CSLT TECHNICAL REPORT-20160024 [Saturday 7th January, 2017]

Decision Making Based on Cohort Scores for Speaker Verification

Lantian Li[1,2]
, Renyu Wang[4]
, Gang Wang[1,2]
, Caixia Wang[1,2]
and Thomas Fang Zheng[1,3]*

*Correspondence:
fzheng@tsinghua.edu.cn
[1]Center for Speech and Language
Technology, Research Institute of
Information Technology, Tsinghua
University, ROOM 1-303, BLDG
FIT, 100084 Beijing, China
Full list of author information is
available at the end of the article

Abstract

Decision making is an important component in a speaker verification system.
For the conventional GMM-UBM architecture, the decision is usually conducted
based on the log likelihood ratio of the test utterance against the GMM of the
claimed speaker and the UBM. This single-score decision is simple but tends to
be sensitive to the complex variations in speech signals (e.g. text content,
channel, speaking style, etc.). In this paper, we propose a decision making
approach based on multiple scores derived from a set of cohort GMMs (cohort
scores). Importantly, these cohort scores are not simply averaged as in
conventional cohort methods; instead, we employ a powerful discriminative model
as the decision maker. Experimental results show that the proposed method
delivers substantial performance improvement over the baseline system, especially
when a deep neural network (DNN) is used as the decision maker, and the DNN
input involves some statistical features derived from the cohort scores.

Keywords: Decision making; Cohort; GMM-UBM

1 Introduction

Speaker verification aims to verify claimed identities of speakers, and has gained
great popularity in a wide range of applications including access control, forensic
evidence provision and user authentication. After decades of research, lots of pop-
ular speaker verification approaches have been proposed, such as Gaussian mixture
model-universal background model (GMM-UBM) [1], joint factor analysis (JFA) [2]
and its 'simplified' version, the i-vector model [3]. Accompanied with these models,
various back-end techniques have also been proposed to promote the discriminative
capability for speakers, such as within-class covariance normalization (WCCN) [4],
nuisance attribute projection (NAP) [5] and probabilistic LDA (PLDA) [6], etc.
These methods have been demonstrated to be highly successful. Recently, deep
learning has been applied to speaker verification and gained much interest [7, 8].

Within a speaker verification system, decision making is an important componen-
t [9]. To make a decision, the verification system first determines a score for the
test utterance that reflects the confidence that the utterance is from the claimed
speaker, and then compares the score with a predefined threshold. In a typical
GMM-UBM system, the score is often computed as the log likelihood ratio that the

test utterance being generated from the GMM of the claimed speaker and the UBM. This single-score decision is simple and efficient, but it tends to be quite sensitive to variations in speech signals, e.g., in terms of text contents, channel conditions and speaking styles. This sensitivity means that choosing an appropriate threshold is rather difficult, or leading to error-pron decisions.

To deal with this score variation, various score normalization techniques have been proposed. Most of the normalization approaches, according to [10], can be explained using the Bayes' theorem. Among these approaches the cohort normalization is particular interesting. This approach chooses a set of cohort speakers who are close to the genuine speaker, and for each test utterance, it computes a set of 'cohort scores' on the models of these speakers. These cohort scores then replace the UBM to normalize the score of the test utterance against the claimed speaker [11, 12]. Using cohort models tends to model the alternative hypothesis more accurately, due to its more flexible structure compared to a single UBM. However, the existing methods based on cohort models do not fully utilize the information involved in the cohort scores: they are just simply averaged to normalize the target score, which is still a single-score approach.

This paper presents a new cohort approach that utilizes the cohort scores in a more effective way. Specifically, we propose to make decisions on the whole cohort scores (formulated as a score vector), and employ a powerful discriminative model to make the decision. Our assumption is that the knowledge involved in the cohort scores is more than a mean average, but as complex as their distributions, their ranks, spanning areas, etc. Fully utilization of these rich information results in a true multi-score decision making, which is expected to be more reliable than the traditional single-score approach.

The technique presented in this paper involves three steps: (1) Firstly, a set of cohort models is constructed by a clustering algorithm; (2) Secondly, for each test utterance, scores are estimated among the claimed speaker GMM, the global UBM and the cohort GMMs; (3) Finally, a classification model (SVM or DNNs) is employed to make the decision based on some features derived from the scores derived above.

The layout of this paper is organized as follows. Section 2 presents the proposed cohort-based decision making framework. The experiments are presented in Section 3, and Section 4 concludes the paper.

2 Cohort-based decision making framework

In a typical GMM-UBM speaker verification system, the score likelihood ratio of a test utterance is computed over the GMM of the claimed speaker model and UBM. Then the likelihood ratio will be compared with a predefined threshold. If it is higher than the threshold, the test utterance will be accepted, else rejected. We argue that this naive decision making approach is unreliable and less robustness because this likelihood ratio only describes the distance between the claimed GMM and UBM, and it does not make use of the world speakers and corresponding score information. Therefore, we design a cohort-based decision making framework, as shown in Fig. 1. This framework is made up of three parts: cohort selection, feature design and discriminative model training.
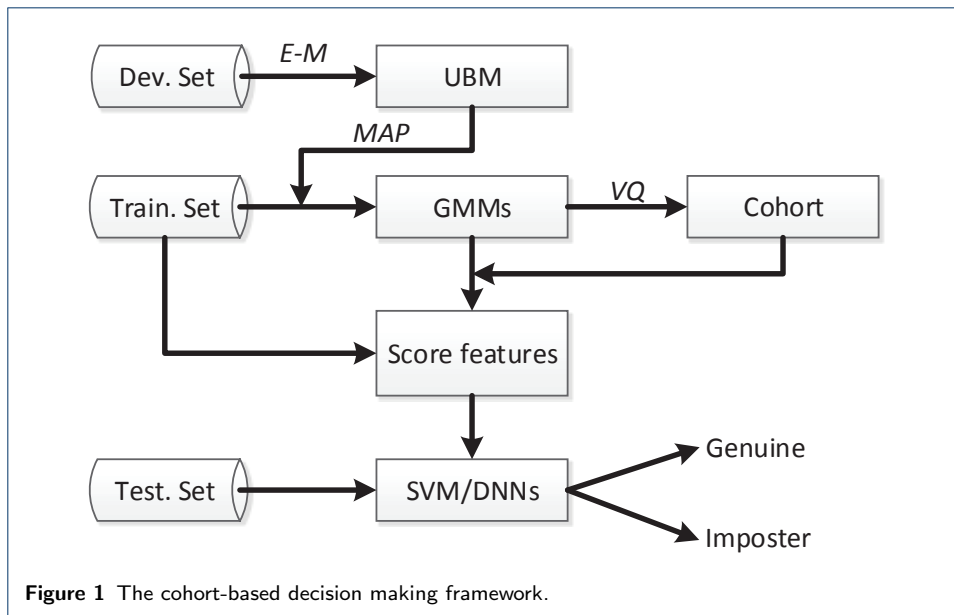
Figure 1 The cohort-based decision making framework.

2.1 Cohort selection

A vector quantization (VQ) method [13] based on the K-means algorithm was utilized to conduct the speaker model clustering. The centroid of each cluster represents a reference speaker, and all the reference speakers build the 'cohort'. We chose a weighted K-L distance to measure the distances among Gaussian mixture models, given by:

$$D(\lambda_1, \lambda_2) = \sum_{i=1}^{M} w_i KL(N_i^1, N_i^2) \tag{1}$$

$$KL(N_1, N_2) = \frac{1}{2}(\mu_1 - \mu_2)^T (\Sigma_1^{-1} - \Sigma_2^{-1})(\mu_1 - \mu_2) \tag{2}$$

where λ_1 and λ_2 are two Gaussian mixture models, and w_i is the weight of ith Gaussian component. Note that, for fast computation, only the mean parameters are adapted in the GMM-MAP process, while the weights and variances of the GMMs are the same as UBM. Equ. 2 is used to measure the distance between two multi-dimensional Gaussian distributions.

Given a set of speaker GMMs $\lambda = (\lambda_1, \lambda_2, ..., \lambda_n)$ and λ_{c^i} that is the cluster centroid where speaker i is assigned to. The optimization objective is to minimize the within-class cost J, and finally each cluster centroid is regarded as one 'cohort' model.

$$J = \frac{1}{N} \sum_{i=1}^{N} D(\lambda_i, \lambda_{c^i}) \tag{3}$$

2.2 Feature design

Once the cohort models (CGMMs) have been determined, a set of cohort scores are calculated on the claimed speaker GMM, UBM and CGMMs respectively for each test utterance. We seek to use these cohort scores to explore some potential knowledge and design more discriminative features on genuine and imposter speaker models. In this part, three cohort-based score features are discussed.

2.2.1 Cohort-based score normalization

The inspiration of this feature comes from the conventional score normalization techniques [10]. For a test feature vector X, the normalized score $\widetilde{L}_\lambda(X)$ is given as follows:

$$\widetilde{L}_\lambda(X) = \frac{L_\lambda(X) - \mu_\lambda}{\sigma_\lambda} \tag{4}$$

where λ represents a claimed speaker model, and μ_λ, σ_λ is estimated from the cohort scores.

2.2.2 Rank position

Assuming the size of cohort is K, for each test trial, a $(1+K)$-dimensional score vector is calculated based on GMM and CGMMs. And we think that the likelihood scores on the genuine speaker GMMs are at the top-rank position in the $(1+K)$-dimensional score vector, while for the imposter speakers, it lies in a random rank position.

2.2.3 Rank of score differences

Similar assumption with the rank position, we also believe that the distribution of cohort scores on the genuine speaker models is different from that on imposter speaker models. For each test utterance, the score feature is calculated by subtracting the likelihood score on the claimed speaker GMM from that on each cohort CGMM. It describes a high-dimensional cohort-based score distribution instead of the UBM space. After ranking it, this score feature also covers the information of rank position, and has strong discriminability on genuine and imposter speaker models. This assumption will be verified in Section 3.3.

2.3 Discriminative model training

Based on these features derived from the cohort scores, discriminative models (e.g., support vector machine (SVM) and deep neural networks (DNNs)) can be directly optimize with respect to the speaker verification task, i.e., the genuine/imposter speaker decision.

3 Experiments

3.1 Database

The experiments are performed on a database called '*CSLT-DSDB*' (Digit String Database) that was jointly created by CSLT (Center for Speech and Language Technologies), Tsinghua University and Beijing d-Ear Technologies, Co. Ltd. The text

of all recordings is the text-prompted digit strings. The recordings were conducted using different mobile microphones, sampled at 16 kHz with 16-bit precision.

- Training set: It contains an approximate size of 1 GB data (about 200 males and 200 females) recorded in an ordinary office environment. And it is used for the UBM training.
- Development set: It contains 280 enrollment utterances covering 145 speakers and $2,874$ test utterances. And it is used for cohort selection and feature design.
- Evaluation set: It involves 92 speakers. For each speaker, there are text-prompted digit strings of about 40 seconds in length for speaker model training; and 8-16 randomly generated digit strings each of which is an 8-digit string for verification. There are overall $1,220$ test utterances and $1,220$ target trials and $111,020$ non-target trials.

3.2 Experimental setup

The acoustic feature was the conventional 39-dimensional Mel frequency cepstral coefficients (MFCC), which involves 13-dimensional static components plus the first and second order derivatives. The UBM consisted of 256 Gaussian components and was trained with the training set. Note that this setting is 'almost' optimal in our experiments, i.e., using more Gaussian components cannot improve system performance. And the baseline of GMM-UBM system on the evaluation set was 1.621% in terms of EER (Equal Error Rate).

Besides, with the maximum a posterior (MAP) algorithm, 280 speaker GMMs were adapted from UBM. And The K-means algorithm was used to cluster the 280 speaker GMMs into a suitable cohort. Fig. 2 presents the function between the number of clusters and the clustering cost J. It can be observed that when the number of clusters exceeds 10, the clustering cost J has already been converged. Therefore, the size of cohort was set to 10.

Figure 2 The function between the number of clusters and the clustering cost.
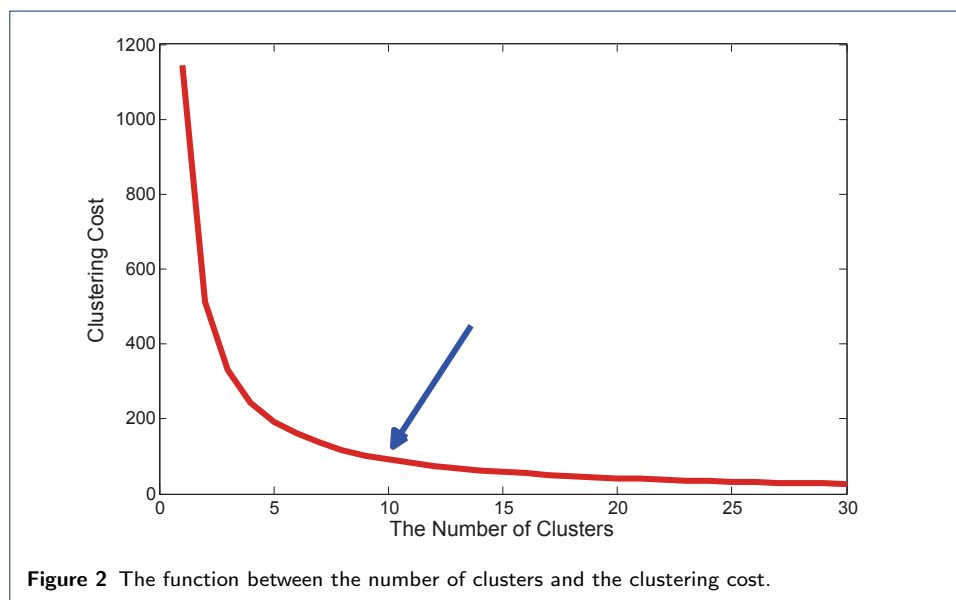
In order to select the discriminative score feature, $6,115$ target trials and $5,748$ imposter trials were selected from the development set. Considering the unbalanced data problem [1], only the top two scores were selected from all the imposter speaker models.

3.3 Feature design

3.3.1 Cohort-based score normalization

According to Equ. 4, the normalized score for each test was calculated, and the system performance was 1.639% in EER. It shows reasonable performance and can be considered as an available score feature.

3.3.2 Rank position

From Fig. 3, we observed that this rank position has certain discriminability. Nearly all the likelihood scores on the genuine speaker GMMs are at the first rank position, while for imposter speaker models, the rank position distribution approximately satisfies a Gaussian distribution with the mean of 5.


Figure 3 The statistical histogram of rank-position distribution.
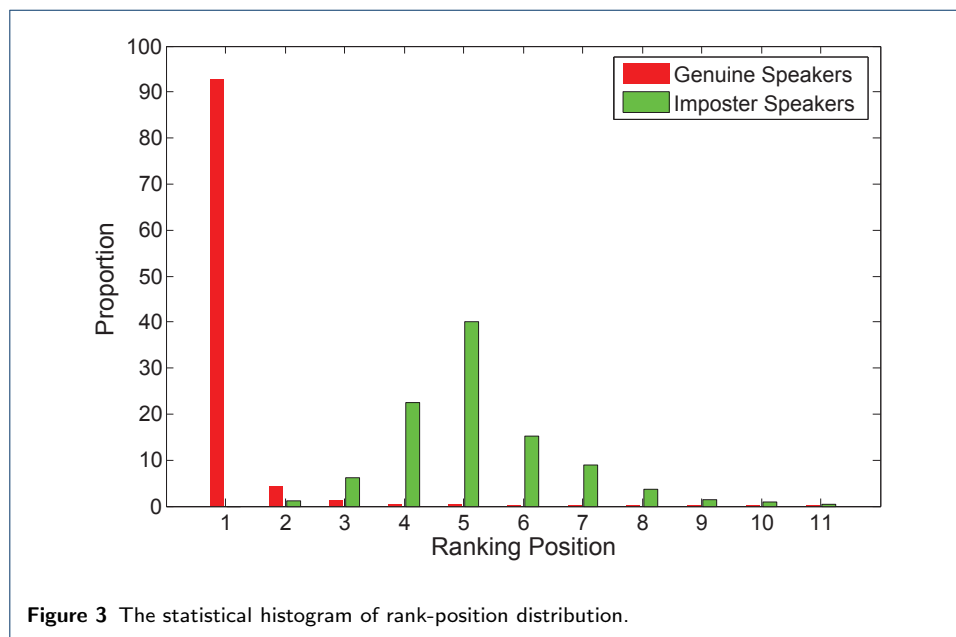
3.3.3 Rank of score differences

To provide an intuitive understanding of the discriminative capability of this feature, the rank of score differences of all the test trials are plotted in a two-dimensional space via T-SNE [14]. As shown in Fig. 4.

It can be seen that there exists a distinct non-linear boundary between genuine speaker models and imposter speaker models. That is to say, this 'rank of score differences' has strong discriminative capability on genuine and imposter speaker models.

[1]The number of target samples and imposter samples will be highly unbalanced, one or some few target samples against large amount of imposter samples. And learning from such unbalanced data will result in biased SVM/DNNs models.

Figure 4 The distribution of the 'sorted score differences' on genuine speakers and imposter speakers (via T-SNE).
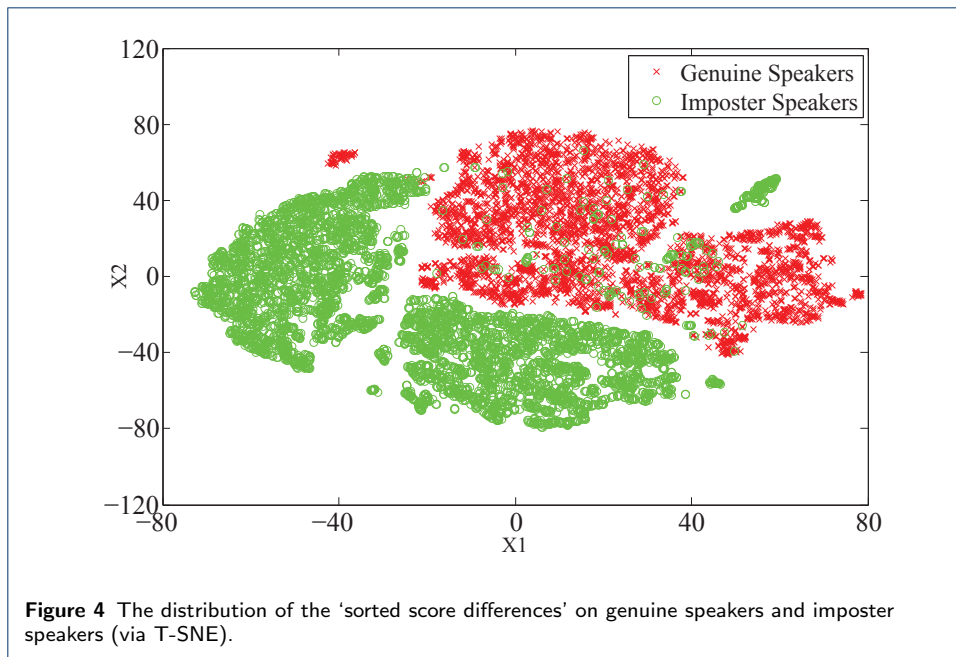
3.4 Discriminative model training

With these cohort-based score features, discriminative models can be optimized with respect to discriminating the genuine/imposter speakers. In this paper, both the SVM and DNNs models were trained as the decision maker for speaker verification system.

3.4.1 SVM-based scoring

The SVMs were trained for each cohort-based score feature with the linear kernel function. Results are shown in Table 1 on condition of C1-C3. Note that 'norm' is the 'Cohort-based score normalization', 'r-pos' is the 'Rank position', 'r-diff' is the 'Rank of score differences' and $\sqrt{}$ represents that related features are chosen as the input of SVMs.

Table 1 The SVM-based discriminative scoring evaluation system.

Condition	score	norm	r-pos	s-diff	EER(%)
C1	$\sqrt{}$	$\sqrt{}$	–	–	1.598
C2	$\sqrt{}$	–	$\sqrt{}$	–	1.574
C3	$\sqrt{}$	–	–	$\sqrt{}$	1.475
C4	$\sqrt{}$	$\sqrt{}$	$\sqrt{}$	–	1.625
C5	$\sqrt{}$	$\sqrt{}$	–	$\sqrt{}$	1.475
C6	$\sqrt{}$	–	$\sqrt{}$	$\sqrt{}$	1.475
C7	$\sqrt{}$	$\sqrt{}$	$\sqrt{}$	$\sqrt{}$	1.479

3.4.2 DNN-based scoring

The DNN models were trained with these cohort-based score features, and the decision was made by logistic regression model at the soft-max layer. Note that for different input feature, the experimental results can be optimized with tuning of the DNNs structure such as the number of hidden units and hidden layers. Whereas, in order to unify the experimental configuration, we just set the number of hidden

layer units 10 times as much as the dimension of input features, and there is only 1 hidden layers. The results are shown in Table 2 on the condition of C1-C3.

Table 2 The DNN-based discriminative scoring evaluation system.

Condition	score	norm	r-pos	r-diff	EER(%)
C1	√	√	–	–	1.556
C2	√	–	√	–	1.639
C3	√	–	–	√	1.148
C4	√	√	√	–	1.639
C5	√	√	–	√	1.230
C6	√	–	√	√	2.049
C7	√	√	√	√	1.077

3.4.3 Feature combination

From Table 1 and Table 2, it can be seen that in condition C1-C3, both the SVM- and DNN-based scoring offer clear performance improvement than the GMM-UBM baseline 1.621%. Therefore, a feature combination scheme was proposed by concatenation these score features together. Experiment results are shown on condition of C4-C7. It can be observed that the performance of this simple feature combination is inconsistent, and we attribute it to the feature redundancy because all these features are embedded from the cohort scores. Besides, the overall feature combination C7 on DNN-based scoring system obtains the best performance.

4 Conclusions

This paper presents a decision making method based on cohort scores instead of the traditional single decision score. Some potential discriminative features are embedded from cohort scores, and then more powerful discriminative models are trained as the decision maker. Experimental results show that the proposed 'rank of score differences' with SVM/DNN-based scoring model can obtain stable and better system performance than the GMM-UBM baseline. Moreover, a feature combination scheme is proposed to further improve system performance. Future work involves designing more robustness score-level discriminative features and more reasonable cohort selection approaches.

Acknowledgment

Author details
[1]Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China. [2]Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, ROOM 1-303, BLDG FIT, 100084 Beijing, China. [3]Department of Computer Science and Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China. [4]School of Linguistic Sciences, Jiangsu Normal University, Jiangsu, China.

References
1. D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.
2. P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.
3. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

4. A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for svm-based speaker recognition," in *Proc. INTERSPEECH'06*, 2006.
5. W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.
6. S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *ICCV'07*. IEEE, 2007, pp. 1–8.
7. P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition," in *Odyseey'2014*. Odyssey, 2014.
8. V. Ehsan, L. Xin, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP),2014*. IEEE, 2014, pp. 4052–4056.
9. J. P. Campbell Jr, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.
10. R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, 2000.
11. R. A. Finan, R. I. Damperb, and A. T. Sapeluk, "Impostor cohort selection for score normalisation in speaker verification," *Pattern Recognition Letters*, vol. 18, no. 9, 1997.
12. A. E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, and F. K. Soong, "The use of cohort normalized scores for speaker verification," in *ICSLP'92*, 1992, pp. 4052–4056.
13. R. Gray, "Vector quantization," *IEEE Assp Magazine*, vol. 1, no. 2, pp. 4–29, 1994.
14. L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Machine Learning Research*, 2008.