

USING SECOND ORDER STATISTICS FOR TEXT INDEPENDENT SPEAKER VERIFICATION

Ran D. Zilca

IBM Research, T. J. Watson Center
P.O. Box 218, Yorktown Heights, NY 10598
Email: ranzilca@ieee.org¹

ABSTRACT

This paper describes a computationally simple method to perform text independent speaker verification using second order statistics. The suggested method, called Utterance Level Scoring (ULS), allows obtaining a normalized score using a single pass through the frames of the tested utterance. The utterance sample covariance is first calculated and then compared to the speaker covariance using a distortion measure. Subsequently, a distortion measure between the utterance covariance and the sample covariance of data taken from different speakers is used to normalize the score. Experimental results from the 2000 NIST speaker recognition evaluation are presented for ULS, used with different distortion measures, and for a GMM system. The results show a relative degradation of 40% in accuracy with respect to GMM, indicating ULS as a viable alternative to GMM whenever computational complexity and verification accuracy needs to be traded. ULS is intended to be used as a first stage of an efficient open set speaker identification system. All speakers will be first scored by ULS and then the top scoring speakers will be scored again using GMM.

1 INTRODUCTION

Most speaker verification systems use likelihood computation at the frame level. Each speech frame is scored separately against the speaker and background models, and frame scores are accumulated to an utterance score. This approach may be related to the use of multi-modal models, where local clusters in the feature space are modeled explicitly (e.g. Vector Quantization (VQ) [1], GMM [2]). Each speech utterance includes only a subset of the phonetic space modeled by a text independent model, and therefore cannot be scored against a model as a whole. Consequently, each frame is scored mostly against local feature clusters. For example, for VQ models, the codebook vector that is the closest to the current frame is explicitly found, and for GMM the frame's score is affected mostly by the closest Gaussian components. The use of multi-modal speaker and background models results in extensive computation since scoring against each model involves going through all the individual frames for likelihood computation. In this paper we use simple second order statistics for speaker and background modeling, i.e. a

uni-modal approach, where the model is simply the sample covariance matrix of the training data. This approach allows going through the frames of a speech utterance only once for computing its sample covariance, and subsequently scoring against speaker and world models with a single operation. Previous work on covariance modeling focused on speaker identification tasks, with no background normalization (e.g. [3]-[5]). We recently suggested a simple extension to such methods, termed ULS, that allows to perform speaker verification with a normalized score [6]. The current contribution describes further experimentation conducted as part of the 2000 NIST speaker recognition evaluation. The experiments show that improved performance is obtained by using a distortion measure described in [5] in conjunction with ULS. In addition, for this distortion measure it is shown that using a single universal covariance for normalization is superior to the use of cohort normalization. Our intended use for ULS is as a rapid scoring stage for an efficient open set speaker identification system.

The rest of this paper is organized as follows: the following section describes the different variants of covariance modeling that we use for speaker verification. Section 3 describes experiments conducted on the 2000 NIST speaker recognition evaluation corpus [7], and the paper ends with conclusions and future work.

2 COVARIANCE MODELS (CM)

2.1 CM with Frame Level scoring

In its simplest form, second order statistics may be used as speaker and background models following the prevalent likelihood ratio approach. This may be viewed as a reduction of GMM, where a single Gaussian models the training data. When cepstral features are used together with cepstral mean subtraction, the mean equals zero. The resulting model therefore includes only a covariance matrix. We refer to this method as Single Gaussian (SG). Verification using SG is obtained by calculating a likelihood ratio (or log likelihood difference) to obtain normalized frame scores, that are then accumulated to an utterance score according to the standard GMM verification framework.

¹ Work performed while with R&D division, Amdocs Israel.

Let us term the claimant speaker covariance model, C_s , which is calculated as follows:

$$C_s = \frac{1}{N_s - 1} \sum_{i=1}^{N_s} (x_i - \mu_s)(x_i - \mu_s)^t \quad (1)$$

where N_s is the number of training feature vectors, x_i is the i^{th} training vector and μ_s is the sample mean of the training data. Suppose we are using a single covariance, C_w , as a background model (i.e. a world model, or Universal Background Model - UBM). C_w is calculated similar to C_s , using vectors gathered from a diverse speaker population. The normalized score of a single frame, s , is obtained by subtracting the two scores:

$$s = s_s - s_w = \log(p_s(x)) - \log(p_w(x)) \quad (2)$$

where x is a feature vector of d components extracted from a speech frame and $p_s(x)$ and $p_w(x)$ are the explicit Gaussian Probability Density Functions (PDF) expressions, e.g.:

$$p_s(x) = \frac{1}{(2\pi)^{d/2} |C_s|^{1/2}} e^{-\frac{1}{2}(x-\mu_s)^t C_s^{-1} (x-\mu_s)} \quad (3)$$

When using SG, the means μ_s and μ_w may be either used directly as described above, or set to zero in order to disregard the exact location in feature space and address only the ‘‘shape’’ of the PDF’s. As mentioned earlier, when using cepstral coefficients with cepstral mean subtraction, the means equal zero by nature regardless of the modeling approach. The utterance score is obtained by averaging the frame scores across the utterance.

2.2 CM with Utterance Level Scoring (ULS)

An alternative approach to SG would be to first calculate the sample covariance of an utterance and then score the utterance with a single operation with respect to the speaker and background covariances, i.e. adapting a model-based approach rather than a frame-by-frame approach [8]. We term this approach Utterance Level Scoring, or ULS. The main computational advantage of ULS over likelihood based methods, is that although the utterance sample covariance is calculated frame by frame, similar to equation (1), each frame calculation is performed regardless of the speaker model. Consequently, once the utterance covariance is obtained, scoring with respect to a model (either speaker or world) involves only a single operation. Speaker verification involves scoring with respect to a speaker model and at least one more additional world or background model, in order to obtain a normalized score. Therefore, ULS allows to obtain a normalized score while going through each frame only once

(unlike likelihood based systems where the process of going through the entire utterance frame by frame needs to be repeated for every model). This computational advantage is more significant when several models rather than a single world model are used for score normalization. For example, when using cohort based normalization on [2] and when gender and handset-type dependent world models are used not only for normalization but also as gender and handset-type detectors, by using the best scoring world model for normalization.

CM may be used together with ULS using distortion measures between Gaussian PDF’s that depend only on the covariance matrices. We suggest calculating the normalized utterance score, S as follows:

$$S = -\frac{D(C_{utt}, C_s)}{D(C_{utt}, C_w)} \quad (4)$$

where C_{utt} is the sample covariance of the utterance to be verified, and $D(\cdot)$ is a distortion measure that depends on two covariance matrices. As seen from this expression, a speech utterance having a small distortion measure with respect to the claimant speaker’s CM, relative to its distortion measure from the world CM, will produce a high score, and vice versa. In this paper we use measures that were previously tested for closed set speaker identification with no score normalization to obtain a normalized score with (4).

2.2.1 Divergence Shape Ratio (DSR)

The *divergence shape* is an information-theoretic measure between two Gaussian classes, introduced by Campbell as a measure of dissimilarity between a reference speaker utterance and a tested utterance [4]. Its derivation is based on calculating the symmetric directed divergence between the two Gaussian classes. Ignoring the Gaussian means, or setting them to zero, results in an expression that depends only on the covariance matrices, termed the Divergence Shape (DS):

$$DS_{1,2} = DS(C_1, C_2) = \frac{1}{2} \text{tr}[(C_1 - C_2)(C_2^{-1} - C_1^{-1})] \quad (5)$$

where $\text{tr}(\cdot)$ is the trace operator and $C_{1,2}$ are the covariance matrices of the two classes.

We suggest using this measure for speaker verification *score normalization*, as part of the approach stated in (4). Also, while the DS was tested previously only on clean speech using Line Spectra Pairs (LSP), the experiments presented herein are performed on conversational telephone-quality speech, using Mel Frequency Cepstral Coefficients (MFCC). We suggest a new normalized ULS method, called the *Divergence Shape Ratio* (DSR), where the utterance score, S , is calculated following equation (4) with the divergence shape as the distortion measure, i.e.:

$$S = -\frac{DS(C_{utt}, C_s)}{DS(C_{utt}, C_w)} \quad (6)$$

2.2.2 Sphericity Measure Ratio (SMR)

The *Sphericity Measure Ratio* (SMR) is computed by applying (4) to a distortion measure termed arithmetic harmonic sphericity measure or the Sphericity Measure in short (SM). The expression for the SM is derived from the analysis in [5]. Same as for DS, this measure was previously used only for closed set speaker identification. Also, all preceding experimentation was performed on TIMIT/NTIMIT data. We used the SM for background normalization in speaker verification in accordance with (4), and tested it with the 2000 NIST speaker recognition evaluation corpus, a subset of the switchboard corpus. A detailed analysis on using second order statistical measures and the derivation of the arithmetic-harmonic sphericity measure may be found in [5].

The SM is computed as follows:

$$SM_{1,2} = SM(C_1, C_2) = \frac{1}{2} tr(C_1 C_2^{-1}) tr(C_2 C_1^{-1}) \quad (7)$$

and the SMR:

$$S = -\frac{SM(C_{utt}, C_s)}{SM(C_{utt}, C_w)} \quad (8)$$

3 EXPERIMENTS

3.1 2000 NIST Speaker Recognition Evaluation Corpus

Speaker verification using the suggested method was evaluated on the 2000 NIST speaker recognition evaluation corpus [7], which is derived from the switchboard corpus. Each conversation is labeled with the telephone number, presumably related to a particular handset. In the 2000 evaluation all verification trials were from a different telephone number than the one used for enrollment, and are therefore assumed to originate from a different handset. In addition, Each speech segment is labeled according to the *type* of handset, either electret or carbon-button. The corpus consists of 926 claimant speakers, tested with 6096 target trials (speech that was uttered by the claimed speaker) and 60,476 imposter trials. There are no cross gender trials. Each speaker model is created from two minutes of training data (taken from a particular phone conversation). Test segments are of varying duration, most of them in the range of 15-45 seconds. A detailed description of the evaluation data and conditions may be found in [7].

3.2 Signal Processing

An identical signal processing procedure was applied to all the systems, in order to allow for a fair comparison. The speech vector was first segmented into 25ms frames every 12.5ms. Each frame was multiplied by a Hamming window and analyzed using a voicing detector. Only voiced frames are selected for further use. Very aggressive voicing detection was applied, filtering about half of the speech data. FFT based 18th order MFCC were then extracted from each speech frame using a triangle filter bank. Cepstral mean subtraction was then performed to compensate for convolutional channel effects; i.e. the mean vector of each test utterance was subtracted from each single MFCC vector.

3.3 Results

In order to compare the suggested covariance modeling methods to the current state of the art, a GMM system was also tested. Each speaker GMM includes 512 Gaussian components, and is adapted from a world GMM using the Bayesian adaptation procedure and adaptation parameters described in [10]. GMM verification was performed using a 5 component approximation [10]. For all the systems (second order and GMM), four different world models were trained according to gender and handset type (electret or carbon button). The world model used for score normalization of tested utterances was chosen according to the gender and training handset type. The world models were trained using the entire test portion of the 1999 NIST speaker recognition evaluation. For the CM systems training the world model simply involved calculating the sample covariance matrix of this data. For the GMM the world model was trained using DB-GMM [9]; The data was clustered using the K-means algorithm, and the GMM parameters were calculated as follows: the weights were given by the relative number of vectors in each cluster, and the means and variances were the sample mean and variance of each cluster. No EM iterations were performed.

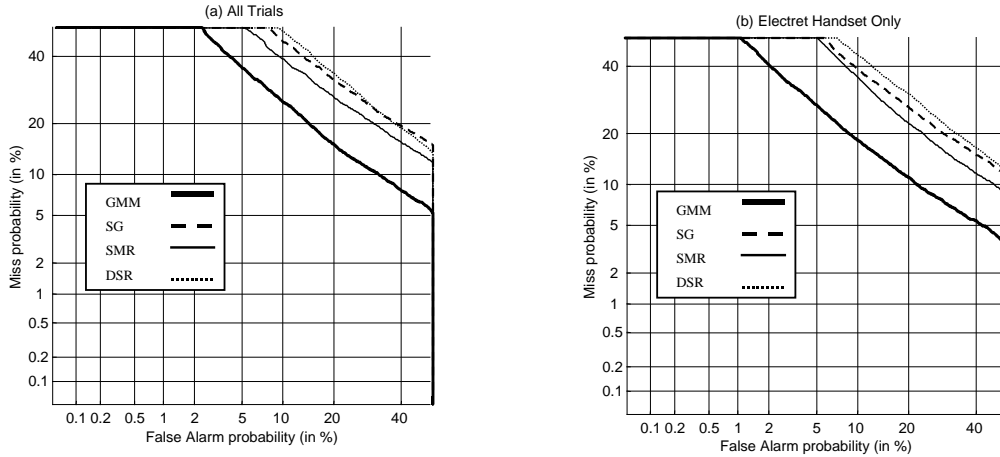


Figure 1. DET curves comparing GMM, SG, SMR and DSR for all trials, and for electret data

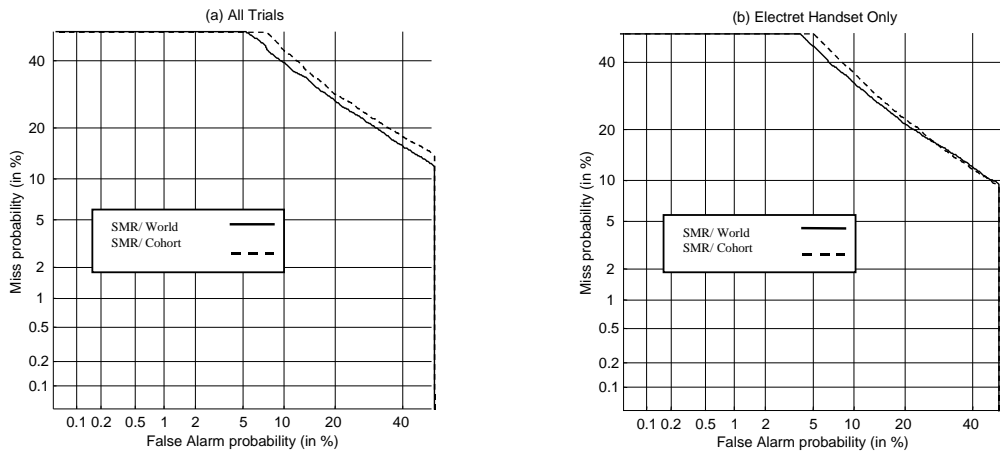


Figure 2. DET curves SMR using UBM and cohort normalization for all trials, and for electret data

The DET curves for the entire evaluation and for electret handset only (i.e. speakers were trained using an electret handset and scored only against electret handset trials) are shown in figure 1. A curve was produced for each of the tested systems: GMM, SG, SMR, and DSR. As expected, GMM accuracy is significantly better than the second order systems. For all trials, The CM systems have Equal Error Rates (EER) in the range of 24% to 28% comparing to 17% for GMM. For electret only, CM performance ranges between 21% and 26% while the GMM EER is 15%. It is evident that SMR outperforms the other CM systems. The relative degradation in EER from GMM to SMR is 40% for both all trials and electret data. Considering the significant alleviation of computational complexity with respect to GMM, SMR may therefore be useful for applications where computational requirements should be traded with verification accuracy. Our experience with computation time is that SMR training (i.e. sample covariance calculation) is four orders of magnitude

faster than GMM, and verification is about three orders of magnitude faster than GMM using a five component approximation [10]. Although this proportion may be different depending on the computing environment, and on the application of various speed-up methods (e.g. [11]), the difference in computational complexity is clearly very significant. It should also be noted that the design of the 2000 NIST evaluation corpus is aimed at applications that require verification to be performed “anywhere”, i.e. from any telephone handset, and that the difference in accuracy is smaller for applications where the same handset is commonly used [6].

3.4 Cohort Normalization Vs. World Normalization

The results presented in figure 1, clearly indicate that among the suggested CM methods, SMR performs best. Since SMR follows the ULS approach, scoring involves a single operation once the utterance covariance has been calculated,

therefore the computational cost of using cohort normalization [2] instead of a single world model is relatively small. When using cohort normalization, no world model is pre-trained, and for the verification procedure the denominator of equation (4) is replaced by an average of scores with respect to few speaker models. Figure 2 shows the performance of SMR for cohort normalization comparing to normalization using a world model. The cohort speakers were taken from the 1999 NIST speaker recognition evaluation. As with the world models, different normalization was used according to gender and handset type. There are 857 male electret cohort speakers, 591 male -carbon button, 1449 female-electret, and 523 female -carbon button. Figure 2 clearly indicates that cohort normalization performs worse than world normalization for SMR. Although the system performance clearly depends on the selection of the cohort speakers, recognition accuracy is generally improved as the number of cohort speakers increases. Therefore, since the experiment was performed with large cohorts, we expect world normalization to perform better than cohort normalization for SMR regardless of the choice of cohort speakers.

4 CONCLUSIONS

This paper describes text independent speaker verification using second order statistics. Based on previous studies on covariance modeling for speaker identification, we suggest an extension that allows to obtain a normalized score for speaker verification. This method, termed Utterance Level Scoring (ULS), uses a normalized distortion measure between covariance matrices. It allows to score the utterance with respect to each model in a single operation after computing the utterance sample covariance. The ULS approach provides substantial computational simplification with respect to likelihood ratio classifiers such as GMM. The degradation in verification accuracy with respect to GMM was evaluated in handset mismatch conditions in the 2000 NIST speaker recognition evaluation. The best performing ULS method, the Sphericity Measure Ratio (SMR), was shown to perform 40% worse relative to GMM using the same feature extraction and world modeling settings. For electret-only conditions SMR obtains 21% EER comparing to 15% for GMM. An additional experiment with SMR has shown that normalization based on single world modeling outperforms cohort based normalization.

ULS/SMR is computationally simpler than GMM by orders of magnitude, and is therefore suitable for applications where severe computational constraints exist. It is intended to be used as part of an efficient open set speaker identification system, with a large number of enrolled speakers. In order to save the significant computation of scoring the utterance against all the speaker GMM's, the utterance covariance may be first scored using ULS to determine the top scoring speakers. Only these speaker GMM's are then-scored to determine the best scoring speaker.

The poor modeling provided by second order statistics alone is more evident for world normalization where a very large amount of data is available. Future work will therefore focus on expanding CM/ULS to support more elaborated world modeling.

This may be obtained by allowing GMM world models and using the distance measure suggested in [8], or by training a large number of world CM's - each on a small amount of data (e.g. particular handset models, speaker accents, etc.).

REFERENCES

- [1] F. K. Soong, A. E. Rosenberg, L. R. Rabiner and B. H. Juang, "A Vector Quantization Approach to Speaker Recognition," *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc.*, Tampa, FL, pp. 387-390, 1985.
- [2] D. A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models," *Speech Communication*, Vol. 17, pp. 91-108, 1995.
- [3] H. Gish, "Robust Discrimination in Automatic Speaker Identification," *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc.*, pp. 289-292, 1990.
- [4] J. P. Campbell, "Speaker Recognition: A Tutorial," *Proceedings IEEE*, Vol. 85, No. 9, pp. 1437-1462, Sept. 1997.
- [5] F. Bimbot, I. M. Magrin-Chagnolleau and L. Mathan, "Second Order Statistical Measures for Text Independent Speaker Identification," *Speech Communication*, Vol. 17, pp. 177-192.
- [6] R. D. Zilca, "Text Independent Speaker Verification Using Covariance Modeling," *IEEE Signal Processing Letters*, April 2001.
- [7] <http://www.nist.gov/speech/tests/spk/2000/index.htm>
- [8] H. S. M. Beigi, S. H. Maes and J. S. Sorensen, "A Distance Measure Between Collections of Distributions and its Application to Speaker Recognition," *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc.*, 1998.
- [9] R. D. Zilca and Y. Bistriz, "Distance-Based Gaussian Mixture Model for Speaker Recognition over the Telephone," *Proc. ICSLP-2000, Beijing, China*, pp. 1001-1003, 2000.
- [10] D. A. Reynolds, "Comparison of Background Normalization Methods for Text-Independent Speaker Verification," *Proc. Eurospeech-97, Rhodes, Greece*, pp. 963-966, 1997.
- [11] J. McLaughlin, D. A. Reynolds, and T. Gleason, "A Study of Computation Speed-ups of the GMM-UBM Speaker Recognition System," *Proc. Eurospeech-99, Budapest, Hungary*, Vol. 3, pp. 1215-1218, 1999.