

A-PRIORI SELECTION OF COHORT SETS FOR A SPEAKER VERIFICATION SYSTEM: ISSUES AND INSIGHTS

Michael Barlow(¥), Brett Watson (*),
Ah Chung Tsoi (§) & Tom Downs (£)
{spike@cs.adfa.edu.au, brett.Watson@adc.com.au,
act@uow.edu.au, td@elec.uq.edu.au }

(¥)School of Computer Science, University of
NSW/ADFA, (*)ADC, (§)Faculty of Informatics,
University of Wollongong, (£)School of Computer
Science & Elec. Engineering, University of Queensland

ABSTRACT - The paper describes a series of speaker verification experiments using the well-known cohort normalisation method. Cohort sets are selected a-priori based only on training data, for a database of 42-speakers uttering digits in isolation, recorded over a period of 18-months. Baseline performance is contrasted with post (at time of verification) selection of cohort set. Further, a-priori set selection is examined along a number of axes: text-independent versus text-dependent, similarity between cohort sets for the different utterances, the significance of speaker ordering in a cohort set, together with the issue of length of verification utterance. This analysis is then used to explore and highlight issues regarding similarity and dissimilarity between speakers.

INTRODUCTION

Cohort set normalisation (Higgins et. al., 1991, Matsui and Furui, 1993) is a well established technique in speaker verification that may be employed as a post-processing adjunct to any number of algorithms (e.g., HMM, ANN, VQ). Based on Bayesian theory, the distance (likelihood) between the input utterance and claimed customer model is contrasted with the distance between the utterance and a number of cohorts (other known speakers) of the claimed identity. In a Bayesian sense, the cohorts (and resulting distance) represent the probability that the utterance was produced by a speaker other than the claimed identity. The method has proved to be very effective in reducing error rates.

The paper describes a series of speaker verification experiments in which a-priori determination of cohort sets is performed: cohort sets are selected on the basis of training data only (remaining fixed from training stage onwards) and not as some function of the input utterance. Such experiments are representative of applications of speaker verification with large customer bases where it is not computationally feasible to measure the input utterance against all customers. Using a speaker population of forty-two adults repeating the digits over a period of eighteen months (hence long-term intra-speaker variability was captured) a number of experiments were conducted examining the cohort normalisation process itself, and through those experiments seeking to gain an insight into more fundamental issues of speaker similarity and dissimilarity. In particular, difference in verification performance between a-priori and post (at time of identity claim) determination of cohort sets, text-dependent versus text-independent selection of cohort sets, significance of ordering and difference between cohort sets for the same speaker, together with the significance of input utterance length are all examined.

THE DATA

For all experiments a single database consisting of 42 adults (36 male and 6 female) repeating the isolated digits over a period of eighteen months (Barlow et. al., 1992) was used. The speaker population was split into two groups: a set of eleven (11) all-male 'customers' from whom thirty (30) repetitions of each of the digits was available; and a set of thirty one (31) 'impostors' from whom nine (9) repetitions of each of the digits were available.

Speakers were recorded with a close-talking microphone onto a DAT recorder in a quiet office environment. Recorded utterances were down-sampled to 16kHz and quantised at 16-bits. Linear predictive analysis with a frame size of 32ms (8ms shift) was used to derive 10th order cepstral coefficients, as well as their first and second derivatives.

EXPERIMENTAL CONDUCT

For each speaker a single 50-element VQ (Vector Quantisation) codebook was constructed using as training data the first nine (9) repetitions of each digit: hence ninety (90) utterances - ten digits by nine repetitions each, were used to compose a codebook. Further, a single combined codebook was used for all thirty (30) acoustic features (cepstra plus first and second derivatives) rather than three separate codebooks. VQ was selected as the mechanism for verification due to its robust high performance, particularly under constraints of low amounts of training data (Matsui & Furui, 1993).

Verification experiments were conducted using the final twenty (20) repetitions (those not used in codebook training) of each digit from the customer set as true speaker samples and all nine (9) repetitions of each digit by each member of the impostor set as impostor samples. Hence each verification experiment for a single digit consisted of 3289 verification trials (11 customers, and for each customer 20 true speaker trials and 279 false speaker trials). Verification experiments were conducted for all ten (10) digits individually. Equal error rates were then determined on an individual digit basis, and the mean across digits taken to derive an average equal error rate. Multi-word error rates (i.e., asking the customer to say more than one word before making a verification decision) were also determined for all possible numbers of words (two to ten) and all possible combinations of the digits (e.g., there are 45 possible pairings of the digits for the two-word case). Once again, as for the single word error rates, a mean equal error rate for the multi-word input was determined by taking the mean of error rates of all possible combinations.

Base speaker verification trials were conducted by accumulating the VQ distortion measure calculated against the VQ codebook for the claimed identity as well as the VQ codebooks for the first N members of the cohort set of the claimed identity. The ratio of distortions between that for the claimed identity and that of its cohorts was then thresholded in order to make a decision whether to reject or accept the identity claim. Equation 1 shows the score to be thresholded:

$$\frac{\sum_{i=1}^{CSS} D_i(x)}{D_c(x)CSS} \quad (1)$$

where x is the utterance, C is the claimed identity, CSS is the Cohort Set Size, and $D_i(x)$ is the accumulated distortion of encoding utterance x using speaker j 's codebook.

A-priori cohort set selection was performed on a per-customer basis. The entire, remaining population (41 speakers) were rank ordered using the algorithms detailed subsequently. This complete ordering then allowed the analysis of the effects of different cohort set sizes (for a normalisation set size of N , the first N speakers from the ordered list were used), importance of order etc. Post cohort set selection was achieved by computing the distortion of the input utterance against that of all speaker's codebooks (41 in the case of a true identity claim and 40 in the case of an impostor) and employing the N smallest values.

A cohort set size of five (5) speakers was used for all experiments. Other set sizes were examined with similar results but are not reported here (Barlow, 1993). Impostors were excluded from the cohort set of the identity they were claiming against: if they were one of the first five members selected on an a-priori basis, they were replaced by the 6th member of the ordered list.

A-PRIORI VS. POST SELECTION

Of prime importance in practical considerations of using cohort set normalisation for verification is how effective is the method (contrasted with verification without normalisation), and, given the practical considerations which force a-priori selection of cohort set, how close is practical performance to the maximum achievable (post selection). Figure 1 shows this three-way contrast.

The a-priori result represents text-dependent (i.e., 10 cohort sets for each speaker: one per digit) determination of cohort sets for all customers. Ordering of speakers within a cohort set was on the basis of maximum distortion, across all training utterances, against the identity in question: those with smaller maximum distortions were ranked higher. This selection algorithm was found to perform consistently better, in terms of verification equal error rate, than others on the basis of either minimum or average distortion.

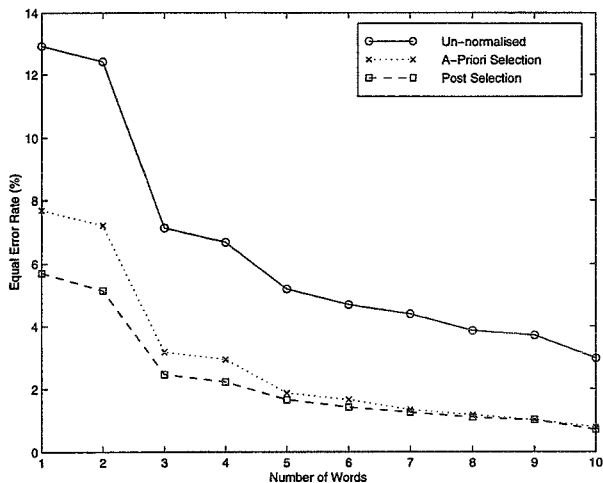


Figure 1: Speaker verification performance as a function of number of words in utterance contrasting un-normalised with a-priori and post selection of cohort sets.

As can be seen from Figure 1, applying a-priori normalisation reduces error rates between 41% (single word case) and 73% (ten word case) with a mean reduction of 61%. The improvement from a-priori to post normalisation is also marked, though not of the same order: a reduction of 29% (two word case) through to 0% (nine word case) with a mean improvement of 15%. In particular the improvements for utterances of more than four words is far less marked (8% on average).

REORDERING OF A-PRIORI SETS

A number of different algorithms were employed in the a-priori selection of cohort sets. All were based on the principle of finding cohort speakers similar (or least dissimilar) to the customer. Amongst these approaches the maximum distortion was found to be consistently, though only marginally, better than all other approaches.

In order to investigate the implicit assumption that selection of cohorts on the basis of similarity to the customer matched their role in the Bayesian equation (probability the utterance came from a different speaker), a set of verification experiments were conducted in which cohort set order was changed. Two reorderings were examined: using the cohorts sets in reverse (i.e., the 37th through 41st candidates, hence the most dissimilar speakers) and taken from the middle of the ordered set (i.e., 21st through 25th candidates, hence an 'average' or unrelated set of speakers). Figure 2 shows the result of those experiments.

As can be seen from Figure 2, choosing the most highly ranked candidates from the ordered set to act as cohorts yields significantly lower error rates than ones from the middle of the set or the tail end. However, even employing the lowest ranked candidates (those most dissimilar to the claimed identity) still improves verification performance over no normalisation: a

reduction in error rates of approximately half that obtained with the highest ranking candidates.

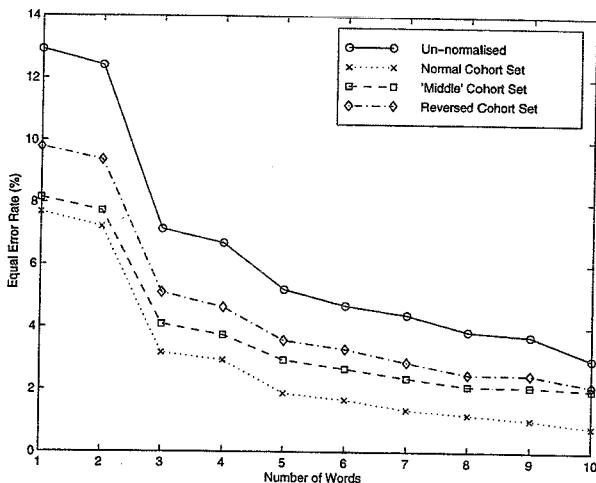


Figure 2: Speaker verification performance as the function of the number of words in an utterance contrasting cohort set ordering on an 'optimal' basis with that where it is reversed or midway between the two extremes.

TEXT-DEPENDENT VS TEXT-INDEPENDENT SELECTION OF COHORT SETS

The a-priori normalisation result of Figure 1 represents performance using text-dependent cohort set selection: each customer has ten cohort sets, one for each of the digits. From an application perspective this may not be desirable or practical and hence text-independent selection of cohort sets is worthy of investigation. At a more fundamental level, differences in performance between the two schemes, and differences between cohort sets for the same customer but different digits, may shed light on the consistency of speaker relationships.

Figure 3 is a contrast of verification performance where a-priori cohort sets are determined in a text-dependent and text-independent fashion. For the text-independent cohort set ordering the ten cohort sets corresponding to each of the digits were combined in an equi-weighted fashion to yield a single cohort set. As can be seen from the figure text-dependent set selection yields a significantly lower verification error rate, particularly as more digits are employed in the utterance.

Clearly, as witnessed by the lower error rate when text-dependent selection is employed, speaker relationships (similarity/dissimilarity) do not remain fixed across different utterances. In order to examine this phenomena, a distance measure between two cohort sets (A and B) was constructed as follows:

$$D = \frac{1}{CSS} \sum_{i=1}^{CSS} |P_A(s_i) - P_B(s_i)| \quad (2)$$

where CSS is the Cohort Set Size (41), s_i is the i 'th speaker in the population and $P_c(s)$ is the position (a value from 1 to 41) of speaker j in cohort set C . Intuitively the distance measures the average separation (in positions) between speakers in the two cohort sets.

This distance was then applied to all pairings of the digit-based cohort sets from a single customer yielding a (10-by-10) distance matrix between the ten cohort sets. Distances ranged

from a minimum of 5.4 (between "one" and "nine") through to a maximum of 14.5 (between "four" and "eight") with a mean of 8.9. In order to visualise the relationship (distance) between the cohort sets a Spring-Embedder (Battista et al., 1994) method was employed to generate a 2D representation that conformed to the table of distances. Figure 4 is a realisation of that 2D space. As is clear from the figure, even the "closest" cohort sets are significantly different, particularly when it is recalled that a cohort set size of 5 was used in verification trials.

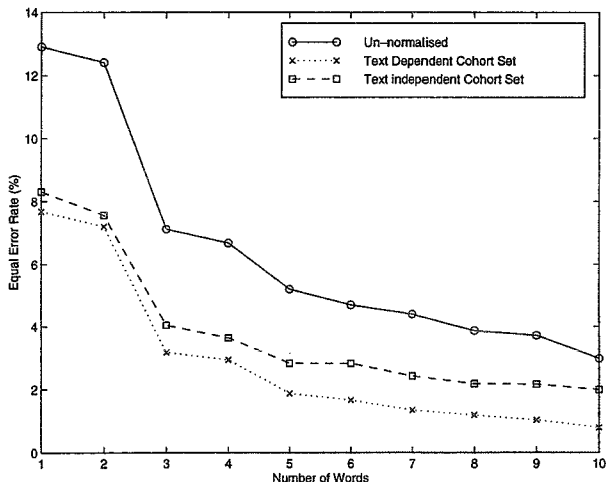


Figure 3: Speaker verification performance as a function of the number of words in the utterance contrasting text-dependent and text-independent selection of the cohort sets.

DISCUSSION

The results have shown that cohort set normalisation, as applied to a VQ distortion based speaker verification system, substantially reduces error rates. Post selection of cohort sets (when the input utterance is known) yielded the lowest error rates with a reduction of over 70%, while a-priori selection of cohort sets (based solely on training data and remaining fixed for each speaker regardless of input utterance) reduced error rates by over 60%.

A-priori selection of speakers for a customer's cohort set was performed on the basis of choosing cohorts "similar" to that of the customer. A number of different quantifications of "similar" were examined (though not directly reported on here; Barlow, 1993), including "average" (across all repetitions) closest, best-case closest, and worst-case closest. The worst-case approach (i.e., largest distortion across all repetitions) performed marginally better in terms of speaker verification error rate than all others, and was used in all experiments reported in the paper, implying that perhaps an upper-bound on speaker dissimilarity is a more useful measure than average or lower-bounds on speaker similarity. When contrasting a-priori cohort selection with post cohort selection it was found that post selection, representing the optimal achievable performance in a Bayesian sense, reduced error rates a further 15%. In a further analysis of the significance of a-priori selecting cohorts on the basis of similarity to the customer, verification trials were conducted in which speakers most dissimilar, as well as neither markedly similar or dissimilar were employed as cohorts. In both cases verification performance was better than not employing cohort normalisation at all, but significantly poorer than selection on the basis of similarity. Clearly, selecting cohort sets on the basis of similarity to the customer is not a bad strategy for the a-priori selection of cohort sets but insufficient to achieve the highest possible performance.

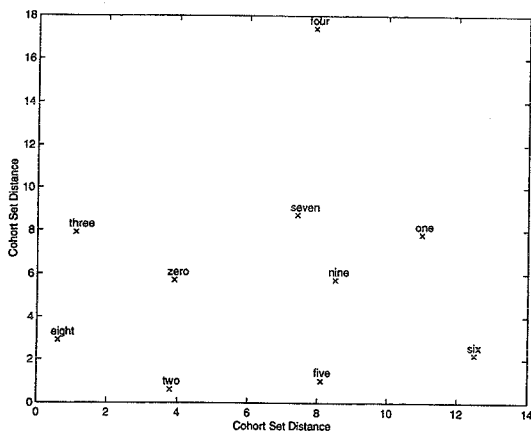


Figure 4: Spring-Embedder derived 2D realisation of the relationship (similarity/dissimilarity) between the text-dependent (one per digit) cohort sets for a single speaker.

The importance of a-priori selection of cohort members on a text-dependent basis was also examined. Speaker verification performance was found to be significantly better when ten cohort sets (one per digit) were employed per customer, rather than just one. The text-dependent cohort sets for a single customer were examined and contrasted one-with-the-other via a distance metric. The relationships between the cohort sets were visualised in two dimensional space using a spring-embedder method. Cohort sets were found to be significantly different, with even the most similar pairing, that between "nine" and "one" (similarity likely because of sharing the nasal /n/) differing substantively. The result appears to show that speaker relationships are highly dynamic: speaker similarity/dissimilarity is not simply a global phenomena, but highly dependent on the utterances in question. The methodology of analysing and visualising the parameters (such as the cohort sets selected) of the verification process itself, with the goal of gaining a deeper understanding of speaker relationships, appears to show promise. The method could equally be applied to confusion or distance matrices between speakers to gain more direct insights.

ACKNOWLEDGEMENT

Portions of this research were conducted as part of the "Speaker Verification Project" at the University of Queensland.

REFERENCES

- Barlow M. (1993) "An Investigation of Normalisation Parameters for VQ-Distortion Based Speaker Verification", Speaker verification Group Tech. Report, University of Queensland.
- Barlow M, Booth I., and Parr A. (1992) "The Collection of Two Speaker Recognition Targeted Speech Databases", Proc. Fourth Aust. Int. Conf. Speech Science and Technology, 706-711.
- Battista G.D., Eades P., Tamassia R and Tollis I.G. (1994) "Algorithms for Drawing Graphs: An Annotated Bibliography", Computer Geometry and Theory Application, 4, 235-282.
- Higgins A, Bahler L., and Porter J. (1991) "Speaker Verification using Randomized Phrase Prompting", Digital Signal Processing, vol. 1, 89-106.
- Matsui T. and Furui S. (1993) "Concatenated Phoneme Models for Text-Variable Speaker Recognition", Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc., vol 2, 391-294.