

ADAPTIVE SCORE NORMALIZATION FOR PROGRESSIVE MODEL ADAPTATION IN TEXT INDEPENDENT SPEAKER VERIFICATION

Shou-Chun Yin^{1,2}, Richard Rose¹, Patrick Kenny²

¹ Department of Electrical and Computer Engineering
McGill University, Montreal, Canada
(syin2, rose)@ece.mcgill.ca

² Centre de recherche informatique de Montréal
(CRIM)
(shouchun.yin, pkenny)@crim.ca

ABSTRACT

This paper deals with the interaction between progressive model adaptation and score normalization strategies which are used for reducing the variation in likelihood ratio scores in making speaker verification decisions. This issue is important in establishing robust decision thresholds for practical speaker verification systems. An adaptive score normalization method is proposed that is designed to reduce the drift in likelihood ratio scores that occurs when speaker models are adapted. This method is investigated and compared with other more well know score normalization methods in the context of a joint factor analysis speaker verification approach. All approaches are evaluated on the progressive adaptation track in the NIST 2005 text independent speaker verification evaluation plan.

Index Terms— Gaussian distribution, speaker recognition

1. INTRODUCTION

Adaptation of speaker models has been considered to be an important part of both text dependent and text independent speaker verification (SV) systems that rely on statistical models of speaker identity. This is mainly because it is very difficult to capture all sources of variability that impact speaker verification performance in a single enrollment session. Progressive speaker model adaptation scenarios have been explored in an attempt to incorporate sources of intersession variability in target speaker models. However, these methods have always introduced difficulties in implementing decision rules for verifying the claimed speaker identity. This is because as target speaker models are adapted over time, the likelihoods computed for successive test utterances with respect to the adapted models tend to drift [1]. As a result, it is difficult to identify an a priori decision threshold to implement a decision rule that is stable over time.

This paper will investigate the behavior of score normalization techniques for dealing with the issue of robust specification of decision thresholds for progressive adaptation in speaker verification. In particular, a new adaptive score normalization procedure will be presented that is designed to reduce the drift in normalized likelihood ratio scores obtained from Gaussian mixture based speaker verification systems. The problem of robust decision thresholds has been addressed by the use of score normalization techniques like the z -norm and t -norm which reduce the variability of the likelihood ratio scores that are used in the speaker verification decision criterion [2].

It will be shown in Section 4 that the t -norm technique does not compensate for the increase in the value of the likelihood ratio scores obtained for target and impostor speaker utterances that occurs after each adaptation update. Text-dependent speaker verification systems often deal with the “drift” in likelihood ratio scores by adapting the decision threshold at the same time the target speaker model is adapted [3, 4]. The adaptive approach presented in this paper introduces an adaptive implementation of the t -norm to facilitate the use of a single fixed decision threshold.

The paper is organized as follows. A progressive speaker adaptation paradigm for text-independent speaker verification (SV) is described in Section 2. It is applied in the context of a joint factor analysis model. The issue of how score normalization strategies interacts with progressive speaker adaptation is presented along a description of the new score adaptive normalization procedure in Section 3. Finally, a description of the experimental configuration and an experimental study performed on the NIST 2005 evaluation set for both supervised and unsupervised adaptation scenarios are provided in Section 4.

2. PROGRESSIVE MODEL ADAPTATION

This section presents the progressive model adaptation scenarios that are applied here for text independent Gaussian mixture based speaker verification. First, the general model adaptation scenarios are discussed. Second, a brief overview of the implementation of the progressive model adaptation scenario with a joint factor analysis model is provided.

2.1. Adaptation Scenario

In this work, it is assumed that speaker verification consists of four parts. First, system initialization involves estimating speaker independent parameter set, Λ_0 , which may characterize a general speaker population as well as various other sources of variability. Second, enrollment for target speaker, s , involves estimation of an initial set of speaker dependent model parameters, $\Lambda_1(s)$, from a single enrollment utterance χ . Third, target speaker adaptation involves estimation of speaker dependent model parameters, $\Lambda_i(s)$, at the i th iteration epoch from adaptation utterance, χ_i , speaker parameters from the previous epoch, $\Lambda_{i-1}(s)$, and speaker independent parameters, Λ_0 . Finally, verification of the claimed identity of speaker, s , is performed by applying a likelihood ratio based decision rule to a given test utterance, χ_{test} , with respect to the most recently updated model parameters, $\Lambda_i(s)$.

This very general update procedure is applied in both supervised and unsupervised adaptation scenarios and implemented on

This work was supported by funding from the National Science and Engineering Research Council of Canada and was performed at the Centre de recherche informatique de Montréal (CRIM) and at the Department of Electrical and Computer Engineering, McGill University.

the NIST 2005 speaker recognition evaluation [5]. In the supervised adaptation scenario, the adaptation utterances for a target speaker s are taken from a set of eight enrollment conversation-sides obtained from that target speaker. Beginning with the initial speaker model, $\Lambda_1(s)$, speaker specific model parameters are progressively updated with the remaining 7 conversation-sides in the enrollment set using the factor analysis based adaptation paradigm outlined below.

In the unsupervised adaptation scenario, progressive speaker model adaptation is performed using selected conversation-side utterances in the NIST 2005 evaluation set for the target speaker which consist of unlabeled target speaker utterances randomly interspersed with impostor speaker utterances. The decision to use a particular unlabeled test utterance to adapt the target speaker model is made by comparing the log likelihood ratio score obtained for that utterance and model to an adaptation threshold. If there is a decision to accept a given test utterance, the model will be adapted and used in subsequent verification trials until another adaptation utterance is identified.

2.2. Joint factor analysis based adaptive model

Factor analysis based SV [6, 1, 7], is based on a maximum a posteriori (MAP) approach for adaptation of Gaussian mixture model (GMM) means from a speaker independent GMM which is referred to as a universal background model (UBM). The UBM is estimated from a collection of utterances obtained from a large independent population of background speakers. A target speaker, s , is represented by the concatenation of the F dimensional component mean vectors of the C component speaker dependent GMM which results in a CF dimensional supervector. It is assumed that this supervector is decomposed into the sum of a speaker-dependent supervector \mathbf{s} and channel-dependent supervector \mathbf{c} . This provides an explicit model of the fact that utterances from a given speaker, s , may be spoken over many different channels. As described in [7], a probability density function (pdf) describing \mathbf{s} is the speaker component of a factor analysis based SV representation, and a pdf describing \mathbf{c} is the channel component. Assuming that both \mathbf{s} and \mathbf{c} are normally distributed, and given a large population of “background speakers” used for training the GMM-UBM, one can jointly estimate the prior distributions of \mathbf{s} and \mathbf{c} [6]. Assuming that \mathbf{s} has a prior distribution with mean \mathbf{m} and diagonal covariance \mathbf{d}^2 , and \mathbf{c} has a prior distribution with zero mean and low rank covariance matrix $\mathbf{u}\mathbf{u}^*$, then \mathbf{s} and \mathbf{c} can be represented as

$$\begin{aligned}\mathbf{s} &= \mathbf{m} + \mathbf{d}\mathbf{z} \\ \mathbf{c} &= \mathbf{u}\mathbf{x}.\end{aligned}\quad (1)$$

In Equation 1, the speaker factors \mathbf{z} and channel factors \mathbf{x} are assumed to have standard normal distribution. The mean \mathbf{m} of the prior distribution is actually taken from the GMM-UBM. Finally, the factor analysis model includes a $CF \times CF$ diagonal covariance matrix, $\mathbf{\Sigma}$, to represent variability that is not captured by \mathbf{s} and \mathbf{c} .

The complete hyperparameter set associated with the model in Equation 1 is given by $\Lambda = \{\mathbf{m}, \mathbf{u}, \mathbf{d}, \mathbf{\Sigma}\}$. After the i th adaptation iteration for target speaker s , the speaker-dependent hyperparameters $\Lambda_i(s) = \{\mathbf{m}_i(s), \mathbf{d}_i(s)\}$ are obtained providing an updated estimate of the speaker-specific supervector \mathbf{s} [1, 7]. When the i th adaptation utterance becomes available, $\Lambda_i(s) = \{\mathbf{m}_i(s), \mathbf{d}_i(s)\}$ are obtained from $\Lambda_{i-1}(s) = \{\mathbf{m}_{i-1}(s), \mathbf{d}_{i-1}(s)\}$ and the speaker-independent hyperparameters $\{\mathbf{u}, \mathbf{\Sigma}\}$. The speaker-independent hyperparameters are not updated during speaker adaptation. A more complete description of the hyperparameter update procedure for the

joint factor analysis model along with the definition of the likelihood ratio test is provided in [1, 7].

3. SCORE NORMALIZATION METHODS

The use of score normalization techniques has become important in GMM based speaker verification systems for reducing the effects of the many sources of statistical variability associated with log likelihood ratio scores [2]. The sources of this variability are thought to include changes in the acoustic environment and communications channel as well as intra-speaker variation that may occur across multiple sessions. The issue of log likelihood ratio (LLR) score variability is further complicated by changes in the likelihood ratio score that may occur as a result of progressive speaker model adaptation. After reviewing some well known score normalization algorithms, this section presents an adaptive score normalization procedure for reducing the variability of the LLR score associated with speaker adaptation.

3.1. Score normalization techniques

For a given target speaker s and a test utterance χ_{test} , speaker normalization is applied to the log likelihood ratio score $LLR(\chi_{test}, s)$. The form of test likelihood function is presented in [6]. It is generally assumed that $LLR(\chi_{test}, s)$ is Gaussian distributed when evaluated over utterances that represent a range of the possible sources of variability. Two well known score normalization techniques, the z -norm and t -norm, form a normalized LLR score by obtaining estimates of the mean μ and standard deviation σ and normalizing as

$$LLR(\chi_{test}, s)_{norm} = \frac{LLR(\chi_{test}, s) - \mu}{\sigma}. \quad (2)$$

The z -norm and t -norm differ in how these normalization parameters are computed. In the z -norm, the parameters μ and σ are estimated as the sample mean and standard deviation of a set of log likelihood ratio scores $LLR(\chi_i, s)$, $i = 1, \dots, N_{imp}$, where χ_i , $i = 1, \dots, N_{imp}$, is a set of N_{imp} impostor speaker utterances. This represents an average of scores obtained by scoring the target speaker model against a set of impostor utterances.

In the t -norm, the parameters μ and σ are estimated as the sample mean and standard deviation of a set of log likelihood ratio scores $LLR(\chi_{test}, s_j)$, $j = 1, \dots, M_{imp}$, where s_j , $j = 1, \dots, M_{imp}$, is a set of M_{imp} impostor speaker models. This represents an average of scores obtained by scoring a set of impostor speaker models against the test utterance.

The z -norm is generally considered to be a means for compensating with respect to inter-speaker variability in the LLR speaker verification scores. It is generally assumed that the t -norm compensates for inter-session variability. The zt -norm, which performs z -normalization followed by t -normalization, was originally proposed to compensate for both effects [8].

3.2. Adaptive t -norm score normalization

The score drifting phenomenon in speaker model adaptation scenarios occurs in many detection problems including telephony based text-dependent speaker verification applications where scores tend to drift as the amount of adaptation data increases [3, 4]. Using t -norm based score normalization, an alternative to adapting decision thresholds to reflect the increases in the log likelihood ratio score was investigated. It is possible to adapt the t -norm speaker models, or the t -norm speaker-dependent hyperparameter sets in our case, so

that the t -norm distribution estimated for score normalization also reflects the effects of adaptation in LLR scores.

Under the adaptive t -norm strategy, whenever a target speaker model is adapted, the t -norm speaker models are also adapted using utterances from t -norm speakers. For M_{imp} t -norm models, we have an adaptation utterance from each of the M_{imp} t -norm speakers for each adaptation epoch. This allows adaptation of the t -norm models for a particular target speaker to be performed off-line resulting in minimal increase in computational complexity during verification trials. The comparison of speaker verification performance obtained using the t -norm and the adaptive t -norm is given in the next section.

4. EXPERIMENTAL STUDY

This section presents an evaluation of the adaptive t -norm performance obtained using a joint factor analysis approach to progressive speaker adaptation under supervised and unsupervised speaker verification scenarios. In all experiments, gender-dependent UBMs were used that consisted of 2048 Gaussians and 26 dimension acoustic feature vectors consisting of 13 Gaussianized cepstral features and their first derivatives. The same feature analysis was used for the entire experimental study.

4.1. NIST 2005 evaluation set

The NIST 2005 evaluation dataset used in the progressive speaker adaptation experiments is summarized in Table 1. The supervised adaptation scenario is based on the ‘8 conversation 2-channel’ condition. The unsupervised adaptation scenario is based on the ‘1 conversation 2-channel’ core condition specifying that only a single conversation-side is used for training.

Table 1. Summary of NIST 2005 speaker recognition evaluation set

NIST 2005 Data Set Summary		
	supervised	unsupervised
Target Speakers	497	644
Enrollment Utt.	8 per spkr.	1 per spkr
Target Utt.	2230 (984 m, 1246 f)	2771 (1231 m, 1540 f)
Nontarget Utt.	21216 (8962 m, 12254 f)	28472 (12317 m, 16155 f)

4.2. Supervised speaker adaptation scenario

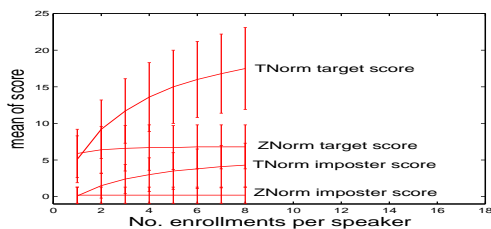


Fig. 1. t -norm and z -norm score distributions

The ‘‘TNorm’’ labeled curves in Figure 1 illustrate how the score drifting issue becomes progressively more pronounced for the t -norm as more adaptation utterances are used in a supervised adaptation scenario. This figure displays the average t -normalized and

z -normalized LLR scores respectively for target speaker and impostor speaker utterances after adaptation using one through eight enrollment utterances. Both the target and impostor utterance scores tend to ‘‘drift’’ upward for the t -norm making it difficult to implement any decision rule based on a fixed threshold. The ‘‘ZNorm’’ labeled curves in Figure 1 describe the evolution of the average z -normalized scores for all target and impostor utterances after adaptation with from one through eight adaptation utterances. Note that the z -norm scores do not exhibit the same ‘‘drift’’ that is associated with the t -norm scores. The reason for this is that, given a target speaker s , both the test utterance score and the z -norm utterance scores are computed against the same adapted speaker model. The error bars given in Figure 1 represent the standard deviation of the LLR scores obtained using the different score normalization techniques. It is apparent that the standard deviation of the average z -normalized scores does not increase with additional target model adaptation. However, this is not the case for the scores normalized using the t -norm.

The test score distribution obtained from the new adaptive t -norm using one through eight enrollment utterances is shown in Figure 2. Comparing the average scores shown in this figure with those plotted for the non-adaptive t -norm, indicated by ‘‘TNorm’’, it is clear that the score drifting problem associated with the t -norm has been removed.

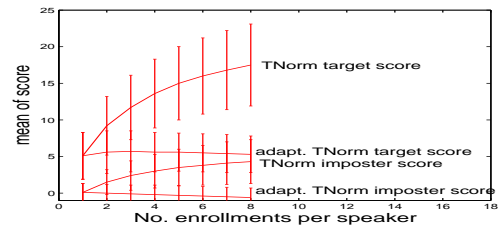


Fig. 2. t -norm and the adaptive t -norm score distributions

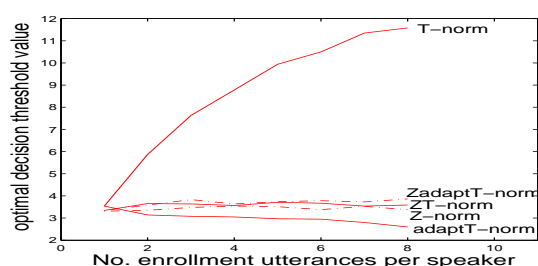
Table 2 displays the SV performance as the equal error rate (EER) and minimal detection cost function (DCF) obtained using supervised adaptation for one through eight adaptation utterances. Three different score normalization strategies were evaluated including t -norm (T), z -norm (Z), and the adaptive t -norm (AT). First, the first row shown in Table 2 corresponds to the non-adaptive performance. While the t -norm and z -norm performance after four or more adaptation utterances are very similar, the relative performance improvement obtained using speaker adaptation with t -norm score normalization is much greater than that obtained using the z -norm. Second, it is clear that both EER and DCF performance measures saturate after approximately 4 adaptation utterances. Finally, the EER and DCF performance obtained using the adaptive t -norm based score normalization methods does not differ significantly from the non-adaptive score normalization methods.

Figure 3 displays the optimum decision thresholds corresponding to the minimum DCF value that are obtained for five different score normalization methods after each enrollment utterance. It is clear from the figure that a fixed decision threshold is not practical in the case of the t -norm. For the adaptive t -norm, the optimal decision threshold does not vary far less with respect to enrollment utterance when compared to the t -norm. Hence, the specification of a fixed decision threshold would be far more practical in this case. In Figure 3, as one would expect, the optimal decision threshold variability with respect to the number of enrollment utterances using z -norm or zt -norm is small. A combination of z -norm and adaptive

Table 2. Comparison between t -, z -, and the adaptive t -norm.

No. enroll.	EER (T)	EER (Z)	EER (AT)	DCF (T)	DCF (Z)	DCF (AT)
1	10.7	6.9	10.7	0.037	0.027	0.037
2	6.3	5.6	6.4	0.024	0.021	0.024
3	5.2	5.2	5.2	0.019	0.018	0.019
4	4.6	4.8	4.8	0.017	0.017	0.016
5	4.5	4.7	4.4	0.016	0.016	0.016
6	4.3	4.5	4.4	0.015	0.016	0.015
7	4.3	4.7	4.4	0.015	0.016	0.015
8	4.4	4.6	4.3	0.015	0.015	0.014

t -norm, indicated by “ZadaptiveT-norm” in Figure 3, appears to have a very small effect of further reducing the variability of optimum decision thresholds across enrollment utterances.

**Fig. 3.** Comparison of optimal decision thresholds

4.3. Unsupervised speaker adaptation scenario

Table 3. Performance for ideal unsupervised adaptation scenario

Normalization	EER	DCF
t -norm	13.0%	0.039
adaptive t -norm	4.2%	0.014
z -norm	3.6%	0.012

Table 3 displays the unsupervised adaptation performance obtained under an idealized adaptation scenario using three score normalization techniques. The scenario is ideal in that all of the 2771 target speaker utterances shown in Table 1 are selected for adapting speaker models and all of the non-target utterances from the test trials are rejected and not used for adaptation. With 644 target speakers, this amounts to an average of approximately 4 adaptation utterances per speaker. According to Table 3, with this number of adaptation utterances, the optimum performance obtainable using the z -norm is about 15% better than the adaptive t -norm in the unsupervised scenario.

Table 4 compares the adaptive t -norm and z -norm performance for three different cases where exactly the same sequence of adaptation utterances are used for both methods. The number of target and imposter utterances used for the three different cases are given in the first two columns of each of the three rows of the table. The first row of Table 4 corresponds to the adaptive t -norm and z -norm performance obtained with a fixed adaptation threshold setting equal to 3.5. There are several observations that can be made. First, comparing rows 1 and 2, incorrect acceptance of 136 adaptation utterances has negligible effect on performance for either score normalization methods. Second, comparing rows 2 and 3, the performance

Table 4. z -norm (Z) vs. adaptive t -norm (AT)

No. targets accepted	No. non-targets accepted	EER (Z)	EER (AT)	DCF (Z)	DCF (AT)
2180	136	6.2	10.1	0.019	0.026
2180	0	6.2	10.1	0.019	0.025
2769	0	3.6	4.2	0.012	0.014

difference between z -norm and adaptive t -norm is far less when an additional 600 utterances are available for adaptation. Clearly, speaker verification performance with the z -norm method has saturated faster than when adaptive t -norm score normalization is used. One can expect that the adaptive t -norm results will give comparable performance with z -norm results in the unsupervised scenario for a fixed adaptation threshold, when the length of trial list is longer so that more target utterances are available for adapting speaker models.

5. CONCLUSION

This paper has addressed the issue of SCOT normalization for progressive model adaptation in text independent speaker verification. The score drifting effect introduced by speaker model adaptation was analyzed and an adaptive t -norm score normalization procedure was proposed. The behavior of the adaptive t -norm was analyzed for both supervised and unsupervised speaker adaptation scenarios and speaker verification results were presented and compared with the results from other score normalization techniques. The best supervised adaptation results obtained using the adaptive t -norm corresponded to an EER of 4.3% and a minimum DCF of 0.014.

6. REFERENCES

- [1] Yin S.-C., P. Kenny, and R. Rose, “Experiments in speaker adaptation for factor analysis based speaker verification,” in *Proc. Odyssey 2006*, San Juan, Puerto Rico, June 2006.
- [2] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, “Score normalization for text-independent speaker verification systems,” *Digital Signal Processing*, vol. 10, pp. 42–52, 2000.
- [3] N. Mirghafori and L. Heck, “An adaptive speaker verification system with speaker dependent a priori decision thresholds,” in *Proc. ICSLP*, Denver, Colorado, Sept. 2002.
- [4] N. Mirghafori and M. Hébert, “Parameterization of the score threshold for a text-dependent adaptive speaker verification system,” in *Proc. ICASSP 2004*, Montreal, Canada, May 2004.
- [5] “The NIST year 2005 speaker recognition evaluation plan,” 2005.
- [6] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 4, May 2007.
- [7] Yin S.-C., R. Rose, and P. Kenny, “A joint factor analysis approach to progressive model adaptation in text independent speaker verification,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, Sept. 2007.
- [8] R. Vogt, B. Baker, and S. Sridharan, “Modeling session variability in text-independent speaker verification,” in *Proc. Eurospeech*, Lisbon, Portugal, Sept. 2005.