Bayesian Speaker Verification with Heavy-Tailed Priors

Patrick Kenny

Centre de recherche informatique de Montréal (CRIM)
Patrick.Kenny@crim.ca

Abstract

We describe a new approach to speaker verification which, like Joint Factor Analysis, is based on a generative model of speaker and channel effects but differs from Joint Factor Analysis in several respects. Firstly, each utterance is represented by a low dimensional feature vector, rather than by a high dimensional set of Baum-Welch statistics. Secondly, heavy-tailed distributions are used in place of Gaussian distributions in formulating the model, so that the effect of outlying data is diminished, both in training the model and at recognition time. Thirdly, the likelihood ratio used for making verification decisions is calculated (using variational Bayes) in a way which is fully consistent with the modeling assumptions and the rules of probability. Finally, experimental results show that, in the case of telephone speech, these likelihood ratios do not need to be normalized in order to set a trial-independent threshold for verification decisions.

We report results on female speakers for several conditions in the NIST 2008 speaker recognition evaluation data, including microphone as well as telephone speech. As measured both by equal error rates and the minimum values of the NIST detection cost function, the results on telephone speech are about 30% better than we have achieved using Joint Factor Analysis.

1. Introduction

Because speech signals are time series of arbitrary duration, and because the order of events in these time series is largely irrelevant for speaker recognition, it has proved difficult to find feature representations which contain the information needed to distinguish between speakers and which are of sufficiently low dimension to be amenable to fully fledged Bayesian methods. A recent advance in this direction was reported in [1, 2], where a type of principal components analysis was used to represent a given speech utterance by a feature vector of low dimension (400), independent of the length of the utterance. A speaker verification system using these features and a simple classifier produced better results on the 2008 NIST speaker recognition evaluation (SRE) data than Joint Factor Analysis (JFA) [2, 3].

This is interesting in its own right but it is also interesting because working with low dimensions makes it possible to develop new methods to study the fundamental problem of speaker recognition, namely how to decompose speech data D into a speaker component S and a channel component C:

$$D \quad = \quad S + C. \tag{1}$$

In tackling this problem, the easiest assumptions to work with are that (i) S and C are statistically independent and (ii) S and C have Gaussian distributions. These are the assumptions underlying both JFA and, *mutatis mutandis*, the independently developed Probabilistic Linear Discriminant Analysis (PLDA) model for face recognition [4]. (PLDA can be viewed as a special case of JFA in which the universal background model has a single Gaussian component.)

Both assumptions are questionable. Channel effects are not speaker-independent (for example, it is well known that gender-dependent eigenchannel modeling is more effective than gender-independent modeling), but the problem of how to improve on the statistical independence assumption (i) seems to be a difficult one and we will not attempt to address it in this paper. As for the Gaussian assumption (ii), it effectively prohibits large deviations from the mean (sometimes referred to as "black swans") but it is clear that these occur both in the case of speaker effects (e.g. speakers whose native language is not English) and channel effects (e.g. gross channel distortions, particularly in the case of microphone speech). Adequate modeling of large deviations requires using heavy-tailed distributions such as Student's *t* distributions [5, 6]. (The Student's *t* distribution is defined in the Appendix; see also [7].)

We aim to show in this paper how to design a speaker recognition system using heavy-tailed assumptions to calculate likelihood ratios for speaker verification. Our approach is similar in spirit to [4] so it is appropriate to use the term *heavy-tailed PLDA* to refer to the model we propose. The feature vectors we use to represent speech segments are similar to the i-vectors in [2], modified to accommodate microphone as well as telephone speech [8]. Our main contributions will be to show that modeling S and C with Gaussian distributions gives results on the 2008 SRE data which are similar to those obtained in [2], and that using heavy-tailed distributions instead leads to substantial gains in speaker recognition accuracy on telephone speech, so that the best error rates we report are generally about 30% lower than we have achieved with JFA.

The multivariate Student's *t* distribution is a convenient choice for heavy tailed modeling as it easy to use variational Bayes to perform probability calculations with it and to estimate its parameters [5, 6]. The Student's *t* distribution is a power law distribution in the sense that density $P(x)$ has the property that there is a positive exponent k such that

$$P(x) \quad = \quad O(\|x\|^{-k})$$

as $\|x\| \rightarrow \infty$. (Contrast this with the exponential decrease at infinity of the Gaussian distribution.) The exponent k depends on a parameter (known as the number of degrees of freedom of the distribution) which can be estimated using the same likelihood criterion as the other parameters of the PLDA model. The smaller the exponent k, the heavier the tails of the distribution. In the extreme case where k is less than 1, none of the moments of the distribution exist; at the other extreme ($k \rightarrow \infty$), the Student's *t* distribution tends to the Gaussian distribution.

As the title of the paper indicates, we will take a Bayesian (that is, fully probabilistic) approach. This is facilitated by the fact that we are using a low dimensional feature representation of speech segments; it is much more difficult with JFA but

Bayesian treatments of JFA have yielded some positive results in speaker recognition [9, 10, 11] and speaker diarization [12].

One of the advantages of a Bayesian approach in speaker recognition is that, in principle, no particular effort is needed to design a classifier — all that is required is to follow the rules of probability consistently in calculating the likelihood ratios for making speaker verification decisions, in the same spirit as [4]. A perfect generative model (that is, one which can simulate speech data perfectly when driven by a random number generator), will produce likelihood ratios that do not need to be normalized or calibrated. This is an important consideration since the requirement for domain-dependent score normalization with zt-norm is a well known weakness of JFA. One of our most interesting findings concerning the model we develop here is that, in the case of telephone speech, there is little or no benefit to be derived from score normalization if S and C are assumed to have heavy-tailed distributions, but this is not the case if Gaussian distributions are assumed instead.

One of our principal motivations for exploring heavy-tailed distributions was to see if we could get a handle on non-Gaussian microphone effects in the NIST SRE interview data. It turns out however that when applied to this problem, Student's *t* modeling degenerates in an interesting way. Microphone effects turn out to be so non-Gaussian that when we attempt to model them with Student's *t* distributions, the variance turns out to be infinite so that arbitrarily large microphone distortions are "normal". This behavior causes speaker recognition to break down and the only way have been unable to avoid it is by artificially flooring the number of degrees of freedom in the Student's *t* distribution. This expedient enables us to produce speaker recognition results which are as good as, but no better than, those obtained under Gaussian assumptions. Thus we have not been able to produce an adequate probabilistic model of microphone effects and, at this writing, it seems that the best hope of achieving this may be to project away some of the troublesome dimensions in the i-vector space using some type of linear discriminant analysis (classical LDA or PLDA).

2. The generative model

We use F to denote the dimension of the feature vectors. We assume that we are given R recordings of a speaker and denote the corresponding feature vectors by $\{D_r : r = 1, \ldots, R\}$.

2.1. Gaussian priors

The model parameters are an $F \times 1$ mean vector m; a matrix U_1 of dimension $F \times N_1$ whose columns are referred to as eigenvoices; a matrix U_2 of dimension $F \times N_2$ whose columns are referred to as eigenchannels; and an $F \times F$ precision matrix Λ.

The generative model is

$$D_r \quad = \quad m + U_1 x_1 + U_2 x_{2r} + \epsilon_r \tag{2}$$

for $r = 1, \ldots, R$. Here x_1 is a vector having a standard normal distribution of dimension N_1; x_{2r} is a vector having a standard normal distribution of dimension N_2; and the residual ϵ_r is an F-dimensional vector having a normal distribution with mean 0 and precision matrix Λ. (We consistently use Greek letters to refer to the residual and its distribution.) See Fig. 1.

The dimensions N_1 and N_2 are the only model parameters that have to manually tuned. (We are following [4] here but Bayesian purists would avoid this by putting priors on N_1 and N_2. See the section on Bayesian PCA in [7].)

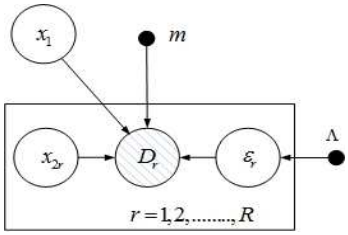

Figure 1: Probabilistic graphical model representing PLDA with Gaussian priors.

In the terminology of [4], x_1 is an "identity variable"; in the terminology of [3], the elements of x_1 are speaker factors and the elements of x_{2r} are channel factors. Referring to (1), we have

$$\begin{aligned} S &= m + U_1 x_1 \\ C_r &= U_2 x_{2r} + \epsilon_r \end{aligned}$$

where C_r is the channel component of D_r. The covariance matrix of S is $U_1 U_1^*$ and the covariance matrix of C_r is $\Lambda^{-1} + U_2 U_2^*$. It is assumed in [4] that Λ is diagonal, but the feature vectors that we use are of sufficiently low dimension that a full precision matrix can be robustly estimated, at least in the case of telephone speech (where large amounts of data are available). Thus the term $U_2 x_{2r}$ in (2) is not needed to model telephone speech (it adds no extra modeling capability) but we retain it because it is useful in modeling microphone speech. (Our strategy for modeling microphone speech is to estimate U_1 and Λ on telephone speech and U_2 on microphone speech. We will show that a model trained in this way gives very good results on both types of speech.)

2.2. Heavy-tailed priors

We leave the form of the model (2) unchanged but assume that the priors on x_1, x_{2r} and ϵ_r are Student's *t* rather than Gaussian. Thus we introduce scalar parameters n_1, n_2 and ν, referred to as numbers of degrees of freedom, and scalar valued hidden variables u_1, u_{2r} and υ_r and assume that,

$$\begin{aligned} &x_1 \sim \mathcal{N}(0, u_1^{-1} I) \text{ where } u_1 \sim \mathcal{G}(n_1/2, n_1/2) \\ &x_{2r} \sim \mathcal{N}(0, u_{2r}^{-1} I) \text{ where } u_{2r} \sim \mathcal{G}(n_2/2, n_2/2) \\ &\epsilon_r \sim \mathcal{N}(0, \upsilon_r^{-1} \Lambda^{-1}) \text{ where } \upsilon_r \sim \mathcal{G}(\nu/2, \nu/2) \end{aligned}$$

for $r = 1, \ldots, R$. Here $\mathcal{N}(\mu, \Sigma)$ indicates a Gaussian distribution with mean μ and covariance matrix Σ and $\mathcal{G}(a, b)$ indicates a Gamma distribution with parameters a and b (as defined in the Appendix). See Fig. 2.

2.3. Supplementary notation

It is convenient to introduce some extra notation. Let x_1^+ be the vector of dimension $(N_1 + 1) \times 1$ obtained by appending 1 to x_1 and similarly for x_{2r}^+. Let U_1^+ be the matrix of dimension $F \times (N_1 + 1)$ obtained by appending a column vector m_1 to U_1 and similarly for U_2^+. By imposing the condition that $m_1 + m_2 = m$, we can write the model as

$$D_r \quad = \quad U_1^+ x_1^+ + U_2^+ x_{2r}^+ + \epsilon_r \tag{3}$$

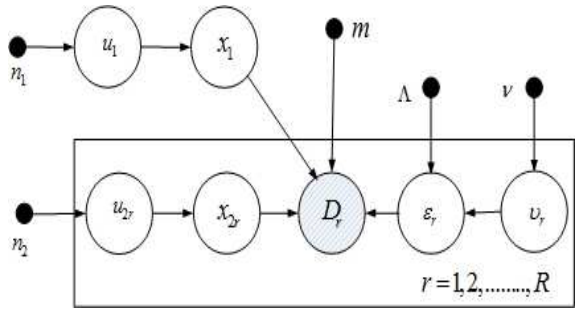

Figure 2: Probabilistic graphical model representing PLDA with Student's *t* priors.

for $r = 1, \ldots, R$. By setting

$$W = \begin{pmatrix} U_1^+ & U_2^+ \end{pmatrix} \quad (4)$$

$$z_r = \begin{pmatrix} x_1^+ \\ x_{2r}^+ \end{pmatrix} \quad (5)$$

we can write this as

$$D_r = Wz_r + \epsilon_r. \quad (6)$$

We use the following shorthands:

$$\begin{aligned} D &= \{D_r : r = 1, \ldots, R\} \\ x_2 &= \{x_{2r} : r = 1, \ldots, R\} \\ u_2 &= \{u_{2r} : r = 1, \ldots, R\} \\ \upsilon &= \{\upsilon_r : r = 1, \ldots, R\} \\ x &= (x_1, x_2) \\ u &= (u_1, u_2) \end{aligned}$$

and we set

$$h = (x, u, \upsilon)$$

so that h represents the entire collection of hidden variables associated with a speaker.

3. Posterior calculations

We continue to assume that we are dealing with a single speaker for which R observation vectors are available and suppose that we wish to calculate the posterior distribution of the hidden variables associated with the speaker. This is the principal calculation that needs to be performed in all phases of implementing the model.

Consider first the simplest case where $U_2^+ = 0$ (so that there are no channel factors) and x_1 has a Gaussian prior. (As we mentioned above, this is a realistic scenario in the case of telephone speech.) Using $\equiv$ to indicate equality up to an additive constant, the posterior distribution $P(x_1|D)$ is given by

$$\begin{aligned} &\ln P(x_1|D) \\ &\quad \equiv \ln P(D|x_1) + \ln P(x_1) \\ &\quad = \sum_{r=1}^{R} \ln N(D_r|U_1^+ x_1^+, \Lambda^{-1}) + \ln N(x_1|0, I) \\ &\quad \equiv -\frac{1}{2}\sum_{r=1}^{R}(D_r - U_1^+ x_1^+)^* \Lambda (D_r - U_1^+ x_1^+) - \frac{1}{2} x_1^* x_1. \end{aligned}$$

Since this expression is quadratic in x_1, the posterior is Gaussian and the posterior mean and covariance can be read off by collecting first and second order terms. Denoting the posterior mean and covariance by $\langle x_1 \rangle$ and $\mathrm{Cov}\,(x_1, x_1)$, this gives

$$\begin{aligned} \mathrm{Cov}\,(x_1, x_1) &= (RU_1^* \Lambda U_1 + I)^{-1} \\ \langle x_1 \rangle &= (RU_1^* \Lambda U_1 + I)^{-1} U_1^* \Lambda \sum_{r=1}^{R} (D_r - m_1). \end{aligned}$$

Thus the posterior calculation reduces to inverting a matrix of dimension $N_1 \times N_1$.

Next consider the case where $U_2^+ \neq 0$ and the prior of x_2 is also Gaussian. As explained in [4] and [13] (Section III-D), a direct calculation of the posterior of the hidden variables is still feasible in this case but it entails inverting an $R \times R$ block matrix. The complication here arises from the fact that, although the hidden variables x_1 and x_2 are independent in the prior, they are correlated in the posterior. This complication can be avoided by assuming a variational approximation of the form

$$\ln P(x_1, x_2|D) \approx \ln Q(x_1) + \ln Q(x_2)$$

and iteratively applying the standard variational Bayes update formulas [7]

$$\begin{aligned} \ln Q(x_2) &\equiv E_{x_1}\left[\ln P(D, x_1, x_2)\right] \\ \ln Q(x_1) &\equiv E_{x_2}\left[\ln P(D, x_1, x_2)\right]. \end{aligned}$$

Evaluating $Q(x_1)$ entails inverting an $N_1 \times N_1$ matrix and it turns out that $Q(x_2)$ factorizes as

$$\ln Q(x_2) = \sum_{r=1}^{R} \ln Q(x_{2r}).$$

Evaluating each term $Q(x_{2r})$ entails inverting a single $N_2 \times N_2$ matrix (common to all terms). So although it is not strictly necessary to appeal to variational Bayes in the Gaussian case, doing so greatly simplifies the calculations.

The posterior is intractable in the case where the priors on x_1 and x_2 are heavy-tailed rather than Gaussian and we have little choice but to use variational Bayes. We assume a variational approximation of the form.

$$\begin{aligned} \ln P(x, u, \upsilon|D) \approx\ & \ln Q(x_1) + \ln Q(x_2) \\ & + \ln Q(u_1) + \ln Q(u_2) + \ln Q(\upsilon). \end{aligned}$$

It turns out that, under this assumption, further factorizations follow, namely

$$\begin{aligned} \ln Q(x_2) &= \sum_{r=1}^{R} \ln Q(x_{2r}) \\ \ln Q(u_2) &= \sum_{r=1}^{R} \ln Q(u_{2r}) \\ \ln Q(\upsilon) &= \sum_{r=1}^{R} \ln Q(\upsilon_r). \end{aligned}$$

Furthermore, the variational posteriors $Q(u_1)$, $Q(u_{2r})$ and $Q(\upsilon_r)$ are Gamma distributions (just like the priors) and $Q(x_1)$ and $Q(x_{2r})$ are Gaussian *even in the case where the priors of x_1 and x_2 are assumed to be heavy-tailed.*

3.1. Updating the Gaussian distributions

The variational posterior distribution $Q(x_1)$ is a Gaussian distribution covariance matrix and mean vector given by

$$\begin{aligned}\mathrm{Cov}\,(x_1, x_1) &= \left(\langle u_1\rangle I + \sum_{r=1}^{R}\langle \upsilon_r\rangle U_1^*\Lambda U_1\right)^{-1}\\ \langle x_1\rangle &= \left(\langle u_1\rangle I + \sum_{r=1}^{R}\langle \upsilon_r\rangle U_1^*\Lambda U_1\right)^{-1}\\ &\quad\times\sum_{r=1}^{R}\langle \upsilon_r\rangle U_1^*\Lambda\left(D_r - m_1 - U_2^+\langle x_{2r}^+\rangle\right).\end{aligned}$$

For $r = 1, \ldots, R$, $Q(x_{2r})$ is a Gaussian distribution with covariance matrix and mean vector given by

$$\begin{aligned}\mathrm{Cov}\,(x_{2r}, x_{2r}) &= \left(\langle u_{2r}\rangle I + \langle \upsilon_r\rangle U_2^*\Lambda U_2\right)^{-1}\\ \langle x_{2r}\rangle &= \left(\langle u_{2r}\rangle I + \langle \upsilon_r\rangle U_2^*\Lambda U_2\right)^{-1}\\ &\quad\times\langle \upsilon_r\rangle U_2^*\Lambda\left(D_r - m_2 - U_1^+\langle x_1^+\rangle\right).\end{aligned}$$

The expectations $\langle u_1\rangle, \langle u_{2r}\rangle$ and $\langle \upsilon_r\rangle$ are calculated from the posteriors $Q(u_1), Q(u_{2r})$ and $Q(\upsilon_r)$ using the formula (13) in the Appendix. (Setting $\langle u_1\rangle = 1, \langle u_{2r}\rangle = 1$ and $\langle \upsilon_r\rangle = 1$ in these equations gives the update formulas for the case of Gaussian priors.)

3.2. Updating the Gamma distributions

The variational posterior distribution $Q(u_1)$ is a Gamma distribution with parameters a_1 and b_1 given by

$$\begin{aligned}a_1 &= \frac{n_1 + N_1}{2}\\ b_1 &= \frac{n_1 + \langle x_1^* x_1\rangle}{2}.\end{aligned}$$

For $r = 1, \ldots, R$, $Q(u_{2r})$ is a Gamma distribution with parameters a_{2r} and b_{2r} given by

$$\begin{aligned}a_{2r} &= \frac{n_2 + N_2}{2}\\ b_{2r} &= \frac{n_2 + \langle x_{2r}^* x_{2r}\rangle}{2},\end{aligned}$$

and $Q(\upsilon_r)$ is a Gamma distribution with parameters α_r and β_r with parameters α_r and β_r given by

$$\begin{aligned}\alpha_r &= \frac{\nu + F}{2}\\ \beta_r &= \frac{\nu + \langle \epsilon_r^*\Lambda\epsilon_r\rangle}{2}\end{aligned}$$

where the term $\langle \epsilon_r^*\Lambda\epsilon_r\rangle$ is calculated as follows. Recall that $\epsilon_r = D_r - Wz_r$ where z_r is defined by (5) so that

$$\langle Wz_r\rangle = U_1^+\langle x_1^+\rangle + U_2^+\langle x_{2r}^+\rangle$$

and

$$\begin{aligned}&\mathrm{Cov}\,(Wz_r, Wz_r)\\ &\quad= U_1\,\mathrm{Cov}\,(x_1, x_1)\,U_1^* + U_2\,\mathrm{Cov}\,(x_{2r}, x_{2r})\,U_2^*.\end{aligned}$$

Thus

$$\begin{aligned}\langle \epsilon_r^*\Lambda\epsilon_r\rangle &= \mathrm{tr}\left(\langle \Lambda\epsilon_r\epsilon_r^*\rangle\right)\\ &= \mathrm{tr}\left(\Lambda\,\mathrm{Cov}\,(\epsilon_r, \epsilon_r)\right) + \langle \epsilon_r^*\rangle\Lambda\langle \epsilon_r\rangle\end{aligned}$$

where

$$\begin{aligned}&\mathrm{tr}\left(\Lambda\,\mathrm{Cov}\,(\epsilon_r, \epsilon_r)\right)\\ &\quad= \mathrm{tr}\left(\Lambda\,\mathrm{Cov}\,(Wz_r, Wz_r)\right)\\ &\quad= \mathrm{tr}\left(U_1^*\Lambda U_1\,\mathrm{Cov}\,(x_1, x_1) + U_2^*\Lambda U_2\,\mathrm{Cov}\,(x_{2r}, x_{2r})\right)\end{aligned}$$

and

$$\langle \epsilon_r\rangle = D_r - U_1^+\langle x_1^+\rangle - U_2^+\langle x_{2r}^+\rangle.$$

4. Likelihood calculations

We continue to assume that we are dealing with a single speaker for which R observation vectors are available. Using h to denote the entire collection of hidden variables (x, u, υ) associated with the speaker, the likelihood $P(D)$ is calculated by marginalizing $P(D, h)$ with respect to h:

$$P(D) = \int P(D, h)\mathrm{d}h$$

and is known in Bayesian circles as the evidence. This integral is intractable in the general (heavy-tailed) case but if $\mathcal{L}$ is defined by

$$\mathcal{L} = E\left[\ln\frac{P(D, h)}{Q(h)}\right]$$

where the expectation is taken with respect to h, then $\mathcal{L} \leq \ln P(D)$ with equality holding iff $Q(h)$ is the exact posterior $P(h|D)$ [7]. We use the lower bound $\mathcal{L}$ as a proxy for $\ln P(D)$.

4.1. Evaluating $\mathcal{L}$

It is convenient to write

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$$

where

$$\begin{aligned}\mathcal{L}_1 &= E\left[\ln P(D|h)\right]\\ \mathcal{L}_2 &= -D\left(Q(h)\|P(h)\right)\end{aligned} \tag{7}$$

and $D\left(\cdot\|\cdot\right)$ denotes the Kullback-Leibler divergence. To evaluate $\mathcal{L}_1$, note that

$$\begin{aligned}&E\left[\ln P(D|x, u, \upsilon)\right]\\ &\quad= \sum_{r=1}^{R} E\left[\ln N(D_r|Wz_r, \upsilon_r^{-1}\Lambda^{-1})\right]\\ &\quad= \sum_{r=1}^{R}\Big(\frac{F}{2}\langle \ln\upsilon_r\rangle + \ln\frac{1}{(2\pi)^{F/2}|\Lambda^{-1}|^{1/2}}\\ &\qquad- \frac{1}{2}\langle \upsilon_r\rangle\langle \epsilon_r^*\Lambda\epsilon_r\rangle\Big).\end{aligned}$$

We explained how to evaluate the expression $\langle \epsilon_r^*\Lambda\epsilon_r\rangle$ in Section 3.2 and, since the posterior distribution of υ_r is a Gamma distribution, the quantities $\langle \upsilon_r\rangle$ and $\langle \ln\upsilon_r\rangle$ can be calculated using equations (13) and (14) in the Appendix.

To evaluate $\mathcal{L}_2$, we decompose the divergence $D\left(Q(x, u, \upsilon)\|P(x, u, \upsilon)\right)$ as

$$\begin{aligned}&D\left(Q(x_1, u_1)\|P(x_1, u_1)\right)\\ &+ \sum_{r=1}^{R} D\left(Q(x_{2r}, u_{2r})\|P(x_{2r}, u_{2r})\right)\\ &+ \sum_{r=1}^{R} D\left(Q(\upsilon_r)\|P(\upsilon_r)\right).\end{aligned}$$

Each divergence can be evaluated using standard formulas for the divergence between two multivariate Gaussian distributions and the divergence between two Gamma distributions given in the Appendix (equations (12) and (15)). For example, writing

$$\begin{aligned} & D\left(Q(x_1)Q(u_1)\|P(x_1,u_1)\right) \\ &= E_{u_1}\left[D\left(Q(x_1)\|P(x_1|u_1)\right)\right] \\ &\quad + D\left(Q(u_1)\|P(u_1)\right), \end{aligned}$$

the second term here is recognized to be the divergence between two Gamma distributions and the first term can be evaluated using the formula (12) for the divergence between two multivariate Gaussians:

$$\begin{aligned} & D\left(Q(x_1)\|P(x_1|u_1)\right) \\ &= -\frac{N_1}{2} - \frac{1}{2}\ln|\operatorname{Cov}(x_1,x_1)|u_1^{N_1} \\ &\quad + \frac{1}{2}\operatorname{tr}\left(\left[\operatorname{Cov}(x_1,x_1) + \langle x_1\rangle\langle x_1^*\rangle\right]u_1\right) \end{aligned}$$

so that

$$\begin{aligned} & E_{u_1}\left[D\left(Q(x_1)\|P(x_1|u_1)\right)\right] \\ &= -\frac{N_1}{2} - \frac{N_1}{2}\langle\ln u_1\rangle - \frac{1}{2}\ln|\operatorname{Cov}(x_1,x_1)| \\ &\quad + \frac{1}{2}\langle u_1\rangle\langle x_1^* x_1\rangle. \end{aligned}$$

4.2. Diagonalization

The computation of $\mathcal{L}$ can be speeded up by observing that the matrices $\Lambda, U_1^*\Lambda U_1$ and $U_2^*\Lambda U_2$ can be simultaneously diagonalized.

If O is any orthogonal matrix, O^*x_1 has the same prior distribution as x_1 so, since $(U_1O)(O^*x_1) = U_1x$, we obtain an equivalent model by replacing U_1 by U_1O. Taking O to be the matrix whose columns are the eigenvectors of $U_1^*\Lambda U_1$ ensures that $(U_1O)^*\Lambda U_1O$ is diagonal. A similar operation can be carried out on U_2.

Let P be the matrix whose columns are the eigenvectors of Λ, so that $P^*\Lambda P$ is diagonal. Transforming the data by

$$D_r \leftarrow P^*D_r$$

and the model by

$$\begin{aligned} \Lambda &\leftarrow P^*\Lambda P \\ U_1^+ &\leftarrow P^*U_1^+ \\ U_2^+ &\leftarrow P^*U_2^+ \end{aligned}$$

enables all of the calculations in Sections 3 and 4.1 to be carried out with a diagonal precision matrix.

An interesting consequence of this diagonalizationn (not pointed out in [4]) is that, in the case of Gaussian PLDA, the term U_2x_{2r} can *always* be eliminated form (2) at recognition time even if the feature vectors are of such high dimension that Λ has to be constrained to be diagonal in order to be robustly estimated. All that is required is to diagonalize the effective covariance matrix $\Lambda^{-1} + U_2U_2^*$, something that can be done off-line. The only computational cost incurred at recognition time is that of transforming the data by

$$D_r \leftarrow P^*D_r$$

where P is the matrix whose columns are the eigenvectors of $\Lambda^{-1} + U_2U_2^*$. (In the heavy tailed case, this trick can only be performed if n_2 and ν are identical.)

4.3. The likelihood ratio for speaker verification

Suppose we are given two observation vectors D_1 and D_2 and we wish to test the hypothesis H_1 that they come from the same speaker against the hypothesis H_0 that they come from different speakers. The log likelihood ratio for this hypothesis test is

$$\ln\frac{P(D_1,D_2|H_1)}{P(D_1|H_0)P(D_2|H_0)}.$$

(This is referred to as the "batch likelihood ratio" in [11, 10].) We have to use the lower bound $\mathcal{L}$ as a proxy for the log likelihood, so we approximate the numerator by evaluating $\mathcal{L}$ as in Section 4.1 with $R = 2$ and data (D_1, D_2). (Taking $R = 2$ in Fig. 1 or Fig. 2 gives the graph for this calculation.) Similarly, we approximate $\ln P(D_1|H_0)$ by evaluating $\mathcal{L}$ with $R = 1$ and data D_1 and likewise for $\ln P(D_2|H_0)$. (Taking $R = 1$ in Fig. 1 or Fig. 2 gives the graph for each of these calculations.)

The approximate log likelihood ratio calculated in this way is symmetric in D_1 and D_2 and the construction generalizes straightforwardly to the situation where the hypothesis test involves comparing one collection of utterances with another (rather than comparing one utterance with another). Similarly it can be extended to the family of speaker recognition problems considered in [14].

4.4. Score normalization

The denominator of the likelihood ratio we have just constructed can be thought of performing score normalization in the sense in which this term is used in speaker verification. So it is not *a priori* clear that standard score normalization methods (z-norm and t-norm) will be helpful. Note that these normalizations break the symmetry between D_1 and D_2, so in experimenting with this question it is natural to seek a score normalization procedure which preserves this symmetry. The simplest choice is to normalize the score s of a the pair (D_1, D_2) using the formula

$$\frac{s-\mu_1}{\sigma_1} + \frac{s-\mu_2}{\sigma_2} \tag{8}$$

where the mean μ_1 and standard deviation σ_1 are calculated by matching D_1 against a given imposter cohort and similarly for μ_2 and σ_2. We refer to this type of score normalization as s-norm. It has been our experience that, in situations where score normalization is helpful, s-norm is more effective than z-norm or zt-norm.

5. Model estimation

We now suppose that we have a training set consisting of data collected from multiple speakers. We index the speakers by s and denote the number of observation vectors for speaker s by $R(s)$. We denote the number of speakers by S and we set $R = \sum_s R(s)$. For each speaker s, we define $\mathcal{L}(s), \mathcal{L}_1(s)$ and $\mathcal{L}_2(s)$ by applying (7) to the speaker's recordings.

We estimate the model parameters by maximizing $\sum_s \mathcal{L}(s)$. We outline two approaches to estimating W and Λ which we refer to as maximum likelihood and minimum divergence. Maximum likelihood estimation consists in maximizing the quantity $\sum_s \mathcal{L}_1(s)$; this is the auxiliary function in Bishop's formulation of the EM algorithm [7]. Minimum divergence estimation consists in maximizing the negative sum of divergences $\sum_s \mathcal{L}_2(s)$ in such a way as to preserve the value of the EM auxiliary function. This is also the objective function used in estimating the numbers of degrees of freedom n_1, n_2 and ν.

5.1. Maximum likelihood estimation

Maximizing $\sum_s \mathcal{L}_1(s)$ with respect to W is equivalent to minimizing

$$\sum_s \sum_{r=1}^{R(s)} \langle \upsilon_r(s) \rangle \langle (D_r(s) - W z_r(s))^* \Lambda (D_r(s) - W z_r(s)) \rangle.$$

Setting the derivative with respect to W equal to 0 leads to the equation[1]

$$W \sum_s \sum_{r=1}^{R(s)} \langle \upsilon_r(s) \rangle \langle z_r(s) z_r^*(s) \rangle = \sum_s \sum_{r=1}^{R(s)} \langle \upsilon_r(s) \rangle D_r(s) \langle z_r^*(s) \rangle$$

which can be readily solved once the first and second moments of $z_r(s)$ have been evaluated. Dropping the reference to s, these are given by

$$\begin{aligned} \langle z_r \rangle &= \begin{pmatrix} \langle x_1^+ \rangle \\ \langle x_{2r}^+ \rangle \end{pmatrix} \\ \text{and } \langle z_r z_r^* \rangle &= \begin{pmatrix} \langle x_1^+ x_1^{+*} \rangle & \langle x_1^+ \rangle \langle x_{2r}^{+*} \rangle \\ \langle x_{2r}^+ \rangle \langle x_1^{+*} \rangle & \langle x_{2r}^+ x_{2r}^{+*} \rangle \end{pmatrix} \end{aligned}$$

where

$$\langle x_1^+ x_1^{+*} \rangle = \begin{pmatrix} \mathrm{Cov}\,(x_1, x_1) + \langle x_1 \rangle \langle x_1^* \rangle & \langle x_1^* \rangle \\ \langle x_1 \rangle & 1 \end{pmatrix}$$

and $\langle x_{2r}^+ x_{2r}^{+*} \rangle$ is evaluated similarly.

The maximum likelihood estimate of the covariance matrix Λ^{-1} is given by

$$\frac{1}{R} \sum_s \sum_{r=1}^{R(s)} \langle \upsilon_r(s) \rangle \langle (D_r(s) - W z_r(s))(D_r(s) - W z_r(s))^* \rangle.$$

This can also be readily evaluated in terms of W and the first and second moments of $z_r(s)$.

5.2. Minimum divergence estimation

The idea here is to identify a class of transformations which apply to both the hidden variables and the model parameters and which preserve the value EM auxiliary function and then select from this class the transformation which minimizes the divergences which make up $\sum_s \mathcal{L}_2(s)$. This type of estimation is generally applicable in the context of variational Bayes EM algorithms involving continuous hidden variables and, when interleaved with maximum likelihood estimation, it helps to speed up convergence [15].

[1] If only U_2^+ is to be estimated (the case of microphone speech), the equation to be solved is

$$\begin{aligned} & U_2^+ \sum_s \sum_{r=1}^{R(s)} \langle \upsilon_r(s) \rangle \langle x_{2r}^+(s) x_{2r}^{+*}(s) \rangle \\ &= \sum_s \sum_{r=1}^{R(s)} \langle \upsilon_r(s) \rangle \left(D_r(s) - U_1^+ \langle x_1^+(s) \rangle \right) \langle x_{2r}^{+*}(s) \rangle. \end{aligned}$$

5.2.1. Eigenvoices

Applying this idea to U_1, the natural class of transformation is

$$\begin{aligned} x_1'(s) &= A(x_1(s) - a) \\ U_1' &= U_1 A^{-1} \\ m_1' &= m_1 + U_1 a \\ u_1'(s) &= k u_1(s) \end{aligned}$$

where A is an $N_1 \times N_1$ matrix, a is an $N_1 \times 1$ vector and k is a scalar. These have to be chosen so as to minimize the sum of divergences

$$\sum_s D\left(Q(x_1'(s), u_1'(s)) \| P(x_1'(s), u_1'(s)) \right).$$

We explained how to evaluate these divergences in Section 4.1. The optimization is straightforward and results in the following equations for A, a and k:

$$\begin{aligned} \frac{1}{k} &= \frac{1}{S} \sum_s \langle u_1(s) \rangle \\ a &= \frac{1}{\sum_s \langle u_1(s) \rangle} \sum_s \langle u_1(s) \rangle \langle x_1(s) \rangle \\ A^{-1} &= L \end{aligned}$$

where L is a lower triangular matrix such that LL^* is the Cholesky decomposition of

$$\frac{1}{\sum_s \langle u_1(s) \rangle} \sum_s \langle u_1(s) \rangle \langle x_1(s) x_1^*(s) \rangle - a a^*.$$

(In the case of a Gaussian prior, set $\langle u_1(s) \rangle = 1$ in these equations. The minimum divergence estimation formulas are readily seen to have the same form as in [3] in this case.)

5.2.2. Eigenchannels

Applied to U_2, the natural class of transformations is

$$\begin{aligned} x_{2r}'(s) &= A(x_{2r}(s) - a) \\ U_2' &= U_2 A^{-1} \\ m_2' &= m_2 + U_2 a \\ u_{2r}'(s) &= k u_{2r}(s) \end{aligned}$$

where A is an $N_2 \times N_2$ matrix, a is an $N_2 \times 1$ vector and k is a scalar. Minimizing the corresponding divergences leads to the following equations for A, a and k:

$$\begin{aligned} \frac{1}{k} &= \frac{1}{R} \sum_s \sum_{r=1}^{R(s)} \langle u_{2r}(s) \rangle \\ a &= \frac{1}{\sum_s \sum_{r=1}^{R(s)} \langle u_{2r}(s) \rangle} \sum_s \sum_{r=1}^{R(s)} \langle u_{2r}(s) \rangle \langle x_{2r}(s) \rangle \\ A^{-1} &= L \end{aligned}$$

where L is a lower triangular matrix such that LL^* is the Cholesky decomposition of

$$\frac{1}{\sum_s \sum_{r=1}^{R(s)} \langle u_{2r}(s) \rangle} \sum_s \sum_{r=1}^{R(s)} \langle u_{2r}(s) \rangle \langle x_{2r}(s) x_{2r}^*(s) \rangle - a a^*$$

(In the case of a Gaussian prior, set $\langle u_{2r}(s) \rangle = 1$ in these equations.)

5.2.3. Precision

Applied to Λ, the natural class of transformations is

$$\begin{aligned} \Lambda' &= \frac{1}{\kappa}\Lambda \\ v_r'(s) &= \kappa v_r(s) \end{aligned}$$

where κ is a scalar. The optimization leads to the equation

$$\frac{1}{\kappa} = \frac{1}{R}\sum_s \sum_{r=1}^{R(s)} \langle v_r(s)\rangle.$$

(There is no analogue of this in the case of a Gaussian prior.)

5.3. Estimating the numbers of degrees of freedom

To estimate n_1, we minimize the sum of divergences

$$\sum_s D\left(\mathcal{G}(a_1(s), b_1(s)) \| \mathcal{G}\left(n_1/2, n_1/2\right)\right)$$

by differentiating with to n_1 and setting the derivative to 0. This leads to the equation

$$\psi\left(\frac{n_1}{2}\right) - \ln\frac{n_1}{2} = 1 + \frac{1}{S}\sum_s \left(\langle \ln u_1(s)\rangle - \langle u_1(s)\rangle\right)$$

which can be solved for n_1 by a line search. Similarly, n_2 and ν are estimated by solving

$$\begin{aligned} \psi\left(\frac{n_2}{2}\right) - \ln\frac{n_2}{2} &= 1 + \frac{1}{R}\sum_s \sum_{r=1}^{R(s)} \left(\langle \ln u_{2r}(s)\rangle - \langle u_{2r}(s)\rangle\right) \\ \psi\left(\frac{\nu}{2}\right) - \ln\frac{\nu}{2} &= 1 + \frac{1}{R}\sum_s \sum_{r=1}^{R(s)} \left(\langle \ln v_r(s)\rangle - \langle v_r(s)\rangle\right). \end{aligned}$$

6. Features

In [2], it was shown how, by extracting a single 400 dimensional feature vector from each recording, a high performance speaker recognition system could be built using a simple classifier. The word feature in the title of this section refers to vectors of this type, rather than to the acoustic feature vectors produced in the front end (which is same as in [3, 2]). These feature vectors are known in the vernacular as i-vectors.

A gender-dependent universal background model and a gender dependent factor analysis model were first trained on a large corpus of background data without distinguishing between speaker and channel effects, so that the assumption is that the supervector M associated with a recording has the form

$$M = M_0 + Tw \tag{9}$$

where M_0 is a mean supervector, T is a rectangular matrix of rank 400 and w is a random vector of dimension 400 having a standard normal distribution. The feature vector associated with a given recording is just the MAP estimate of w, calculated as in [16]. We denote this by $\hat{w}$.

The experiments reported in [2] were conducted on telephone speech and only telephone data was used for factor analysis training. In the current paper and in the companion paper [8], we aim to use similar features to perform speaker recognition on microphone as well as telephone speech. It is not entirely straightforward to construct a feature extractor for this purpose as the amount of microphone data at our disposal for factor analysis training is ten times less than the amount of telephone data. We encountered a similar problem in estimating eigenchannels for microphone speech in joint factor analysis in [3], where a method of estimating supplementary eigenchannels on microphone development data was proposed. We use a similar method here.

Thus we estimated a (gender dependent) matrix T' of rank 200, using only microphone data, by assuming the supervector M associated with a microphone recording has the form

$$M = M_0 + T\hat{w} + T'w' \tag{10}$$

where w' is a random vector having a standard normal distribution.

We then combined (9) and (10) to produce a feature extractor which could be used for both microphone and telephone speech. That is, we assumed that the supervector M associated with a recording has the form

$$M = M_0 + Sx$$

where $S = \begin{pmatrix} T & T' \end{pmatrix}$ and x is a random vector of dimension 600 having a standard normal distribution. The feature vector associated with the recording is the MAP estimate of x.

The universal background model and the matrix T were trained using telephone speech from the Switchboard and Fisher corpora and from the NIST 2004 and 2005 SRE's, as described in [2]. For estimating T', we used microphone data from the 2005 and 2006 evaluations as well as the interview development data from the 2008 evaluation.

7. Experiments

We performed experiments on the short2-short3, 8conv-short3 and 10sec-10sec conditions of the NIST 2008 speaker recognition evaluation [17]. We restricted ourselves to English language data (the 2010 SRE is English-only) and to female speakers (the error rates are higher than for males). We will report results on trials involving English language telephone speech (det7) in enrollment and testing, on trials involving interview speech (det1), and on mixed trials (det4 and det5).

For all experiments (whether on microphone or telephone data), we used the 600-dimensional feature vectors described in the previous section.

7.1. Telephone speech

For our first series of experiments, we trained Gaussian and heavy-tailed models using only telephone speech and we tested on telephone speech. The results are summarized in Table 1 (EER refers to the equal error rate and DCF to the minimum value of the NIST detection cost function). For model training, we used the same telephone data as for the factor analysis training described in the previous section, except that we added telephone data from the 2006 SRE. We set $N_1 = 400$ and $N_2 = 0$ (that is, 400 eigenvoices and no eigenchannels) and we took the precision matrix Λ to be full rather than diagonal. For score normalization we used s-norm (8) with 200 imposters taken from the 2004 SRE data (this works better than taking 100 from 2004 and 100 from 2005).

We note firstly that heavy-tailed priors produce much better results than Gaussian priors. Secondly, score normalization is uniformly helpful in the Gaussian but harmful in the heavy-tailed case (with one exception, namely the equal error rate in the 8conv-short3 condition). However, even when score normalization is applied, heavy-tailed priors consistently outperform Gaussian priors.

Table 1: *EER/DCF on telephone speech conditions (det7) obtained with Gaussian and heavy-tailed priors (i) without score normalization and (ii) with score normalization*

	Gaussian	heavy-tailed
short2-short3	3.6% / 0.014	**2.2**% / **0.010**
8conv-short3	3.7% / 0.009	1.3% / **0.005**
10sec-10sec	16.4% / 0.070	**10.9**% / **0.053**

	Gaussian	heavy-tailed
short2-short3	2.7% / 0.013	2.4% / 0.012
8conv-short3	1.5% / 0.009	**0.8**% / 0.007
10sec-10sec	13.3% / 0.063	12.8% / 0.066

7.2. Microphone speech

We used the heavy-tailed model just described and augmented it by adding 200 eigenchannels, estimated on microphone data from the 2005 and 2006 evaluations as well as the interview development data from the 2008 evaluation, using the equation given in the footnote to Section 5.1 and the minimum divergence procedure described in Section 5.2.2. Depending on whether the channel factors x_{2r} are assumed to be Gaussian or heavy-tailed, the resulting model is either "partially heavy-tailed" or "fully heavy-tailed". Results obtained in the two cases are summarized in Table 2; det1 refers to the subset of the core condition where interview speech was used for enrollment and testing; det4 to the case where interview speech was used for enrollment and telephone speech for testing; and det5 to the case where telephone speech was used for enrollment and non-interview microphone speech for testing. We used the speech recognition transcripts of the microphone data provided by NIST as a proxy for voice activity detection. Files for which speech recognition broke down were excluded in drawing up the results.

Note that score normalization is consistently effective in these conditions, producing major improvements in equal error rates. (We used an s-norm cohort consisting of 100 recordings from the 2005 SRE microphone data and 100 from the 2006 microphone data.) The "partially heavy-tailed" model generally performs better than the "fully heavy-tailed model".

Table 2: *EER/DCF on microphone speech obtained with partially and fully heavy-tailed priors (i) without score normalization and (ii) with score normalization*

	partially heavy-tailed	fully heavy-tailed
det1	5.0% / 0.017	5.4% / 0.019
det4	5.2% / 0.020	6.2% / 0.028
det5	4.0% / 0.021	6.2 % / 0.026

	partially heavy-tailed	fully heavy-tailed
det1	**3.3**% / **0.017**	3.4% / **0.017**
det4	**2.8**% / **0.016**	3.1% / 0.018
det5	4.0% / **0.020**	**3.8**% / **0.020**

Listening to the interview data suggests that the channel effects are quite non-Gaussian (there are lots of black swans): this was a primary motivation for experimenting with heavy-tailed priors. In fact, it turns out that these effects are so non-Gaussian that, if left to their own devices, the maximum likelihood and minimum divergence training algorithms exhibit degenerate behavior in the heavy-tailed case. The estimate produced for the number of degrees of freedom n_2 is 0.5 (whereas values in the range 10—15 are typical for n_1 and ν). With such a low value for n_2, channel effects of large magnitude are lightly penalized so that any speaker can be matched with any other and speaker verification breaks down (see the Appendix). To prevent this happening, we had to floor n_2 at 10.0 in the course of training U_2. The "fully heavy-tailed" results in Table 2 were produced in this way and they are not as good as the "partially heavy-tailed" results in which the channel factors are assumed to be normally distributed. (Flooring at 3.0 gives slightly worse results than flooring at 10.0.)

An interesting question is whether adding eigenchannels trained on microphone speech degrades performance on telephone speech. Counter-intuitively, it turns out that doing so leads to some improvements in performance (cf. [18]). We used the "partially heavy-tailed" model just described to replicate our experiments on telephone speech, producing the results in Table 3 which can be compared with those those in the heavy-tailed column of Table 1. In particular, the improvement in the 10sec-10sec case is substantial.

Table 3: *EER/DCF on telephone speech conditions (det7) obtained with partially heavy-tailed priors (i) without score normalization and (ii) with score normalization*

short2-short3	**2.2** % / **0.010**
8conv-short3	1.2 % / **0.004**
10sec-10sec	11.3 % / **0.051**

short2-short3	2.3 % / 0.011
8conv-short3	**0.7**% / 0.005
10sec-10sec	**9.9**% / **0.051**

Inspecting the boldface entries in the Table shows that the question of whether score normalization is effective in this situation is less clear cut than in the case of Table 1. The mismatch between the partially heavy-tailed model and the telephone test data may account for this.

The best results in Table 3 are about 30% better than we have achieved with JFA; improvements in the case of microphone speech are consistent but less dramatic. See the companion paper [8] and [2] for JFA benchmarks on 2008 SRE microphone and telephone speech respectively.

8. Conclusion

We have shown how using low dimensional feature representations for speech data enables JFA to be recast so as to encompass heavy-tailed modeling of speaker and channel effects and Bayesian decision making, resulting in a 30% improvement in error rates on the NIST 2008 SRE telephone data. A particularly interesting aspect of this work is that in the case of telephone speech, score normalization is not needed.

In theory, a good generative model for speech should produce likelihood ratios which do not need to be normalized (or even calibrated) in order to set a trial-independent decision threshold for speaker verification. Score normalization is typically fragile and computationally expensive but it is generally needed in practice because outlying recordings tend to produce exceptionally low scores for all of the trials in which they are involved. It appears that the additional hidden variables u_1, u_{2r}

and v_r in heavy-tailed PLDA are able model these outliers adequately.

In the case of microphone speech, we found that Student's *t* modeling of channel effects degenerates if it is left to its own devices. This indicates extreme non-Gaussian behavior in the data. We have attempted to deal with this using other probabilistic methods not reported here, notably multimodal distributions (discrete and continuous mixtures), but with little success. At this writing, a fully probabilistic approach to this problem seems to be beyond our grasp and the best way to proceed may be to project away the troublesome dimensions in the i-vector space using some type of linear discriminant analysis (classical LDA or PLDA).

Recall that, in the introduction, we summarized PLDA as follows. Speech data D is assumed to be decomposable into a speaker component S and a channel component C:

$$D = S + C \tag{11}$$

in such a way that (i) S and C are statistically independent and (ii) S and C have Gaussian distributions. In this paper we have shown how the Gaussian assumption can be relaxed but we have not examined the statistical independence assumption at all. The success of cosine distance scoring in speaker recognition [2] and speaker clustering [19] provides strong empirical evidence that the statistical independence assumption is flawed. It seems likely that modifying the heavy-tailed PLDA model in such a way as to incorporate a cosine distance type metric will prove to be a fruitful line of investigation although it may not be easy to achieve this in a fully probabilistic framework. For some speculative remarks on this subject, see the accompanying slides.[2]

9. References

[1] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proc. Interspeech*, Brighton, UK, Sept. 2009.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," to appear in *IEEE Trans. Audio, Speech, Lang. Process.*

[3] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 16, no. 5, pp. 980–988, July 2008. [Online]. Available: http://www.crim.ca/perso/patrick.kenny

[4] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. 11th International Conference on Computer Vision*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1–8.

[5] M. Svensén and Bishop, "Robust Bayesian mixture modelling," *Neurocomputing*, vol. 64, pp. 235–252

[6] C. Archambeau and M. Verleysen, "Robust Bayesian clustering," *Neural Networks*, vol. 20, pp. 129–138, 2007.

[7] C. Bishop, *Pattern Recognition and Machine Learning*. New York, NY: Springer Science+Business Media, LLC, 2006.

[2] Available at http://www.crim.ca/perso/patrick.kenny

[8] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech," in *Proc. Odyssey Speaker and Language Recogntion Workshop*, Brno, Czech Republic, June 2010.

[9] X. Zhao, Y. Dong, J. Zhao, L. Lu, J. Liu, and H. Wang, "Variational Bayesian joint factor analysis for speaker verification," in *Proc. ICASSP 2009*, Taipei, Taiwan, Apr. 2009.

[10] X. Zhao and Y. Dong, "Variational Bayesian joint factor analysis models for speaker verification," submitted to *IEEE Trans. Audio, Speech, Lang. Process.*

[11] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1448–1460, May 2007.

[12] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE J. of Selected Topics in Signal Processing*, Dec. 2010.

[13] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms, Tech. Report CRIM-06/08-13," 2005. [Online]. Available: http://www.crim.ca/perso/patrick.kenny

[14] N. Brummer and E. de Villiers, "The speaker partitioning problem," in *Proc. Odyssey Speaker and Language Recogntion Workshop*, Brno, Czech Republic, June 2010.

[15] J. Luttinen, A. Ilin, and T. Raiko, "Transformations for variational factor analysis to speed up learning," in *Proc. of the 17th European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning (ESANN 2009)*, Bruges, Belgium, Apr. 2009, pp. 77–82.

[16] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 3, pp. 345–359, May 2005.

[17] (2008) The NIST year 2008 speaker recognition evaluation plan. [Online]. Available: http://www.nist.gov/speech/tests/spk/2008/

[18] L. Burget, P. Matejka, V. Hubeika, and J. Cernocky, "Investigation into variants of joint factor analysis for speaker recognition," in *Proc. Interspeech 2009*, no. 9. International Speech Communication Association, 2009, pp. 1263–1266.

[19] H. Tang, S. M. Chu, and T. S. Huang, "Spherical discriminant analysis in semi-supervised speaker clustering," in *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Morristown, NJ, USA: Association for Computational Linguistics, 2009, pp. 57–60.

Appendix: Normal, Gamma and Student's *t* distributions

We denote the N-dimensional Gaussian distribution with mean μ and covariance matrix Σ by $\mathcal{N}(\mu, \Sigma)$ and the corresponding probability density function by $N(x|\mu, \Sigma) \quad (x \in \mathbb{R}^N)$. The

formula for the Kullback-Leibler divergence between two N-dimensional Gaussians is

$$\begin{aligned} & D\left(\mathcal{N}(\mu,\Sigma)||\mathcal{N}(\mu_0,\Sigma_0)\right) \\ &= -\frac{N}{2} - \frac{1}{2}\ln|{\Sigma_0}^{-1}\Sigma| \\ &\quad + \frac{1}{2}\mathrm{tr}\left({\Sigma_0}^{-1}\left[\Sigma + (\mu-\mu_0)(\mu-\mu_0)^*\right]\right). \end{aligned} \quad (12)$$

The Gamma and Digamma functions are defined by

$$\begin{aligned} \Gamma(a) &= \int_0^\infty x^{a-1}e^{-ax}\mathrm{d}x \\ \psi(a) &= \frac{\Gamma'(a)}{\Gamma(a)} \end{aligned}$$

for $a > 0$. We denote the Gamma distribution with shape parameter a and rate parameter b by by $\mathcal{G}(a,b)$ and the corresponding probability density function by $G(u|a,b)$. This is defined for $u > 0$ by

$$G(u|a,b) = \frac{b^a}{\Gamma(a)}u^{a-1}e^{-bu}.$$

The Gamma distribution has the properties that

$$\langle u\rangle = ab^{-1} \quad (13)$$

$$\langle \ln u\rangle = \psi(a) - \ln b \quad (14)$$

and the Kullback-Leibler divergence between two Gamma distributions is given by

$$\begin{aligned} & D\left(\mathcal{G}(a_0,b_0)||\mathcal{G}(a,b)\right) \\ &= \ln\frac{\Gamma(a)}{\Gamma(a_0)} + a_0\ln b_0 - a\ln b \\ &\quad + (a_0-a)(\psi(a_0)-\ln b_0) + a_0\frac{b-b_0}{b_0} \end{aligned} \quad (15)$$

A random vector X is said to have a multivariate Student's *t* distribution with ν degrees of freedom, mean μ and scale parameter Λ if there is a hidden variable u such that $X \sim \mathcal{N}(\mu,(u\Lambda)^{-1})$ where $u \sim \mathcal{G}(\nu/2,\nu/2)$. There is a well known closed form expression for the probability density function of the Student's *t* distribution which can be found in [7] but we don't need it.

In the case where $\nu < 2$, the second order moments of X do not exist, in the sense that $E\left[\|X-\mu\|^2\right] = \infty$. (The reason for this is that the probability density function $G(u|\nu/2,\nu/2)$ blows up in the neighborhood of $u = 0$ if $\nu < 2$.) If such a Student's *t* distribution is used to model channel effects, then channel distortions of arbitrarily large magnitude will be treated as "normal". Such a model is clearly unreasonable for speaker recognition. This explains why Student's *t* modeling breaks down on microphone data (where maximum likelihood estimation produces an estimate of 0.5 for the number of degrees of freedom, n_2).