Efficient Online Cohort Selection Method for Speaker Verification

Tomi Kinnunen, Evgeny Karpov, and Pasi Fränti

University of Joensuu, Joensuu, Finland
Department of Computer Science
{tkinnu,ekarpov,franti}@cs.joensuu.fi

Abstract

Cohort normalization is a method for normalizing the scores in speaker verification in order to reduce undesirable variation arising from acoustically mismatched conditions. A particular form of cohort normalization, *unconstrained cohort normalization* (UCN) is addressed in this study. The UCN method has been shown to give excellent results but its major drawback is the huge computational load arising from the search of the cohort speakers. In this paper, we propose a fast cohort search algorithm, that quantizes the test vector sequence and uses the quantized data for both impostor and claimant scoring. Results on the NIST-1999 corpus show a speed-up factor of 23:1 compared to full search. Furthermore, the equal error rates are *decreased* from those of the full search.

1. Introduction

Speaker verification [1] is a task of deciding whether an unknown speech utterance was produced by a claimed identity. The output of a verification system is binary: either the speaker is accepted or rejected.

The verification task is a *statistical hypothesis testing* problem that can be formulated as follows. Suppose that the unknown speaker produced an utterance X and claims to be a person S. The two opposite hypotheses are

$$\begin{cases} H_0 & : \quad X \text{ was produced by } S \\ H_1 & : \quad X \text{ was } \textit{not} \text{ produced by } S, \end{cases}$$

and the verification engine must decide which one of these two hypotheses is true.

Suppose for a moment that the likelihoods of both hypotheses are known. In this case, the *likelihood ratio* [2] gives the optimal decision in Bayes sense (minimum risk classification) [2, 3]. The decision rule is then

$$\text{Decide} \begin{cases} H_0 & , \quad \text{if } LR_{H_0,H_1} > \Theta_S \\ H_1 & , \quad \text{if } LR_{H_0,H_1} \leq \Theta_S, \end{cases} \tag{1}$$

where LR_{H_0,H_1} is the ratio of the likelihoods of the two hypotheses, and Θ_S is a decision threshold for speaker S. The thresholds Θ_i are determined from the training data so that a desired balance between false acceptances (FA) and false rejections (FR) is obtained. The threshold can be global for all speakers, or it can be speaker-depended [4].

The data for speaker verification is obtained in the form of *acoustic feature vectors* $\{\boldsymbol{x}_i\}$ extracted from real speech utterances. In the enrollment phase, a *speaker model* is trained from the training vectors. In the verification phase, the input utterance is first converted into feature vectors, which are then used for estimating the likelihoods of the two hypotheses H_0 and H_1.

The likelihood of the *null hypothesis* H_0 is estimated by matching the vectors $X = \{\boldsymbol{x}_i\}$ against the claimed speaker's model S, which is intuitively reasonable. Suppose that the probability density of the claimant's feature vectors $p(\boldsymbol{x}|S)$ is known; in this case, the matching is carried out by computing the likelihood $p(X|S)$ under some simplifying assumptions (independence of the test vectors). In reality, due to finite amount of training data, the densities are only estimates of the true underlying distributions.

The estimation of the likelihood of the *alternative hypothesis* H_1 is considerably much harder. Estimating this is equivalent to solving what is the likelihood that *anyone else in the world* (except S) produced X. In speaker recognition community, there have emerged two main approaches for modeling the alternative hypothesis [5], so-called *world model* and *cohort model* approaches.

The world model W (*background model*, *universal background model*, *global speaker model*) is a large model that aims characterizing all possible speakers and speaking contexts of the "world". It is trained from a large amount (several hours) of speech data from a variety of speakers. Estimating the likelihood of H_1 then translates simply to computing the likelihood $p(X|W)$ similarly as with the client model.

The second approach for estimating the likelihood of H_1 uses the concept of *cohort models*. Rather than modeling the whole world, cohort approach uses a small representative set of models, called *cohort set*. Individual cohort models' scores are obtained and combined e.g. by averaging.

The world-model and cohort approaches have been combined successfully in [6]. In this approach, a coarse matching based on world model is first carried out, and this is followed by detailed matching using the cohort models. The combined approach was reported to give smaller error than neither of the two approaches alone.

There is no general consensus whether the world model or cohort approach is more accurate [4, 7], and results supporting both hypotheses exist. The reason for this is that the studies are done in rather diverse conditions (different corpora, acoustic features, speaker models, number of speakers, ...), so one approach might be better in certain situation than the other.

An argument in favor of the world model is that the cohort approach is more complex. It is true that in the cohort approach the following issues must be addressed: (1) the cohort selection method, (2) the size of the cohort set, and (3) the method of combining the individual cohort scores.

An argument in favor of the cohort approach originates from its dynamic nature: for each client speaker, there is a personal set of "impostors", which might simulate the case of an intruder attack, and potentially to decreases the false acceptance rate.

However, it is recognized that the cohort approach is com-

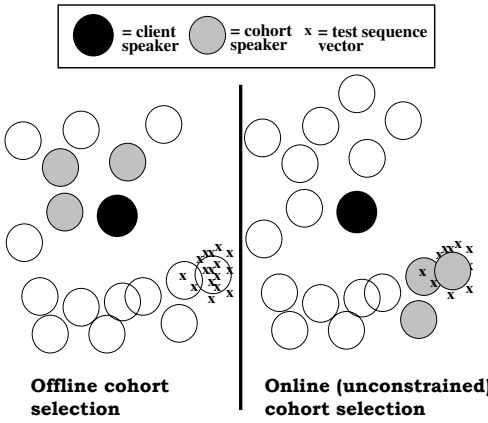

Figure 1: A case where offline cohort selection fails. In the offline case, the client score is higher than any of the cohort speakers, whereas in the online case the cohort scores are much higher, preventing an obvious intrusion.

putationally more expensive than the world model approach, especially if the number of cohort speakers is high or the cohort models are complex. Furthermore, it is often argued that the cohort approach requires a lot of storage, since the cohort models need to be stored for each client speaker. However, this argument is incorrect if the cohort speakers are selected among the client speakers. In this case, the cohort approach takes less space than the world model approach since only a look-up table of the cohort indices must be stored in addition to the client models. In fact, it was noticed in [8] that a smaller equal error rate (EER) is obtained if the cohort models are selected among the client speakers instead of an external cohort.

In this study, we propose a computationally efficient algorithm for cohort normalization, called *pre-quantization based cohort search* (PQS). The PQS method reduces computational load by quantizing the input vector sequence using the LBG clustering algorithm [9]. The experimental results on the NIST-1999 speaker recognition evaluation corpus indicate that a speed-up of 23:1 can be obtained, and at the same time the equal error rate is decreased.

2. Review of Cohort Normalization

Regarding the *selection* of cohort speakers, the most common approach is to select the closest speakers to the claimant speaker, since intuitively these are the best "impostors" for the claimant speaker [5, 8, 10–12]. This method is referred to as the *closest impostors* method in this paper.

There are several ways of selecting the cohort speakers [5,8, 10, 13–18]. The methods can be divided into two classes: those that select the cohort speakers *offline* in the training phase, and those that select the cohort *online* based on the input utterance. In offline selection, a set of cohort models is searched based on their closeness to the claimant model. Therefore, the database consists of the speaker models and a look-up table of cohort speaker indices. In the verification phase, the cohort models of the client speaker are scored. Most of the computation is spent on scoring the cohort models, especially if the cohort size is large.

The online cohort selection, also known as *unconstrained cohort normalization* [8] (UCN), selects the cohort speakers during the matching phase, based on the closeness to the input utterance. This method is expected to give better results than the offline selection, since it adapts to the current verification trial. In [8, 18] the superiority of the UCN method was verified experimentally. Another feature in favor of the UCN is that it does not require updating the cohort pointers when new clients are enrolled.

Fig. 1 shows a case where the offline selection fails. The three closest impostors to the client speaker are on the "wrong side" of the client speaker in respect to the test vectors. Consequently, the score for the client speaker will be higher than for any of the cohort speaker, thus probably accepting the claimant. However, it can be seen that the vectors are much more probable produced by the rightmost speaker. The problem is solved in the online case; the closest impostor to the *test sequence* are selected and scored, thus giving high likelihood for the alternative hypothesis and rejecting the speaker as it should be. In the light of this example, it is expected that the number of false acceptances is reduced by using the online approach.

The computational overhead of the online selection, however, limits it's usability in practise. The cohort models must be first searched which requires N matching function evaluations for a candidate set of N speakers. In the next Section, we propose an efficient online cohort search method.

The optimal cohort size depends on the cohort selection method, the task, as well as the rule for combining the cohort speaker scores. For instance, in [8, 18] it was observed that for offline cohort selection, the EER decreases with increasing cohort size. However, for online selection, a cohort size of one speaker was optimal according to Ariyaeeinia and Sivakumaran [8].

There are several approaches for combining the cohort speakers' scores in the likelihood ratio context (see [18] for a comparitive study). The two most commonly used methods [14] are (1) *averaging* the cohort scores, and (2) taking the *maximum* of the cohort scores, i.e. the score of the most competiting speaker. Fuzzy integration of the cohort scores has been proposed in [11]. Finally, we note that there are alternatives to the likelihood ratio based score normalization. In [19], the claimant and cohort scores were combined with a multilayer perceptron (MLP) neural network.

3. Efficient Cohort Scoring for Verification

For the sake of simplicity, we will formulate the methods for vector quantization (VQ) based recognition [20]. The methods are straightforward to generalize to GMM and other models, as will be demonstrated.

In VQ-based verification, the match score between sequence of vectors $X = \{\boldsymbol{x}_i\}$ and a speaker codebook $C = \{\boldsymbol{c}_j\}$ is computed as the average quantization distortion defined as follows:

$$D(X, C) = \frac{1}{|X|} \sum_i \min_j \|\boldsymbol{x}_i - \boldsymbol{c}_j\|. \qquad (2)$$

The smaller the distortion is, the better X and C match.

Let $C = \{C_1, \ldots, C_N\}$ be the set of all client models. Furthermore, let $FindCohort(X, C, K)$ denote a procedure that returns the indices of top K best scoring models for vector sequence X from the model set C (cohort candidates).

The pseudocode of the proposed method is given in Algorithm 1. First, the input sequence X is quantized with the LBG-clustering algorithm, producing a reduced vector sequence $\hat{X}$ of size M. The LBG algorithm gives reasonable good results for

Algorithm 1 Pre-Quantization Based Cohort Search (PQS)

Let X be the feature vectors and I the claimed identity ;
$\hat{X} :=$ LBG-Clustering(X, M) ;
Let $Coh := FindCohort(\hat{X}, C \setminus \{C_I\}, K)$;
Return $Score(X, C_I) := \frac{D(\hat{X}, C_I)}{\frac{1}{K}\sum_{i \in Coh} D(\hat{X}, C_i)}$;

quantizing speech data. It reduces effectively redundant vectors and outliers while retaining the characteristics of the original distribution.

Cohort models are searched from the set of all models, excluding the client model. The client score is then divided by the average cohort score, giving an estimate of the inverse likelihood ratio. For GMM, the procedure is exactly the same, but the ratio of the GMM likelihoods is used instead.

Notice that only the quantized data is used in scoring, both the target speaker and the impostors. This is reasonable, since *both* the client and cohort scores are degraded, but their ratio is expected to remain the same - this is the fundamental rationale behind score normalization. In other words, we assume:

$$\frac{D(X, C_I)}{\sum_i D(X, C_i)} \approx \frac{D(\hat{X}, C_I)}{\sum_i D(\hat{X}, C_i)}, \tag{3}$$

where the sums go through the indices of the cohort models selected based on X and $\hat{X}$, respectively. The approximation (3) is good when X and $\hat{X}$ follow the same probability distribution. This assumption is reasonable, if the intermediate codebook is created properly and the codebook is large enough. The control parameters of the algorithm are the cohort size (K) and the size of the quantized test set (M).

4. Experiments

For the experiments, we used the *NIST 1999 speaker recognition evaluation corpus* [21]. We selected the male subset containing 230 speakers for the verification experiments. Both the "a" and "b" files for each speaker were used for training. We selected to use the 1-speaker test segments from the same telephone line with mixed handsets. There are $N = 692$ genuine speaker trials and $N(N-1)/2 = 239086$ impostor trials. The model size is fixed to 128 for both VQ and GMM. For VQ codebook training, we use the LBG algorithm, and for GMM training, we use the Expectation Maximization (EM) algorithm [22].

We use the standard MFCCs as the features [23]. A preemphasiz filter $H(z) = 1 - 0.97z^{-1}$ is used before framing. Each frame is multiplied by a 30 ms Hamming window, shifted by 10 ms. The FFT spectrum is computed, followed by bandpass filtering with 27 triangular filters spaced linearly on the mel-scale. The log-compressed filter outputs are converted into 12 cepstral coefficients by DCT, and the 0th cepstral coefficient is ignored. On average, there is about 2 minutes of training data for each speaker, and approximately 30 seconds of test data, giving on average 2900 feature vectors per verification trial.

First, we studied the effect of the cohort size (K) and the size of the PQ codebook (M). The cohort size was varied from 1 to 20, and the PQ codebook from 4 to 512. The equal error rates (EER) for VQ and GMM are shown in Figures 2 and 3, respectively. We observe that the cohort size is less important parameter than the pre-quantization factor, as expected. The absolute recognition performance is rather poor, which might be because we did not apply any channel compensation method nor added the delta/double-delta coefficients.

It is surprising how much the test sequence can be quantized without noticeable degradation; the equal error rates for

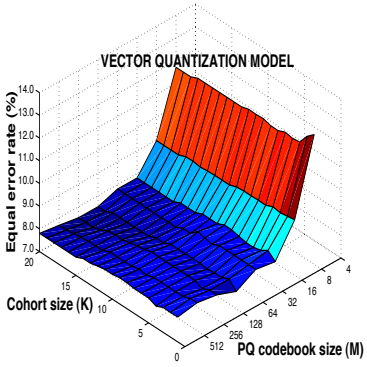

Figure 2: Accuracy of the PQS method (VQ-128).

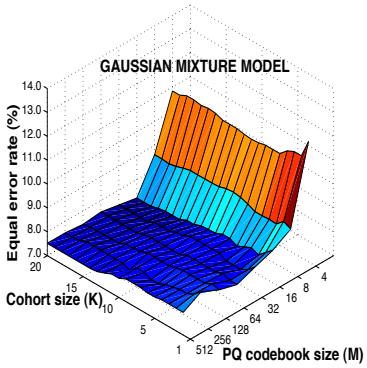

Figure 3: Accuracy of the PQS method (GMM-128).

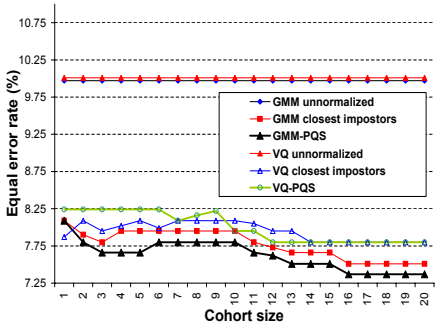

Figure 4: Comparison of normalization methods (VQ-128, GMM-128; M = 32).

both methods converge to minimum EERs (about 7.4 %) around $M = 32$. In other words, the test sequence can be compressed by a factor of 2900/32 = 90:1 without noticeable degradation in the accuracy. In their experiments, McLaughlin & al. [24] observed that the test sequence could be decimated up to factor 20:1 without degradation. However, they simply decimated the vector sequence (taking every Mth vector), whereas we aimed at retaining the original density by clustering the test sequence.

Next, we fixed the PQ codebook size to 32, and compared the following methods by varying the cohort size: (1) no normalization, (2) closest impostors, (3) pre-quantization based cohort search. The results are shown in Fig. 4. We observe that increasing cohort size improves verification accuracy, which is consistent with the results reported in [12, 18].

For GMM, the PQS method gives systematically better re-

sults than the closest impostors method. For VQ, the ordering is not as consistent; nevertheless the PQS method reaches the accuracy of the closest impostors method when cohort size is increased. The GMM modeling gives slightly better results than VQ, and both modeling techniques reduce the EER from about 10 % of the unnormalized case to about 7.4 %.

Table 1: Summary of the scoring methods (cohort size $K = 20$).

Method	Model	EER (%)	Avg. verif. time (s)	Speed-up factor
Closest impostors	VQ-128	7.80	5.79	1:1
	GMM-128	7.51	18.94	1:1
PQS	VQ-128	7.80	0.65	9:1
	GMM-128	7.37	0.84	23:1

The average verification times along with the error rates for the best cohort size ($K = 20$) are summarized in Table 1. We observe that VQ achieves a speed-up of 9:1 compared to the full-search unconstrained cohort normalization (closest impostors), whereas GMM achieves a speed-up factor of 23:1.

5. Conclusions

We have proposed a computationally efficient method for unconstrained cohort normalization problem in speaker verification. The results with NIST-1999 corpus indicate that the method can decrease both the error rate and the running time. In future, we plan to extend the experiments by comparing the other various existing cohort selection and score normalization methods. In particular, we are interested whether the cohort or world model normalization works better in practise since no one has compared this systematically in large scale to our knowledge.

6. Acknowledgements

The work of E. Karpov was supported by the National Technology Agency of Finland (TEKES).

7. References

[1] J. Campbell. Speaker recognition: a tutorial. *Proc. of the IEEE*, 85(9):1437–1462, 1997.

[2] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, London, second edition, 1990.

[3] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley Interscience, New York, second edition, 2000.

[4] F. Bimbot, M. Blomberg, L. Boves, D. Genoud, H.-P. Hutter, C. Jaboulet, J. Koolwaaij, J. Lindberg, and J.-B. Pierrot. An overview of the CAVE project research activities in speaker verification. *Speech Communications*, 31:155–180, 2000.

[5] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digit. Sign. Proc.*, 10:42–54, 2000.

[6] W.D. Zhang, M.W. Mak, and M.X. He. A two-stage scoring method combining world and cohort models for speaker verification. In *Proc. ICASSP 2000*, volume II, pages 1193–1196.

[7] R. Bolle, J. Connell, S. Pankanti, N. Ratha, and A. Senior. *Guide to Biometrics*. Springer Verlag, 2004.

[8] A.M. Ariyaeeinia and P. Sivakumaran. Analysis and comparison of score normalization methods for text dependent speaker verification. In *Proc. Eurospeech 1997*, pages 1379–1382, Rhodes,Greece, 1997.

[9] Y. Linde, A. Buzo, and R.M. Gray. An algorithm for vector quantizer design. *IEEE Trans. on Comm.*, 28(1):84–95, 1980.

[10] R.A. Finan, A.T. Sapeluk, and R.I. Damper. Impostor cohort selection for score normalization in speaker verification. *Pattern Recognition Letters*, (18):881–888, 1997.

[11] T. Pham and M. Wagner. Fuzzy-integration based normalization for speaker verification. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 1998)*, pages 3273–3276, Sydney, Australia, 1998.

[12] P. Sivakumaran, J. Fortuna, and A.M. Ariyaeeinia. Score normalization applied to open-set, text-independent speaker identification. In *Proc. Eurospeech 2003*, pages 2669–2672, Geneva, Switzerland, 2003.

[13] D.A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech Communications*, 17:91–108, 1995.

[14] C.-S. Liu, H.-C. Wang, and C.-H. Lee. Speaker verification using normalized log-likelihood score. *IEEE Trans. on Speech and Audio Processing*, 4(1):56–60, 1996.

[15] T. Isobe and J. Takahashi. Text-independent speaker verification using virtual speaker based cohort normalization. In *Proc. Eurospeech 1999*, pages 987–990, Budapest, Hungary, 1999.

[16] E. H. C. Choi and J. Song. Successive cohort selection (scs) for text-independent speaker verification. In *Proc. ICSLP 2000*, pages 442–445, Beijing, China, 2000.

[17] N. Mirghafori and L. Heck. An adaptive speaker verification system with speaker dependent a priori decision thresholds. In *Proc. ICSLP 2002*, pages 589–592, Denver, Colorado, USA, 2002.

[18] Y. Zigel and A. Cohen. On cohort selection for speaker verification. In *Proc. Eurospeech 2003*, pages 2977–2980, Geneva, Switzerland, 2003.

[19] W.R. Belfield and R.P. Mikkilineni. Speaker verification based on a vector quantization approach that incorporates speaker cohort models and a linear discriminator. In *IEEE Int. Conf. on Systems, Man, and Cybernetics*, pages 4525–4529, 1997.

[20] F.K. Soong, A.E. Rosenberg A.E., B.-H. Juang, and L.R. Rabiner. A vector quantization approach to speaker recognition. *AT & T Tech. J.*, 66:14–26, 1987.

[21] A. Martin and M. Przybocki. The NIST 1999 speaker recognition evaluation - an overview. *Digit. Sign. Proc.*, 10(1-18):1–18, 2000.

[22] C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, New York, 1996.

[23] J.R. Jr. Deller, J.H.L. Hansen, and J.G. Proakis. *Discrete-Time Processing of Speech Signals*. IEEE Press, New York, second edition, 2000.

[24] J.McLaughlin, D.A. Reynolds, and T. Gleason. A study of computation speed-ups of the GMM-UBM speaker recognition system. In *Proc. Eurospeech 1999*, pages 1215–1218, Budapest, Hungary, 1999.