A Joint Factor Analysis Approach to Progressive Model Adaptation in Text-Independent Speaker Verification

Shou-Chun Yin, Richard Rose, *Senior Member, IEEE*, and Patrick Kenny

***Abstract*—This paper addresses the issue of speaker variability and session variability in text-independent Gaussian mixture model (GMM)-based speaker verification. A speaker model adaptation procedure is proposed which is based on a joint factor analysis approach to speaker verification. It is shown in this paper that this approach facilitates the implementation of a progressive unsupervised adaptation strategy which is able to produce an improved model of speaker identity while minimizing the influence of channel variability. The paper also deals with the interaction between this model adaptation approach and score normalization strategies which act to reduce the variation in likelihood ratio scores. This issue is particularly important in establishing decision thresholds in practical speaker verification systems since the variability of likelihood ratio scores can increase as a result of progressive model adaptation. These adaptation methods have been evaluated under the adaptation paradigm defined under the NIST 2005 Speaker Recognition Evaluation Plan, which is based on conversation sides derived from telephone speech utterances. It was found that when target speaker models were trained from a single conversation, an equal error rate (EER) of 4.5% was obtained under the NIST unsupervised speaker adaptation scenario.**

***Index Terms*—Factor analysis, Gaussian mixture model (GMM), speaker adaptation, speaker verification.**

I. Introduction

PROGRESSIVE speaker adaptation has been considered to be an important component of commercial telephony based text-dependent speaker verification systems [10], [11]. This is largely because it is extremely difficult to formulate a single practical speaker enrollment scenario that will result in the capture of all sources of variability that impact speaker verification performance. The important sources of variability in speaker verification are generally considered to be the acoustic environment and communications channel as well as intraspeaker variation that may occur across multiple sessions. It is reasonable to believe that any implementations of text-independent speaker verification tasks operating in telephony environments will suffer from the same fundamental limitation in their ability to characterize all sources of variability in a single enrollment session. As a result, the National Institute of Standards and Technology (NIST) in the U.S. has formulated an unsupervised speaker adaptation task as part of the annual NIST organized text-independent speaker verification campaigns to encourage research laboratories to investigate progressive speaker adaptation techniques [1].

This paper presents a factor analysis approach to implementing progressive speaker adaptation in Gaussian mixture model (GMM)-based speaker verification [8]. There are several important issues that are addressed by this method which are similar to the issues that are being considered when implementing progressive adaptation in text-dependent speaker verification. A first issue is intersession variability which is known to be particularly problematic when unsupervised speaker adaptation is used to progressively update target speaker models for each potential trial utterance during verification. Care must be taken to prevent the adaptation update for a trial utterance from simply providing a better representation of the channel characteristics associated with that particular utterance rather than improving the model representation of the speaker. The important aspect of the factor analysis approach from the standpoint of model adaptation is that it is based on a model that accounts for speaker and channel variability using two sets of latent variables that will be referred to as speaker factors and channel factors in Section II. As a result of this decomposition, speaker adaptation can be performed by updating a set of speaker-dependent hyperparameters and minimizing the influence of channel effects in an utterance.

A second issue involves 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 [12]. These techniques are important in establishing decision thresholds in practical speaker verification systems. It will be shown in Section V 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 this "drift" in likelihood ratio scores by adapting the decision threshold at the same time the target speaker model is adapted [10], [11]. An alternative approach is described in Section V, where an adaptive implementation of the t-norm is introduced to facilitate the use of a single fixed decision threshold. Another more practical issue concerns the question of when the effects of progressive adaptation begin to "saturate" as the number of adaptation utter-

Manuscript received February 16, 2007; revised June 5, 2007. This work was supported by 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 associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jean-François Bonastre.

The authors are with the Electrical and Computer Engineering Department, McGill University, Montréal, QC H3A 2A7, Canada, and also with the Centre de Recherche Informatique de Montréal (CRIM), Montréal, QC H3A 1B9, Canada (e-mail: syin2@ece.mcgill.ca; rose@ece.mcgill.ca; patrick.kenny@crim.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2007.902410

ances increases. Determining the point at which adaptation updates cease to affect speaker verification performance and other empirical questions are addressed in Section VI.

Both supervised and unsupervised speaker adaptation scenarios for text-independent speaker verification are investigated in this paper. The verification tasks are based on the NIST 2005 speaker recognition evaluation plan, which uses the conversational telephone speech data collected for the Mixer Corpus by the Linguistic Data Consortium [1]. The unsupervised adaptation experiments were performed on the core test data according to the specifications given in the NIST 2005 plan for this scenario [1]. This involved using a single 5 min conversation to train target speaker models. Speaker verification and progressive speaker model adaptation was then performed on conversation length trial utterances from the target speaker that were randomly interspersed with impostor speaker utterances.

The paper is organized as follows. The basic factor analysis model for speaker verification is briefly summarized in Section II. Section III describes how the factor analysis model is applied to updating speaker-dependent hyperparameters for progressive speaker adaptation. Section IV describes how the supervised and unsupervised adaptation strategies are implemented and provides a summary of the data sets used for the experimental study. The issue of how score normalization strategies interact with progressive speaker adaptation, along with a new score normalization procedure, is presented in Section V. Finally, a description of the experimental study and a summary of the results for both supervised and unsupervised adaptation scenarios are provided in Section VI.

II. Introduction to Joint Factor Analysis for Speaker Verification

This section provides a brief introduction to the joint factor analysis model for GMM-based speaker verification [4], [6], [7], [9]. First, the joint factor analysis model is summarized in Section II-A as a means for describing speaker- and channel-dependent GMMs using hidden variables. Then, Section II-B discusses the estimation of both the speaker-independent and speaker-dependent hyperparameter sets that form the factor analysis model.

A. Speaker Factors and Channel Factors

GMMs have become the most commonly used representation for text-independent speaker recognition and speaker verification. The work described in this paper relies on a GMM-based speaker verification system where speakers are represented by the means, covariance matrices, and weights of a mixture of C multivariate diagonal-covariance Gaussian densities defined over an F dimensional feature space. Based on the maximum *a posteriori* (MAP) approach [2], the GMM parameters for a particular target speaker are estimated by adapting the parameters of a universal background model (UBM) using utterances from the target speaker. The UBM is a C component GMM trained from a large speaker population. In practice, speaker-specific GMM mean vectors are obtained from the UBM, and it is assumed that weights and covariance matrices are not adapted when the enrollment utterances are limited.

Assuming that a C component GMM in an F dimensional feature space is used to characterize a speaker s, it is convenient to describe the speaker by concatenating the GMM mean vectors into a CF dimensional supervector which we denote by $\mathbf{s}$. In order to incorporate channel effects into the model, allowing for the fact that there will be many utterances from speaker s taken from many different channels, we will use the notation $\mathbf{M}$ to refer a speaker- and channel-dependent, or utterance-dependent, supervector. We assume that $\mathbf{M}$ can be decomposed as

$$\mathbf{M} = \mathbf{s} + \mathbf{c}. \tag{1}$$

In (1), $\mathbf{s}$ is the speaker-dependent supervector which is independent of session variations, and $\mathbf{c}$ is a channel-dependent supervector. Both $\mathbf{s}$ and $\mathbf{c}$ are assumed to be normally distributed.

Kenny *et al.* have described a factor analysis model for speaker verification where both the speaker and channel supervectors can be represented in separate low-dimensional subspaces [7]. A simplified version of this model is used here to facilitate speaker adaptation within the factor analysis framework for speaker verification. The simplified factor analysis model assumes that only the channel-dependent supervector is represented in a low-dimensional channel space. The speaker-dependent supervector is represented as

$$\mathbf{s} = \mathbf{m} + \mathbf{dz} \tag{2}$$

where a vector $\mathbf{m}$ and a diagonal matrix $\mathbf{d}$ are estimated from a large ancillary training set, and a random vector $\mathbf{z}$ is assumed to have a standard normal distribution. In (2), one can show that $\mathbf{s}$ is normally distributed with mean $\mathbf{m}$ and diagonal covariance $\mathbf{d}^2$. In our case, $\mathbf{m}$ corresponds to the concatenated mean vectors of the universal background model that is trained from a large population of "background speakers." If enrollment data for a speaker is given and channel effects are ignored, then a point estimate of $\mathbf{z}$ and hence of $\mathbf{s}$ can be obtained from the enrollment data by classical MAP estimation.

The supervector $\mathbf{c}$ which represents the channel effects in an utterance, is assumed to be distributed according to

$$\mathbf{c} = \mathbf{ux} \tag{3}$$

where $\mathbf{u}$ is a rectangular matrix of low rank and random vector $\mathbf{x}$ has a standard normal distribution. The entries of $\mathbf{x}$ are known as channel factors. This is equivalent to saying that $\mathbf{c}$ is normally distributed with mean $\mathbf{0}$ and covariance matrix $\mathbf{uu}^*$. Given an utterance by a speaker whose supervector $\mathbf{s}$ is known, a point estimate of $\mathbf{x}$ and hence of $\mathbf{c}$ can be obtained by eigenchannel MAP estimation.

In practice, channel effects cannot be ignored in estimating $\mathbf{s}$ from (1), and $\mathbf{s}$ is not known in estimating $\mathbf{c}$. To get around this difficulty, a Gauss–Seidel type iterative procedure is proposed in [13]. The solution used here involves calculating the joint posterior distribution of the hidden variables $\mathbf{z}$ and $\mathbf{x}$ [4] and is briefly described in Appendix A.

The supervector given in (1) models GMM mean vectors. In order to fully specify a GMM model, we also need speaker- and channel-independent GMM diagonal covariances $\mathbf{\Sigma}_c$, $c = 1, \ldots, C$. For convenience, a $CF \times CF$ diagonal covariance matrix $\mathbf{\Sigma}$ is defined whose cth diagonal block is given by $\mathbf{\Sigma}_c$.

The role of $\mathbf{\Sigma}$ is to model the variability which is not captured by $\mathbf{s}$ and $\mathbf{c}$.

B. Speaker-Independent and Speaker-Dependent Hyperparameter Estimation

The difference between two kinds of hyperparameters is based on the availability of previous data from the target speaker. If no previous data are available from the target speaker, the target speaker model is obtained from speaker-independent hyperparameters. If previous training data are available from the target speaker, the target model is obtained from speaker-dependent parameters. Specifically, the hyperparameters $\mathbf{m}, \mathbf{u}$, and $\mathbf{d}$ model the prior distribution of a GMM supervector $\mathbf{M}$: $\mathbf{s}$ is normally distributed with expectation $\mathbf{m}$ and covariance matrix $\mathbf{d}^2$, and $\mathbf{c}$ is normally distributed with expectation zero and covariance matrix $\mathbf{uu}^*$. We will refer to them as *speaker-independent* hyperparameters, which are fixed in estimating any utterance-specific posterior distribution of $\mathbf{M}$.

In order to describe our progressive adaptation algorithm, we have to introduce *speaker-dependent* hyperparameters $\mathbf{m}(s)$ and $\mathbf{d}(s)$ which are estimated using the prior distribution of $\mathbf{s}$ and some enrollment data for a speaker s and used to model the posterior distribution of a speaker-specific supervector $\mathbf{s}$. The assumption is that

$$\mathbf{s} = \mathbf{m}(s) + \mathbf{d}(s)\mathbf{z}. \tag{4}$$

Thus, in the posterior distribution, we assume that $\mathbf{s}$ is normally distributed with expectation $\mathbf{m}(s)$ and covariance matrix $\mathbf{d}^2(s)$. Whereas in (2), $\mathbf{d}$ models the variability of the speaker population as a whole, and $\mathbf{d}(s)$ models the residual uncertainty in the point estimate of speaker-specific supervector $\mathbf{s}$ that arises from the fact that the enrollment data is of limited duration.

In order to estimate the speaker-independent hyperparameter set $\mathbf{\Lambda} = \{\mathbf{m}, \mathbf{u}, \mathbf{d}, \mathbf{\Sigma}\}$, we use a large ancillary training set described in Section IV and the EM algorithms described in [5] and [7]. In estimating the speaker-independent hyperparameters, we skipped the "adaptation to the target speaker population" step in [7, Sec. 3] in order to follow the NIST evaluation protocol. However, this step *applied in the case of a single speaker* is the fundamental idea used to estimate the speaker-dependent hyperparameters $\mathbf{\Lambda}(s)$ for each target speaker s. Simply stated, the initial hyperparameters come from $\mathbf{\Lambda}$. We fix the speaker-independent $\mathbf{u}$ and $\mathbf{\Sigma}$, and re-estimate the speaker-dependent $\mathbf{m}$ and $\mathbf{d}$ using the incoming utterance of the target speaker s. The enrollment procedure which we use to estimate the posterior distribution of a target speaker's supervector $\mathbf{s}$ and the likelihood function that we use to make verification decisions are the same as in [9]. The posterior calculation needed for enrollment is summarized in Appendix A, and the likelihood function is described in Appendix B. In progressive speaker adaptation, we use the speaker-dependent hyperparameters as the starting point for enrolling a target speaker.

III. Progressive Adaptation for Speaker Verification

Given an enrollment utterance and the speaker-independent hyperparameters $\mathbf{m}$ and $\mathbf{d}$ modeling the prior distribution of supervector $\mathbf{s}$, as described in Section II, the speaker-dependent hyperparameters $\mathbf{m}(s)$ and $\mathbf{d}(s)$ used to model the posterior distribution of a speaker-specific supervector $\mathbf{s}$ can be adapted from $\mathbf{m}$ and $\mathbf{d}$. The algorithm used for progressive speaker adaptation is based on the speaker-dependent hyperparameter estimation algorithm from [4]. That is, a speaker-dependent hyperparameter set $\mathbf{\Lambda}(s) = \{\mathbf{m}(s), \mathbf{u}, \mathbf{d}(s), \mathbf{\Sigma}\}$ is updated whenever a new utterance by the speaker becomes available. The algorithm is summarized in the following theorem which is a special case of [4, Theorem 10]. The likelihood function $P_{\mathbf{\Lambda}}$ in the statement of this theorem is the factor analysis likelihood function defined in [4].

Theorem: Suppose we are given a speaker s, a hyperparameter set $\mathbf{\Lambda_0}$ where $\mathbf{\Lambda_0} = \{\mathbf{m_0}, \mathbf{u_0}, \mathbf{d_0}, \mathbf{\Sigma_0}\}$, and a recording χ uttered by s. Let $\mathbf{\Lambda}(s)$ be the hyperparameter set $\{\mathbf{m}(s), \mathbf{u_0}, \mathbf{d}(s), \mathbf{\Sigma_0}\}$ where

$$\begin{aligned}\mathbf{m}(s) &= \mathbf{m_0} + \mathbf{d_0}\boldsymbol{\mu}_\mathbf{z}\\ \mathbf{d}(s) &= \mathbf{d_0}\mathbf{K}_\mathbf{zz}^{1/2}\end{aligned} \tag{5}$$

and

$$\begin{aligned}\boldsymbol{\mu}_\mathbf{z} &= E[\mathbf{z}]\\ \mathbf{K}_\mathbf{zz} &= \mathrm{diag}(\mathrm{Cov}(\mathbf{z}, \mathbf{z})).\end{aligned} \tag{6}$$

$\boldsymbol{\mu}_\mathbf{z}$ and $\mathbf{K}_\mathbf{zz}$ are the posterior expectation and covariance matrix of the speaker-specific latent variable $\mathbf{z}$ calculated using $\mathbf{\Lambda_0}$ as described in Appendix A. Then, $P_{\mathbf{\Lambda}(s)}(\chi) \geq P_{\mathbf{\Lambda_0}}(\chi)$ [4].

Taking $\mathbf{\Lambda_0}(s)$ as the speaker-independent hyperparameter set $\mathbf{\Lambda}$ and applying this algorithm recursively whenever a new utterance uttered by the speaker s becomes available, a sequence of speaker-dependent hyperparameters in the sets $\mathbf{\Lambda_1}(s), \mathbf{\Lambda_2}(s), \ldots$ are obtained for a given speaker s as follows:

$$\begin{aligned}\mathbf{m_i}(s) &= \mathbf{m_{i-1}} + \mathbf{d_{i-1}}\boldsymbol{\mu}_\mathbf{z}\\ \mathbf{d_i}(s) &= \mathbf{d_{i-1}}\mathbf{K}_\mathbf{zz}^{1/2}.\end{aligned} \tag{7}$$

In (7), the posterior expectation $\boldsymbol{\mu}_\mathbf{z}$ and covariance $\mathbf{K}_\mathbf{zz}$ are calculated using $\mathbf{\Lambda_{i-1}}(s)$, where $\mathbf{\Lambda_{i-1}}(s) = \{\mathbf{m_{i-1}}(s), \mathbf{u_0}, \mathbf{d_{i-1}}(s), \mathbf{\Sigma_0}\}$, for $\mathbf{i} = 1, 2, \ldots,$. The vector $\mathbf{m_i}(s)$ represents the speaker-specific prior expectation of $\mathbf{s}$ after the $\mathbf{i}$th speaker-specific recording has been collected and used in the progressive adaptation. Similarly, ${\mathbf{d_i}}^2(s)$ represents the $\mathbf{i}$th speaker-specific prior diagonal covariance matrix of $\mathbf{s}$. According to (7) and Appendix A, it can be shown that the posterior covariance ${\mathbf{d_i}}^2(s) \to \mathbf{0}$ as either the amount of adaptation $\mathbf{i}$ tends to infinity or the total amount of acoustic observations tends to infinity, as one would expect.

When this speaker adaptation algorithm is implemented and the adapted speaker-dependent hyperparameter set is estimated, the same log likelihood function used in the nonadaptive case [9] and described in Appendix B could be used to perform the log likelihood ratio (LLR) test. Comparing the adaptive case and nonadaptive case, the speaker-independent prior distribution of $\mathbf{s}$, which provides $E[\mathbf{s}] = \mathbf{m}$ and $Cov(\mathbf{s}, \mathbf{s}) = \mathbf{d}^2$, can be used to represent the alternative hypothesis appearing in the denominator term of LLR in both cases. The only difference is that, in the adptive case, the posterior distribution of $\mathbf{s}$ representing the null hypothesis appearing in the numerator term of LLR is estimated using $\mathbf{m_i}(s)$ and $\mathbf{d_i}(s)$ from the speaker-dependent hyperparameter set $\mathbf{\Lambda}(s)$, rather than $\mathbf{\Lambda}$ used in the nonadaptive case

$$\begin{aligned} E[\mathbf{s}] &= \mathbf{m_i}(s) \\ Cov(\mathbf{s}, \mathbf{s}) &= \mathrm{diag}(\mathbf{d_i}(s) Cov(\mathbf{z}, \mathbf{z}) \mathbf{d_i}(s)) \end{aligned} \tag{8}$$

where $\mathbf{i}$ counts the adaptation iteration for each target speaker s. Based on the adaptation theorem presented in this section, one would expect that the log likelihood, $\log P(\chi|\mathbf{s})$ presented in (22), is not decreased if the utterance χ is uttered by speaker s and the iteration counter $\mathbf{i}$ is increased.

Instead of progressively adapting $\mathbf{m}(s)$ and $\mathbf{d}(s)$ using the last estimates obtained in the previous adaptation epoch as *a priori* models, an alternative approach could be to assume that all previous target speaker utterances were available of each adaptation step. The hyperparameters $\mathbf{m}(s)$ and $\mathbf{d}(s)$ could then be estimated using $\mathbf{\Lambda}$ and all the adaptation utterances available at each adaptation epoch. The complexity associated with retraining the speaker-dependent $\mathbf{m}(s)$ and $\mathbf{d}(s)$ from scratch in each adaptation epoch could be huge, especially as more and more speaker sessions are available for adaptation.

IV. Implementation of Progressive Adaptation Scenarios

The joint factor analysis-based progressive speaker adaptation procedure described in Section III was applied to the NIST 2005 speaker recognition evaluation scenario specified for text-independent speaker verification using unsupervised speaker adaptation [1]. This section describes this unsupervised adaptation scenario. A supervised adaptation scenario for providing a perspective of the best case performance that is achievable under the unsupervised case is also described. Next, a summary of the datasets used for UBM training, target speaker model enrollment, and evaluation test in the supervised and unsupervised speaker adaptation scenarios is provided.

A. Supervised and Unsupervised Speaker Adaptation

For a given target speaker s, both adaptation scenarios involve progressively adapting the speaker-dependent hyperparameter set $\mathbf{\Lambda}(s)$ described in Section III using a set of conversation-side adaptation utterances. The conversation-sides range from approximately 3–5 min in length. For supervised adaptation, the adaptation utterances for a target speaker s are taken from a set of eight enrollment conversation sides obtained from that target speaker. The initial speaker model, the speaker-specific posterior distribution of supervector $\mathbf{s}$, is trained using the first enrollment utterance. It is then updated with the remaining seven conversation sides in the enrollment set using the progressive adaptation algorithm presented in Section III. After each adaptation epoch, the speaker verification performance is evaluated on the target and impostor speaker utterances contained in the verification test set specified for speaker s in the NIST evaluation protocol.

Unsupervised adaptation is performed according to the scenario specified by the NIST 2005 core condition “1 conversation 2-channel” which is summarized in Section IV-B. Under this scenario, a single conversation-side utterance is used to train the initial speaker model for each target speaker. Progressive speaker model adaptation is then performed using selected conversation-side utterances in the NIST 2005 evaluation set for that speaker which consists 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.

Section VI-B will discuss the impact of false acceptance and false rejection of potential adaptation utterances in this unsupervised scenario on speaker verification performance. It will be demonstrated that the adaptation threshold should not necessarily be the same as the decision threshold which is used to accept or reject a claimant’s identity. In general, the decision threshold for speaker verification is largely determined by a detection cost function which may depend on the application. The setting of the adaptation threshold will be dictated by the tradeoff between the effects of insufficient adaptation data occurring when too many target utterances are rejected and the corrupting influence of including impostor utterances for adaptation when too many impostor utterances are accepted.

B. Experimental Configurations

There are several sources of speech data used for training universal background models, training target speakers, evaluating speaker verification performance, and implementing the score normalization procedures discussed in Section V. Both UBM training and speaker verification evaluation trials are performed in a gender-dependent mode. All results reported in Section VI represent the average of separate gender-dependent trials.

In all experiments, gender-dependent UBMs were used that consisted of 2048 Gaussians and 26-dimensional acoustic feature vectors consisting of 13 Gaussianized cepstral features and their first derivatives. The same feature analysis was used for the entire experimental study. The data set used for UBM training consisted of the LDC releases of phase 1 through 3 of the Switchboard II corpus, parts 1 and 2 of the Switchboard Cellular corpus, the Fisher English corpus part 1 and the NIST 2004 evaluation data. We chose only those speakers for whom multiple recordings (six or more) were available in order to model channel variability properly. The female training data contains 612 speakers and 6764 conversation sides, and the male training data contains 463 speakers and 5254 conversation sides.

The dataset used in the progressive speaker adaptation experiments is summarized in Table I. The speaker verification tasks are based on the NIST 2005 speaker recognition evaluation plan which uses conversational telephone speech data collected for the Mixer Corpus [1]. The supervised adaptation scenario is based on the “eight-conversation two-channel” condition, specifying that eight conversation-side enrollment utterances, where the two channels of the conversation are labeled separately, are used for training target speaker models. The unsupervised adaptation scenario is based on the “one-conversation two-channel” core condition specifying that only a single conversation side is used for training. Table I displays

TABLE I
SUMMARY OF THE NUMBER OF CONVERSATION-SIDE UTTERANCES USED FOR SUPERVISED AND UNSUPERVISED ADAPTATION SCENARIOS IN TEXT-INDEPENDENT SPEAKER VERIFICATION

NIST 2005 Data Set Summary		
	Supervised Adaptation	Unsupervised Adaptation
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)

the number of target speakers, total number of target utterances, and the total number of nontarget utterances used for both supervised and unsupervised adaptation scenarios.

The score normalization techniques discussed in Section V require the estimation of a set of parameters that are used to reduce the statistical variability of the log likelihood ratio scores described in Appendix B. The estimation of these parameters will be discussed in the next section. In all cases, the utterances that are used to estimate these score normalization parameters are taken from the dataset described above used for training the UBM.

V. Score Normalization in Progressive Adaptation

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 [12]. The sources of this variability are thought to include changes in the acoustic environment and communications channel as well as intraspeaker variation that may occur across multiple sessions. The issue of log likelihood ratio 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 discusses how speaker adaptation can affect the variability of LLR score distributions and suggests a normalization procedure to reduce this source of variability.

A. *t-Norm, z-Norm, and zt-Norm-Based Score Normalization*

For a given target speaker s, the corresponding speaker specific supervector $\mathbf{s}$ and a test utterance χ_{test} speaker normalization is applied to the log likelihood ratio score $\text{LLR}(\chi_{\text{test}}, s)$. The definition of the score itself and its use in forming the decision rule for accepting or rejecting the claimed identity of speaker s is provided in Appendix B. It is generally assumed that $\text{LLR}(\chi_{\text{test}}, s)$ 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

$$\text{LLR}(\chi_{\text{test}}, s)_{\text{norm}} = \frac{\text{LLR}(\chi_{\text{test}}, s) - \mu}{\sigma}. \tag{9}$$

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 $\text{LLR}(\chi_i, s)$, $i = 1, \ldots, N_{\text{imp}}$, where s is the target speaker providing a speaker-specific posterior distribution of supervector $\mathbf{s}$ and χ_i, $i = 1, \ldots, N_{\text{imp}}$, 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. A set of $N_{\text{imp}} = 120$ impostor speaker utterances were used for all of the z-norm results given in Section VI.

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 $\text{LLR}(\chi_{\text{test}}, s_j)$, $j = 1, \ldots, M_{\text{imp}}$, where s_j, $j = 1, \ldots, M_{\text{imp}}$, is a set of M_{imp} impostor speakers providing M_{imp} different posterior distributions of $\mathbf{s}$. This represents an average of scores obtained by scoring a set of impostor speaker models against the test utterance. A set of $M_{\text{imp}} = 120$ impostor speaker models were used for computing the t-norm results in Section VI.

The z-norm is generally considered to be a means for compensating with respect to interspeaker variability in the LLR speaker verification scores. It is generally assumed that the t-norm compensates for intersession variability. The zt-norm, which performs z-normalization followed by t-normalization, was originally proposed to compensate for both effects [13]. The form of the zt-norm is similar to the t-norm; however, both the test score $\text{LLR}(\chi_{\text{test}}, s)$ and impostor scores $\text{LLR}(\chi_{\text{test}}, s_j)$, $j = 1, \ldots, M_{\text{imp}}$ used in computing the t-norm distribution are first normalized using the z-norm prior to implementing the t-norm. It will be shown in Section VI that the zt-norm was found to be more effective than the z-norm or t-norm for most of the factor analysis-based speaker verification systems.

B. *Applying Score Normalization in Speaker Adaptation Scenarios*

Progressive speaker adaptation is known to introduce additional variability in likelihood ratio score variability in a range of text-dependent and text-independent speaker verification tasks [10], [11]. After each adaptation epoch, the LLR for the adapted target model, computed against either the target speaker or impostor speaker utterances, has a tendency to increase in comparison to nonadapted models. As a result, the average t-normalized LLR scores will also tend to increase.

The "TNorm" labeled curves in Fig. 1 illustrates how this effect becomes progressively more pronounced for the t-norm as more adaptation utterances are used in a supervised adaptation scenario. Figs. 1 and 2 display 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 Fig. 2 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, $\text{LLR}(\chi_{\text{test}}, s)$, and the z-norm utterance scores, $\text{LLR}(\chi_i, s)$, $i = 1, \ldots, N_{\text{imp}}$, are computed against the same

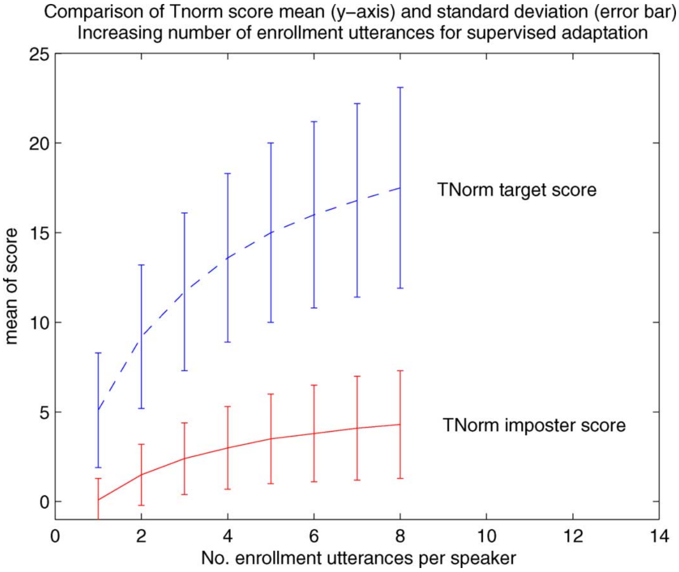

Fig. 1. Eight-conversation two-channel condition of NIST 2005 evaluation. Comparison of target score mean (given by 2230 target trials) and nontarget score mean (given by 21 216 target trials) obtained using various conversations in enrolling each speaker-specific supervector $\mathbf{s}$, based on t-norm. Joint factor analysis with 25 channel factors. All trials, male and female.

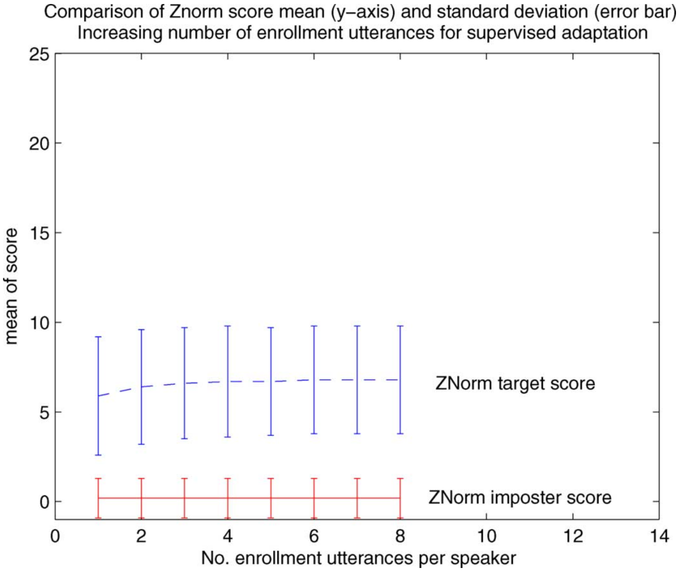

Fig. 2. Eight-conversation two-channel condition of NIST 2005 evaluation. Comparison of target score mean (given by 2230 target trials) and nontarget score mean (given by 21 216 target trials) obtained using various conversations in enrolling each speaker-specific supervector $\mathbf{s}$, based on z-norm. Joint factor analysis with 25 channel factors. All trials, male and female.

adapted speaker model, or the supervector $\mathbf{s}$ specified using the adapted hyperparameter set $\mathbf{\Lambda}(s)$. The error bars given in Figs. 1 and 2 represent the standard deviation of the LLR scores obtained using the different score normalization techniques. It is apparent from Figs. 1 and 2 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 score drifting problem associated with the t-normalized LLR scores that is illustrated in Fig. 1 is even more pronounced if a fixed adaptation threshold has to be used. This is because the decision to accept or reject a test utterance for use in speaker adaptation is based on a fixed adaptation threshold applied to the normalized LLR scores. Since impostor t-normalized scores increase with additional target model adaptation, the scores for impostor speaker utterances will eventually exceed the adaptation threshold and be incorrectly accepted as adaptation utterances. A solution to this problem that is commonly used in text-dependent speaker verification is to adapt the decision threshold at the same time the target speaker model is adapted [10], [11]. A consequence of this strategy is that it is often difficult to obtain a robust mechanism for threshold adaptation. An alternative strategy based on adaptation of the t-norm impostor speaker models is proposed in the next subsection.

C. *Adaptive t-Norm Score Normalization*

The score drifting phenomenon discussed in Section V-B for 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 [10], [11]. 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 estimated parameters μ and σ in (9) also reflect the effects of adaptation in LLR scores.

Under the adaptive t-norm strategy, whenever a target speaker supervector $\mathbf{s}$ is adapted, the t-norm speaker supervectors 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 offline resulting in minimal increase in computational complexity during verification trials. The test score distribution obtained from this new adaptive t-norm using one through eight enrollment utterances is shown in Fig. 3. Comparing the average scores shown in Fig. 1 with those plotted for the nonadaptive t-norm, indicated by "TNorm," it is clear that the score drifting problem associated with the t-norm has been removed. The comparison of speaker verification performance obtained using the t-norm and the adaptive t-norm is given in Section VI-A.

VI. Experimental Study

This section presents an evaluation of the joint factor analysis approach to progressive speaker adaptation under the supervised and unsupervised speaker verification scenarios outlined in Section IV. There are three major issues that will be investigated. First, the performance of the procedure for updating the speaker-dependent hyperparameter set $\mathbf{\Lambda}(s) = \{\mathbf{m}(s), \mathbf{u}, \mathbf{d}(s), \mathbf{\Sigma}\}$ given in Section III under a supervised adaptation scenario is evaluated. Second, the performance of score normalization strategies presented in Section V are compared under supervised adaptation scenario. Third, speaker verification performance under the unsupervised adaptation scenario will be evaluated to investigate the impact of the tradeoff between incorrectly rejecting evaluation utterances

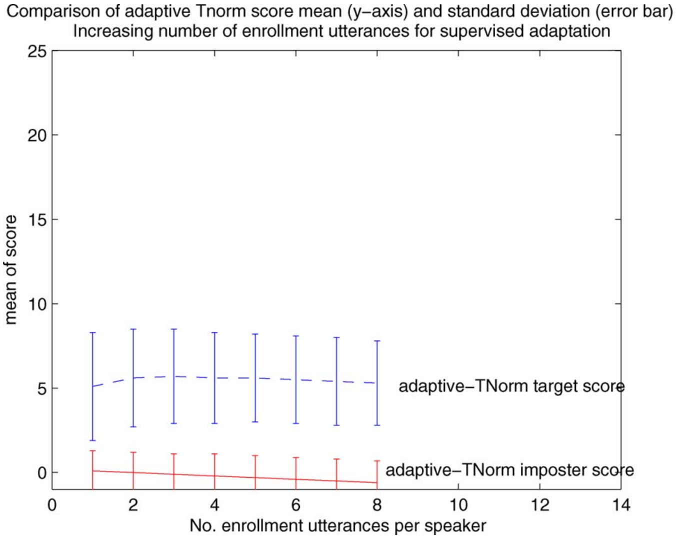

Fig. 3. Eight-conversation two-channel condition of NIST 2005 evaluation. Comparison of target score mean (given by 2230 target trials) and nontarget score mean (given by 21 216 target trials) obtained using various conversations in enrolling each speaker-specific supervector $\mathbf{s}$, based on the adaptive t-norm. Joint factor analysis with 25 channel factors. All trials, male and female.

TABLE II
RESULTS OBTAINED WITHOUT ADAPTATION (ONE ENROLLMENT UTTERANCE) AND WITH SUPERVISED SPEAKER ADAPTATION (EIGHT ENROLLMENT UTTERANCES). zt-NORM SCORE NORMALIZATION

Channel Factors	Adaptation	EER	DCF
25	supervised	3.4%	0.011
25	non-adaptive	6.9%	0.022
75	supervised	3.4%	0.010
75	non-adaptive	6.6%	0.019
100	supervised	3.1%	0.010
100	non-adaptive	6.4%	0.019

from the target speaker and incorrectly accepting impostor utterances for use in progressive adaptation.

All experiments are carried out using the training and test utterance designations given in the NIST 2005 evaluation plan which is summarized in Section IV. All of the trials (male trials and female trials) are included in the evaluation. Speaker verification performance will be reported using both the equal error rate (EER) and the minimum detection cost function (DCF) obtained from detection error tradeoff (DET) curves as specified in the NIST 2005 evaluation plan [1]. The unsupervised adaptation results that are reported were obtained on the core condition of the NIST evaluation using all of the trials in this condition. The supervised adaptation results were obtained using the eight conversation-side condition.

A. Supervised Speaker Adaptation

The performance of the supervised hyperparameter update procedure is presented in Table II. Table II displays the speaker verification performance as the EER and DCF obtained using supervised adaptation with eight enrollment utterances (supervised) and with no adaptation where the speaker specific hyperparameters were trained with a single enrollment utterance (nonadaptive) which is specified in the "one-conversation two-channel" core condition given by NIST. The number of channel factors given in Table II represents the assumed rank of the low-dimensional channel subspace and ranges from 25 to 100. There are two points that can be observed from this table. First, the supervised adaptation procedure reduces both the EER and the minimum DCF by a factor of 2 relative to the nonadapted performance for all three assumed channel factors. Second, the performance after adaptation does not increase significantly when the number of channel factors is increased beyond 25. As a result, all of the remaining experiments were performed with 25 channel factors. The supervised adaptation results obtained using 100 channel factors and zt-norm, shown in Table II, were comparable with the results obtained from SRI International using long-term acoustic features, namely an EER of 3.02% and *a priori* DCF of 0.0097, which show the best performance in NIST 2005 Speaker Recognition Evaluation using eight conversation sides.

TABLE III
COMPARISON OF EER AND DCF PERFORMANCE OBTAINED USING INCREASING NUMBER OF CONVERSATIONS FOR ENROLLING EACH SPEAKER-SPECIFIC SUPERVECTOR $\mathbf{s}$, BASED ON t-NORM, z-NORM, AND zt-NORM

No. enroll.	EER (T)	EER (Z)	EER (ZT)	DCF (T)	DCF (Z)	DCF (ZT)
1	10.7	6.9	5.9	0.037	0.027	0.022
2	6.3	5.6	4.4	0.024	0.021	0.015
3	5.2	5.2	4.3	0.019	0.018	0.013
4	4.6	4.8	3.8	0.017	0.017	0.012
5	4.5	4.7	3.7	0.016	0.016	0.012
6	4.3	4.5	3.6	0.015	0.016	0.012
7	4.3	4.7	3.5	0.015	0.016	0.012
8	4.4	4.6	3.4	0.015	0.015	0.011

TABLE IV
COMPARISON OF OPTIMAL DECISION THRESHOLD USED TO MINIMIZE THE DCF OBTAINED USING INCREASING NUMBER OF CONVERSATIONS FOR ENROLLING EACH SPEAKER-SPECIFIC SUPERVECTOR $\mathbf{s}$, BASED ON t-NORM AND ADAPTIVE t-NORM

No. enroll.	optimal decision threshold (T)	optimal decision threshold (adaptive T)
1	3.54	3.54
2	5.87	3.14
3	7.63	3.08
4	8.78	3.05
5	9.94	2.97
6	10.50	2.95
7	11.35	2.80
8	11.58	2.60

The behavior of the hyperparameter update procedure with respect to the number of adaptation epochs and the score normalization strategies is presented in Tables III–VI. Table III displays the speaker verification performance as the EER and DCF obtained using supervised adaptation according to the update procedure given in Section III for one through eight adaptation utterances. Three different score normalization strategies were evaluated including t-norm (T), z-norm (Z), and zt-norm (ZT). There are several observations that can be made from this table. First, it is clear that both EER and DCF performance measures saturate after approximately four adaptation utterances. Second, 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

TABLE V
COMPARISON OF EER AND DCF PERFORMANCE OBTAINED USING INCREASING NUMBER OF CONVERSATIONS FOR ENROLLING EACH SPEAKER-SPECIFIC SUPERVECTOR $\mathbf{s}$, BASED ON t-NORM AND ADAPTIVE t-NORM

No. enroll.	EER (T)	EER (adaptive T)	DCF (T)	DCF (adaptive T)
1	10.7	10.7	0.037	0.037
2	6.3	6.4	0.024	0.024
3	5.2	5.2	0.019	0.019
4	4.6	4.8	0.017	0.016
5	4.5	4.4	0.016	0.016
6	4.3	4.4	0.015	0.015
7	4.3	4.4	0.015	0.015
8	4.4	4.3	0.015	0.014

TABLE VI
COMPARISON OF EER AND DCF PERFORMANCE OBTAINED USING INCREASING NUMBER OF CONVERSATIONS FOR ENROLLING EACH SPEAKER-SPECIFIC SUPERVECTOR $\mathbf{s}$, BASED ON zt-NORM AND ADAPTIVE zt-NORM

No. enroll.	EER (ZT)	EER (adaptive ZT)	DCF (ZT)	DCF (adaptive ZT)
1	5.9	5.9	0.022	0.022
2	4.4	4.3	0.015	0.015
3	4.3	4.0	0.013	0.013
4	3.8	3.7	0.012	0.012
5	3.7	3.6	0.012	0.012
6	3.6	3.5	0.012	0.012
7	3.5	3.5	0.012	0.011
8	3.4	3.5	0.011	0.011

z-norm. Finally, the performance obtained with zt-norm-based score normalization was always better than obtained using the z-norm or t-norm.

It is important to note that evaluating performance using the EER and DCF measures does not address the issue of statistical robustness of decision thresholds. The DCF values shown in Table III correspond to choosing an optimum decision threshold at each adaptation step. In practice, it is necessary to define a single common decision threshold when the system is initialized. LLR scores must be compared to this threshold both to accept or reject an evaluation utterance to be used for updating speaker-dependent hyperparameters and also for accepting or rejecting claimed speaker identity.

All of the results presented in Tables II and III are computed using optimum *a posteriori* decision thresholds. In order to provide an anecdotal comparison with performance obtained using an *a priori* selected decision threshold, the actual and optimal DCF are given here for the "eight-conversation two-channel" condition of NIST 2006 speaker recognition evaluation, which consists of 33 973 test trials. The actual DCF obtained using a preselected threshold for this condition was 0.020, and the optimum DCF obtained using an *a posteriori* threshold was 0.015. The EER was 3.1%. This performance was obtained using 75 channel factors and zt-norm. Note that, while this performance for the *a priori*-selected decision threshold was obtained on a different evaluation set, it is clear that the performance given for the same condition in Table II presents only a small degradation.

If there is a large variation in LLR scores, it is difficult to specify an appropriate decision threshold that will achieve performance approaching the optimum values shown in Table III. In the t-norm case, it has been shown in Section V that progressive adaptation of speaker models can introduce a "drift" in LLR scores. Table IV shows that an appropriate decision threshold can be chosen within a fairly small range for the adaptive t-norm case. On the other hand, the decision threshold used in the nonadaptive t-norm case has to be increased to compensate for the score drifting phenomenon.

Table V compares t-norm and adaptive t-norm score normalization strategies and Table VI compares zt-norm and adaptive zt-norm score normalization, where the term "adaptive zt-norm" indicates the combination of z-norm and adaptive t-norm. In both cases, the same supervised adaptation scenario as described for Table III is used. It is clear from these comparisons that the EER and DCF performance obtained using the adaptive t-norm-based score normalization methods does not differ significantly from the nonadaptive score normalization methods.

B. *Unsupervised Speaker Adaptation*

Speaker verification performance was evaluated under the NIST 2005 core condition unsupervised adaptation scenario. In particular, the impact of rejecting target speaker utterances and accepting impostor utterances for use in adaptation was investigated. Verification of claimed speaker identity was performed by comparing normalized likelihood ratio scores with a predefined fixed adaptation threshold. Evaluation utterances were presented in a prespecified order with target speaker utterances interspersed with impostor speaker utterances that was defined by the NIST evaluation protocol.

Our speaker verification (SV) simulations were performed using a prespecified sequence of randomly ordered target and imposter utterances defined by NIST. However, it is important to note that varying the order in which target utterances and imposter utterances are presented for adaptation may have the potential to affect SV performance. For example, increasing the number of target utterances that occur early in the test sequence for speaker adaptation may facilitate faster adaptation and result in a significant performance improvement. A quantitative analysis of this effect would require multiple repetitions of each simulation using different orderings of test utterances. It was felt that this was not practical due to the computational overhead associated with NIST evaluation paradigm.

First, Table VII displays the ideal unsupervised adaptation performance using three score normalization techniques. These simulations are referred to as the optimal test of the unsupervised adaptation scenario. The speaker adaptation utterances are selected by accepting all the target utterances and rejecting all the nontarget utterances from the test trials. This cannot be achieved in practice using a predefined, common adaptation threshold. According to Table VII, it can be expected that z-norm gives better performance than adaptive t-norm in the unsupervised scenario. Comparing t-norm and adaptive t-norm, Table VII shows that the score drifting phenomenon seriously degrades the performance obtained using the nonadaptive t-norm.

Fig. 4 displays DET curves computed for unsupervised adaptation using adaptation threshold values ranging from $T = 1.5$

TABLE VII
PERFORMANCE FOR IDEAL UNSUPERVISED ADAPTATION SCENARIO WHERE ONLY TARGET SPEAKER UTTERANCES ARE USED FOR SPEAKER ADAPTATION

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

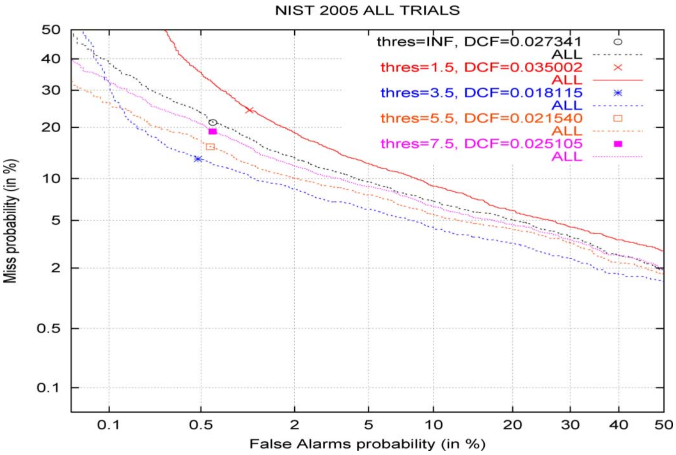

Fig. 4. DET curves over a range of adaptation thresholds used in the unsupervised speaker adaptation for z-norm score normalization.

TABLE VIII
PERFORMANCE OVER A RANGE OF ADAPTATION THRESHOLDS USED IN UNSUPERVISED SPEAKER ADAPTATION FOR z-NORM SCORE NORMALIZATION

Adaptation Threshold	EER	DCF
∞	7.8%	0.027
7.5	7.2%	0.025
5.5	6.5%	0.022
3.5	5.9%	0.018
1.5	9.2%	0.035

TABLE IX
COMPARISON OF THE NUMBER OF TARGET AND NONTARGET UTTERANCES ACCEPTED FOR SPEAKER ADAPTATION USING DIFFERENT VALUES OF ADAPTATION THRESHOLD FOR z-NORM SCORE NORMALIZATION

Adaptation Threshold	No. targets accepted	No. non-targets accepted
∞	0	0
7.5	966	4
5.5	1764	10
3.5	2443	205
1.5	2565	4066

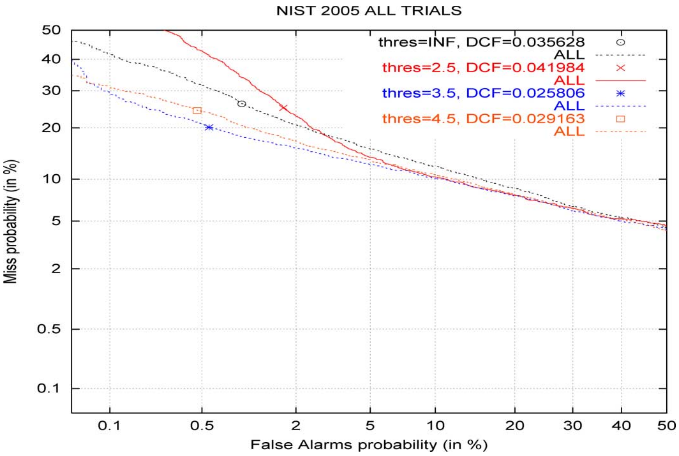

Fig. 5. DET curves over a range of adaptation thresholds used in the unsupervised speaker adaptation for adaptive t-norm score normalization.

TABLE X
PERFORMANCE OVER A RANGE OF ADAPTATION THRESHOLDS USED IN UNSUPERVISED SPEAKER ADAPTATION FOR ADAPTIVE t-NORM SCORE NORMALIZATION

Adaptation Threshold	EER	DCF
∞	11.4%	0.036
4.5	10.5%	0.029
3.5	10.1%	0.026
2.5	10.3%	0.042

to $T = \infty$. Using an adaptation threshold of $T = \infty$ corresponds to the nonadaptive case since no test utterances are accepted for use in adaptation. The EER and DCF measures obtained for these five curves are given in Table VIII. All results were obtained using 25 channel factors and z-norm score normalization. To give insight into the tradeoffs associated with adjusting the adaptation threshold, Table IX displays the number of target speaker utterances and the number of impostor speaker utterances that are accepted for each value of the adaptation threshold. It is clear from Table IX that decreasing the adaptation threshold increases the number of test utterances from both target and nontarget speakers that are used to adapt the speaker models. This initially improves the values for both EER and DCF. However, Table VIII also shows that the performance degrades once a significant number of nontarget utterances are accepted for adaptation.

Of the four values chosen for the adaptation threshold in Table VIII, $T = 3.5$ yielded the lowest EER of 5.9%. The optimum value for the decision threshold used to accept or reject claimed speaker identity for this same system is the threshold value associated with the minimum decision cost function (DCF) shown on the fourth row of Table VIII. This optimum value was found to be equal to 3.75. It is interesting to note that the values of the best empirically chosen adaptation threshold and the optimum speaker verification decision threshold are very similar. This is true in this case despite the fact that these two thresholds should not necessarily have the same value.

When using the adaptive t-norm, the DET curves, the speaker verification performance, and the number of target and nontarget utterances accepted for adaptation obtained using adaptation threshold values ranging from $T = 2.5$ to $T = \infty$ are described in Fig. 5 and Tables X and XI, respectively. Similar to the z-norm results, Fig. 5 and Tables X and XI show that decreasing the adaptation threshold from ∞ to a low value initially improves the performance, but hurts the performance when using a low adaptation threshold value which allows too many nontarget utterances for adaptation. The best adaptation threshold obtained in these adaptive t-norm experiments is 3.5, which yields the lowest EER of 10.1% and the minimum DCF of 0.026.

In order to make a fair comparison between z-norm and adaptive t-norm results, the same adaptation sequences should be applied in the simulations, as shown in Table XII. The adaptation sequence presented in the first row is given by the best

TABLE XI
COMPARISON OF THE NUMBER OF TARGET AND NONTARGET UTTERANCES ACCEPTED FOR SPEAKER ADAPTATION USING DIFFERENT VALUES OF ADAPTATION THRESHOLD FOR ADAPTIVE t-NORM SCORE NORMALIZATION

Adaptation Threshold	No. targets accepted	No. non-targets accepted
∞	0	0
4.5	1793	12
3.5	2180	136
2.5	2446	2032

TABLE XII
COMPARISON OF UNSUPERVISED ADAPTATION PERFORMANCE USING z-NORM (Z) AND ADAPTIVE t-NORM (ADAPT. T). THE SAME SEQUENCE OF ADAPTATION UTTERANCES IS USED FOR BOTH TYPES OF SCORE NORMALIZATION

No. targets accepted	No. non-targets accepted	EER (Z)	EER (adapt. T)	DCF (Z)	DCF (adapt. T)
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

adaptive t-norm results using an adaptation threshold of 3.5. The sequence of target utterances used to adapt speaker models for the simulations presented in the first and second rows of Table XII are identical. However, all 136 nontarget utterances are not used for adaptation in the simulations given in the second row of Table XII. The last simulation shown in the third row of Table XII corresponds to the optimal test of the unsupervised adaptation scenario. There are two observations that can be made from these simulation results. First, comparing the first two rows, the small number of nontarget utterances accepted for adaptation has a negligible impact on performance. Second, comparing the last two rows, the performance difference using z-norm and adaptive t-norm is smaller when an additional 600 target utterances are allowed for adaptation. This is reasonable according to the z-norm and adaptive t-norm results shown in Tables III and V, since these results show that the adaptive t-norm does not give a better performance than z-norm when the amount of adaptation utterances is not sufficient. However, one can expect that the adaptive t-norm results will give comparable performance with z-norm results in the unsupervised scenario using an adaptation threshold of 3.5, when the length of trial list is longer, which makes more target utterances to be allowed for adapting each speaker model.

Speaker verification using unsupervised speaker adaptation was also performed using zt-norm score normalization. Table XIII displays the EER and DCF for unsupervised adaptation with $T = 3.5$ (unsupervised) and the case with no adaptation where $T = \infty$ (nonadaptive) using z-norm and zt-norm score normalization. There are two observations that can be made from the results shown in Table XIII. First, it can be seen from the table that the relative reduction with respect to z-norm performance in EER obtained using the zt-norm is approximately twice as large for the unsupervised adaptation case as it is for the nonadaptive case. The last and most important observation can be made by comparing the EER shown for unsupervised adaptation with zt-norm score normalization in Table XIII and the supervised adaptation performance obtained using the zt-norm shown in Table III. The fact that the EER increases by only approximately 20%, from 3.5% to 4.5%, when going from a supervised adaptation scenario to a more difficult unsupervised adaptation scenario indicates that the factor analysis-based approach for speaker adaptation is very robust.

TABLE XIII
NONADAPTIVE RESULTS (OBTAINED WITH A SINGLE ENROLLMENT UTTERANCE) AND THE BEST UNSUPERVISED SPEAKER ADAPTATION RESULTS OBTAINED USING AN ADAPTATION THRESHOLD OF 3.5

Adaptation	Normalization	EER	DCF
unsupervised	z-norm	5.9%	0.018
non-adaptive	z-norm	7.8%	0.027
unsupervised	zt-norm	4.5%	0.013
non-adaptive	zt-norm	6.9%	0.022

VII. CONCLUSION

A procedure for progressive speaker adaptation in a joint factor analysis based speaker verification system has been proposed. Results for both supervised and unsupervised speaker adaptation scenarios were presented. The issue of how score normalization algorithms interact with progressive speaker adaptation was also investigated and an adaptive t-norm score normalization procedure was proposed.

The best supervised adaptation results obtained using 100 channel factors and zt-norm-based score normalization corresponded to an EER of 3.1% and a minimum detection cost of 0.010. The best EER achieved for the unsupervised adaptation scenario was 4.5% corresponding to a minimum DCF of 0.013. This was obtained using 25 channel factors and zt-norm-based score normalization.

The behavior of likelihood ratio scores during progressive speaker adaptation obtained using different score normalizations was also analyzed in Sections V and VI of the paper. It was shown in the paper that the use of an adaptive t-norm, whether applied independently or as part of a "z-norm adaptive t-norm" strategy resulted in normalized likelihood ratio scores that exhibited less variability than likelihood ratio scores normalized using the nonadaptive score normalization strategies. However, all speaker verification results were reported using the NIST standard EER and DCF both of which assume an optimal threshold setting. Further research will investigate the use of the adaptive t-norm in systems relying on fixed empirical decision thresholds in unsupervised speaker adaptation scenarios.

APPENDIX A
POSTERIOR CALCULATION OF THE LATENT VARIABLES $\mathbf{x}$ AND $\mathbf{z}$

In the joint factor analysis model, a CF-dimensional supervector $\mathbf{M}$ can be decomposed as

$$\mathbf{M} = \mathbf{s} + \mathbf{c} \tag{10}$$

where $\mathbf{s}$ is the speaker-dependent supervector given by

$$\mathbf{s} = \mathbf{m} + \mathbf{d}\mathbf{z} \tag{11}$$

and $\mathbf{c}$ is the channel-dependent supervector given by

$$\mathbf{c} = \mathbf{u}\mathbf{x}. \tag{12}$$

We are given the prior distribution of $\mathbf{s}$, $p(\mathbf{s}) \sim N(\mathbf{m}, \mathbf{d}^2)$, the prior distribution of $\mathbf{c}$, $p(\mathbf{c}) \sim N(0, \mathbf{u}\mathbf{u}^*)$, and the $CF \times CF$ diagonal matrix $\mathbf{\Sigma}$ used to capture the uncertainty in the utterance that is independent of $\mathbf{s}$ and $\mathbf{c}$. In this appendix, we summarize the calculations needed to jointly evaluate the posterior distribution of the hidden variables $\mathbf{x}$ and $\mathbf{z}$. The theorem is presented in [8].

For each mixture component c, let N_c be the total number of acoustic observation vectors o_t in the training utterance for the given mixture component, which is given by

$$N_c = \sum_{t=1}^{T} P(c|o_t, \lambda) \tag{13}$$

and set F_c to be the sum which extends over all acoustic observations o_t aligned with the given mixture component c, which is given by

$$F_c = \sum_{t=1}^{T} P(c|o_t, \lambda) o_t, \tag{14}$$

where $P(c|o_t, \lambda)$ refers to the *a posteriori* probability of occupying mixture component c in the GMM-based universal background model λ at time t. Let $\mathbf{N}$ be the $CF \times CF$ diagonal matrix whose diagonal blocks are $N_c I$ (for $c = 1, \ldots, C$) where I is the $F \times F$ identity matrix. Let $\mathbf{F}$ be the $CF \times 1$ vector obtained by concatenating F_c (for $c = 1, \ldots, C$). Set

$$S_c = \mathrm{diag}\left(\sum_{t=1}^{T} P(c|o_t, \lambda) o_t {o_t}^*\right). \tag{15}$$

This can be used to compute the log likelihood function described in the Appendix B.

Theorem: If

$$\mathbf{X} = \begin{pmatrix} \mathbf{x} \\ \mathbf{z} \end{pmatrix}$$

then the posterior distribution of $\mathbf{X}$ is Gaussian of the same form as the posterior distribution described in Proposition 1 of [6]. Specifically, if $\mathbf{I}$ is the identity matrix and $\mathbf{V}$ and $\mathbf{L}$ are the matrices defined by

$$\mathbf{V} = (\,\mathbf{u} \quad \mathbf{d}\,) \tag{16}$$

$$\mathbf{L} = \mathbf{I} + \mathbf{V}^*\mathbf{\Sigma}^{-1}\mathbf{N}\mathbf{V} \tag{17}$$

then the posterior distribution of $\mathbf{X}$ has covariance matrix $\mathbf{L}^{-1}$ and mean $\mathbf{L}^{-1}\mathbf{V}^*\mathbf{\Sigma}^{-1}(\mathbf{F} - \mathbf{N}\mathbf{m})$. Thus, calculating the posterior distribution of $\mathbf{X}$ is essentially a matter of inverting the matrix $\mathbf{L}$.

A straightforward calculation shows that $\mathbf{L}$ can be written as

$$\begin{pmatrix} \mathbf{I} + \mathbf{u}^*\mathbf{\Sigma}^{-1}\mathbf{N}\mathbf{u} & \mathbf{u}^*\mathbf{\Sigma}^{-1}\mathbf{N}\mathbf{d} \\ \mathbf{d}\mathbf{N}\mathbf{\Sigma}^{-1}\mathbf{u} & \mathbf{I} + \mathbf{\Sigma}^{-1}\mathbf{N}\mathbf{d}^2 \end{pmatrix}. \tag{18}$$

So, $\mathbf{L}^{-1}$ can be calculated by using the identity

$$\begin{pmatrix} \boldsymbol{\alpha} & \boldsymbol{\beta} \\ \boldsymbol{\beta}^* & \boldsymbol{\gamma} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\zeta}^{-1} & -\boldsymbol{\zeta}^{-1}\boldsymbol{\beta}\boldsymbol{\gamma}^{-1} \\ -\boldsymbol{\gamma}^{-1}\boldsymbol{\beta}^*\boldsymbol{\zeta}^{-1} & \boldsymbol{\gamma}^{-1} + \boldsymbol{\gamma}^{-1}\boldsymbol{\beta}^*\boldsymbol{\zeta}^{-1}\boldsymbol{\beta}\boldsymbol{\gamma}^{-1} \end{pmatrix}$$

where

$$\boldsymbol{\zeta} = \boldsymbol{\alpha} - \boldsymbol{\beta}\boldsymbol{\gamma}^{-1}\boldsymbol{\beta}^*$$

with

$$\begin{aligned} \boldsymbol{\alpha} &= \mathbf{I} + \mathbf{u}^*\mathbf{\Sigma}^{-1}\mathbf{N}\mathbf{u} \\ \boldsymbol{\beta} &= \mathbf{u}^*\mathbf{\Sigma}^{-1}\mathbf{N}\mathbf{d} \\ \boldsymbol{\gamma} &= \mathbf{I} + \mathbf{\Sigma}^{-1}\mathbf{N}\mathbf{d}^2. \end{aligned}$$

Appendix B
Log Likelihood Function Used to Make Speaker Verification Decisions

Some notation is required before describing the test likelihood function used in the joint factor analysis-based speaker verification system. We are given the speaker-independent hyperparameter set $\mathbf{\Lambda} = (\mathbf{m}, \mathbf{d}, \mathbf{u}, \mathbf{\Sigma})$. Let $\mathbf{\Sigma}_c$ be the cth $F \times F$ diagonal covariance matrix in $\mathbf{\Sigma}$, where $c = 1, \ldots, C$. Recall that $\mathbf{\Sigma}$ captures the speaker-independent and channel-independent uncertainty, which is fixed in the test likelihood function for any target speaker. Let N_c, F_c, and S_c be the zero-order, first-order, and second-order sufficient statistics for the given mixture component c, which are obtained from the test utterance χ as described in (13)–(15). Use the same definition described in Appendix A for the notations $\mathbf{N}$ and $\mathbf{F}$. In addition, let $\mathbf{S}$ be the $CF \times CF$ diagonal matrix whose diagonal blocks are S_c (for $c = 1, \ldots, C$).

Given the distribution of supervector $\mathbf{s}$, either the prior distribution or the posterior distribution, the expectations of the first- and second-order moments of χ around $\mathbf{s}$, $E[\mathbf{F}_\mathbf{s}]$ and $E[\mathbf{S}_\mathbf{s}]$, are given by

$$\begin{aligned} E[\mathbf{F}_\mathbf{s}] &= \mathbf{F} - \mathbf{N}E[\mathbf{s}] \\ E[\mathbf{S}_\mathbf{s}] &= \mathbf{S} - 2\mathrm{diag}(\mathbf{F}E[\mathbf{s}^*]) \\ &\quad + \mathrm{diag}(\mathbf{N}(E[\mathbf{s}]E[\mathbf{s}^*] + \mathrm{Cov}(\mathbf{s}, \mathbf{s}))). \end{aligned} \tag{19}$$

Finally, let

$$\mathbf{l} = \mathbf{I} + \mathbf{u}^*\mathbf{\Sigma}^{-1}\mathbf{N}\mathbf{u} \tag{20}$$

and let $\mathbf{l}^{1/2}$ be an upper triangular matrix such that

$$\mathbf{l} = \mathbf{l}^{1/2}\mathbf{l}^{1/2*}. \tag{21}$$

Mathematically, $\mathbf{l}^{1/2}$ is given by the Cholesky decomposition of the symmetric square matrix $\mathbf{l}$.

As presented in [8] and [9], the test log likelihood function has a closed-form representation such that

$$\begin{aligned} \log P(\chi|\mathbf{s}) = &\sum_{c=1}^{C} N_c \log \frac{1}{(2\pi)^{F/2}|\mathbf{\Sigma}_c|^{1/2}} \\ &- \frac{1}{2} tr(\mathbf{\Sigma}^{-1}E[\mathbf{S}_\mathbf{s}]) - \frac{1}{2}\log|\mathbf{l}| \\ &+ \frac{1}{2}||\mathbf{l}^{-1/2}\mathbf{u}^*\mathbf{\Sigma}^{-1}E[\mathbf{F}_\mathbf{s}]||^2 \end{aligned} \tag{22}$$

provided that supervector $\mathbf{s}$ is known, or $E[\mathbf{s}]$ and $Cov(\mathbf{s}, \mathbf{s})$ given by (19) are known. $tr(\cdot)$ denotes the trace of the matrix.

Given a target speaker s providing a posterior distribution of supervector $\mathbf{s}$, the verification score of a test utterance χ_{test} is given by an LLR using the function defined in (22). The supervector $\mathbf{s}$ shown in the numerator of LLR is specified by the speaker-dependent posterior distribution, and $\mathbf{s}$ shown in the denominator term is specified by the speaker-independent prior distribution. This LLR score, denoted by LLR(χ_{test}, s), can be normalized using score normalization techniques such as z-norm or t-norm and then be compared with a decision threshold for make a speaker verification decision.

References

[1] *The NIST Year 2005 Speaker Recognition Evaluation Plan*, [Online]. Available: http://www.itl.nist.gov/iad/894.01/tests/spk/2005

[2] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[3] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey*, Crete, Greece, Jun. 2001, pp. 213–218.

[4] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms." [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

[5] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification." [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

[6] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 345–354, May 2005.

[7] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," in *Proc. ICASSP'05*, Philadelphia, PA, Mar. 2005, pp. 637–640.

[8] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[9] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Improvements in factor analysis based speaker verification," in *Proc. ICASSP'06*, Toulouse, France, May 2006, pp. 113–116.

[10] N. Mirghafori and L. Heck, "An adaptive speaker verification system with speaker dependent a priori decision thresholds," in *Proc. ICSLP*, Denver, CO, Sep. 2002, pp. 589–592.

[11] N. Mirghafori and M. Hébert, "Parameterization of the score threshold for a text-dependent adaptive speaker verification system," in *Proc. ICASSP'04*, Montreal, QC, Canada, May 2004, pp. 361–364.

[12] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Process.*, vol. 10, pp. 42–52, 2000.

[13] R. Vogt, B. Baker, and S. Sridharan, "Modeling session variability in text-independent speaker verification," in *Proc. Eurospeech*, Lisbon, Portugal, Sep. 2005, pp. 3117–3120.

Shou-Chun Yin received the B.S. and M.S. degrees in electrical and computer engineering from McGill University, Montreal, QC, Canada, in 2005 and 2007, respectively. His M.S. thesis was titled "Speaker adaptation in joint factor analysis based text independent speaker verification." He is currently pursuing the Ph.D. degree in electrical and computer engineering at McGill University under a McGill Engineering Doctoral Award (MEDA) fellowship.

He has participated in the NIST Speaker Recognition Evaluation (SRE) 2006. His research interests include speaker verification and statistical modeling and machine learning as applied to continuous speech recognition applications.

Richard Rose (SM'00) received the B.S. and M.S. degrees in electrical and computer engineering from the University of Illinois, Urbana, and the Ph.D. degree in electrical engineering from the Georgia Institute of Technology, Atlanta.

From 1980 to 1984, he was with Bell Laboratories working on signal processing and digital switching systems. From 1988 to 1992, he was with Massachusetts Institute of Technology (MIT) Lincoln Laboratory, working on speech recognition and speaker recognition. He was with AT&T Bell Laboratories from 1992 to 1996 and with AT&T Labs–Research, Florham Park, NJ, from 1996 to 2003. Currently, he is an Associate Professor of Electrical and Computer Engineering at McGill University, Montreal, QC, Canada.

Prof. Rose served as a member of the IEEE Signal Processing Society Technical Committee on Digital Signal Processing (DSP) from 1990 to 1995. He was elected as an at large member of the Board of Governors for the Signal Processing Society during the period from 1995 to 1997. He served as an Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING from 1997 to 1999. He served as a member of the IEEE Signal Processing Society Speech Technical Committee (STC) from 2002 through 2005, and was the founding editor of the STC Newsletter. He also served as one of the general chairs of the 2005 IEEE Automatic Speech Recognition and Understanding Workshop. He is a member of Tau Beta Pi, Eta Kappa Nu, and Phi Kappa Phi.

Patrick Kenny received the B.A. degree in mathematics from Trinity College, Dublin, U.K., and the M.Sc. and Ph.D. degrees, also in mathematics, from McGill University, Montreal, QC, Canada.

He was a Professor of Electrical Engineering at INRS-Télécommunications, Montreal, from 1990 to 1995 when he started up a company (Spoken Word Technologies) to spin off INRS's speech recognition technology. He joined he Centre de Recherche Informatique de Montréal (CRIM) in 1998, where he now holds the position of Principal Research Scientist. His current research interests are concentrated on Bayesian speaker and channel adaptation for speech and speaker recognition.