Uncertain LDA: Including observation uncertainties in discriminative transforms

Rahim Saeidi*, *Senior Member, IEEE,* Ramón Fernandez Astudillo, *Member, IEEE*
and Dorothea Kolossa, *Senior Member, IEEE,*

Abstract—Linear discriminant analysis (LDA) is a powerful technique in pattern recognition to reduce the dimensionality of data vectors. It maximizes discriminability by retaining only those directions that minimize the ratio of within-class and between-class variance. In this paper, using the same principles as for conventional LDA, we propose to employ uncertainties of the noisy or distorted input data in order to estimate maximally discriminant directions. We demonstrate the efficiency of the proposed *uncertain LDA* on two applications using state-of-the-art techniques. First, we experiment with an automatic speech recognition task, in which the uncertainty of observations is imposed by real-world additive noise. Next, we examine a full-scale speaker recognition system, considering the utterance duration as the source of uncertainty in authenticating a speaker. The experimental results show that when employing an appropriate uncertainty estimation algorithm, uncertain LDA outperforms its conventional LDA counterpart.

1 Introduction

LINEAR discriminant analysis (LDA) is one of the simplest and most used transforms to enhance class separability for multidimensional observations [1], [2]. Conventional LDA assumes that each class follows a *normal distribution* and classes share the same covariance structure (are homoscedastic) [3]. Although these assumptions do not generally hold in practice, this conventional approach, briefly reviewed in Section 2, and its variants have been found useful in many applications including automatic speech and speaker recognition [4]–[7].

Dimensionality reduction is a usual pre-processing stage to make the input data more suitable for the modeling stage. Though statistical modeling techniques, like Gaussian mixture models, are quite successful in modeling arbitrary distributions, they become unstable when the data dimensionality is high. The main reason is that the number of variables required for covariance structure estimation grows exponentially with data dimension. When the dimensionality of the data becomes comparable with the number of samples per class, the sample covariance estimation becomes unstable, a problem known as *small sample size* [8]. Dealing with ill-posed covariance estimation in finding discriminant directions has been addressed as a challenging problem [9]. Regularization [10] and Bayesian estimation [11] of covariance models have been discussed in the literature to overcome this issue. It is also possible to obtain non-linear class separation using subclass discriminant analysis and the kernel trick in LDA [12]. When each class is composed of several partitions, subclass discriminant analysis [13] aims to maximize the distance between class means and the subclass means in the same class at the same time.

Compared to *principal component analysis*, class-dependent dimensionality reduction is expected to aid in modeling the classes more effectively. The extension of linear discriminant analysis to unequal covariance matrices and non-normal distributions for classes leads to heteroscedastic LDA [14], quadratic discriminant analysis [2], [15] and mixture discriminant analysis [16].

A *distance preserving dimensionality reduction* transform maps the D-dimensional data samples to a d-dimensional space ($d < D$) subject to the constraint that nearby data samples are mapped to nearby low-dimensional representations [17]. Considering K as the number of the classes in a dataset, the selection of less than $K - 1$ dimensions in LDA for data projection does not guarantee to preserve the distance between classes from a classification perspective for $K > 2$ [18].

The application of the *direct distance matrix* (DDM) as a generalization of the between-class scatter matrix has been suggested to address this issue [19]. In LDA, the distance function used for obtaining the DDM could be chosen as the Chernoff distance [20] or as its multi-class generalization, Matusita's separability measure [21], [22].

In this paper we are addressing the task of finding linear discriminant directions when instead of a point estimate for an observation, a probabilistic description is available. We achieve such a probabilistic description by using so-called *observation uncertainties*. In this approach, the feature extraction process outputs the point estimate of a feature vector along with an uncertainty. The point estimate is assumed to form the mean of a Gaussian, while the corresponding variance is set to the estimated uncertainty. Throughout this paper we call this

• *R. Saeidi is with the Department of Signal Processing and Acoustics, Aalto University, Finland (e-mail: rahim.saeidi@aalto.fi).*
• *R. F. Astudillo is with the Spoken Language Systems Lab, INESC-ID, Lisbon, Portugal (email: ramon@astudillo.com)*
• *D. Kolossa is with the Institute of Communication Acoustics, Ruhr-Universität Bochum, Germany (email: dorothea.kolossa@rub.de).*

an *uncertain observation* and propose *uncertain LDA* to account for the observation uncertainties in estimating scatter matrices for LDA.

The remainder of this paper is organized as follows: After introducing standard LDA in Section 2 and briefly discussing the concept of observation uncertainties in Section 3, we introduce the new approach of uncertain LDA (ULDA) in Section 4. We present results on real-world data for speech recognition and speaker identification in Sections 5 and 6, respectively. After a brief discussion in Section 7, conclusions are drawn in Section 8.

2 Conventional LDA

Let $\mathbf{X} = \mathbf{x}_1, \ldots, \mathbf{x}_L$ be a set of L samples (features) each belonging to one of K classes, partitioning the data into clusters $C_1, \ldots, C_K$. The conventional LDA aims at finding a linear transformation of those features that would maximize the separability of the clusters assuming each class is Gaussian distributed and Gaussians share the same covariance structure. In order to find the discriminant directions, we first calculate the sample mean, $\boldsymbol{\mu}$, and class mean, $\boldsymbol{\mu}_k$, as

$$\boldsymbol{\mu} = \frac{1}{L} \sum_{l=1}^{L} \mathbf{x}_l, \tag{1}$$

$$\boldsymbol{\mu}_k = \frac{1}{|C_k|} \sum_{l \in C_k} \mathbf{x}_l, \tag{2}$$

where $|C_k|$ is the cardinality of class k. Next, the within-class and between-class scatters (indicated by $\mathbf{S}_W$ and $\mathbf{S}_B$, respectively) are given as

$$\mathbf{S}_W = \sum_{k=1}^{K} \sum_{l \in C_k} |C_k| (\mathbf{x}_l - \boldsymbol{\mu}_k)(\mathbf{x}_l - \boldsymbol{\mu}_k)^T, \tag{3}$$

$$\mathbf{S}_B = \sum_{k=1}^{K} (\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^T. \tag{4}$$

The optimization problem is then solved by maximizing the Fisher-Rao criterion [1], [3] as

$$\{\hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_{K-1}\} = \arg\max_{\mathbf{w}} \left\{ \frac{|\mathbf{w}^T \mathbf{S}_B \mathbf{w}|}{|\mathbf{w}^T \mathbf{S}_W \mathbf{w}|} \right\}. \tag{5}$$

where $\hat{\mathbf{w}}_i, i = 1, \ldots, K-1$ is the i-th eigenvector corresponding to i-th eigenvalue λ_i, found by solving $(\mathbf{S}_B - \lambda_i \mathbf{S}_W)\hat{\mathbf{w}}_i = 0$. The optimal projection matrix $\mathbf{W}^*$ is formed by putting d eigenvectors ($d \leq K-1$) corresponding to the largest eigenvalues together and the new representation of features is given by

$$\mathbf{Y} = \mathbf{W}^{*T} \mathbf{X}, \tag{6}$$

where $\mathbf{Y} = \mathbf{y}_1, \cdots, \mathbf{y}_L$.

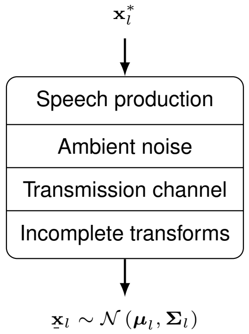

Fig. 1: Assuming there exists an optimal $\mathbf{x}_l^*$ to represent an acoustic event in low dimensional space, there are several factors leading to arrive at uncertain description of one observation (indexed by l in this example) represented as $\underline{\mathbf{x}}_l \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$.

3 Uncertainty-of-Observation Techniques

The concept of uncertainty exists in many branches of science and there are different techniques employed to deal with uncertainty [23]. Uncertainty-of-observation techniques concern the application of machine learning algorithms to situations in which the input signal x can not be directly observed but a probabilistic model relating x and some observed variable y is available. This model can be in the form of a likelihood distribution $p(y|x)$ as e.g. in uncertainty decoding [24], a posterior distribution $p(x|y)$ as e.g. in uncertainty propagation [25], or a joint distribution $p(x,y)$ as e.g. in joint uncertainty decoding [26]. See [27] for review of the topic. As it is shown in Figure 1, the uncertainty in an observation could be a result of several factors which collectively result in deviating from optimal representation.

The focus of this paper is developing a linear discriminant analysis algorithm able to deal with observation uncertainties in the form of a posterior distribution. In such a scenario, the input signal is a sequence of independent random variables each described by a Gaussian posterior distribution. In real world conditions, these posteriors can be attained e.g. from a previous signal enhancement step [28] or errors in estimation [29]. In this work we perform robust automatic speech recognition and speaker recognition. In noise robust speech recognition, the additive noise is considered as the source of the uncertainty. By assuming that the Fourier coefficients of speech and noise are circularly symmetric complex Gaussian distributed, it is possible to use for example Wiener filtering to arrive at an uncertain spectral representation [30], [31].

In modern speaker recognition systems [32], each speaker is represented by a so-called i-vector, which is the mean of the *a posteriori* distribution considering the available speech material for a given speaker. Following the same principle as in [29], [33], [34], we carry out speaker recognition experiments in which we take into account the uncertainty in mean statistics estimation.

In this case we employ the estimated uncertainty of the i-vectors in finding ULDA transform. In the context of speaker recognition experiments, an ideal utterance representation would be the estimated mean statistics with sufficiently many samples and the source of uncertainty in $\mathbf{x}_l$, with limited observations, will be the *incomplete data*. In the speaker recognition paradigm, a long recording ($\sim$ 3 minutes) of speech presents the underlying speaker much more accurately than a short recording ($\sim$ 5 seconds). We consider the duration of the provided speech for authentication as the source of uncertainty in our experiments; where an i-vector uncertainty increases as the duration of utterance gets shorter.

4 LDA with a Posterior description of Uncertain Features

In this section, we extend the conventional LDA for the case that the input data is available in the form of posterior distributions as $\underline{\mathbf{X}} = \underline{\mathbf{x}}_1, \cdots, \underline{\mathbf{x}}_L$ where each $\underline{\mathbf{x}}_l \in \mathbb{R}^D$ is described by a respective probability density function $f_{\underline{\mathbf{x}}_l|\mathrm{I}}(\underline{\mathbf{x}}_l)$, with I as the available information. The conventional LDA would deal with this type of data by using only first-order statistics as $\mathbf{x}_l = \boldsymbol{\mu}_l = \mathbb{E}(\underline{\mathbf{x}}_l)$ and continue to calculate between- and within-class scatter matrices; $\mathbf{S}_B$ and $\mathbf{S}_W$. We develop the concept of *uncertain LDA* in such a way that we use the second-order statistics of $\underline{\mathbf{x}}_l|\mathrm{I}$, namely $\boldsymbol{\Sigma}_l = \mathbb{Cov}(\underline{\mathbf{x}}_l)$ in calculating *expected scatter matrices* $\widehat{\mathbf{S}_B}$ and $\widehat{\mathbf{S}_W}$. The eigenvectors calculated by employing expected scatter matrices are deemed to be more representative of variability in $\underline{\mathbf{X}}$ compared to the ones obtained by considering $\mathbf{X} = \mathbb{E}(\underline{\mathbf{X}}|\mathrm{I})$. This claim is tested with two different applications in this paper. In the following, we describe how to find expected scatter matrices.

4.1 Uncertain LDA

For the sake of tractability, we assume that posterior distributions of observations are Gaussian as $\underline{\mathbf{x}}_l|\mathrm{I} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$, which can be fully described by first- and second-order statistics. In the following derivations, we rely on the fact that sum of Gaussian variables is another Gaussian variable. By applying this principle to (1) and (2) we obtain

$$\underline{\boldsymbol{\mu}}|\mathrm{I} \sim \mathcal{N}\left(\frac{1}{L}\sum_{l=1}^{L}\boldsymbol{\mu}_l, \frac{1}{L^2}\sum_{l=1}^{L}\boldsymbol{\Sigma}_l\right), \tag{7}$$

$$\underline{\boldsymbol{\mu}}_k|\mathrm{I} \sim \mathcal{N}\left(\frac{1}{|C_k|}\sum_{l\in C_k}\boldsymbol{\mu}_l, \frac{1}{|C_k|^2}\sum_{l\in C_k}\boldsymbol{\Sigma}_l\right). \tag{8}$$

Next, we find the distribution for the sample mean deviation $\underline{\boldsymbol{\delta}}_{lk}$, and the class mean deviation $\underline{\boldsymbol{\delta}}_k$ as

$$\begin{aligned}
&\underline{\boldsymbol{\delta}}_{lk} = \underline{\mathbf{x}}_l - \underline{\boldsymbol{\mu}}_k \Rightarrow \underline{\boldsymbol{\delta}}_{lk}|\mathrm{I} \sim \mathcal{N}\left(\mathbb{E}(\underline{\boldsymbol{\delta}}_{lk}), \mathbb{Cov}(\underline{\boldsymbol{\delta}}_{lk}, \underline{\boldsymbol{\delta}}_{lk})\right)\\
&\mathbb{E}(\underline{\boldsymbol{\delta}}_{lk}) = \boldsymbol{\mu}_l - \frac{1}{|C_k|}\sum_{l\in C_k}\boldsymbol{\mu}_l\\
&\mathbb{Cov}(\underline{\boldsymbol{\delta}}_{lk}, \underline{\boldsymbol{\delta}}_{lk}) = \boldsymbol{\Sigma}_l - \frac{2}{|C_k|}\boldsymbol{\Sigma}_l + \frac{1}{|C_k|^2}\sum_{l\in C_k}\boldsymbol{\Sigma}_l
\end{aligned} \tag{9}$$

$$\begin{aligned}
&\underline{\boldsymbol{\delta}}_{k} = \underline{\boldsymbol{\mu}}_k - \underline{\boldsymbol{\mu}} \Rightarrow \underline{\boldsymbol{\delta}}_{k}|\mathrm{I} \sim \mathcal{N}\left(\mathbb{E}(\underline{\boldsymbol{\delta}}_{k}), \mathbb{Cov}(\underline{\boldsymbol{\delta}}_{k}, \underline{\boldsymbol{\delta}}_{k})\right)\\
&\mathbb{E}(\underline{\boldsymbol{\delta}}_{k}) = \frac{1}{|C_k|}\sum_{l\in C_k}\boldsymbol{\mu}_l - \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{\mu}_l\\
&\mathbb{Cov}(\underline{\boldsymbol{\delta}}_{k}, \underline{\boldsymbol{\delta}}_{k}) = \frac{1}{|C_k|^2}\sum_{l\in C_k}\boldsymbol{\Sigma}_l - \frac{2}{L|C_k|}\sum_{l\in C_k}\boldsymbol{\Sigma}_l + \frac{1}{L^2}\sum_{l=1}^{L}\boldsymbol{\Sigma}_l
\end{aligned} \tag{10}$$

where we need to take into account the correlation between the mean and each sample. To complete computing $\widehat{\mathbf{S}_W}$ and $\widehat{\mathbf{S}_B}$, we just need to apply the linearity of the expectation operator so that

$$\begin{aligned}
\widehat{\mathbf{S}_W} &= \mathbb{E}\left\{\sum_{k=1}^{K}\sum_{l\in C_k}(\underline{\mathbf{x}}_l - \underline{\boldsymbol{\mu}}_k)(\underline{\mathbf{x}}_l - \underline{\boldsymbol{\mu}}_k)^T\right\}\\
&= \sum_{k=1}^{K}\sum_{l\in C_k}\mathbb{E}\left\{\underline{\boldsymbol{\delta}}_{lk}\underline{\boldsymbol{\delta}}_{lk}^T\right\}\\
&= \sum_{k=1}^{K}\sum_{l\in C_k}\mathbb{Cov}(\underline{\boldsymbol{\delta}}_{lk}, \underline{\boldsymbol{\delta}}_{lk}) + \mathbb{E}\left\{\underline{\boldsymbol{\delta}}_{lk}\right\}\mathbb{E}\left\{\underline{\boldsymbol{\delta}}_{lk}^T\right\}
\end{aligned} \tag{11}$$

$$\begin{aligned}
\widehat{\mathbf{S}_B} &= \mathbb{E}\left\{\sum_{k=1}^{K}|C_k|(\underline{\mu}_k - \underline{\boldsymbol{\mu}})(\underline{\mu}_k - \underline{\boldsymbol{\mu}})^T\right\}\\
&= \sum_{k=1}^{K}|C_k|\mathbb{E}\left\{\underline{\boldsymbol{\delta}}_{k}\underline{\boldsymbol{\delta}}_{k}^T\right\}\\
&= \sum_{k=1}^{K}|C_k|(\mathbb{Cov}(\underline{\boldsymbol{\delta}}_{k}, \underline{\boldsymbol{\delta}}_{k}) + \mathbb{E}\left\{\underline{\boldsymbol{\delta}}_{k}\right\}\mathbb{E}\left\{\underline{\boldsymbol{\delta}}_{k}^T\right\}),
\end{aligned} \tag{12}$$

leading to

$$\begin{aligned}
\widehat{\mathbf{S}_W} &= \mathbf{S}_W + \sum_{k=1}^{K}\sum_{l\in C_k}\mathbb{Cov}(\underline{\boldsymbol{\delta}}_{lk}, \underline{\boldsymbol{\delta}}_{lk})\\
&= \mathbf{S}_W + \sum_{k=1}^{K}\frac{|C_k|-1}{|C_k|}\sum_{l\in C_k}\boldsymbol{\Sigma}_l
\end{aligned} \tag{13}$$

$$\begin{aligned}
\widehat{\mathbf{S}_B} &= \mathbf{S}_B + \sum_{k=1}^{K}|C_k|\mathbb{Cov}(\underline{\boldsymbol{\delta}}_{k}, \underline{\boldsymbol{\delta}}_{k})\\
&= \mathbf{S}_B + \sum_{k=1}^{K}\left(\frac{|C_k|}{L^2}\sum_{l=1}^{L}\boldsymbol{\Sigma}_l + \frac{L-2|C_k|}{L|C_k|}\sum_{l\in C_k}\boldsymbol{\Sigma}_l\right).
\end{aligned} \tag{14}$$

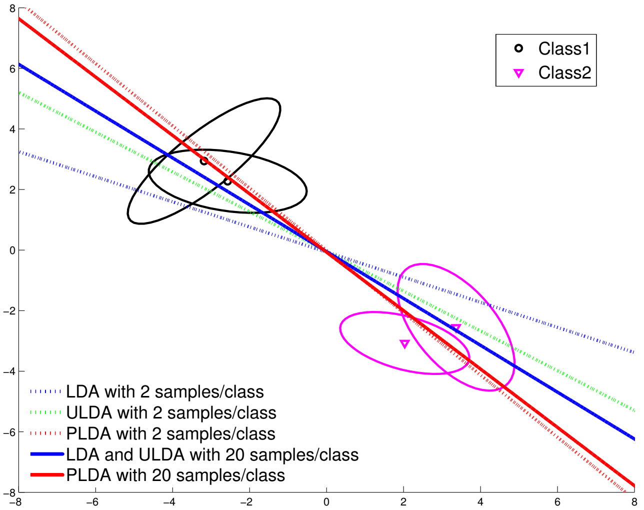

Fig. 2: An example of comparing LDA, ULDA and PLDA in terms of finding the most discriminant projection. The dotted lines represent discriminant direction found by availability of only 2 samples (with respective uncertainty shown as ellipse around each data point) per class. The full lines indicate the projections found using 10 times more data from the same classes. The LDA projection can be considered optimal when many samples are available for estimating scatter matrices. Having many samples per class, the ULDA transform coincides with LDA (blue full line). When the training data is scarce, ULDA (dotted green line) provides a better resemblance of optimal LDA (blue full line).

By removing uncertainties, i.e. setting $\mathbf{\Sigma}_l = 0$, the posterior description of features becomes a Dirac delta function centered on $\mathbf{x}_l$ and hence $\widehat{\mathbf{S}_B} = \mathbf{S}_B$ and $\widehat{\mathbf{S}_W} = \mathbf{S}_W$.

By using $\widehat{\mathbf{S}_W}$ and $\widehat{\mathbf{S}_B}$ in Equation 5, the ULDA transform is found as $\widehat{\mathbf{W}^*}$ and the low-dimensional uncertain observations $\underline{\mathbf{Y}} = \underline{\mathbf{y}}_1, \cdots, \underline{\mathbf{y}}_L$ are obtained via (6), with $\widehat{\mathbf{W}^*}$ replacing $\mathbf{W}^*$. All $\underline{\mathbf{y}}_l \in \mathbb{R}^d$, $d < D$ are then passed to the next steps of the recognizer.

4.2 Relation to Probabilistic LDA

It has been shown [16] that LDA maximizes the likelihood of Gaussian classes and is equivalent to linear regression using the class labels. Probabilistic LDA (PLDA) is a general method that can accomplish a wide variety of recognition tasks [35], [36]. PLDA is formulated by a generative model, where the j-th observation from i-th class is expressed as

$$\mathbf{x}_{ij} = \mathbf{m} + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{c}_{ij} + \epsilon_{ij}. \qquad (15)$$

In Equation 15, $\mathbf{m}$ stands for the expected value of all data. The $\mathbf{F}$ and $\mathbf{G}$ matrices are analogous to the between-class and within-class subspaces in the conventional LDA, but are estimated probabilistically. The $\mathbf{h}_i$ and $\mathbf{c}_{ij}$ are latent variables with normal Gaussian prior distribution to position a data point $\mathbf{x}_{ij}$ in between-class and within-class subspaces. The term ϵ_{ij} is a zero-mean Gaussian noise to capture any remaining variation. The term $\mathbf{m} + \mathbf{F}\mathbf{h}_i$ represents the i-th class in general and by adding $\mathbf{G}\mathbf{c}_{ij} + \epsilon_{ij}$ a full description of $\mathbf{x}_{ij}$ is provided.

It is important to note that methods such as LDA (and ULDA) do not solve recognition and hypothesis testing problems directly, but they are used as a pre-processing stage to reduce dimensionality. On the other hand, PLDA it mostly targeted to accomplish the recognition tasks taking place in the latent variable space. In particular cases of face recognition [36] or speaker recognition [37], the PLDA model parameters are learned from a disjoint set of training data, and the trained system forwarded to deal with matching examples of novel classes. As the Equations 13 and 14 are implying, the uncertainty in the training data are employed in ULDA to extend the expected between-class and within-class scatters. In standard PLDA, the $\mathbf{F}$ and $\mathbf{G}$ matrices are initialized randomly and iteratively refined by an expectation maximization algorithm. However, the uncertainty of the input space is not considered in the training. When the number of available representative points per class is limited, by employing ULDA, the estimated discriminant directions can resemble the optimal discriminant projection better than conventional LDA and standard PLDA. A demonstration of this comparison is provided in Figure 2 where data points are sampled from two Gaussian distributions representing two classes. The uncertainties shown as ellipses are drawn from a Wishart distribution. In plotting PLDA directions, the estimated $\mathbf{F}$ and $\mathbf{G}$ matrices are treated as between-class and within-class scatters.

5 Experiments for Automatic Speech Recognition

For automatic speech recognition, *hidden Markov models* have proven to be highly successful [38], [39]. In HMM-based speech recognition, all relevant phonetic or subphonetic units are described by left-to-right Markov models, which describe the evolution of acoustic observations within the phonetic unit at hand. For example, HMMs might be used to model phones, syllables or entire words, depending on the complexity of the recognition vocabulary and other constraints of the application.

An HMM consists of a set of states q, each with an associated output probability distribution $p(\mathbf{o}_t = \mathbf{x}_l | q_t = i)$ describing observation vector $\mathbf{o}_t$ at time index t, and of a transition matrix $\mathbf{A}$, where the element in the j-th column of row i describes the probability of moving from state i to j, so $[\mathbf{A}]_{i,j} = a_{i,j} = p(q_{t+1} = j | q_t = i)$. HMM-based speech recognition then amounts to the task of finding the best admissible sequence of HMM-states, where a state sequence is admissible if it passes through a sequence of words that is allowed by the task grammar. Thus, this best state sequence will also correspond to the most likely admissible sequence of words, which is then deemed recognized. If necessary, an additional statistical grammar can re-weight word sequences according to their likelihood of occurrence, but this was not needed for the currently considered task, which has a deterministic grammar as shown in Figure 3.

Within the word models, output probability distributions, describing the probability of making a certain observation $\mathbf{x}_l$ in state i, $p(\mathbf{o}_t = \mathbf{x}_l | q_t = i)$, are typically Gaussian mixture models (GMMs). By dropping state index i for the sake of simplicity, the GMM is expressed as

$$\begin{aligned} b_q(\mathbf{x}_l) = p(\mathbf{x}_l|q) &= \sum_{m=1}^{M} w_{qm} \cdot b_{qm}(\mathbf{x}_l) \\ &= \sum_{m=1}^{M} w_{qm} \mathcal{N}\left(\mathbf{x}_l; \boldsymbol{\mu}_{qm}, \boldsymbol{\Sigma}_{qm}\right), \end{aligned} \tag{16}$$

where w_{qm}, $\boldsymbol{\mu}_{qm}$ and $\boldsymbol{\Sigma}_{qm}$ are the parameters (weight, mean and covariance, respectively) of the m^{th} Gaussian mixture component of the acoustic model ($1 \leq m \leq M$). The Gaussian component densities are evaluated according to

$$\mathcal{N}\left(\mathbf{x}_l; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \frac{1}{\sqrt{(2\pi)^d \left|\boldsymbol{\Sigma}\right|}} \mathrm{e}^{\left(-\frac{1}{2}(\mathbf{x}_l-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_l-\boldsymbol{\mu})\right)}. \tag{17}$$

In order to take into consideration the uncertainties of the feature vectors, many techniques have been proposed [40]. One of the best-known of these *uncertainty-of-observation*-methods is uncertainty decoding (UD). This method replaces the likelihood at each state by the expected likelihood according to [41]

$$\begin{aligned} b_q(\underline{\mathbf{x}}_l)^{\mathrm{UD}} &= \mathbb{E}(b_q(\underline{\mathbf{x}}_l) \\ &= \sum_{m=1}^{M} w_{qm} \cdot \mathcal{N}\left(\boldsymbol{\mu}_l; \boldsymbol{\mu}_{qm}, \boldsymbol{\Sigma}_{qm} + \boldsymbol{\Sigma}_l\right). \end{aligned} \tag{18}$$

This is equivalent to adding the variance of the feature posterior to the variance of the currently considered state output probability distribution.

An alternative uncertainty-of-observation-method, termed modified imputation (MI) [42], in effect splits the likelihood computation into two steps: In the first step, the most likely value of the hidden variable $\mathbf{x}_l^*$ for the m-th Gaussian of state q is found via

$$\hat{x}_{qml}^{\mathrm{MI}} = (\boldsymbol{\Sigma}_l^{-1} + \boldsymbol{\Sigma}_{qm}^{-1})^{-1}(\boldsymbol{\Sigma}_{qm}^{-1}\boldsymbol{\mu}_{qm} + \boldsymbol{\Sigma}_l^{-1}\boldsymbol{\mu}_l). \tag{19}$$

In the second step, this estimate $\hat{\mathbf{x}}_{qml}^{\mathrm{MI}}$ is used in the likelihood computation of state q by

$$b_q(\underline{\mathbf{x}}_l)^{\mathrm{MI}} = \sum_{m=1}^{M} w_{qm} \cdot \mathcal{N}\left(\hat{\mathbf{x}}_{qml}^{\mathrm{MI}}; \boldsymbol{\mu}_{qm}, \boldsymbol{\Sigma}_{qm}\right). \tag{20}$$

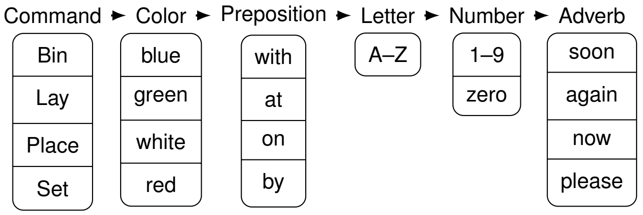

Fig. 3: Task grammar defined in the CHiME speech recognition challenge [43]. An example sentence would be *Lay blue at A five please.*

5.1 Experimental Setup

Experiments were carried out using the CHiME multi-channel robust ASR task [46]. This task simulates human-machine interaction in home environments using a small set of commands. The training data is binaural and reverberant. The test data is also reverberant and corrupted with common household noises at SNRs of $-6dB$, $-3dB$, $0dB$, $3dB$, $6dB$ and $9dB$. The noises have different directions of arrival while the speaker is situated in front of the microphones. Since we are dealing with small vocabularies in this task, word models were used, so we trained one HMM for each of the 50 words in the vocabulary as shown in Figure 3.

As the multi-channel pre-processing stage, a delay-and-sum beamformer with a Wiener filter was used. The noise was estimated from a fixed blocking matrix, nulling the broadside direction [31]. To further improve performance, an uncertainty-propagation-based MMSE-MFCC estimator [25], [47] was used for the feature extraction. Cepstral mean subtraction was applied as a final pre-processing stage. The posterior distribution associated to a Wiener filter in the short-time Fourier transform (STFT) domain was used as a measure of uncertainty [31]. This uncertainty is then transformed into the feature domain using STFT uncertainty propagation (STFT-UP) [25], yielding the desired uncertain feature description.

The hidden Markov model toolkit (HTK) [48] is used to implement the ASR experiments. The training and test scripts were provided for the CHiME challenge [46], and these were also used for the presented experiments. Multiconditional training [30] is employed to train the models. Speaker-dependent HMMs are estimated using a set of trained speaker-independent HMMs as the starting point. A modified version of HTK, extended to allow for likelihood computation by modified imputation or uncertainty decoding, was used for recognition.

As indicated in Figure 4, the new ULDA dimensionality reduction can be applied either alone, or jointly with one of the above uncertainty-of-observation techniques. The original feature space composed of 12 MFCCs appended with Δ and $\Delta\Delta$ resulting in 39-dimensional features. We used LDA and ULDA to transform the features to a 36-dimensional space. The associated uncertainty for each feature vector is also passed through either the

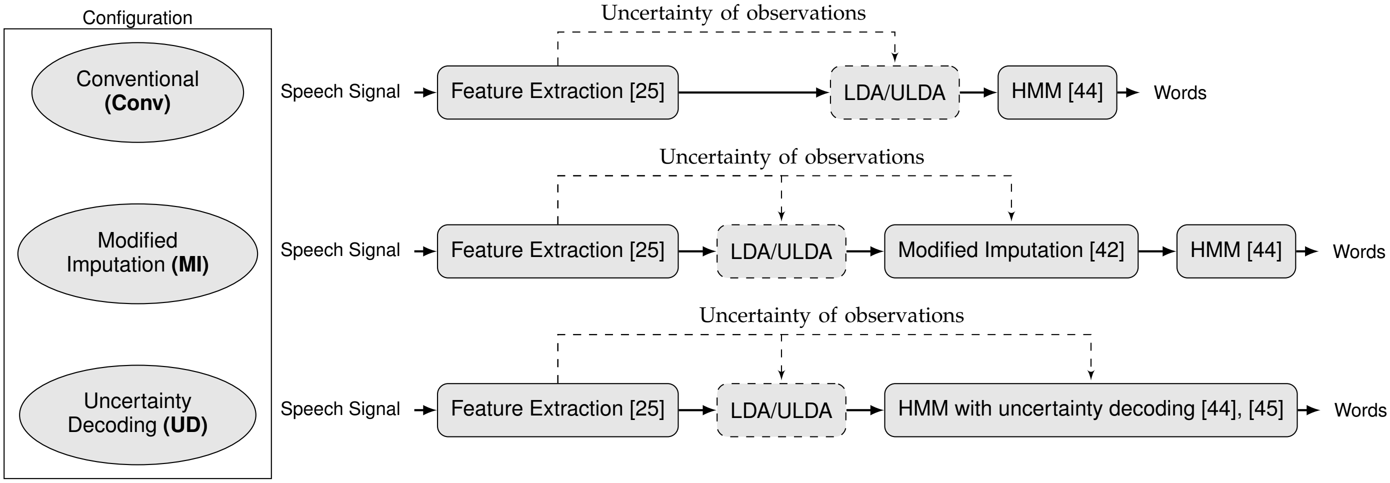

Fig. 4: A block diagram of a typical state-of-the-art automatic speech recognition (ASR) system. Three flows are indicating three different configurations for employing uncertainty of observations in ASR. HMM stands for hidden Markov model and the abbreviations Conv, MI and UD are used in Table 1 to report word recognition error rates for each configuration using LDA or ULDA for dimensionality reduction.

LDA or the ULDA transform before using uncertainty decoding or modified imputation in the test stage [49]. In line with the HTK models, we use 250 speech states (4–10 states per word) to label speech basis atoms [50]. In estimating the LDA and ULDA transforms, we associate the training acoustic features of each speaker to 250 classes to arrive at speaker-dependent discriminative transforms.

5.2 Experimental Results

Tab. 1 shows the keyword recognition error rates for the test set. The keywords are the letter and the number in each utterance.

	Exact LDA			Uncertain LDA		
SNR	Conv	UD	MI	Conv	UD	MI
-6	31.0	30.8	29.7	31.7	31.7	30.4
-3	24.8	24.8	24.8	24.6	23.8	23.6
0	17.9	17.7	17.4	17.0	16.1	16.0
3	13.5	13.5	13.0	13.0	12.2	12.2
6	8.4	8.0	8.5	8.2	7.9	8.2
9	8.6	8.5	8.1	8.8	8.5	8.6
Average	17.4	17.2	16.9	17.2	16.7	16.5
Improvement	0%	1%	3%	1%	4%	5%

TABLE 1: Keyword recognition error rate in percent for a range of approaches, computed on the CHiME [46] test set. The configurations conventional (Conv), modified imputation (MI) and uncertainty decoding (UD) are explained in Figure 4.

Using the conventional configuration of ASR, ULDA brings 1% improvement over using LDA (average recognition error rate of 17.2% compared to 17.4%). Employing conventional LDA, the application of uncertainty decoding and modified imputation improves the recognition performance by 1% and 3%, respectively. Applying uncertain LDA in conjunction with uncertainty decoding and modified imputation further reduces the average recognition error of 17.2% to 16.7% and 16.5%, respectively.

6 Experiments for Automatic Speaker Verification

Automatic speaker verification, the task of accepting or rejecting an identity claim given an utterance of a speaker, has received lots of attention in the last 20 years [55]. One of the main reasons is the support of the National Institute of Standards and Technology (NIST) by organizing a series of benchmarks, the speaker recognition evaluations (SREs) [56] starting in 1996. For each SRE, the task, the data and the evaluation metrics are supplied by NIST and after submission of recognition scores by participating sites, researchers share thoughts in a follow up workshop. Before NIST SRE'12, the task was solely defined as *speaker detection*, whereas in SRE'12, the performance metric and evaluation condition resembles an *open set* speaker recognition task [57], [58].

Most of the research in the speaker recognition area was devoted to finding robust modeling techniques, capable of handling channel and inter-session variability [32], [53], [59]. The state-of-the-art method is now using a low-rank vector, the so-called *i-vector*, to represent an utterance based on *total variability* subspace modeling [32] and *probabilistic linear discriminant analysis* (PLDA) [36] to obtain a likelihood ratio in comparing an enrolment and test utterance. The acoustic feature distribution is captured by a *universal background model* (UBM) [52] and the subspace modeling techniques developed in the joint factor analysis approach [60] are utilized. A schematic block diagram of the state-of-the-art speaker recognition system as used for experiments in this paper is shown in Figure 5.

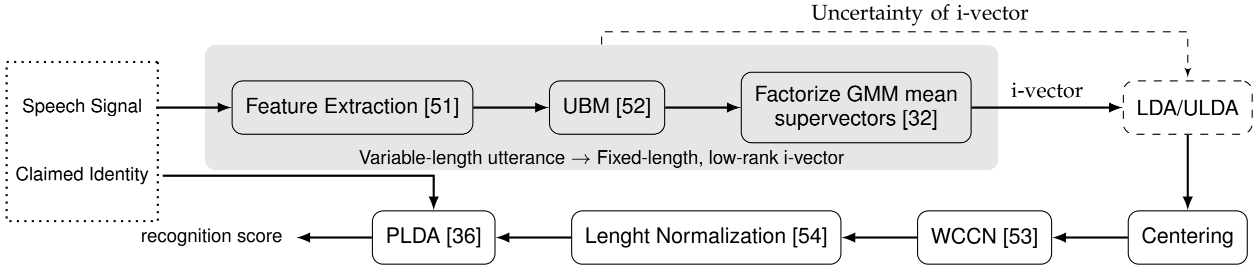

Fig. 5: Schematic block diagram of a typical state-of-the-art speaker verification system [51], [55]. Abbreviations stand for; UBM: universal background model, WCCN: within-class covariance normalization and PLDA: probabilistic LDA.

The uncertainty in i-vector extraction is a result of several factors, such as having noisy or very short speech signals, to name a few important ones. The effect of noise on acoustic features can be modeled by an uncertainty of short-time features and then propagated to the following modeling stages [61]–[63]. The problem of *incomplete observations* in speaker recognition is posed as dealing with variable (short) utterance duration and several techniques have been proposed recently to compensate for this factor in the context of i-vector extraction [33], [34], [64]–[67]. The *i-vectors* extracted using a sufficient amount of speech follow a standard normal distribution. However, as studied recently in [33], [34], this is not the case any more when a considerable amount of uncertainty exists in the i-vector extraction. Considering utterance duration as the source of uncertainty in i-vectors, we assume that a posterior distribution of i-vectors is provided. Next, we examine the speaker verification system to assess the system performance using conventional and uncertain LDA. We use multi-condition training in finding LDA and PLDA parameters with multiple durations (truncated versions) of the same utterance [58], [64].

6.1 Uncertainty Estimation

By dropping the state index in Equation 16, we can explain the UBM as $\sum_{m=1}^{M} w_m \mathcal{N}\left(\mathbf{x}_l; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right)$. Let's show the stacked $\boldsymbol{\mu}_m$s as the mean supervector $\mathbf{u}$ and form $\tilde{\boldsymbol{\Sigma}}$ as a block diagonal matrix with $\boldsymbol{\Sigma}_m$s as its entries. The factor analysis approach [32], [68] represents an utterance with respective acoustic features denoted by $\mathcal{X}$ by a location in high-dimensional space of $\mathbf{u} + \mathbf{T}\phi$. The rectangular matrix $\mathbf{T}$ characterize the low-rank subspace including inter- and intra-speaker variabilities in supervector space and ϕ is the i-vector as a realization of latent variable $\boldsymbol{\Phi}$. The zero- and first-order statistics $N_{\mathcal{X}}^{(m)}$ and $\mathbf{f}_{\mathcal{X}}^{(m)}$ for an utterance can be calculated with respect to m-th component of UBM and as it is shown in [33], [34], [68], [69], assuming a normal distribution for the i-vector $\phi \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\phi}}, \boldsymbol{\Sigma}_{\boldsymbol{\phi}})$, the point estimate $\boldsymbol{\mu}_{\boldsymbol{\phi}}$ and related uncertainty $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}$ is given by

$$\boldsymbol{\mu}_{\boldsymbol{\phi}} = \boldsymbol{\Sigma}_{\boldsymbol{\phi}} \mathbf{T}^T \tilde{\boldsymbol{\Sigma}}^{-1} \mathbf{f}_{\mathcal{X}}, \tag{21}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\phi}} = (\mathbf{I} + \sum_{m=1}^{M} N_{\mathcal{X}}^{(m)} \mathbf{T}^{(m)^T} \boldsymbol{\Sigma}_m^{-1} \mathbf{T}^{(m)})^{-1}, \tag{22}$$

respectively. In the Equations 21 and 22, $\mathbf{f}_{\mathcal{X}}$ is formed by concatenating $\mathbf{f}_{\mathcal{X}}^{(m)}$s and $\mathbf{T}^{(m)}$ is the m-th sub-matrix of $\mathbf{T}$ corresponding to the m-th UBM component. In the following experiments, we employ estimated uncertainties for i-vectors as in Equation 22 for finding ULDA directions. We carried out uncertainty propagation through i-vector post-processing steps and included uncertainty decoding in PLDA [33], [34], [49], [69], [70].

TABLE 2: Number of speakers, speech segments and trials in the modified I4U file list for male speakers in DEV set.

Number of speakers		Number of segments		Number of trials	
Train	Test	Train	Test	True	False
680	828	5475	6501	4801	4415879

6.2 Experimental Setup

In feature extraction stage, 19 Mel-frequency cepstral coefficients are extracted from frames of 30 ms windowed speech every 15 ms, appended with the frame energy and concatenated with delta and delta-delta coefficients, resulting in 60-dimensional feature vectors [51], [71]. The *speech activity detector* in [72] is employed to discard non-speech frames and *Quantile-based cepstral dynamics normalization* [73] with low-pass temporal filtering adopted from RASTA [74] is applied on the final features.

A gender-dependent *universal background model* (UBM) with 1024 components is trained using a subset of NIST SRE 2004–2006, Switchboard I and II, Switchboard cellular phase 1 and 2, Callfriend and Fisher English corpora. We train our i-vector extractor with 450-dimensions using the same data as for the UBMs excluding the Callfriend data. In post-processing the utterance-level i-vectors, we use an LDA or ULDA projection to enhance separability of classes (speakers) and to reduce the i-vector dimension to 200. Prior to PLDA modeling, we remove the mean, perform whitening using within-class covariance normalization (WCCN) [53] and normalize the length of i-vectors to lie on unit sphere [54]. The enrolment i-vectors are not truncated and averaged over

TABLE 3: Speaker recognition results for uncertainty caused by duration. Equal error rate ($E_=$ in percent) and minimum of detection cost function (C_{det} in SRE'10 shown here as $C_{det}^{\min}$) are used to report the performance.

	Conventional LDA				Uncertain LDA			
Test segment	PLDA scoring		+ Uncertainty decoding		PLDA scoring		+ Uncertainty decoding	
duration	$E_=$(%)	$C_{det}^{\min}$	$E_=$(%)	$C_{det}^{\min}$	$E_=$(%)	$C_{det}^{\min}$	$E_=$(%)	$C_{det}^{\min}$
Full	1.42	0.046	1.42	0.045	1.31	0.043	1.29	0.42
40 sec	1.51	0.050	1.54	0.050	1.42	0.047	1.41	0.44
20 sec	1.79	0.058	1.82	0.059	1.67	0.056	1.62	0.054
10 sec	2.36	0.079	2.31	0.077	2.29	0.075	2.17	0.072
5 sec	3.51	0.117	3.47	0.111	3.59	0.114	3.31	0.105
Pooled	2.42	0.080	2.17	0.070	2.35	0.077	2.02	0.066
Improvement	0%	0%	10.3%	12.5%	2.9%	3.7%	16.5%	17.5%
Improvement			0%	0%			6.9%	5.7%

multiple i-vectors per speaker. The truncation is applied by using the first N features of each test segment to produce effective durations of $5, 10, 20$ and 40 seconds. We employed the file lists of I4U developed in preparations during the NIST SRE'12 evaluation period [51]. The file lists are modified by excluding the segments that are less than 40 seconds long, in order to accommodate the truncation experiments in this paper. For the sake of tractability, the experiments in this paper are performed only on male speakers. The i-vector extractor and subsequent transformations are trained using 25441 utterances from 3193 speakers. In *multicondition* training scheme, the truncated version of utterances are utilized in training stage as well. The number of speakers, speech segments and recognition trials for each of the *test segment duration* conditions (reported in Table 3) is provided in Table 2.

6.3 Experimental Results

The results for speaker verification are presented in terms of equal error rate ($E_=$) and minimum of detection cost function $C_{det}^{\min}$. The detection cost function C_{det} is computed using

$$C_{det} = C_{miss} \times P_{tar} \times P_{miss} + C_{fa} \times (1 - P_{tar}) \times P_{fa} \quad (23)$$

with $C_{miss} = C_{fa} = 1$ and $P_{tar} = 1/1000$ as used in NIST SRE'10 [56]. In Equation 23, C_{miss}, C_{fa} and P_{tar} stand for cost of a miss, false alarm and prior probability of a target trial, respectively. The prior probability of a target trial is set according to the likelihood of target speaker presence in the system evaluation phase. The equal error rate ($E_=$) is that point on the *receiver operating characteristic* (ROC) curve where the probabilities of missed detection P_{miss} and false alarm P_{fa} become equal. The C_{det} and $E_=$ values are computed using the BOSARIS[1] toolkit via Bayes error rate computation.

In Table 3, we summarize the results for the experiments on speaker recognition. The results are presented for using LDA or ULDA for i-vectors post-processing and to demonstrate the gain attained by employing uncertainty decoding in PLDA scoring. By using ULDA instead of LDA, and by employing uncertainties in finding the discriminative dimensions, a relative performance improvement of 2.9% in $E_=$ and 3.7% in $C_{det}^{\min}$ is achieved. A considerable gain of 10.3% in $E_=$ and 12.5% in $C_{det}^{\min}$ is attained by using uncertainty decoding. By using uncertainties both in ULDA and in PLDA scoring, the performance gain is extended to 16.5% in $E_=$ and 17.5% in $C_{det}^{\min}$. In the uncertainty decoding scheme, employing ULDA provides improvements of 6.9% in $E_=$ and 5.7% in $C_{det}^{\min}$, respectively.

7 Discussion

The experiments show improvements in all considered cases, both for automatic speaker recognition and for speech recognition. In the case of speech recognition, the improvements are observed specifically for matched training and testing, which means that it is advantageous to take into account observation uncertainties of training data in deriving the optimal linear feature transform. This is likely due to the algorithm's added capability to disregard some of the noise-related components of the covariance matrices. That robustness is achieved by working on expected covariances, rather than on the direct covariance estimates that tend to be even more affected by noise than the first-order moments.

The methods proposed in this paper to find expected scatter matrices can be directly applied in principal component analysis (PCA) to find an *uncertain PCA* transform. The conventional PCA can be viewed as finding the total variability $\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W$ and solving the eigenvalue decomposition problem $(\mathbf{S}_T - \lambda_i \mathbf{I})\hat{\mathbf{w}}_i = 0$ subject to orthonormality of $\hat{\mathbf{w}}_i$s. In uncertain PCA, an expected total variability $\widehat{\mathbf{S}_T} = \widehat{\mathbf{S}_B} + \widehat{\mathbf{S}_W}$ is used to find eigenvectors representing the directions of largest variance irrespective of class labels.

Additionally, we have recently proposed a noise-adaptive LDA (NALDA) [75], a computationally less expensive approximation to the predictive heteroscedastic LDA (HLDA) approach described in [76]. NALDA can take the uncertainty of the acoustic features into account in the feature extraction process. This is achieved by

[1] Software is available at https://sites.google.com/site/bosaristoolkit/

adapting the LDA transform *online* in such a way as to achieve maximum discrimination under observation uncertainty. The approach was developed under the assumption of the source of uncertainty being additive noise in the feature domain, and it requires the recognition model parameters to be updated for every feature vector. While recognition performance is improved notably, NALDA is computationally rather demanding and it cannot be applied in those cases, where other feature post-processing needs to be applied after LDA. The results presented in Table 1 are directly comparable with the ones presented in [75].

8 Conclusion

We have presented a new approach for coping with observation uncertainties in deriving an optimal linear discriminant feature transform. This so-termed uncertain LDA takes into account the probabilistic description of observations in finding the most discriminant directions. We assumed a Gaussian distribution for uncertain observations and performed experiments with real-world data for two challenging problems in speech technology. Our experiments indicate that by employing an appropriate uncertainty definition and, correspondingly, a reliable uncertainty estimator, the state-of-the-art solutions for automatic speech and speaker recognition can be improved further when equipped with uncertain LDA. Since the uncertainty of measurements is a common problem in many pattern recognition problems, and the uncertain LDA framework is derived in a general manner, we believe that other researchers working in pattern recognition and machine vision will find the application of uncertain LDA advantageous in their respective applications as well.

Acknowledgment

This work was supported in part by Academy of Finland (project numbers 256961 and 284671) and Portuguese Foundation for Science and Technology (grant number SFRH/BPD/68428/2010 and project PEst-OE/EEI/LA0021/2013). We acknowledge the computational resources provided by the Aalto Science-IT project.

References

[1] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[2] T. Hastie, R. Tibshirani, and J. J. H. Friedman. *The elements of statistical learning*, volume 1. Springer New York, 2001.

[3] C. R. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203, 1948.

[4] R. Haeb-Umbach and H. Ney. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *Proc. ICASSP*, volume 1, pages 13 –16, 1992.

[5] N. Kumar and A. G. Andreou. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech communication*, 26(4):283–297, 1998.

[6] M. J. Gales. Semi-tied covariance matrices for hidden Markov models. *Speech and Audio Processing, IEEE Transactions on*, 7(3):272–281, 1999.

[7] L. Burget, P. Matějka, P. Schwarz, O. Glembek, and J. Černocký. Analysis of feature extraction and channel compensation in a GMM speaker recognition system. *IEEE Trans. Audio, Speech and Language Processing*, 15(7):1979–1986, September 2007.

[8] S. J. Raudys and A. K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on pattern analysis and machine intelligence*, 13(3):252–264, 1991.

[9] T. Hastie, R. Tibshirani, and J. J. H. Friedman. *The elements of statistical learning*, volume 10. Springer New York, 2013.

[10] J. H. Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989.

[11] S. Srivastava, M. R. Gupta, and B. A. Frigyik. Bayesian quadratic discriminant analysis. *Journal of Machine Learning Research*, 8(6):1277–1305, 2007.

[12] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural computation*, 12(10):2385–2404, 2000.

[13] M. Zhu and A. M. Martinez. Subclass discriminant analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(8):1274–1286, 2006.

[14] M. Loog and R. Duin. Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(6):732–739, 2004.

[15] S. Srivastava, M. R. Gupta, and B. A. Frigyik. Bayesian quadratic discriminant analysis. *Journal of Machine Learning Research*, 8(6):1277–1305, 2007.

[16] T. Hastie and R. Tibshirani. Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 155–176, 1996.

[17] H. Park, H. Zha, and H. Kim. *Distance Preserving Dimension Reduction for Manifold Learning*, chapter 56, pages 527–532.

[18] K. Fukunaga. *Introduction to statistical pattern recognition*. Access Online via Elsevier, 1990.

[19] M. Loog, R. P. W. Duin, and R. Haeb-Umbach. Multiclass linear dimension reduction by weighted pairwise Fisher criteria. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(7):762–766, 2001.

[20] C. Chen. On information and distance measures, error bounds, and feature selection. *Information Sciences*, 10(2):159–173, 1976.

[21] J. Peng and G. Seetharaman. Chernoff distance and relief feature selection. In *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, pages 3493–3496. IEEE, 2012.

[22] M. S. Mahanta and K. N. Plataniotis. A heteroscedastic extension of LDA based on multi-class matusita affinity. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, pages 1921–1924, 2012.

[23] Y. Li, J. Chen, and L. Feng. Dealing with uncertainty: A survey of theories and practices. *Knowledge and Data Engineering, IEEE Transactions on*, 25(11):2463–2482, Nov 2013.

[24] J. Droppo, A. Acero, and L. Deng. Uncertainty decoding with splice for noise robust speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 57–60, 2002.

[25] R. F. Astudillo and R. Orglmeister. A MMSE estimator in mel-cepstral domain for robust large vocabulary automatic speech recognition using uncertainty propagation. In *Proc. Interspeech*, pages 713–716, 2010.

[26] H. Liao and M. Gales. Adaptive training with joint uncertainty decoding for robust recognition of noisy data. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, volume 4, pages 389–392, April 2007.

[27] D. Kolossa and R. Haeb-Umbach, editors. *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*. Springer, 2011.

[28] R. F. Astudillo, D. Kolossa, and R. Orglmeister. Accounting for the uncertainty of speech estimates in the complex domain for minimum mean square error speech enhancement. In *Proc. Interspeech*, pages 2491–2494, 2009.

[29] S. Cumani, O. Plchot, and P. Laface. On the use of i-vector posterior distributions in probabilistic linear discriminant analysis. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(4):846–857, April 2014.

[30] T. Virtanen, R. Singh, and B. Raj. *Techniques for Noise Robustness in Automatic Speech Recognition*. Wiley, 2012.

[31] R. F. Astudillo, et al. Integration of beamforming and uncertainty-of-observation techniques for robust ASR in multi-source environments. *Computer Speech and Language*, 27(3):837 – 850, 2013.

[32] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Trans. Audio, Speech and Language Processing*, 19(4):788–798, May 2011.

[33] P. Kenny, T. Stafylakis, P. Ouellet, J. Alam, and P. Dumouchel. PLDA for speaker verification with utterances of arbitrary duration. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, 2013.

[34] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P. Dumouchel. Text-dependent speaker recognition using PLDA with uncertainty propagation. In *Proc. Interspeech 2013*, 2013.

[35] S. Ioffe. Probabilistic linear discriminant analysis. In *Computer Vision – ECCV 2006*, volume 3954 of *Lecture Notes in Computer Science*, pages 531–542. Springer Berlin Heidelberg, 2006.

[36] S. J. D. Prince and J. H. Elder. Probabilistic linear discriminant analysis for inferences about identity. In *11th International Conference on Computer Vision*, pages 1–8, 2007.

[37] P. Kenny. Bayesian speaker verification with heavy-tailed priors. In *Proc. Odyssey: the Speaker and Language Recognition Workshop (Odyssey 2010)*, Brno, Czech Republic, June 2010.

[38] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989.

[39] M. Gales and S. Young. The application of hidden Markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3):195–304, 2008.

[40] R. Haeb-Umbach. *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*, chapter Uncertainty Decoding and Conditional Bayesian Estimation. Springer, 2011.

[41] N. Yoma and M. Villar. Speaker verification in noise using a stochastic version of the weighted Viterbi algorithm. *IEEE Trans. Speech, Audio Processing*, 10 (3):158–166, 2002.

[42] D. Kolossa, A. Klimas, and R. Orglmeister. Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques. In *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 82–85, Oct. 2005.

[43] J. P. Barker, E. V. N. Ma, H. Christensen, and P. D. Green. The PASCAL CHiME speech separation and recognition challenge. *Computer Speech and Language*, 2013.

[44] D. Kolossa, S. Zeiler, A. Vorwerk, and R. Orglmeister. Audiovisual speech recognition with missing or unreliable data. In *Proc. AVSP*, 2009.

[45] L. Deng, J. Droppo, and A. Acero. Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion. *IEEE Trans. Speech and Audio Processing*, 13(3):412–421, May 2005.

[46] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni. The second chime speech separation and recognition challenge: Datasets, tasks and baselines. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, pages 126–130, May 2013.

[47] R. F. Astudillo. *Integration of Short-Time Fourier Domain Speech Enhancement and Observation Uncertainty Techniques for Robust Automatic Speech Recognition*. PhD thesis, TU Berlin, 2010.

[48] S. Young. *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department., 2006.

[49] R. Saeidi and P. Alku. Accounting for uncertainty of i-vectors in speaker recognition using uncertainty propagation and modified imputation. In *Proc. Interspeech 2015*, 2015.

[50] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen. Modelling non-stationary noise with spectral factorisation in automatic speech recognition. *Computer Speech and Language*, 27(3):763–779, 2013.

[51] R. Saeidi, et al. I4U submission to NIST SRE 2012: a large-scale collaborative effort for noise-robust speaker verification. In *INTERSPEECH*, pages 1986–1990, 2013.

[52] D. Reynolds, T. Quatieri, and R. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1):19–41, January 2000.

[53] A. O. Hatch, S. Kajarekar, and A. Stolcke. Within-class covariance normalization for SVM-based speaker recognition. In *Proc. Interspeech 2006 (ICSLP)*, pages 1471–1474, Pittsburgh, Pennsylvania, USA, September 2006.

[54] D. Garcia-Romero and C. Y. Espy-Wilson. Analysis of i-vector length normalization in speaker recognition systems. In *Proc. Interspeech 2011*, pages 249–252, 2011.

[55] T. Kinnunen and H. Li. An overview of text-independent speaker recognition: from features to supervectors. *Speech Comm.*, 52(1):12–40, January 2010.

[56] NIST speaker recognition evaluations. http://www.nist.gov/itl/iad/mig/sre.cfm.

[57] NIST 2012 SRE, October 2012. http://www.nist.gov/itl/iad/mig/sre12.cfm.

[58] D. A. van Leeuwen and R. Saeidi. Knowing the non-target speakers: The effect of the i-vector population for PLDA training in speaker recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, 2013.

[59] A. Solomonoff, W. Campbell, and I. Boardman. Advances in channel compensation for SVM speaker recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, pages 629–632, Philadelphia, USA, March 2005.

[60] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Speaker and session variability in GMM-based speaker verification. *IEEE Trans. Audio, Speech and Language Processing*, 15(4):1448–1460, May 2007.

[61] C. Yu, G. Liu, S. Hahm, and J. H. L. Hansen. Uncertainty propagation in front end factor analysis for noise robust speaker recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, 2014.

[62] Y. Shao, S. Srinivasan, and D. Wang. Incorporating auditory feature uncertainties in robust speaker identification. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, volume 4, pages IV–277, 2007.

[63] X. Zhao, Y. Wang, and D. Wang. Robust speaker identification in noisy and reverberant conditions. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(4):836–845, April 2014.

[64] T. Hasan, R. Saeidi, J. H. L. Hansen, and D. A. van Leeuwen. Duration mismatch compensation for i-vector based speaker recognition systems. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, 2013.

[65] V. Hautamäki, Y.-C. Cheng, P. Rajan, and C.-H. Lee. Minimax i-vector extractor for short duration speaker verification. In *Proc. Interspeech 2013*, 2013.

[66] A. Kanagasundaram, D. Dean, J. Gonzalez-Dominguez, S. Sridharan, D. Ramos, and J. Gonzalez-Rodriguez. Improving short utterance based i-vector speaker recognition using source and utterance-duration normalization techniques. In *Proc. Interspeech 2013*, pages 2465–2469, 2013.

[67] A. Kanagasundaram, D. Dean, S. Sridharan, J. Gonzalez-Dominguez, J. Gonzalez-Rodriguez, and D. Ramos. Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques. *Speech Communication*, 59(0):69 – 82, 2014.

[68] P. Kenny. Joint factor analysis of speaker and session variability: Theory and algorithms. *CRIM, Montreal,(Report) CRIM-06/08-13*, 2005.

[69] S. Cumani, O. Plchot, and P. Laface. Probabilistic linear discriminant analysis of i-vector posterior distributions. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, pages 7644–7648, May 2013.

[70] L. Lu and S. Renals. Probabilistic linear discriminant analysis for acoustic modeling. *IEEE Signal Processing Letters*, 21(6):702–706, June 2014.

[71] M. I. Mandasari, R. Saeidi, M. McLaren, and D. A. van Leeuwen. Quality measure functions for calibration of speaker recognition systems in various duration conditions. *IEEE Trans. Audio, Speech, and Language Processing*, 21:2425–2438, 2013.

[72] T. Kinnunen and P. Rajan. A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, pages 7229–7233, 2013.

[73] H. Boril and J. Hansen. Unsupervised equalization of lombard effect for speech recognition in noisy adverse environments. *IEEE Trans. Audio, Speech and Language Processing*, 18(6):1379–1393, Aug 2010.

[74] H. Hermansky and N. Morgan. RASTA processing of speech. *Speech and Audio Processing, IEEE Trans.*, 2(4):578–589, 1994.

[75] D. Kolossa, S. Zeiler, R. Saeidi, and R. Astudillo. Noise-adaptive LDA: A new approach for speech recognition under observation uncertainty. *IEEE Signal Processing Letters*, 20(11):1018–1021, 2013.

[76] R. van Dalen. *Statistical Models for Noise-Robust Speech Recognition (2011)*. PhD thesis, University of Cambridge, 2011.

Rahim Saeidi received the Ph.D. degree in computer science from the University of Eastern Finland (UEF, formerly Univ. of Joensuu) in 2011. He was a Marie Curie post-doctoral fellow in EU project Bayesian biometrics for forensics working at Radboud University Nijmegen in Netherlands from 2011 to 2013. Currently, he is a postdoctoral researcher at Department of Speech Processing and Acoustics, Aalto University, Finland. His research interests include robust speaker and speech recognition, machine learning and speech enhancement.

Ramón Fernandez Astudillo received the Dipl.-Ing. degree in electronics and automatics from the University of Oviedo, Oviedo, Spain, in 2005 and the Ph.D. degree with distinction from the Technische Universität Berlin, Berlin, Germany. In 2005, he worked on a research project at PeikerAcustic GmbH in the field of speech enhancement and speech reconstruction. In 2006, he received a DAAD and “La Caixa” Foundation scholarship for research towards the Ph.D. degree at the Electronics and Medical Signal Processing Group, Technische Universität Berlin. His research interests include pattern recognition, signal processing and Bayesian statistics.

Dorothea Kolossa received the Dipl.-Ing. degree in computer engineering and the Ph.D. degree in electrical engineering from Technische Universität Berlin, Germany, in 1999 and 2007, respectively. From 1999 until 2000, she worked on control systems design at DaimlerChrysler Research and Technology, Hennigsdorf, Germany. She was employed as a research assistant at Technische Universität Berlin from 2000 until 2004, and as a senior researcher from 2004 until 2010, also staying at UC Berkeleys Parlab as visiting faculty in 2009/2010. Since 2010, she has been working in a faculty position at the Institute of Communication Acoustics, Ruhr-Universität Bochum, Germany, where currently, she is heading the Cognitive Signal Processing group as an associate professor. Her research interests include speech recognition in adverse environments and robust classification methods for communication and technical diagnostics. She has authored and co-authored more than 80 scientific papers and book chapters. Dr. Kolossa is a member of the IEEE Audio and Acoustic Signal Processing Technical Committee.