as 10dB. The weakly significant improvement found when using the hybrid normalisation suggests that additive multi-speaker babble noise creates shifts in the speaker space whose effects can be partially neutralised by an additional threshold, but which distort the relationship between a client and his/her cohort derived from the clean signal. Such relationships are being explored further via a more detailed analysis of the cohort mechanism. It should be noted that in this experiment no attempt was made to increase the additional threshold applied in the hybrid threshold approach (Chen et al., 1994a) beyond that used for clean speech. Such adaptation of the hybrid algorithm is an obvious direction for future work.

Our third experiment showed that the use of cohort normalisation was able, on average, to reduce the increase in EER, brought about by the inclusion of 'out-of-training-data' phonemic sequences, back to levels achieved for 'within-training-data' sequences, although the overall significance of this result was rather weak.

CONCLUSIONS

This paper has reported on the impact of cohort normalisation in the presence of three types of variance in the test data that was not present in the training data. Two forms of cohort normalisation have been applied: simple cohort normalisation using a 'minimum' statistic and hybrid cohort normalisation. The latter, which restricts the range of application of the cohort method, using an absolute threshold to exclude inputs which are very distant from the client model, has been shown to perform more effectively in every situation.

Cohort normalisation has shown significant benefits in coping with variance due to a limited delay between the building of a speaker model and testing it, and in coping with phonetic sequences that do not occur in the training data of a text-independent model. Only in its enhanced form of 'hybrid cohort normalisation' is it able, and then only weakly, to cope with moderate levels of additive multi-speaker babble noise.

The analysis of the impact of specific forms of speech variance is ongoing, as is the analysis of the precise mechanisms of cohort normalisation. It is expected that these joint areas of study will suggest new and more effective methods for enhancing the performance of speaker verification.

ACKNOWLEDGEMENT

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

REFERENCES

Chen,F., Millar,J.B., Wagner,M. (1994a) 'Hybrid threshold approach in text-independent speaker verification', ***Proc. ICSLP-94***, 1855-1858.

Chen,F., Macleod,I., Millar,J.B., Wagner,M. (1994b) 'Optimal cohort design in VQ-distortion based text-independent speaker verification', (these proceedings).

Furui,S (1974 'An analysis of long-term variation of feature parameters of speech and its application to talker recognition', ***Electronic Communications Japan***, Vol.57a, 34-42.

Higgins,A., Bahler,L., Porter,J. (1991) Speaker verification using randomised phrase prompting, ***Digital Signal Processing***, Vol.1, 89-106.

Matsui,T. and Furui,S. (1994) Similarity normalisation method for speaker verification based on A Posteriori probability, ***Proc. ESCA workshop on speaker recognition, identification, and verification***, Martigny, 59-62.

Millar,J.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, (these proceedings)

Rosenberg,A.E., DeLong,J., Lee,C-H., Juang,B-H., Soong,F.K. (1992) The use of cohort normalised scores for speaker verification, ***Proc ICSLP-92***, Banff, 599-602.

Soong,F.K. Rosenberg,A.E. Rabiner,L.R. Juang,B.-H (1987 'A vector quantisation approach to speaker recognition', ***AT& Technical Journal***, Vol.66, 14-26.

Varga,A. (1993) Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems, ***Speech Communication***, Vol.12, 247-251.

Zhu,X., Gao,Y., Ran,S., Chen,F., Macleod,I., Millar,J.B., Wagner,M. (1994) Text-independent speaker recognition using VQ, mixture Gaussian VQ and ergodic HMMs, ***Proc. ESCA workshop on speaker recognition, identification, and verification***, Martigny, 55-58.