Unsupervised Speaker Adaptation based on the Cosine Similarity for Text-Independent Speaker Verification

Stephen Shum[1], Najim Dehak[1], Reda Dehak[2], James R. Glass[1]

[1] MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, MA 02139, USA
[2] Laboratoire de Recherche et de Developpement de l'EPITA (LRDE), Paris, France
{sshum,najim,jrg}@csail.mit.edu, reda@lrde.epita.fr

Abstract

This paper proposes a new approach to unsupervised speaker adaptation inspired by the recent success of the factor analysis-based Total Variability Approach to text-independent speaker verification [1]. This approach effectively represents speaker variability in terms of low-dimensional *total factor vectors* and, when paired alongside the simplicity of cosine similarity scoring, allows for easy manipulation and efficient computation [2]. The development of our adaptation algorithm is motivated by the desire to have a robust method of setting an adaptation threshold, to minimize the amount of required computation for each adaptation update, and to simplify the associated score normalization procedures where possible. To address the final issue, we propose the Symmetric Normalization (S-norm) method, which takes advantage of the symmetry in cosine similarity scoring and achieves competitive performance to that of the ZT-norm while requiring fewer parameter calculations. In subsequent experiments, we also assess an attempt to replace the use of score normalization procedures altogether with a Normalized Cosine Similarity scoring function [3].

We evaluated the performance of our unsupervised speaker adaptation algorithm under various score normalization procedures on the 10sec-10sec and core conditions of the 2008 NIST SRE dataset. Using results without adaptation as our baseline, it was found that the proposed methods are consistent in successfully improving speaker verification performance to achieve state-of-the-art results.

1. Introduction

In recent years, factor analysis-based approaches have achieved the state of the art for text-independent speaker detection tasks. In an effort to enhance the classical method of modeling speakers using Gaussian Mixture Models (GMMs), methods developed in Joint Factor Analysis (JFA) present powerful tools to better represent speaker variabilities and compensate for channel and, more generally, session inconsistencies [4]. It is, nevertheless, extremely difficult to capture and characterize every source of variability in just a single enrollment session [5]. Indeed, having multiple enrollments of the same speaker over different sessions would help average out these sources of noise and, ultimately, provide a better representation of the speaker model. This motivates the ongoing investigation of speaker model adaptation. Furthermore, it would be even better if these additional enrollments could occur automatically - that is, without a priori knowledge that the utterance actually belongs to the target speaker. The setting in which we update speaker models based on utterances processed during testing is the problem of unsupervised speaker adaptation [6].

As it stands, JFA produces highly variable scores that require the application of score normalization techniques, such as the ZT-norm, to show its performance gains [4]. Previous work has shown that the application of these normalization techniques in the unsupervised speaker adaptation scenario requires a significant amount of additional computation with each adaptation update [5]. Recently, a factor analysis-based approach to speaker recognition using just a cosine similarity metric between low-dimensional vectors proved highly effective [2]. Unlike traditional JFA, this "Total Variability Approach" avoids the joint estimation of separate speaker and session spaces and factors, and is less reliant on the application of score normalizations [1].

In this paper, we utilize the speed and convenience of the cosine similarity metric and develop an algorithm for unsupervised speaker adaptation. We further propose a new score normalization strategy that reduces the need for additional computation after each adaptation update and, ultimately, simplifies the entire procedure. The rest of this paper is organized as follows. In Section 2, we describe the "Total Variability Approach" and its use of cosine similarity scoring. A new algorithm for unsupervised speaker adaptation is developed in Section 3, while Section 4 presents the simplified score normalization strategy. In Section 5 we present the results of our experiments, and then discuss the latest direction of progress in Section 6 before concluding.

2. The Total Variability Approach

Classical JFA modeling defines respective subspaces for the speaker and the channel factors, then estimates them jointly [2]. A more recent approach represents all the factors in a (single) total variability space with no distinction made between speaker and session subspaces [1]. The speaker- and session-dependent supervector[1] M is defined as

$$M = m + Tw \tag{1}$$

where m is the speaker- and session-independent supervector commonly taken from a Universal Background Model (UBM)[2], T is a rectangular matrix of low rank that defines the total variability space, and w is a random vector with a normally distributed prior $\mathcal{N}(0, I)$. The components of w are referred to as

[1] A supervector is composed by stacking the mean vectors from a GMM.

[2] A UBM is a large GMM trained to represent the speaker-independent distribution of features [7].

the "total factors", and w will be referred to as a "total factor vector."

2.1. Parameter Training and Estimation

We begin from scratch with a UBM Ω consisting of C Gaussian mixture components defined in some feature space of dimension F. In this space, we train the total variability matrix T by following a similar process to that of learning the eigenvoice matrix of JFA and is fully detailed in [1]. The main difference between the two is that in training the eigenvoice of JFA, all recordings of a given speaker are considered to belong to the same person, whereas in training T, each instance of a given speaker's set of utterances is regarded as having been produced by a different speaker.

The total factor vector w is a latent variable whose posterior distribution can be determined using Baum-Welch statistics from the UBM [2]. Suppose our given utterance u is represented as a sequence of L frames $u = \{y_1, y_2, ..., y_L\}$. Then the relevant Baum-Welch statistics are

$$N_c\left(u\right) = \sum_{t=1}^{L} P\left(c|y_t, \Omega\right) \tag{2}$$

$$F_c\left(u\right) = \sum_{t=1}^{L} P\left(c|y_t, \Omega\right) y_t \tag{3}$$

where $c = 1, ..., C$ is the index of the corresponding Gaussian component and $P\left(c|y_t, \Omega\right)$ corresponds to the posterior probability of generating the frame y_t by mixture component c. Now for our purposes, we define the centralized first order Baum-Welch statistics based on the mean of the mixture components in the UBM:

$$\tilde{F}_c\left(u\right) = F_c\left(u\right) - N_c\left(u\right) m_c \tag{4}$$

$$= \sum_{t=1}^{L} P\left(c|y_t, \Omega\right)\left(y_t - m_c\right) \tag{5}$$

where m_c is the mean of mixture component c. The total factors vector for utterance u can be obtained using the following equation:

$$w = \left(I + T^t \Sigma^{-1} N(u) T\right)^{-1} \cdot T^t \Sigma^{-1} \tilde{F}(u) \tag{6}$$

where $N(u)$ is the diagonal matrix of dimension $CF \times CF$ whose diagonal blocks are $N_c(u)I$, $(C = 1, ..., C)$ and $\tilde{F}(u)$ is a supervector of dimension $CF \times 1$ obtained by concatenating all the centralized first order Baum-Welch statistics $\tilde{F}_c(u)$. Here, Σ is a diagonal covariance matrix of dimension $CF \times CF$ that is estimated during the training of T. It models the residual variabilities not captured by the total variability matrix T [8].

2.2. Inter-session Compensation

One marked difference between the total variability representation and JFA is that there is no explicit compensation for intersession variability. Once the data has been projected into the lower dimensional space, however, standard compensation techniques can still be applied in a straightforward and computationally efficient manner. Upon experimentation with a variety of different methods, it was found that the best performance can be achieved with a combination of Linear Discriminant Analysis (LDA) followed by Within-Class Covariance Normalization (WCCN). The following paragraphs will briefly summarize this work; more details can be found in [2].

In order to better discriminate between classes, LDA looks to define a new orthogonal basis (rotation) within the feature space. In this case, different speakers correspond to different classes, and a new basis is sought to simultaneously maximize the between-class variance (inter-speaker discrimination) and minimize the within-class variance (intra-speaker variability). We define these axes using a projection matrix A composed of the eigenvectors corresponding to the highest eigenvalues of the general equation

$$\Sigma_b \nu = \lambda \Sigma_w \nu \tag{7}$$

where λ is the diagonal matrix of eigenvalues. The matrices Σ_b and Σ_w correspond to the between-class and within-class covariance matrices, respectively, and are calculated as follows:

$$\Sigma_b = \sum_{s=1}^{S} \left(w_s - \bar{w}\right)\left(w_s - \bar{w}\right)^t \tag{8}$$

$$\Sigma_w = \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} \left(w_s^{(i)} - \bar{w}_s\right)\left(w_s^{(i)} - \bar{w}_s\right)^t \tag{9}$$

where $\bar{w}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} w_s^{(i)}$ is the mean of the total factor vectors $w_s^{(\cdot)}$ for each speaker s with n_s corresponding to the number of utterances for that speaker, and S is the total number of speakers.

The idea behind WCCN [9] is to scale the total variability space by a factor that is inversely proportional to an estimate of the within-class covariance matrix. This has the effect of de-emphasizing directions of high intra-speaker variability and thus makes for a more robust scoring operation. The within-class covariance matrix is estimated using the total factor vectors from a set of development speakers as

$$W = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} \left(A^t w_s^{(i)} - \tilde{w}_s\right)\left(A^t w_s^{(i)} - \tilde{w}_s\right)^t \tag{10}$$

where A is the LDA projection matrix as described previously, $\tilde{w}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} A^t w_s^{(i)}$ is the mean of the LDA-projected total factor vectors $w_s^{(\cdot)}$ for each development speaker s, n_s corresponds to the number of utterances for the respective speaker, and S is the total number of development speakers. We use the Cholesky decomposition of the inverted matrix, $W^{-1} = BB^t$, whose application can be viewed as scaling the total variability space by B. The result of applying both LDA and WCCN is a new vector w', denoted

$$w' = \frac{B^t A^t w}{\|B^t A^t w\|} \tag{11}$$

where w is extracted from (6) and the normalization operation $\|B^t A^t w\|$ is performed in anticipation of the cosine similarity scoring to be discussed in the following section. From now on, we will use w' exclusively to refer to the total factor vectors.

2.3. Cosine Similarity Scoring

The simple cosine similarity metric has been applied successfully in the total variability space to compare two supervectors for making a speaker detection decision [2]. Given two total factor vectors generated by (6) via the projection of two supervectors in the total variability space and the compensation for inter-session variabilities as in (11), a target w'_{target} from a

known speaker and a test w'_{test} from an unknown speaker, the cosine similarity score is given as

$$\text{score}\left(w'_{\text{target}}, w'_{\text{test}}\right) = \frac{\left(w'_{\text{target}}\right)^t \left(w'_{\text{test}}\right)}{\|w'_{\text{target}}\| \|w'_{\text{test}}\|} \quad (12)$$

$$= \left(w'_{\text{target}}\right)^t \left(w'_{\text{test}}\right) \quad (13)$$

$$\gtrless \theta \quad (14)$$

where θ is the decision threshold and (13) is the simple dot product. We can neglect the normalization in (12) since the $w'_{(\cdot)}$ were already normalized in (11). This scoring function is considerably less complex than the log likelihood ratio (LLR) scoring operations used in JFA [10].

3. Unsupervised Adaptation

Introducing additional data in a model adaptation procedure has the effect of averaging out sources of noise, such as session variabilities, that might otherwise have an adverse effect on the model itself. In the single-utterance speaker enrollment scenario, the initial representation of a speaker may be strongly affected by the channel's characteristics; thus, an adaptation procedure that can incorporate additional data of the same speaker (from either a different channel or the same one) would, for the most part, improve and strengthen the model's representation of the speaker. In the case of unsupervised adaptation, however, even before an adaptation procedure can be carried out, a decision must be made as to whether or not the test utterance belongs to the hypothesized speaker. Care must be taken in making this decision, as the inclusion or exclusion of the utterance in adaptation will affect all subsequent decisions and results for the hypothesized speaker model.

3.1. Previous Work

A recursive procedure for progressive model adaptation in a JFA-based speaker verification system was proposed in [5]. Whenever new adaptation data is available, the proposed algorithm computes the posterior distribution of the speaker-dependent hyperparameters using the current hyperparameters as a prior. Using LLR scoring, the method sets a fixed, predefined adaptation threshold to decide whether or not to adapt the speaker model using a given test utterance. Additionally, an "adaptive T-norm score normalization" method was introduced to combat the observed drifting of T-normalized scores caused by additional adaptation updates.

While the work in [5] saw many promising results, the increase in computational complexity associated with both calculating the posterior distributions of the hyperparameters and recomputing the adapted speaker model's Z-norm parameters after each adaptation update is not negligible. Furthermore, even though most of the parameters for the adaptive T-norm procedure can be pre-computed offline, the overhead cost is still rather significant, especially in the case of ZT-normalization where each T-norm model requires the computation of its respective Z-norm parameters. Lastly, the success of the results in [5] was largely dependent on the choice of adaptation threshold, which is difficult to determine a priori.

Our proposed approach is motivated by the following desiderata:

- To have a simple and robust method of setting the adaptation threshold; for example, setting it to be the same as the optimal decision threshold[3] of some development dataset.
- To minimize the amount of computation necessary during each unsupervised adaptation update.
- To simplify the score normalization procedures wherever possible.

3.2. A New Approach

The recent "Total Variability Approach" and the simplicity of its cosine similarity score motivates the development of an algorithm for unsupervised speaker adaptation that is simple, direct, and efficient: Let $W_s = \{w'_s\}$ denote the set of of total factor vectors pertaining to the identity of known speaker s such that w'_s denotes the vector extracted from the initial enrollment utterance. Let $t'_i, i = 1, ..., k$ denote the total factor vector extracted from the test utterance for each of k tests. We define a fixed decision threshold θ (based on the optimal decision threshold as previously mentioned) and the following averaging function based on the cosine similarity score defined in (13):

$$\widetilde{\text{score}}_N\left(W_s, t'_i\right) = \frac{1}{\|W_s\|} \sum_{w'_s \in W_s} \text{score}_N\left(w'_s, t'_i\right) \gtrless \theta \quad (15)$$

where score_N is some normalized version (e.g. Z-, T-, ZT-norm, etc.) of the cosine similarity score in (13) and $\|W_s\|$ denotes the cardinality of the set W_s. If $\widetilde{\text{score}}_N \geq \theta$, then the test utterance t'_i, as well as its relevant score normalization parameters, is added to W_s and the decision is confirmed.

The low dimensionality of the total factor vectors in the total variability space allows for convenient storage in memory and a lower computation cost, while the symmetry in the function $\widetilde{\text{score}}_{(\cdot)}$ nicely allows for a fixed threshold. Section 6 will further discuss the ideas for future work in the design of this function.

4. Score Normalization

This section discusses the details of score normalization pertaining to unsupervised speaker adaptation. In particular, it has been observed that application of ZT-norm in a JFA-based system achieves the best results [4]. The details of this implementation are described further in [1] and [11], but in summary, ZT-norm can be seen as the combination of applying the Z-norm followed by the T-norm, thus respectively compensating first for interspeaker variability and then for intersession variability.

To make these notions more explicit, we will briefly review each score normalization procedure, discuss the parameter updates that need to occur during unsupervised adaptation, then describe our proposed simplifications to this problem.

4.1. Procedures

Score normalization methods are applied to reduce the variability of the decision scores. These techniques are applied based on the assumption that the distribution of target speaker and impostor scores follow two distinct normal distributions [1]. That is, given a total factor vector w'_s obtained from the enrollment of speaker s and a set of total factor vectors $S_s = \{w'_j\}$ corresponding to target speaker s, the scores between w'_s and the elements $\{w'_j\}$ are assumed to follow a normal distribution with mean μ_s and variance σ_s^2. Similarly, given a set of total factor

[3] The optimal decision threshold is usually set based on the results of tests without speaker model adaptation.

vectors $I = \{u'_k\}$ extracted from a set of impostor utterances, the scores between w'_s and the elements of I are assumed to follow a normal distribution $\mathcal{N}\left(\mu_Z(s), \sigma_Z^2(s)\right)$. Both of these score distributions are dependent on the target speaker, however, which means that without some form of global normalization scheme, a decision threshold will need to be set for each individual speaker.

The purpose of zero normalization (Z-norm) is to scale and shift the distribution of scores between a target speaker w'_s and a set of impostor utterances I to the standard normal distribution. For some arbitrary test utterance t'_i and target speaker w'_s, we have

$$\text{score}_Z\left(w'_s, t'_i\right) = \frac{\text{score}\left(w'_s, t'_i\right) - \mu_Z(s)}{\sigma_Z(s)} \tag{16}$$

where $\text{score}\left(w'_s, t'_i\right)$ is calculated as in (13). The parameters $\mu_Z(s)$ and $\sigma_Z(s)$ for a given speaker s can be computed prior to testing as follows:

$$\mu_Z(s) = \frac{1}{\|I\|} \sum_{u'_k \in I} \text{score}\left(w'_s, u'_k\right) \tag{17}$$

$$\sigma_z(s) = \sqrt{\frac{1}{\|I\|} \sum_{u'_k \in I} \left(\text{score}\left(w'_s, u'_k\right) - \mu_z(s)\right)^2} \tag{18}$$

where $\|I\|$ denotes the cardinality of the set of impostor utterances. The Z-norm procedure allows for the finding of a universal decision threshold that is independent of the target speaker s.

As Z-norm applies to every speaker model, we can perform an analogous procedure on every test utterance. Known as test normalization (T-norm) [11], the total factor vector t'_j extracted from test utterance j is scored against a set of impostor models $M = \{v'_k\}$ to obtain the parameters $\mu_T(j)$ and $\sigma_T(j)$ below:

$$\mu_T(j) = \frac{1}{\|M\|} \sum_{v'_k \in M} \text{score}\left(v'_k, t'_j\right) \tag{19}$$

$$\sigma_T(j) = \sqrt{\frac{1}{\|M\|} \sum_{v'_k \in M} \left(\text{score}\left(v'_k, t'_j\right) - \mu_T(j)\right)^2} \tag{20}$$

The T-normalized score between some arbitrary target speaker w'_r and the test utterance t'_j is similar to (16):

$$\text{score}_T\left(w'_r, t'_j\right) = \frac{\text{score}\left(w'_r, t'_j\right) - \mu_T(j)}{\sigma_T(j)} \tag{21}$$

The T-norm addresses the problem of session variability and acoustic mismatch between speaker enrollment and testing conditions [11].

Combining the Z- and T-norm procedures has been shown to improve performance in speaker verification tasks [1]. To do so, we compute Z-norm parameters for all speakers, including those in the set of impostor models $M = \{v'_k\}$. Then to compute the T-norm parameters for test utterance t'_j, we continue to use (19) and (20) but change the function $\text{score}\left(v'_k, t'_j\right)$ to $\text{score}_Z\left(v'_k, t'_j\right)$ so as to use the Z-norm parameters of v'_k.

4.2. ZT-Norm Parameter Updates

In the unsupervised adaptation procedure described in Section 3.2, if a test utterance t'_i is accepted into the set of speaker total factor vectors W_s (renamed $t_i^{\prime(s)} = t'_i$ for clarity), then in the subsequent trial involving t'_j, the evaluation of $\widetilde{\text{score}}_{ZT}\left(W_s, t'_j\right)$ will involve the calculation of $\text{score}_{ZT}\left(t_i^{\prime(s)}, t'_j\right)$, and the corresponding ZT-normalization will require an appropriate Z-norm parameter corresponding to $t_i^{\prime(s)}$. As such, whereas the approach in [5] requires us to precompute Z-norm parameters for each adapted T-norm model and then re-compute the Z-norm parameters of the updated speaker model after each adaptation step, our procedure described above only requires the offline computation of Z-norm parameters for each test utterance t_i in addition to the standard Z-norm parameters for each speaker model and the standard T-norm parameters for each test utterance.

To summarize, we can simply pre-compute the Z-norm parameters for each potential element of the set of speaker total factor vectors $W_s = \{w'_s, t_a^{\prime(s)}, t_b^{\prime(s)}, ...\}$ in addition to precomputing the T-norm parameters for each test utterance t'_i. After these pre-computations, no other normalization-related computations are necessary during verification trials.

4.3. Symmetric Normalization

The nature of the Total Variability Approach is to apply factor analysis as a method to extract speaker-relevant features. Indeed, we can see that the extraction of total factors from an enrollment or test utterance follows the exact same process. Furthermore, by the nature of the cosine similarity metric, the scoring between any two total factor vectors is symmetric, which suggests that there is really no distinction to be made between a Z-norm impostor utterance u' and a T-norm impostor model v'. As such, for any given speaker model or test utterance, we should be able to associate a set of score normalization parameters that are determined by some universal procedure. After defining this procedure, we will propose a slight modification to the cosine similarity scoring function that can apply these parameters in symmetric fashion.

In the symmetric normalization procedure (S-norm), we define $\Lambda_{Imp} = \{I_{Imp} \bigcup M_{Imp}\}$ to be the union of the original list of impostor utterances and a list of impostor models. For each test utterance $t'_i, i = 1, ..., k$ and each initial speaker enrollment utterance $w'_j, j = 1, ..., n$, its respective S-norm parameters are the mean and standard deviation of the scores between the given utterance (i.e. t'_i or w'_j) and all the elements of Λ_{Imp}. Now, just as ZT-norm applies the Z-norm to the speaker model w'_s and a T-norm to the test utterance t'_i, the application of S-norm similarly applies the normalization parameters from both the speaker model w'_s and the test utterance t'_i as follows:

$$\text{score}_S\left(w'_s, t'_i\right) = \frac{\text{score}\left(w'_s, t'_i\right) - \mu_{ws}}{\sigma_{ws}} + \frac{\text{score}\left(w'_s, t'_i\right) - \mu_{ti}}{\sigma_{ti}} \tag{22}$$

where μ_{ws}, σ_{ws} are the S-norm parameters of w'_s and μ_{ti}, σ_{ti} are the S-norm parameters of t'_i. In the implementation of the unsupervised adaptation algorithm, the function (15) used to combine scores, $\widetilde{\text{score}}_S(\cdot, \cdot)$, is modified by replacing every instance of $\text{score}_N(\cdot, \cdot)$ with $\text{score}_S(\cdot, \cdot)$ as above.

The simplicity introduced by the S-normalization exploits the use of total factor vectors as features that describe the speaker for any given utterance, enrollment or test. The result is a universal procedure for calculating S-norm parameters and a correspondingly simple method for score normalization.

4.4. Normalized Cosine Similarity

We have nearly come full circle with the story of score normalization. It began with the introduction of Z- and T-norm procedures, which were combined into a more powerful and complex ZT-norm procedure. In the spirit of simplifying this procedure for the sake of parameter updates during unsupervised adaptation, we introduced the S-norm procedure. Thus the only remaining step in the puzzle is the ultimate simplification: to remove the need for score normalization procedures altogether.

In [3], the authors analyze the effect of score normalization techniques in the cosine similarity metric and obtain a new, extended cosine similarity scoring function that does not require normalization parameters to be pre-computed for any utterance. The score normalization is instead incorporated into the cosine similarity function itself by the following form:

$$\text{score}\left(w_s', w_t'\right) = \frac{\left(w_s' - \bar{w'}_{imp}\right)^t \left(w_t' - \bar{w'}_{imp}\right)}{\|Cw_s'\| \cdot \|Cw_t'\|} \qquad (23)$$

where the calculation of the matrix C and vector $\bar{w'}_{imp}$ are detailed in [3].

5. Experiments

5.1. Setup

Our experiments were run on cepstral features extracted every 10ms using a 25ms Hamming window. We used 19 mel-frequency cepstral coefficients along with the log energy to create 20-dimensional feature vector, which was then subjected to feature warping [12] using a sliding window three seconds in length. From here, delta and delta-delta coefficients were calculated every five frames to finally produce 60-dimensional feature vectors.

Table 1 shows the list of corpora and their respective roles in the creation of our system. Our gender-dependent UBM consisted of 2048 Gaussians (1024 per gender) and the rank of the total variability matrix T was chosen to be 400, while the LDA projection matrix A was of rank 200. We used 1200 impostor utterances to determine the relevant Z-norm parameters and 250 impostor models for T-norm. As previously described, the parameters for S-norm and the normalized cosine similarity were estimated using the union of the Z- and T-norm impostor sets.

The columns of Table 1 denote the LDC releases of the following corpora:

- **S-2** - Switchboard-2, Phases II and III;
- **Cell** - Switchboard Cellular, Parts 1 and 2;
- **NIST2004** - 2004 NIST[4] SRE[5] data;
- **NIST2005** - 2005 NIST SRE data;
- **Fisher** - Fisher English Database.

5.2. Results

We carried out our experiments using the female part of the 2008 NIST SRE dataset, focusing on the condition where training and testing are done on 10 seconds of speech (10sec-10sec condition). We used the optimal a posteriori decision threshold from the 2006 NIST SRE dataset as our fixed decision/adaptation threshold for testing. To determine this threshold, we computed all relevant test scores from the 2006 NIST SRE dataset without the use of speaker model adaptation, then picked the decision boundary that minimized the Detection Cost Function (minDCF). This threshold was fixed as both the decision and adaptation threshold during the testing of our system on the 2008 NIST SRE data.

[4]NIST: National Institute for Standards and Technology

[5]SRE: Speaker Recognition Evaluation

	S-2	Cell	NIST 2004	NIST 2005	Fisher
UBM	X	X	X	X	
T	X	X	X	X	X
A (LDA)	X	X	X	X	
W (WCCN)			X	X	
Z-norm	X	X	X		
T-norm				X	
S-norm	X	X	X	X	
Norm-Cos	X	X	X	X	

Table 1: *List of corpora and their respective uses.*

Our baseline systems used LDA, WCCN, and cosine similarity scoring as described in Section 2, as well as some form of normalization, but without model adaptation of any kind. We compared the performances of the proposed unsupervised adaptation algorithm (Section 3.2) under the use of ZT-norm, S-norm, and normalized cosine similarity with those of the baselines. The testing procedure and unsupervised adaptation updates were done in accordance with the NIST SRE protocol [6].

	English Trials		All Trials	
	EER (%)	minDCF	EER (%)	minDCF
Baseline, ZT-norm	12.45%	0.0575	16.55%	0.0726
Adapted, ZT-norm	12.01%	**0.0534**	15.83%	0.0709
Baseline, S-norm	12.01%	0.0585	16.96%	0.0708
Adapted, S-norm	11.13%	0.0563	16.16%	0.0701
Baseline, Norm-Cos	11.42%	0.0573	15.83%	0.0673
Adapted, Norm-Cos	**10.68%**	0.0560	**15.42%**	**0.0660**

Table 2: *Comparison of results (10sec-10sec condition) between our proposed method of unsupervised adaptation with various normalization procedures and the respective "baseline" approach (without adaptation).*

Table 2 shows the results of the initial experiment on the 10sec-10sec condition with the female part of the 2008 NIST SRE data. These results are given according to an Equal Error Rate (EER), which corresponds to the point at which the False Rejection rate is equal to the False Acceptance rate, and the minimum value of the Detection Cost Function (minDCF) that was set by NIST during the 2008 SRE. We can see that, for the English Trials, the best EER of 10.68% is obtained by unsupervised adaptation using the normalized cosine similarity, while the best minDCF of 0.0534 is obtained by the the use of ZT-norm with unsupervised adaptation. In evaluating All Trials, the best EER of 15.42% and minDCF of 0.0660 are both obtained via unsupervised adaptation using the normalized cosine similarity.

The volatile nature of the 10sec-10sec condition can, in

some cases, lead to results that inaccurately represent a system's capabilities. To ensure the reliability of our results and to further confirm the effectiveness of our unsupervised adaptation algorithm and score normalization methods, we tested the systems on the standard core condition (1conv-1conv) as offered by the 2008 NIST SRE, in which training and testing are each done on an entire conversation (approx. 2.5min) of telephone speech. Table 3 details the results of this experiment.

	English Trials		All Trials	
	EER (%)	minDCF	EER (%)	minDCF
Baseline, ZT-norm	3.17%	0.0129	5.72%	0.0316
Adapted, ZT-norm	3.17%	0.0129	5.34%	0.0287
Baseline, S-norm	3.44%	0.0148	5.71%	0.0285
Adapted, S-norm	3.17%	0.0130	5.44%	0.0260
Baseline, Norm-Cos	3.41%	0.0127	5.21%	0.0248
Adapted, Norm-Cos	**3.17%**	**0.0107**	**4.83%**	**0.0229**

Table 3: *Comparison of results (1conv-1conv core condition) between our proposed method of unsupervised adaptation with various normalization procedures and the respective "baseline" approach (without adaptation).*

Indeed, we can see once again that our best results are achieved using the normalized cosine similarity with unsupervised adaptation. A notable observation can be made for English Trials under ZT-norm, where our system achieves the exact same results with and without adaptation. While this may seem a bit odd, we rest assured realizing that the system with unsupervised adaptation does, at the very least, perform better under All Trials.

6. Discussion

While the performance of the cosine normalization method rather eclipses all the other results, we can, nevertheless, make a number of interesting observations:

- The unsupervised adaptation method we propose is successful in improving performance, regardless of the normalization procedure. Our results are consistent with the notion that unsupervised adaptation (with an appropriately chosen threshold) should be at least as good as - though hopefully better than - the baseline method without adaptation.
- The simplified S-norm approach performs competitively with the more complicated, traditional ZT-norm approach.
- That the best result was ultimately obtained using the cosine normalization demonstrates that we have indeed come full circle in our study of score normalization procedures. At first, these procedures were introduced and enhanced to improve performance, then simplified and finally replaced with a normalization method that does not require any model-/utterance-dependent precomputations or parameter updates after each adaptation.

When we introduced the unsupervised adaptation algorithm in Section 3.2, we mentioned how the symmetry in the function $\widetilde{\text{score}}_{(\cdot)}$, as seen in (15), nicely allows for a fixed threshold. This can be a bit limiting, however, as such a score function treats every total factor vector in the set of vectors W_s as equally important; yet in reality, the only vector that unequivocally belongs to speaker s is the initial enrollment w'_s. Indeed, the presence of a false-alarm decision (where a test utterance t'_j is incorrectly admitted into $\{W'_s\}$) will have an adverse effect on all subsequent tests. It would be better if $\widetilde{\text{score}}_{(\cdot)}$ could combine cosine similarity scores in a way that takes into account the proximity of each admitted test utterance to the initial enrollment vector w'_s. Thus, we have begun experimenting with the following weighted average:

$$\widehat{\text{score}}_N\left(W_s, t'_i\right) = \frac{1}{\sum_i a_i} \sum_{i=1}^{\|W_s\|} a_i \cdot \text{score}_N\left(w'_i, t'_i\right) \quad (24)$$

where the score weight $a_i = \text{score}\left(w'_i, w'_s\right)$ is the cosine similarity score without the application of any normalization procedures. We can see that $a_i \in [-1, 1]$ and, by definition of the original cosine similarity score in (13), $a_i = 1$ if $w_i = w_s$.

There are other ways to determine the respective a_i. One possibility is to weight the scores by the order in which test utterances are admitted: let $a_i = f(i)$, where i denotes the i^{th} utterance accepted for adaptation. In the beginning, having admitted no other utterances, it would make sense for $\widehat{\text{score}}_N(\cdot, \cdot)$ to be conservative with the amount of weight placed on the scores produced by the first few test utterances adapted into the model. As more test utterances are admitted, however, the adapted model is, in theory, obtaining a better and better representation of the speaker. Thus, we would be more and more likely to believe in the model's decisions and, subsequently, increase the weight of the scores produced by test utterances admitted later. At some point during testing, we could even go back and refine our set of vectors W_s and discard the total factor vectors that were incorrectly admitted.

Aside from the score-combining function $\widetilde{\text{score}}_{(\cdot)}$, another parameter that can be tuned is the decision/adaptation threshold. We need not restrict our choosing of the threshold to be the optimal a posteriori decision threshold of the development data set (without adaptation). The threshold could instead be chosen as the one that provides the best unsupervised adaptation performance in the development data. It is also possible to have different thresholds for decision (θ_D) and adaptation (θ_A), where it might make sense to set θ_D as the optimal a posteriori decision threshold (without adaptation), and then choose a more conservative $\theta_A > \theta_D$. That is, a test utterance t'_j may generate a score θ_j such that $\theta_D \leq \theta_j < \theta_A$. The system, being forced to make a decision, might choose to confirm the hypothesis speaker, but then decide not to adapt t'_j into the speaker's set of total factor vectors for fear of a false acceptance adversely affecting the outcome of subsequent trials.

Our experiments with these ideas have not yet yielded results that are significant improvements from those shown in Tables 2 and 3. We intend for this line of work to be continued in the future; there are plenty of opportunities and possibilities. For now, we stand by the methods proposed in this paper as effective measures to both simplify and improve the approaches to text-independent speaker verification.

7. Conclusion

In this paper, we tackled the problem of unsupervised speaker adaptation in the context of Total Variability and the cosine similarity metric. By taking advantage of the low dimensionality of total factor vectors as well as the simplicity and symmetry of cosine similarity scoring, we described an algorithm for unsupervised speaker adaptation that is simpler and more efficient than a previous approach using JFA. In an effort to keep unsupervised adaptation procedures as straightforward as possible, we also proposed the S-norm score normalization method that is a direct simplification of the traditional ZT-norm procedures. Ultimately, the best solution requires no score normalization parameters or any additional procedures; the normalization is simply integrated into a Normalized Cosine Similarity score function. This method of score normalization, used alongside the unsupervised adaptation algorithm, achieves state-of-the-art results in both the 10sec-10sec and 1conv-1conv conditions of the 2008 NIST SRE.

8. Acknowledgements

We would like to thank Patrick Kenny (Centre de Recherche d'Informatique de Montreal) and Douglas Reynolds (MIT Lincoln Laboratory) for their valuable insights and contributions to this work.

9. References

[1] Najim Dehak, *Discriminative and Generative Approaches for Long- and Short-Term Speaker Characteristics Modeling: Application to Speaker Verification*, Ph.D. thesis, Ecole de Technologie Superieure de Montreal, June 2009.

[2] Najim Dehak, Patrick Kenny, Reda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, 2009, Submitted.

[3] Najim Dehak, Reda Dehak, James Glass, Douglas Reynolds, and Patrick Kenny, "Cosine similarity scoring without score normalization techniques," in *Odyssey*, 2010, Submitted.

[4] Patrick Kenny, Pierre Ouellet, Najim Dehak, Vishwa Gupta, and Pierre Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, July 2008.

[5] Shou-Chun Yin, Richard Rose, and Patrick Kenny, "A joint factor analysis approach to progressive model adaptation in text-independent speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, July 2007.

[6] "The nist year 2004 speaker recognition evaluation plan," Online, January 2004, Available: http://www.itl.nist.gov/iad/mig/tests/spk/2004/SRE-04_evalplan-v1a.pdf.

[7] Douglas Reynolds, Thomas Quatieri, and Robert Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, 2000.

[8] Patrick Kenny, Gilles Boulianne, and Pierre Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions on Speech and Audio Processing*, May 2005.

[9] Andrew Hatch, Sachin Kajarekar, and Andreas Stolcke, "Within-class covariance normalization for svm-based speaker recognition," in *Proceedings of ICSLP*, 2006.

[10] Ondrej Glembek, Lukas Burget, Najim Dehak, Niko Brummer, and Patrick Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.

[11] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, 2000.

[12] Jason Pelecanos and Sridha Sridharan, "Feature warping for robust speaker verification," in *Proceedings of A Speaker Odyssey*, June 2001.