On Cohort Selection for Speaker Verification

Yaniv Zigel and Arnon Cohen

Electrical and Computer Engineering Department,
Ben-Gurion University, Beer-Sheva, Israel
yaniv(arnon)@ee.bgu.ac.il

Abstract

Speaker verification systems require some kind of background model to reliably perform the verification task. Several algorithms have been proposed for the selection of cohort models to form a background model. This paper proposes a new cohort selection method called the Close Impostor Clustering (CIC). The new method is shown to outperform several other methods in a text-dependent verification task. Several normalization methods are also compared. With three cohort models and the best score-normalization method, the CIC yielded an average Equal Error Rate (EER) of 0.8%, while the second best method (Maximally-Spread Close, MSC) yielded average EER of 1.1%.

1. Introduction

The goal of speaker verification systems is to determine whether a given utterance is produced by the claimed speaker or not. This is done by comparing a score, which reflects the match of the given utterance and the claimed speaker's model, with a threshold.

In verification systems based on stochastic models (such as HMM and GMM) the simplest score is the likelihood of the utterance given the claimed speaker's model. This score is very sensitive to variations in text, speaking behavior, and recording conditions, especially from the non-speaker (impostors) utterances, in both text-independent and text-dependent tasks. This sensitivity causes wide variations in scores, and makes the task of threshold determination a very difficult one.

In order to overcome this score's sensitivity, the use of normalized score, based on cohort speakers (impostors) has been proposed [1 – 6]. Several issues arise with the use of cohorts, among them, the selection of the impostors' models (the cohort set), the number of the impostors in the cohort set, and the score normalization technique (the normalization function).

In this paper, a new method for cohort selection based on speaker clustering is introduced. This selection method is compared with other reported methods. The problem of the order of the cohort set, namely the number of impostors, is also examined. The problem of verification-score normalization is discussed, and results of several score normalization methods are presented. For these, a text-dependent speaker verification based on Hidden Markov Model (HMM) system has been implemented.

Supported in part by the EC project: MOUMIR

2. Score Normalization using Cohort Models

In verification systems, the decision to accept or reject an identity claim, *T*, is based on the comparison of a *score*, $s(\mathbf{O})$, with a threshold, τ:

$$s(\mathbf{O}) \begin{cases} \geq \tau & \rightarrow accept \\ < \tau & \rightarrow reject \end{cases} \quad (1)$$

The simplest score for stochastic model based verification systems is the log likelihood, which is the log probability of the (utterances) observations, **O**, given the target's (claimed speaker) model, λ_T:

$$s(\mathbf{O}) = \log p(\mathbf{O} \mid \lambda_T). \quad (2)$$

As was mentioned in the previous section, normalized scores are preferred over the un-normalized score of (2).

The most obvious normalization term is probably that of the background model likelihood:

$$s_n(\mathbf{O}) = \log\left(\frac{p(\mathbf{O} \mid \lambda_T)}{p(\mathbf{O} \mid \lambda_B)}\right) = \log p(\mathbf{O} \mid \lambda_T) - \log p(\mathbf{O} \mid \lambda_B) \quad (3)$$

$p(\mathbf{O} \mid \lambda_B)$, known as the normalization term, is the likelihood of the observed vector sequence for a background (filler or "garbage") model. The background model is trained by speakers, other than the target, uttering general text-independent utterances (text-independent tasks) or the *T* user's phrase (text-dependent tasks). In other words, $p(\mathbf{O} \mid \lambda_B)$ represents a dynamic threshold [2], which is sensitive to variations in **O** from trial to trial.

The main problem with the above normalization term is how to construct a good background model λ_B. Rather than averaging a group of speakers into one "wide" model, it may be better to construct several models from speakers who are close to the claimed speaker in the feature space (these are called "cohort").

In the cohort models idea, the normalization term is estimated only from a group of speakers, $C(T)$, whose models are somehow determined to be most "competitive" with the model of the target (claimed) speaker *T*.

2.1. Cohort Normalized Scores

Several score-normalization techniques may be considered. Maybe the most intuitive one is the normalization with the "closest" impostor model:

$$s_2(\mathbf{O}) = \log p(\mathbf{O} \mid \lambda_T) - \max_{c \in C(T)} \left[\log p(\mathbf{O} \mid \lambda_c)\right] \quad (4)$$

where λ_c is the model of the impostor "closest" to the target. Note that the close impostor is defined here as the one having the highest conditional probability.

In the literature, several different normalization methods have been proposed, such as [1 – 6]:

$$s_1(\mathbf{O}) = \frac{\log p(\mathbf{O} \mid \lambda_T)}{\max_{c \in C(T)} \left[\log p(\mathbf{O} \mid \lambda_c) \right]} \quad (5)$$

$$s_3(\mathbf{O}) = \log p(\mathbf{O} \mid \lambda_T) - \log \left\{ \frac{1}{C} \sum_{c=1}^{C} p(\mathbf{O} \mid \lambda_c) \right\} \quad (6)$$

$$s_4(\mathbf{O}) = \log p(\mathbf{O} \mid \lambda_T) - \frac{1}{C} \sum_{c=1}^{C} \log p(\mathbf{O} \mid \lambda_c) \quad (7)$$

$$s_5(\mathbf{O}) = \frac{\log p(\mathbf{O} \mid \lambda_T)}{\frac{1}{C} \sum_{c=1}^{C} \log p(\mathbf{O} \mid \lambda_c)} \quad (8)$$

$$s_6(\mathbf{O}) = \frac{\log p(\mathbf{O} \mid \lambda_T)}{\log \left\{ \frac{1}{C} \sum_{c=1}^{C} p(\mathbf{O} \mid \lambda_c) \right\}} \quad (9)$$

where C is the number of models in the cohort set.

3. The cohort selection

Two main issues that arise with the use of cohort speakers: (1) What procedure should be used for choosing the cohorts from the given database, (2) How many speakers should be included in the cohort set, C.

Choosing the "closest" (to the target) models for the cohort set is very logical, since these models provide samples of impostors' scores close to the true speaker (target) scores, and thus provide protection against acceptance of impostors. Such a technique of cohort selection has been proposed in [2] and [4].

3.1. Criterion for Cohort Models "Closeness"

In order to measure the "closeness" between two speakers, a suitable criterion is needed. Rosenberg et al. [2] have used an average score obtained from a pair-wise comparison:

$$d_K(\lambda_T, \lambda_j) = \frac{1}{2} \left(\log p(\mathbf{O}_T \mid \lambda_j) + \log p(\mathbf{O}_j \mid \lambda_T) \right) \quad (10)$$

where λ_j and $\mathbf{O}_j$ are the tested jth impostor model and its observations (sequence of feature vectors from the training data) respectively, and λ_T and $\mathbf{O}_T$ are the target's model and its observations respectively. The closest cohort models are those that maximize equation (10).

Reynolds [3] have used a symmetric, divergence-like criterion:

$$d_D(\lambda_T, \lambda_j) = \log \frac{p(\mathbf{O}_T \mid \lambda_T)}{p(\mathbf{O}_T \mid \lambda_j)} + \log \frac{p(\mathbf{O}_j \mid \lambda_j)}{p(\mathbf{O}_j \mid \lambda_T)} \quad (11)$$

The "closer" the models are, the smaller the criterion becomes.

3.2. Cohort Selection Methods

Different cohort selection techniques have been suggested in the literature, such as, random impostors, closest imposters (CI) [2], Maximally-Spread Close (MSC) [3], and Maximally-Spread Far (MSF) [3].

In this paper, a new cohort selection method is introduced, called "Close Impostors Clustering" (CIC) method.

The main disadvantage of the Closest Impostors (CI) method is that it may leave the target exposed from a certain "angle" in the feature space, if the training database did not include a "close" impostor there. The CIC method is designed to solve this problem.

3.2.1. The Close Impostors Clustering (CIC) method

The goal of the CIC is to select the best C impostor models from the complete impostor set. The algorithm consists of three main steps: (1) *Outliers removal* - the initial step of the algorithm is to select a subset of N impostors ($N \geq 2C$) from the complete impostors community. The subset of N impostors consists of the candidates for cohort. The impostors excluded from this set are outliers and impostors that are very un-similar to the target that may obscure the correct selection of cohort. (2) *Clustering* - the subset of N impostors is clustered into C clusters. Any one of several clustering methods may be used. (3) *Cohort selection procedure* – One impostor is selected from each cluster as a representative of the given cluster. Any one of several selection methods may be used, for example, select the "closest" (to the target) member of the given cluster.

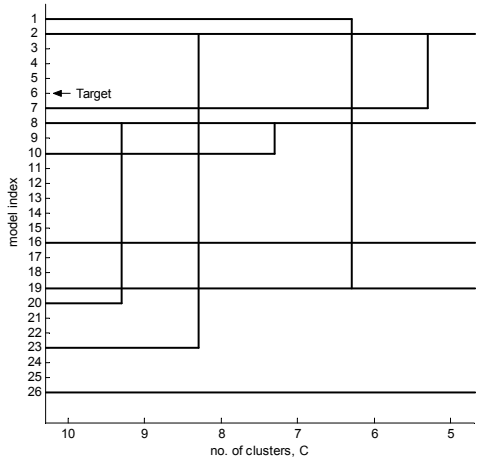

Figure 1: Dendogram example of the clustering process for target #6 from the database used for this paper with 25 impostors (C = 5).

The clustering algorithm used here is a simple single-link hierarchical [7] clustering algorithm. A simple example is depicted in figure 1. Here the target is speaker #6, there are 25 impostors and C = 5. For this example, we start with the 10 closest impostors to the target, $\mathcal{A}(T) = \{1,2,7,8,10,16,19,20,23,26\}$, then, merging the two closest impostors ({8} and {20}) into one cluster, where the representant of the cluster is the one that is closest to the target {8}. This process is repeated till C = 5.

4. Experimental Setup

The experiment was set for text-dependent speaker-verification task. The model for each speaker was trained as a left-to-right Continuous Density Hidden Markov Model (CD-HMM) [8], with 5 states and 2 Gaussians per state.

24-features have been extracted from each 30 msec window (50% overlapping). The features were 12 Mel-Frequency-Cepstral-Coefficients (MFCC) and 12 Δ-MFCC (the time derivatives) [9].

For this paper, four cohort selection methods have been implemented: (see 3.2):

1. The Closest Impostors (CI) method,
2. Maximally Spread Close (MSC) method.
3. Maximally Spread Far (MSF) method.
4. Close Impostors Clusters (CIC) method.

And six normalization methods: $s_1(\mathbf{O}), s_2(\mathbf{O}), s_3(\mathbf{O}), s_4(\mathbf{O}), s_5(\mathbf{O}), s_6(\mathbf{O})$ (eq. (4 - 9)):

These cohort selection methods and normalization methods have been explored and compared to each other and to the un-normalized score. The determination of the cohort size has also been examined.

The algorithm was evaluated with utterances of the Hebrew word /hamesh/ (five), taken from the Hebrew Isolated Digits (HID) database. The database contains high quality speech; sampled at 16KHz with 12 bits resolution.

The 26 male speakers, with the largest number of utterance repetitions, have been selected. Each one of these speakers was used as a target, with 25 impostors. The number of utterances (repetitions of the word 'five') for each target speaker is between 45 - 120, and the number of utterances for each impostor is 45. The first 20 utterances for each speaker are used for training, and the rest utterances for testing.

There are two types of errors to be consider in verification systems: the False Acceptance of an invalid user (FA) and the False Rejection of a valid user (FR or miss). The FA and FR errors are not independent. It is usually possible to reduce one by increasing the other. The well-known Detection Error Trade-off (DET) curve is used to show the relations between these two errors in a given system.

The equal-error rate (EER) is a commonly accepted scalar overall measure of speaker verification system performance. It corresponds to the threshold at which the false acceptance rate is equal to the false rejection (miss) rate.

5. Results and Discussion

The average Equal Error Rate (EER) values from verification results using the four cohort selection methods with the three best score-normalization methods ($s_1(\mathbf{O}), s_3(\mathbf{O}), s_6(\mathbf{O})$) and different cohort size (C = 1, 2,…,8), are depicted in figure 2.

The verification results included speakers chosen as cohorts in the impostors bank. This is undesirable. It was done due to the relatively small database and only after experiments showed that except for a very slight change in overall error rate, results were similar.

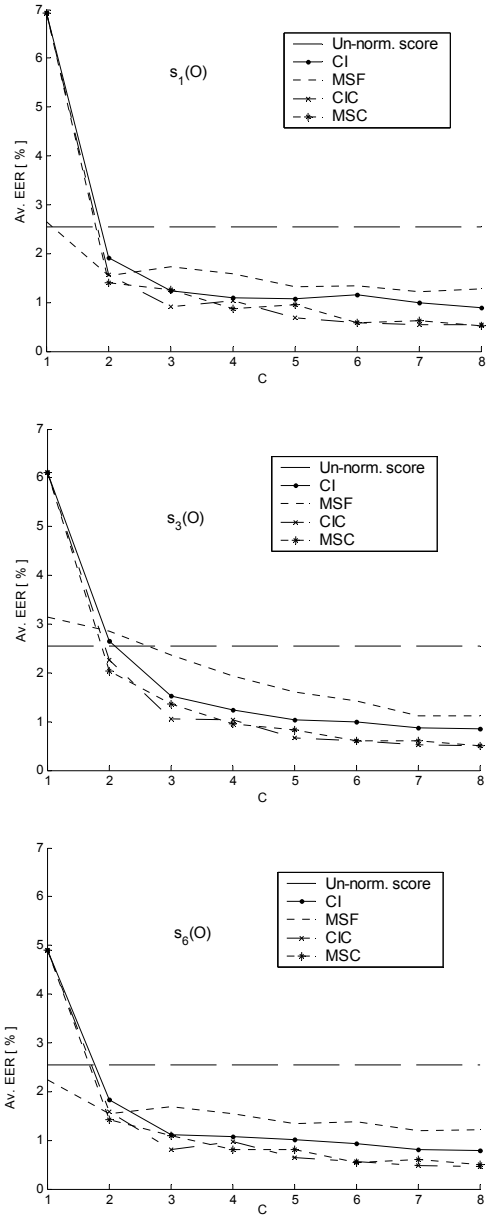

Figure 2: Average EER using different cohort selection methods with three best score-normalization methods ($s_1(\mathbf{O}), s_3(\mathbf{O}), s_6(\mathbf{O})$) as a function of cohort size, C.

Several conclusions may be drawn from the results of figure 2: (a) all normalization methods improve the un-normalized results (average EER of 2.55%) when sufficient cohort size is used. This is a well-known conclusion. (b) The MSF method, consistently provides the worst results for $C >$ 2. This is also a logical conclusion since the method was originally designed for dissimilar population, such as opposite sex impostors [3], which is not the case here. (c) The two best methods are the MSC and the proposed CIC. Table 1 shows that on the average, the CIC with an average EER of 0.79% (with $s_6(\mathbf{O})$ - the best normalization method) is slightly better than the MSC with an average EER of 0.82%.

(d) From all the six normalized methods, $s_6(\mathbf{O})$ was shown to be the best. The best results were achieved with CIC, $s_6(\mathbf{O})$ and $C = 8$. The EER of that system was 0.46%.

Table 1: Average EER (average on $C = 2 \div 8$) of the different cohort selection methods and different score-normalization methods

Cohort selection method	Score normalization method					
	$s_1(\mathbf{O})$	$s_2(\mathbf{O})$	$s_3(\mathbf{O})$	$s_4(\mathbf{O})$	$s_5(\mathbf{O})$	$s_6(\mathbf{O})$
CI	1.19	1.43	1.30	1.50	1.14	1.08
MSF	1.42	1.86	1.77	2.34	1.50	1.41
CIC	0.83	1.03	0.94	1.21	0.98	0.79
MSC	0.88	1.06	0.98	1.15	1.04	0.82

Generally, increasing the cohort size (*C*), provides better verification results. However, in practical systems, it is obvious that increasing the cohort size has its disadvantages. Each cohort model requires memory for the storage of parameters. In addition, increasing the cohort size increases verification time. Therefore it is logical to modify the EER criterion by introducing a punishment for increased cohort size. Thus a criterion, *J*, is suggested:

$$J = E + \nu C \tag{12}$$

where E is the average EER, and ν is a regularization constant. Figure 3 shows the criterion *J* values vs. cohort size, *C*, for each one of the four cohort selection methods with the best cohort normalization technique, $s_6(\mathbf{O})$. For this figure we used $\nu = 0.2$.

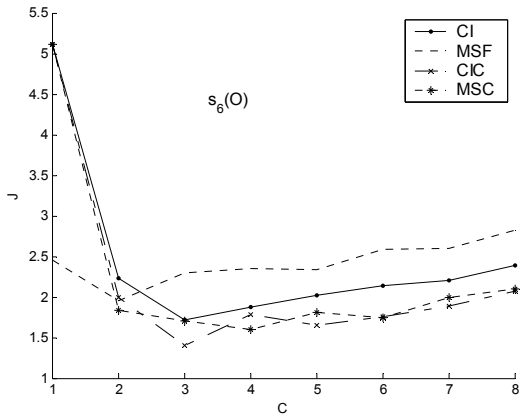

Figure 3: The criterion *J* values vs. cohort size, *C*, for each one of the four cohort selection methods with cohort normalization $s_6(\mathbf{O})$ and $\nu = 0.2$.

From figure 3 one can see that the optimal cohort size is $C = 3$, and the best cohort selection method is the proposed CIC. With the best score-normalization method ($s_6(\mathbf{O})$) and the optimal cohort size ($C = 3$), the best selection method (CIC) has (average) EER of 0.8%, while the second best selection method (MSC) has EER of 1.1%.

Figure 4 shows an average DET curves from verification results. Each curve corresponds to one cohort selection technique (and the un-normalized score) in the best score normalization method ($s_6(\mathbf{O})$) and for the chosen cohort size, $C = 3$. In this case the CIC outperforms all other methods in the range $P_{miss} < 2\%$. The CIC and the MSC are similar in the range $P_{miss} \geq 2\%$.

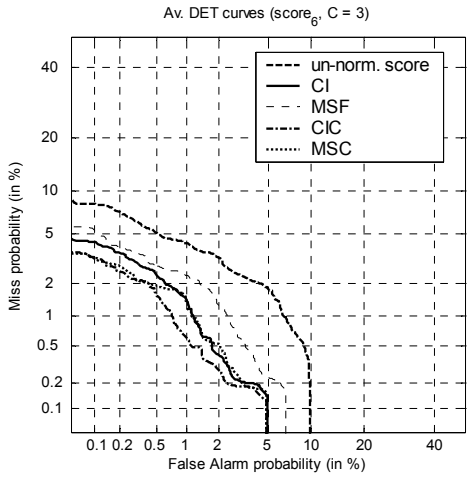

Figure 4: Average DET curves for different cohort selection methods with $s_6(\mathbf{O})$ and $C = 3$.

It has been shown that under the given system and database the CIC cohort selection method outperforms all the tested methods. With these encouraging results, work is underway to improve the clustering algorithm and evaluate the CIC on an extended database and with text-independent tasks.

6. References

[1] A. Higgins, L. Bahler, and J. Porter, "Speaker Verification Using Randomized Phrase Prompting," *Digital Signal Processing*, Vol. 1, pp. 89-106, 1991.

[2] A. E. Rosenberg, J. DeLong, C-H Lee, B-H Juang, and F. K. Soong, "The Use of Cohort Normalized Scores for Speaker Verification," *Proc. ICSLP 92*, pp. 599-602, Nov. 1992.

[3] D. A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models," *Speech Communication*, vol. 17, pp. 91-108, 1995.

[4] D. Tran and M. Wagner, "A Proposed Likelihood Transformation for Speaker Verification," *in Proc. ICASSP 2000*, Turkey, 2000.

[5] D. Tran and M. Wagner, "A Generalized Normalization Method for Speaker Verification," *A Speaker Odyssey 2001 workshop*, Crete, pp. 73-76, 2001.

[6] J. M. Colombi, J. S. Reider, and J. P. Campbell, "Allowing Good Impostors to Test," *Conference Record of the Thirty-First Asilomar Conference on Signals, Systems & Computers, 1997*, pp. 296-300, 1998.

[7] B. D. Ripley, *Pattern Recognition and Neural Networks,* Cambridge University Press, NY, 1996.

[8] L. R. Rabiner, B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[9] S. B. Davis and P. Marmelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28, No. 4, pp. 357-366, Aug. 1980.