A Cohort Methods for Score Normalization in Speaker Verification System, Acceleration of On-line Cohort Methods

Zbyněk Zajíc, Jan Vaněk, Lukáš Machlica, Aleš Padrta
University of West Bohemia, Department of Cyberneticsscow,
Univerzitní 8, 306 14 Plzeň, Czech Republic

Abstract

This paper deals with several cohort methods for score normalization in speaker verification systems. At first, the reasons for score normalization are provided. Next, the principle of score normalization techniques based on Bayesian theorem are explained. The world, cohort, and unconstraint cohort normalization techniques are presented. A new normalization technique, unconstraint cohort extrapolated normalization, is introduced. Experiments on NIST 2002 corpus were performed in order to find which of the normalization methods give the best result. All experiments show that on-line cohort methods (especially unconstraint cohort extrapolated normalization) have outperformed the others. Finally, there is a discussion about time consumption of the on-line methods. An improvement for acceleration of these methods are proposed. The results of experiments on NIST 2002 data set showed same significant improvements of cohort and proposed cohort extrapolated normalization in comparison with the standard world normalization. The acceleration (4-times) on-line cohort methods is shown in another experiment.

1. Introduction

The goal of the verification system is to decide, if the test utterance was spoken by the speaker with claimed identity. The verification system determines a score of the given utterance for the claimed speaker identity. The final decision is given by comparing the score and the threshold. However, the score depends on the operating conditions. It makes the task of threshold determination a very difficult one [1]. To overcome this problem, several score normalization methods were proposed. These methods allow to set a threshold for different operation conditions [2]. The cohort approach is based on a subset of impostor models which represents the closest population to the claimed speaker . The threshold is given as a score of the impostor model. This approach is based on the Bayesian theorem.

The paper is organized in the following order. In the next section, the theory of score normalization in speaker verification task is presented. Individual normalization techniques are described in Section 3. Section 4 describes data used for the experiment. The experiments and appropriate results are discussed in Section 5. Time consumption of the method based on cohorts is discussed in Section 6. In Section 6, the experiment aims at finding the time consumption of the on-line methods. In the next section, an experiment focused on time consumption inspection is introduced. Section 8 deals with results of previous experiments. Finally, in Section 9 the overall conclusions are discussed.

2. Speaker Verification Procedure

Suppose that there is a test speaker utterance O represented by the I feature vectors, i.e. $O = \{o_1, o_2, ..., o_I\}$, and a Gaussian mixture model λs for the claimed identity speaker s. The matching of the test data with the model is denoted as a verification score. In the verification systems based on stochastic models (such as hidden Markov models and Gaussian mixture models), the common score is equal to the likelihood of the utterance given by the claimed speaker model[3].

Final verification decision is determined by the comparison of the score with the threshold:

$$p(O/\lambda_s) \geq T \rightarrow O \in s \quad else \quad O \notin s$$

where O is feature vector of the test utterance, λs is the model of claimed identity of speaker s, $p(O/\lambda s)$ is likelihood of utterance O and model λs, T stands for a threshold.

Absolute value of the threshold is very sensitive to variations in text, speaker behavior, and the operating conditions, especially concerning impostor utterances. The sensitivity causes wide variations in scores, and makes the task of threshold determination a very difficult one. Actually, a final decision evaluates if the verification score is large enough. To overcome the score sensitivity, a score normalization is used.

3. Score normalization using Bayesian approach

In the probabilistic view [4], the final decision can be achieved in the following way:

$$p(\lambda_S/O) > p(\lambda_U/O) \rightarrow O \in \lambda_S \quad else \quad O \in \lambda_U .$$

where $p(O/\lambda s)$ is the score given by claimed speaker model and $p(O/\lambda u)$ is the score given by another speaker model. In the verification task, λs stands for hypohtesis that the test data were spoken by the claimed speaker. By contrast, λu stands for hypohtesis that test data were not from the claimed speaker, i.e. somebody else is the author of the utterance.

By application of Bayesian theorem, it can be shown that

$$\frac{p(O/\lambda_S)}{p(O/\lambda_{\bar{U}})} > \frac{P(\lambda_U)}{P(\lambda_{\bar{S}})} \rightarrow O \in \lambda_S \quad else \quad O \in \lambda_U ,$$

where $\frac{p(O/\lambda_S)}{p(O/\lambda_{\bar{U}})} = l(O)$ is the verification score and $\frac{P(\lambda_U)}{P(\lambda_{\bar{S}})} = T$ is a priori determined threshold.

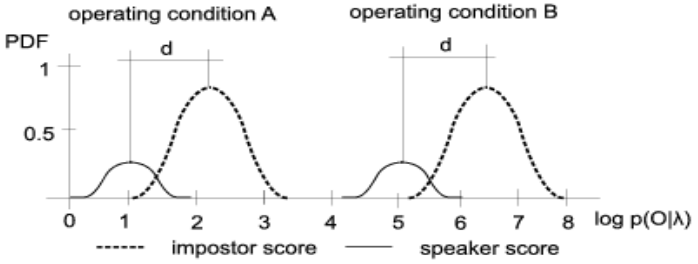
The threshold value can be determined a priori, e.g. based on estimated ratio of authorized and unauthorized accesses. The verification score has to be computed. A frequently used form to represent the verification decision is the following [3]:

$$L(O) = \log p(O/\lambda_S) - \log p(O/\lambda_U)$$

The reason for using this equation is that, division is numerically more demanding then addition (or subtraction) and division is numerically labile for small numbers.

The score $p(O/\lambda s)$ is affected by the operating conditions as well as the score $p(O/\lambda u)$. Thus, the distance between $p(O/\lambda s)$ and $p(O/\lambda u)$ stands constant for various operating conditions. The likelihood $p(O/\lambda u)$ represents a dynamic threshold, which is sensitive to changes of operating condition of utterance O, see Fig.1.

Figure 1.: An illustration of Bayesian approach, d is a distance between two scores for different operating conditions A and B.



There is no problem with computing a score $p(O/\lambda s)$, because speaker model is known. The evaluation of a normalization score $p(O/\lambda u)$ is a problem, because the speaker model λu is not available. This problem can be solved in several ways as follows.

3.1. World (background) model (WM)

One of the possible solution is that, the unknown model λu can be aproximated by the world model λ_{WM} [2], which is created by using utterances from very huge speaker population:

$$p(O/\lambda_U) = p(O/\lambda_{WN})$$

3.2. Cohort normalisation (CN)

In this method [6], every speaker model has a cohort of N similar models $C = \left\{\overline{\lambda}_S^1,...,\overline{\lambda}_S^K\right\}$. These models in the cohort are the most competitive models with the reference speaker model, i.e. these are closest to the reference speaker. The selection of the cohort is made in advance (off-line). The score normalization is computed as:

$$\log p(O/\lambda_U) = \log p(O/\lambda_{CN}) = \frac{1}{K}\sum_{k=1}^{K} \log p(O/\overline{\lambda}_S^k)$$

3.3. Unconstraint cohort normalization (UCN)

The unconstraint cohort model [5] combines the approaches of the world and the cohort model. The cohort consists of the closest models to the claimed speaker model like in CN, but models in the cohort are not chosen a priori (off-line), they are found in the verification process (on-line). This is the advantage over the CN method, because the actual utterance *O* is known by now. The selection of the cohort is more accurate. The normalization score is given by:

$$\log p(O/\lambda_U) = \log p(O/\lambda_{UCN}) = \frac{1}{K}\sum_{k=1}^{K} \log p(O/\overline{\lambda}_S^{\phi(k)})$$

where $\phi(i) \neq \phi(j)\ pro\ i \neq j$ and $\overline{\lambda}_S^{\phi(1)}, \overline{\lambda}_S^{\phi(2)} \ldots \overline{\lambda}_S^{\phi(K)}$ are models, which bring the best score (excluding the claimed speaker model λs) for utterance O.

3.3. Unconstraint cohort extrapolated normalization (UCEN)

A new method is introduced here, which uses the same principles for cohort selection like the method mentioned above. In contrast to UCN, the normalization score is computed using extrapolation of model scores from the chosen cohort. All models in cohort are sorted by score value and then they are extrapolated by straight line. A score $log\ p(O|\ \lambda_{UCEN})$ is a first point on extrapolated straight line. Extrapolation utilizes the fact, that the claimed speaker model λs (if speaker *s* is real author of the utterance *O*) gives the best score.

4. Experimental investigation

4.1. Speech Data

The male part of NIST 2002 corpus [8] is used in these experiments. The data in NIST are taken from the second release of LDC's cellular switchboard corpus.
Actual durations of training data for each speaker are constrained to lie within the range of 110-130 seconds. Each test segment is extracted from a 1-minute excerpt of a single conversation and is the concatenation of all speech from the subject speaker during the excerpt. The duration of the test segment is between 15 and 45 seconds. Evaluation trials for cellular data include both "same number" and "different number" tests. Both speakers are of the same sex (male).
In NIST corpus, there are about 400 target speakers and about 3500 test segments. Each test segment is evaluated against 11 hypothesized speakers of the same sex as the speaker of the segment. Data to create the world model are from NIST~2002 Two speaker detection task. Data to create the cohort model are from the male part training set of NIST~2002 One speaker detection task.

4.2. Feature Vectors, Acoustic Modeling

A modified version of MFCC extended by a voice activity detector is used [7]. The Hamming window has 25 millisecond length and 15 millisecond overlap. A power spectrum is computed by FFT. 25 triangular band filters are set up linearly in the mel-scale between 200 and 4000 Hz. The logarithms of band-filters outputs are decorrelated by the discrete cosinus transformation (DCT). The computed cepstrum has 20 coefficients without the zero-th (log-energy) coefficient. A time sequence of each coefficient is smoothed by 11 frames long Blackman window. Then the delta coefficients are added [8]. A downsampling with factor 3 is applied to the final features for the reduction of amount of data. At the end, frames which were marked as non-speech event are removed. The non-speech event detector estimates the noise level and the speech level independently in each band. If the estimated speech level is lower than the estimated noise level then the actual feature vector is marked as non-speech event and is discarded.
Each speaker is represented by GMM model with 64 mixtures trained by DB-EM algorithm [7]. World model is represented by GMM model with 128 mixtures.

4.3. Description of Experiments

In the experiments, efficiency of the normalization method and the cohort size *N* were examined. For comparison, efficiency of the world normalization method was used (this method is independent of the cohort size).

5. Experimental results and discussions

Table 1 shows the results of all cohort methods which depend on the cohort size. For comparison, the WM method is displayed too.

Table 1. EER [%] of the NIST 2002 corpus.

Cohort size N	Results [EER] UCN	UCEN	CN	WM
2	15.21%	15.05%	18.10%	16.94%
4	15.11%	15.21%	17.53%	16.94%
5	15.11%	**14.63%**	16.90%	16.94%
6	16.01%	**14.63%**	16.15%	16.94%
8	16.60%	**14.63%**	16.05%	16.94%

The on-line cohort methods UCEN have the best results for a small cohort size (5 speakers). UCEN have better results then UCN and the UCEN method is more independent of the cohort size.
The disadvantage of on-line methods (like UCN, UCEN) is their time consumption. It is necessary to find the whole cohort for the claimed identity speaker during the verification trial.

6. The Time-consumption

The speaker verification task is very time-consuming. The verification results have to be obtained in real-time conditions (this is one of the obvious requirement on the verification task). The normalization process decelerates a calculation, because comparisons are largely performed during the normalization (especially in on-line methods). The verification time is strongly dependent on the number of different speakers in the system, which can be used for the cohort. The nearest cohort speakers are selected according to their score for the test utterance *O*. The score is computed for all potential cohort models. Then the models with the biggest score are inserted in the cohort. For accelerating the selection of the cohort models in on-line methods following procedures were proposed (it is possible to combine these two principles of accelerating).

6.1. Downsampling for cohort selection

The scores of all cohort speakers are computed only for each *Z*-th frame of the test utterance (accuracy of calculation sinks along with expansion of *Z*). Then, *M* nearest cohort speakers are found (*M* > *N*, where *N* is size of final cohort). These models create a pre-selected cohort set. The *N* best models are selected from pre-selected cohort using a whole utterance *O*.

6.1. Regression tree of closest speaker models

For another acceleration of the cohort selection, *M* best speaker models are pre-selected off-line (*M* > *N*). Then *N* best models are selected out of *M* models using the test utterance on-line. At first, the uneven binary tree must be constructed, see Fig.3.

Figure 3.: Regression tree for 6 speakers, *sp_i* is the speaker *i* and *λij* is the model for the speaker *i,jj*

Each node represents a set of the nearest speakers, which is divided into two sets of the nearest models. A divergence-like criterion Ds_j [6] is used to obtain a distance between the two models λs and λj, the closest models minimalize a criterion:

$$d^P(\lambda_S, \lambda_j) = \frac{1}{2}(\log p(O/\lambda_S) - \log p(O/\lambda_j))$$

Each node is determined by model, which is created using the train utterance from all node speakers. In the verification process, the tree is browsed from above. In the actual node, the score is computed for two branch models and the better branches score is chosen. The passing of the tree is stopped, when the node has the required number of speakers (*M*). Afterwards, the *N* best speaker models are selected for these *M* speakers.

7. Description of Experiments of time consumption

In this experiment, time consumption of the classic UCEN method and the accelerated UCEN methods are investigated. For comparison, time consumption of the world method is used. The size of pre-selection cohort *M*=30, the size of finally cohort *N*=5 and downsampling *Z*=10 (this is the biggest number without a worsening an EER).
All tests were performed on PC Intel Pentium 4, 1.9GHz, 512MB RAM.

8. Experimental results and discussions about time consumption

Table 2 shows time consumption of the verification using UCEN and WM methods for one trial in the NIST corpus. UCEN means the classic UCEN method with cohort size 5 speakers. UCEN-R is the UCEN method with pre-selection using regression tree, size of pre-selected cohort is 30 speakers and cohort size is 5 speakers. UCEN-D is the UCEN method with preselected cohort size *M=30* speaker using downsampling *Z=10*, final cohort size is 5 speakers too.
Table 2. EER [%] of the male part of NIST 2002 corpus and time-consumption [s] of one trial.

	WM	UCEN	UCEN-R	UCEN-D
EER [%]	16.94%	**14.63%**	15.21%	**14.93%**
Time [s]	**0.32s**	4.10s	1.13s	1.06s

As the results show, the UCEN method is better than the WM method, but it is more time-consuming. Pre-selection with regress tree is less time-consuming then the WM method, but EER is

a little bigger then the classic UCEN method. If proposed acceleration is used, time consumption is rapidly lowered. Thus, time consumption of the UCN method is acceptable for the system used in real-time.

9. Conclusions

This paper describes a different cohort methods for score normalization techniques. There is a lot of disturbances affecting the verification task; the score value depends on the recording conditions, so it is very difficult to reach a final decision. This problem can be solved by normalization. Experiments were performed on NIST 2002 corpora. In comparison with the classic approach base on WM, cohort normalization methods improved the verification, as the experiment results have shown. The on-line methods (especially proposed UCEN method) showed the best results. Moreover UCEN method is more independent of the cohort size.
The main problem of the on-line normalization is time consumption. In this paper, two types of the acceleration of cohort models pre-selection were proposed. The suggested algorithms bring a significant improvement. The verification time for one trial was reduced to 1 second, which represents about 2-3% of the test utterance length.

Acknowledgement

This research was supported by the Grant Agency of Academy of Sciences of the Czech Republic, project No.1ET101470416
and the Ministry of Education of the Czech Republic, project No. MŠMT LC536

References

1. *Schalk H., Reininger H., Euler S.* : A System for Text Dependent Speaker Verification - Field Trial Evaluation and Simulation Results, Eurospeech 2001, pp.783-786, 2001.

2. *Reynolds D., Quatieri T., Dunn R.* : Speaker verification using adapted Gaussian mixture models, Digital Signal Process. 10, pp.19-41, 2000.

3. *Sivakumaran P., Fortuna J., Ariyaeeinia A. M.* : Score Normalisation Applied to Open-Set, Text-Independent Speaker Identification, In EUROSPEECH-2003, pp.2669-2672, 2003.

3. *Rosenberg A. E., DeLong J., Lee C. H., Juang B. H., Soong F. K.* : The use of cohort normalised scores for Speaker Verification, in Proc. ICSLP'92, pp.599-602, 1992.

4. *Auckenthaler R., Carey M., Lloyd-Thomas H.* : Score Normalization for Text-Independent Speaker Verification Systems, Digital Signal Processing 10, pp.42–54, 2000.

5. *Zigel Y., Cohen A.* : On Cohort Selection for Speaker Verification, EUROSPEECH 2003 Geneva, pp.2977-2980, 2003.

6. *Vaněk J., Padrta A.* : Introduction of Improved UWB Speaker Verification System, Proc. of Text Speech and Dialogue 2005, Karlovy Vary, Czech Republic, pp.364-370, 2005.

7. *Vaněk J., Padrta A.* : Optimization of Features for Robust Speaker Recognition, In Speech processing. Prague: Academy of Sciences of the Czech Republic, pp. 140-147.

8. The NIST Year 2002 Speaker Recognition Evaluation Plan},2002, http://www.nist.gov/speech/tests/spk/2002/doc/.