ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/3457787

Joint Factor Analysis Versus Eigenchannels in Speaker Recognition

Article *in* IEEE Transactions on Audio Speech and Language Processing · June 2007

DOI: 10.1109/TASL.2006.881693 · Source: IEEE Xplore

CITATIONS
436

READS
585

4 authors, including:

Patrick Kenny
Centre de recherche informatique de Montréal
136 PUBLICATIONS **6,297** CITATIONS
SEE PROFILE

Gilles Boulianne
Centre de recherche informatique de Montréal
68 PUBLICATIONS **2,835** CITATIONS
SEE PROFILE

Pierre Dumouchel
École de Technologie Supérieure
103 PUBLICATIONS **4,571** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

LIve closed-captioning View project

Human stress measurement scale from ubiquitous sensors View project

All content following this page was uploaded by Gilles Boulianne on 11 December 2012.

The user has requested enhancement of the downloaded file.

Joint Factor Analysis versus Eigenchannels in Speaker Recognition

Patrick Kenny, G. Boulianne, P. Ouellet and P. Dumouchel

***Abstract*— We compare two approaches to the problem of session variability in GMM-based speaker verification, eigenchannels and joint factor analysis, on the NIST 2005 speaker recognition evaluation data. We show how the two approaches can be implemented using essentially the same software at all stages except for the enrollment of target speakers. We demonstrate the effectiveness of zt-norm score normalization and a new decision criterion for speaker recognition which can handle large numbers of t-norm speakers and large numbers of speaker factors at little computational cost. We found that factor analysis was far more effective than eigenchannel modeling. The best result we obtained was a detection cost of 0.016 on the core condition (all trials) of the evaluation.**

***Index Terms*— Speaker verification, Gaussian mixture model, speaker factors, channel factors, eigenchannels**

I. Introduction

This article compares two approaches to dealing with the problem of session variability in GMM-based speaker recognition. We use the term session variability to refer to all of the phenomena which cause two recordings of a given speaker to sound different from each other. This type of variability is usually thought of as being attributable to channel effects although other, more mysterious, factors may be at play such as the well known aging phenomenon whereby the performance of speaker models degrades over time.

Our principal effort in the direction of solving this problem has been to develop a model which we refer to as a joint factor analysis of speaker and channel variability [1]. This is based on similar assumptions to feature mapping [2] but it treats channel effects as continuous rather than discrete and it exploits correlations between Gaussians in modeling speaker variability. This model has two major drawbacks. Firstly, although we have made substantial progress in simplifying it [3], it is mathematically and computationally demanding in many respects. Secondly, it seems to require a well balanced training set in which a typical training speaker is recorded under a variety of channel conditions that is sufficiently broad to cover most of the channel variation that is likely to be encountered at recognition time.

This is not an easy requirement to meet. In fact it is not easy even to find speakers who have been recorded over both landline and cellular transmission channels. As a rule, each speaker in the Switchboard collections was recorded over either landline channels or cellular channels but not both. Only a small fraction of the speakers in the Fisher English Part I database were recorded more than once (and furthermore the speaker identifications in this database are not reliable). So if a joint factor analysis model is trained on the union of the telephone speech corpora currently available through the Linguistic Data Consortium, the model could be misled into believing that some speakers are 'landline speakers' and others are 'cellular speakers'. Thus, despite the success we have with the joint factor analysis model when it is trained and tested on Switchboard data [4], it is not *a priori* clear that it can be trained so as to perform well on a test set drawn from the Mixer corpus such as the NIST 2005 evaluation set [5].

These considerations led us to re-implement another approach to the problem of session variability which we had previously developed and which we refer to as eigenchannel modeling [6]. This stands in the same relationship to speaker model synthesis [7] as the joint factor analysis model does to feature mapping. It is a cruder model than the joint factor analysis model in that it only attempts to deal with channel effects at recognition time and not at enrollment time but it is easier to implement and it appears to make less stringent demands for training data.

We will report the results of experiments carried out on the NIST 2005 test set using both of these approaches. Our principal conclusion will be that, notwithstanding our concerns about mismatched training data, the joint factor analysis model is capable of far better performance than eigenchannel modeling. Moreover we will show that the joint factor analysis model can perform very well in speaker verification using a computationally inexpensive decision rule that steers a middle course between the 'exact' and 'simplified' scoring rules in [3]. This decision criterion can handle large numbers of t-norm speakers at little cost and large numbers of common speaker factors at no extra cost.

II. Models of Session Variability

Our approach to speaker recognition is GMM-based in that we estimate a GMM for each target speaker but we use these GMM's to make speaker verification decisions in non-traditional ways. The issue here is how we attempt to compensate for inter-session variability and for channel mismatches between enrollment and test conditions in particular. By way of introduction, in this section we will briefly sketch

The authors are with the Centre de recherche informatique de Montréal (CRIM), 550 Sherbrooke Street West, Suite 100, Montreal, Quebec, Canada H3A 1B9. Phone (514) 840 1234, Fax (514) 840 1244.

email: patrick.kenny@crim.ca, gilles.boulianne@crim.ca, pierre.ouellet@crim.ca, pierre.dumouchel@crim.ca

This work was supported in part by the Natural Science and Engineering Research Council of Canada and by the Ministère du Développement Économique et Régional et de la Recherche du gouvernement du Québec

0000–0000/00$00.00 © IEEE

the model adaptation techniques that have been developed to tackle the problem of channel compensation in GMM-based speaker recognition.

The method of speaker model synthesis [7] assumes that channel effects are discrete and that channel detection can be performed in a pre-processing step in which utterances are classified as, say, 'GSM cellular', 'landline electret' and so forth. Suppose we are given an enrollment utterance and a test utterance and it is hypothesized that they are uttered by the same speaker. Denote the corresponding speaker- and channel-dependent supervectors by $\boldsymbol{M}$ and $\boldsymbol{M}'$. The basic assumption is that $\boldsymbol{M}'$ can be synthesized from $\boldsymbol{M}$ by adding a supervector $\boldsymbol{c}$ which depends only on the enrollment and test channel conditions (and not on the speaker) as in Fig. 1 so that

$$\boldsymbol{M}' \quad = \quad \boldsymbol{M} + \boldsymbol{c}. \tag{1}$$

There is one supervector $\boldsymbol{c}$ for every possible pair of microphone/channel conditions.

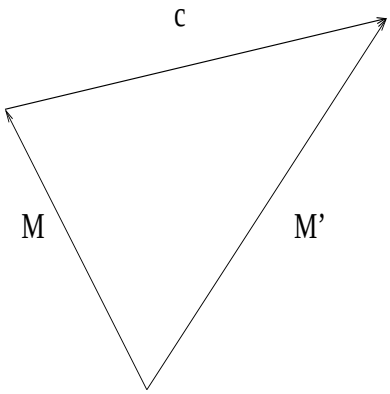

Fig. 1. *In speaker model synthesis and eigenchannel modeling it is assumed that if $\boldsymbol{M}$ and $\boldsymbol{M}'$ are the speaker- and channel-dependent supervectors for two recordings of a given speaker then the difference between them, namely $\boldsymbol{c}$, depends only on channel effects.*

The eigenchannel model in [6] is a continuous version of speaker model synthesis which dispenses with the need for channel detection. The idea is that just as most speaker variability is low dimensional (the premise of eigenvoice modeling) the same is probably true of channel variability so that similar methods can be brought to bear on modeling both types of variability. (The idea of using eigenvoice methods to model channel effects seems to have been first mooted in [8].) So the assumption in eigenchannel modeling is that (1) holds with the channel compensation supervector $\boldsymbol{c}$ being normally distributed in a low dimensional subspace of the supervector space. Put another way, it is assumed that there is a rectangular matrix $\boldsymbol{u}$ of low rank such that

$$\boldsymbol{c} \quad = \quad \boldsymbol{u}\boldsymbol{x} \tag{2}$$

where $\boldsymbol{x}$ has a standard normal distribution so that all of the channel compensation supervectors in Fig. 1 can be expressed as linear combinations of the columns of $\boldsymbol{u}$. In the terminology of [6], the non zero eigenvectors of $\boldsymbol{u}\boldsymbol{u}^*$ are the eigenchannels.

This distribution on $\boldsymbol{c}$ can be used as a prior for MAP adaptation of a speaker GMM to the channel conditions in a test utterance. This type of MAP adaptation is formally identical to eigenvoice MAP [9] so we dubbed it eigenchannel MAP in [6]. Eigenchannel MAP provides a mechanism for synthesizing a new model for each speaker whenever a new test utterance is encountered. Speaker recognition can then be performed by evaluating these channel-adapted models with the standard GMM likelihood function in the usual way. Similar types of model synthesis were used in the systems submitted by Spescom DataVoice in the NIST 2004 and 2005 evaluations, by Queensland University of Technology in the 2005 evaluation [10] and also in [11]. This seems to be quite an effective way to proceed but it has the disadvantage of being very computationally expensive (particularly if there is a large number of t-norm speakers). It is also a rather dubious procedure from a purely mathematical point of view, since adapting a model to data and then evaluating the likelihood of data with the adapted model results in a 'likelihood function' which integrates to something bigger than 1. In this article we will propose another scoring mechanism which does not suffer from these drawbacks.

The most widely used method of GMM adaptation for channel compensation is feature mapping [2]. Although this is usually thought of as a front-end compensation scheme it can equally well be viewed as a model adaptation technique if it is assumed that GMM supervectors can be decomposed into speaker- and channel-dependent parts as illustrated in Fig. 2. Specifically, the assumptions are that (i) for each speaker there is a speaker-dependent supervector $\boldsymbol{s}$ such that if $\boldsymbol{M}$ is the supervector corresponding to a given recording of the speaker then

$$\boldsymbol{M} = \boldsymbol{s} + \boldsymbol{c} \tag{3}$$

where $\boldsymbol{c}$ depends only on the channel effects in the recording and that (ii) channel effects can be treated as discrete and identified in a pre-processing step, just as in speaker model synthesis. Given an enrollment utterance for a target speaker, the speaker supervector $\boldsymbol{s}$ can be estimated by subtracting the appropriate channel supervector for the enrollment utterance from the data; adding the appropriate channel supervector for a given test utterance to $\boldsymbol{s}$ gives a channel-adapted supervector which can be used to test the hypothesis that the speaker in the test utterance is the target speaker. The joint factor analysis model in [1] also takes (3) as its starting point but, as in eigenchannel modeling, it treats channel supervectors as continuous rather than discrete and does away with the need for channel detection in a pre-processing step. Thus we assume in this case as well that the channel compensation supervectors $\boldsymbol{c}$ in Fig. 2 can be written in the form

$$\boldsymbol{c} \quad = \quad \boldsymbol{u}\boldsymbol{x} \tag{4}$$

where $\boldsymbol{u}$ is a rectangular matrix of low rank and $\boldsymbol{x}$ has a standard normal distribution. In the terminology of [1], $\boldsymbol{u}$ is a factor loading matrix and the components of $\boldsymbol{x}$ are channel factors.

There are several ways of constructing a likelihood function for using joint factor analysis models in speaker recognition [4]. We will propose a new likelihood function in this article

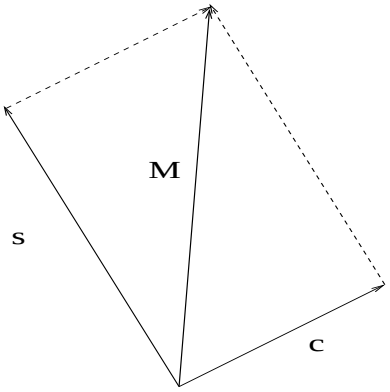

Fig. 2. *Feature mapping and joint factor analysis of speaker and channel variability are based on a decomposition of the form* $\boldsymbol{M} = \boldsymbol{s} + \boldsymbol{c}$*.* $\boldsymbol{M}$ *is the speaker- and channel-dependent supervector for a given recording,* $\boldsymbol{s}$ *depends only on the speaker and* $\boldsymbol{c}$ *depends only on the channel.*

which can be viewed as a compromise between the 'exact' and 'simplified' scoring rules presented in [3].

Comparing the eigenchannel approach to channel compensation with the joint factor analysis model we see that the eigenchannel model is weaker in that it only compensates for channel effects in test utterances whereas the joint factor analysis model handles channel effects in enrollment utterances as well. On the other hand it is much easier to implement. (Similar remarks apply to speaker model synthesis *vis-à-vis* feature mapping.) Although it is relatively straightforward to carry out the decomposition (3) in feature mapping (because the channel supervector $\boldsymbol{c}$ can only take a discrete range of values), the problem of disentangling speaker and channel effects in the joint factor analysis model is much more difficult (since both $\boldsymbol{s}$ and $\boldsymbol{c}$ are assumed to be continuous). A solution derived by formulating the problem as one of calculating the posterior distribution of the hidden variables in the joint factor analysis model is presented in Section III-C of [1] and summarized in the appendix.

Finding an appropriate training set for the joint factor analysis model also seems to be problematic if the model is to be evaluated on the NIST 2005 test set. Like the NIST 2004 test set, this was specifically designed to evaluate the robustness of speaker recognition systems to gross channel mismatches whereas previous test sets had been taken from individual Switchboard corpora where speakers were recorded either over landline or cellphone channels but not both. Speakers in the 2004 and 2005 test sets were recorded using different types of microphone (speaker phone, head mounted and ear bud as well as regular and cellphone handsets) and transmission channel (cordless, landline and cellular). As we mentioned in the introduction, neither the Fisher corpus nor any of the Switchboard corpora is really suitable for modeling this type of variability; the only publicly available data that seems to fit the bill is the 2004 evaluation data but this is a relatively small set (consisting of just 310 target speakers). On the other hand, eigenchannel modeling imposes less stringent requirements on the training set; all that seems to be required is that for any given pair of channel conditions there should be at least some training speakers recorded under both conditions.

Note that in this brief discussion we have only touched on model-based methods for channel compensation of GMM's and not on score or feature normalization methods or the parallel developments in SVM-based speaker recognition [12]. It is interesting to note that although the approach in SVM-based speaker recognition is discriminative rather than generative and it is concerned with finding feature representations which are immune to channel variability rather than modeling this type of variability, the key algorithm in 'nuisance attribute projection' is also formulated as an eigenvalue problem.

III. Models of Speaker and Utterance Variability

In the case of the joint factor analysis model, the first term in the right hand side of (3) is modeled by assuming that if $\boldsymbol{s}$ is the speaker supervector for a randomly chosen speaker then

$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{v}\boldsymbol{y} + \boldsymbol{d}\boldsymbol{z} \tag{5}$$

where $\boldsymbol{m}$ is the speaker- and channel-independent supervector, $\boldsymbol{v}$ is a rectangular matrix of low rank, $\boldsymbol{d}$ is a diagonal matrix and $\boldsymbol{y}$ and $\boldsymbol{z}$ are random vectors having standard normal distributions. This is a factor analysis in the sense of [13]. It is usual to refer to the elements of $\boldsymbol{y}$ simply as factors but the terminology of [14] is also useful: the elements of $\boldsymbol{y}$ are 'common factors' (because each of them serves to account for the variance in all of the elements of $\boldsymbol{s}$) and the elements of $\boldsymbol{z}$ are 'specific factors'. In the absence of the specific factors, (5) implies that all supervectors are contained in the linear span of $\boldsymbol{m}$ and the columns of $\boldsymbol{v}$ which is the assumption of eigenvoice modeling. In practice, the common factors account for most of the variance in the data and the term $\boldsymbol{d}\boldsymbol{z}$ serves as a residual to compensate for the fact that the eigenvoice assumption may be unrealistic and it may be difficult to find enough training speakers to estimate $\boldsymbol{v}$ reliably.

In the case of eigenchannel modeling, we model the first term in the right hand side of (1) by assuming that if $\boldsymbol{M}$ is the speaker- and channel-dependent supervector for a randomly chosen utterance then

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{v}\boldsymbol{y} + \boldsymbol{d}\boldsymbol{z} \tag{6}$$

where the terms on the right hand side of this equation are subject to the same conditions as in (5). Note that (6) is a model of *utterance* variability whereas (5) is a model of *speaker* variability since $\boldsymbol{M}$ is sensitive to channel effects whereas $\boldsymbol{s}$ is not. (In the core condition of the NIST evaluations 'utterances' are whole conversation sides so using the term utterance here is a bit of a stretch but it is convenient to do so.) Put another way, we treat all utterances by a given speaker as being statistically independent in (6) but not in (5).

The role of these models of speaker and utterance variability in our work is to provide a prior distribution for MAP estimation of GMM supervectors for target speakers. This type of MAP estimation combines classical MAP [15] and eigenvoice MAP [9] whose strengths and weaknesses complement each other. Classical MAP estimation of GMM's requires large amounts of enrollment data and because the matrix $\boldsymbol{d}$ is of full rank it is guaranteed to be asymptotically equivalent to speaker-dependent training; because $\boldsymbol{v}$ is of low rank there is

no such guarantee for eigenvoice MAP but, by the same token, eigenvoice MAP can use small amounts of enrollment data to good advantage. The earliest instance of factor analysis MAP adaptation in the literature on speaker recognition is [16]. Our treatment is different from [16] in that we use a likelihood criterion similar to that of [9] to estimate the hyperparameters that specify the prior. (The estimation procedure is summarized in the appendix.)

IV. Fitting the Models

In the previous sections we outlined two ways of modeling the two types of variability which are of interest in speaker recognition, namely speaker variability and session variability. One approach, known as joint factor analysis, is to model speaker variability by (5) and session variability by (4). The other is to model utterance variability by (6) and session variability by (2). It will be convenient henceforth to use the term 'eigenchannel modeling' to refer to the latter combination (rather than using this term to refer merely to equation (2)).

We now describe how we estimated eigenchannel and joint factor analysis models using about 1000 hours of speech (exclusive of silences) consisting of whole conversation sides extracted from the following databases: the LDC releases of Switchboard II, Phases 1, 2 and 3; Switchboard Cellular, Parts 1 and 2; the Fisher English Corpus, Part 1 and the NIST 2004 evaluation data. For most of our experiments we took these models to be gender-dependent (that is, one model for male speakers and another for female speakers) rather than gender-independent. We will describe the quantities of data that we used for eigenchannel modeling in the female case; the figures for the other cases (female factor analysis, female eigenchannel modeling and male eigenchannel modeling) are similar (but somewhat smaller).

Estimating eigenchannels and the parameters of the corresponding model of utterance variability (6) is relatively straightforward; most of the mathematics needed is presented in [9]. Exact algorithms for training the joint factor analysis model are presented in [1] (Theorems 4, 5 and 7) but these are not easy to implement and they are computationally demanding. Fortunately, these difficulties can be avoided at the cost of a very slight degradation in performance by some simple approximations at least if the training set is well balanced in the sense that the recordings for each training speaker are sufficiently numerous and diverse that channel effects can be averaged out by pooling the recordings for each speaker. (Compare the results obtained with 'exact' and 'simplified' training in [3].[1]) These approximations lead to the simplified training algorithms described in the appendix which enable the joint factor analysis model to be trained using the same software that serves for eigenchannel modeling. We begin by describing the acoustic features that we used and then we sketch the training algorithms for the eigenchannel and joint factor analysis models.

[1]In implementing the simplified training algorithm for the experiments reported here, we skipped the 'adaptation to the target speaker population' step mentioned in Section 3 of [3] in order to adhere strictly to the NIST evaluation protocol.

A. Feature extraction

Using a 25 ms Hamming window, 12 mel frequency cepstral coefficients together with a log energy feature are calculated every 10 ms. These 13-dimensional feature vectors are subjected to feature warping [17] using a 3 s sliding window. (There is very strong synergy between feature warping and our techniques for modeling session variability [4].) Delta coefficients are then calculated using a 5 frame window giving a 26-dimensional feature vector.

Where available, ASR transcripts containing time stamps are used to suppress silence intervals. In other cases the ISIP voice activity detector is used [18].

First and second order Baum-Welch statistics are extracted from the non-silence portions of the speech signal using a standard universal background model (UBM). We regard this as a pre-processing step since we use no information about the speech signal other than that which is encoded in these statistics.

We used 5719 conversation sides (278 hours of data after removing silences) from as many speakers to train a female GMM with 2,048 mixture components and diagonal covariance matrices which serves as a universal background model. Let C denote the number of mixture components in the GMM and F the dimensionality of the acoustic feature vectors (so that $C = 2048$ and $F = 26$).

B. Speaker and utterance variability

Note first that the speaker variability model (5) reduces to the eigenvoice model in [9] if $\boldsymbol{d} = \boldsymbol{0}$. So if $\boldsymbol{m}$ is given (the UBM supervector is a natural choice), $\boldsymbol{v}$ can be estimated by the algorithm described in Proposition 3 of [9]. (This is a version of probabilistic principal components analysis designed to work with Baum-Welch statistics rather than with point estimates of utterance supervectors as in conventional probabilistic principal components analysis [19].) The extensions needed to handle the general case ($\boldsymbol{d} \neq \boldsymbol{0}$) are described in the appendix. (The training algorithms presented there are derived from Theorems 4, 5 and 7 of [1] by taking $\boldsymbol{u} = \boldsymbol{0}$ in the statement of each theorem).

In implementing this training procedure, the Baum Welch statistics extracted from all recordings of each training speaker are pooled together (as if there were just one recording per speaker). In order to estimate the hyperparameters $\boldsymbol{m}, \boldsymbol{v}$ and $\boldsymbol{d}$ in the utterance variability model (6), the only modification that is needed is to apply the procedure *without* pooling the Baum-Welch statistics.

The training set that we used to fit the utterance variability model (6) consisted of 9,291 conversation sides (393 hours of speech exclusive of silences) in the female case. The results of fitting a model with 25 common factors to this training set are illustrated in Fig. 3 which is a plot of the eigenvalues corresponding to the non-zero eigenvectors of $\boldsymbol{v}\boldsymbol{v}^*$. Note that the eigenvalues decrease exponentially. The relative importance of the special and common factors can be measured by comparing the expected values of $\|\boldsymbol{v}\boldsymbol{y}\|^2$ and $\|\boldsymbol{d}\boldsymbol{z}\|^2$. Since $\boldsymbol{y}$ and $\boldsymbol{z}$ have standard normal distributions,

these expected values are given by the following matrix traces:

$$\mathrm{tr}\left(\boldsymbol{d}^2\right) \quad = \quad 77.26 \tag{7}$$

$$\mathrm{tr}\left(\boldsymbol{v}\boldsymbol{v}^*\right) \quad = \quad 730.48. \tag{8}$$

This suggests that essentially all of the utterance variability in the training data could be accounted for by increasing the number of common factors which motivated us to experiment with some model configurations having a very large number of common factors. (The eigenvalue plots and traces in the male case are similar.)

We note in passing that it has been our experience that in the case where $\boldsymbol{v}$ is set to $\mathbf{0}$ (the situation in classical MAP), estimating $\boldsymbol{d}$ by a maximum likelihood criterion does not give a better estimate than the empirical method in [15]. This is consistent with the observation in [15] that the effectiveness of relevance MAP is insensitive to the value of the relevance factor.

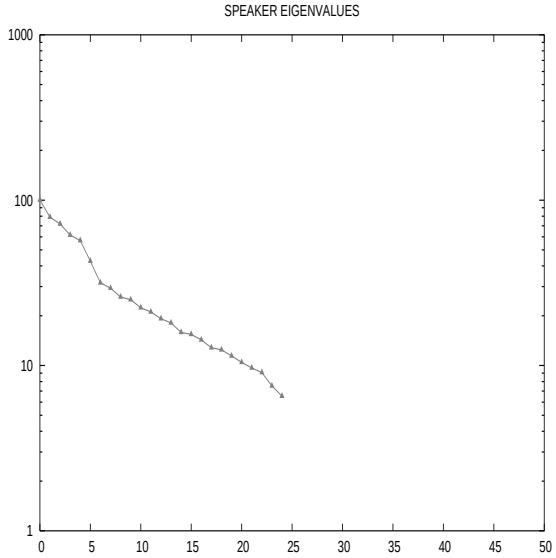

Fig. 3. *Eigenvalues obtained by modeling utterance variability with 25 common factors. Female data. Compare with Fig. 4. The graph in the male case is similar.*

C. Session variability

We now turn to the problem of estimating the rectangular matrix $\boldsymbol{u}$ in (2) and (4) which serves as our model of session variability, both in the joint factor analysis model and in the eigenchannel model.

Combining (4) and (5), we can summarize the joint factor analysis model by writing

$$\boldsymbol{M} \quad = \quad \boldsymbol{s} + \boldsymbol{u}\boldsymbol{x}. \tag{9}$$

This is formally almost the same as (5) with $\boldsymbol{d}$ set to $\mathbf{0}$ which suggests that $\boldsymbol{u}$ can be estimated by the same methods as those used to estimate $\boldsymbol{v}$ in [9]. Two complications arise however. Firstly, the speaker supervector $\boldsymbol{s}$ in (9) varies from one speaker to another whereas the supervector $\boldsymbol{m}$ in (5) is speaker-independent. This can be dealt with by converting the Baum-Welch statistics for each utterance by a given speaker into first and second moments about $\boldsymbol{s}$ rather than first and second moments about $\boldsymbol{m}$ as was done in [9]. Secondly, for a given speaker, the speaker supervector $\boldsymbol{s}$ has to be estimated from the data available for the speaker. Under these conditions a point estimate of $\boldsymbol{s}$ may be unreliable and, as we showed in [3] it may be advantageous to take account of the uncertainty entailed in estimating $\boldsymbol{s}$ with limited amounts of data. We will explain in Section V-B how this complication can be dealt with by a slight modification of the Baum-Welch statistics for each of the speaker's utterances.

In the case of eigenchannel modeling, the rectangular matrix $\boldsymbol{u}$ in (2) can be estimated in much the same way. Recall that the basic assumption here is that if we have a pair of conversation sides for a given speaker with corresponding speaker- and channel-dependent supervectors $\boldsymbol{M}$ and $\boldsymbol{M}'$ then

$$\boldsymbol{M}' \quad = \quad \boldsymbol{M} + \boldsymbol{u}\boldsymbol{x} \tag{10}$$

where $\boldsymbol{x}$ has a standard normal distribution. Suppose we are given a training set consisting of a suitably large collection of pairs of utterances by different speakers. We can use one of the utterances in each pair (we chose the longer of the two) to estimate the supervector $\boldsymbol{M}$ appearing in the right hand side of (1). This plays the same role as the estimate of the speaker supervector $\boldsymbol{s}$ in (9) so the procedure for estimating $\boldsymbol{u}$ is formally the same as in the case of the joint factor analysis model.

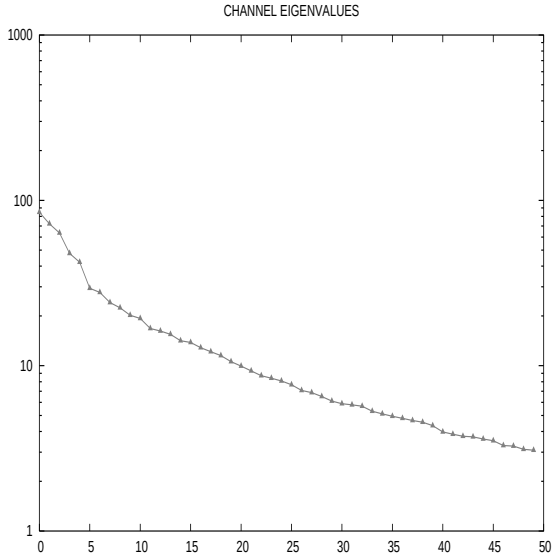

Fig. 4. *Eigenvalues obtained by fitting an eigenchannel model of rank 50. Female data. Note that the decrease in the eigenvalues is approximately exponential. The graph in the male case is similar.*

The training set that we used to estimate $\boldsymbol{u}$ in (2) consisted of 27,399 utterance pairs in the female case. (The training set that we used to estimate the utterance variability model in Section IV-B was obtained by choosing the longer utterance in each of these pairs.) The results of fitting a model of rank 50 to this training set are illustrated in Fig. 4 which is a plot of the eigenvalues corresponding to the non-zero eigenvectors of $\boldsymbol{u}\boldsymbol{u}^*$ sorted in decreasing order. (The eigenvalue plot in the male case is similar.) The fact that the decrease is approximately

exponential means that only a small fraction of the channel variability will be lost if channel supervectors are expressed in terms of the eigenchannels and the expansion is cut off after a finite number of terms. This provides empirical justification for the assumption that channel variability is intrinsically low dimensional.

Estimating the matrix $\boldsymbol{u}$ in (4) (that is, the joint factor analysis case) yields a similar picture with one critical difference, namely that the sum of the eigenvalues in this case is about half of the sum of the eigenvalues in Fig. 4. This is undoubtedly the reason why joint factor analysis outperformed eigenchannel modeling in our experiments.

Comparing Figs. 1 and 2 it is easy to see why this difference should arise. The supervectors $\boldsymbol{M}$ and $\boldsymbol{M}'$ in Fig. 1 are both contaminated by channel effects whereas the supervector $\boldsymbol{s}$ in Fig. 2 is not. Thus the variance of the channel compensation supervectors (in other words, the sum of the eigenvalues of $\boldsymbol{u}\boldsymbol{u}^*$) in Fig. 1 is twice as large as the variance of the channel compensation supervectors in Fig. 2. Of course this argument assumes that an uncontaminated speaker supervector $\boldsymbol{s}$ can be estimated for each speaker in the case of the factor analysis model. We used a very crude approximation to achieve this goal in training the joint factor analysis model — we simply pooled the Baum-Welch statistics over all recordings of each speaker. This works (compare the results obtained with the 'exact' and 'simplified' training algorithms in [3]) because we chose our training set so that there were sufficiently many recordings of each training speaker that channel effects could be averaged out in this way. However the situation is quite different when it comes to enrolling a target speaker for speaker recognition because there is just *one* recording for each target speaker in the core condition of the NIST evaluations. In fact it is *only* in the enrollment phase that the implementations of the joint factor analysis model and the eigenchannel model require different software.

V. Using the Models to Build Speaker Verification Systems

In this section we explain how we use the joint factor analysis model and the eigenchannel model to construct speaker verification systems. We have to describe how we estimate a GMM supervector for each target speaker, how we evaluate the likelihood of a test utterance using a target speaker GMM and how we normalize likelihoods calculated in this way so that a common decision threshold can be used in all speaker verification trials. We will concentrate on the joint factor analysis model, indicating the modifications that need to be made in the case of the eigenchannel model as we go along. The decision criterion that we will propose for speaker verification with the joint factor analysis model in this article is a compromise between the 'exact' and 'simplified' scoring methods in [3]. (In the terminology of [4], the 'exact' scoring method refers to the sequential likelihood ratio statistic.)

A. Enrolling a target speaker

In using the joint factor analysis model for speaker recognition, the key calculation in enrolling a target speaker is to disentangle the speaker and channel effects in the enrollment utterance, that is, to estimate the speaker's supervector $\boldsymbol{s}$ by carrying out the decomposition (2). In [1] we showed how to formulate this problem as one of calculating the posterior distribution of the hidden variables in the factor analysis model, namely $\boldsymbol{x}$ in (4) and $\boldsymbol{y}$ and $\boldsymbol{z}$ in (5). This calculation is described in detail in Section III of [1]. (The treatment is general enough to handle extended data tasks where there are multiple enrollment recordings for each target speaker, rather than just a single enrollment recording as in the core condition of the 2005 evaluation).

Since the amount of enrollment data for a target speaker is limited, a point estimate of the speaker's supervector $\boldsymbol{s}$ derived in this way may not be reliable. Comparing the results obtained with the 'exact' and 'simplified' scoring methods in [3] shows that ignoring this type of uncertainty can degrade speaker recognition performance. Unfortunately the exact scoring method that we used in [4] is very computationally expensive. To deal with this problem we propose a new way of addressing the uncertainty issue in this article which is no more computationally expensive than the simplified scoring method. This entails calculating not only the posterior expectation of the target speaker's supervector $\boldsymbol{s}$ at enrollment time (which, following the notation in [1], we denote by $E[\boldsymbol{s}]$) but also the diagonal of posterior covariance matrix of $\boldsymbol{s}$ (which we denote by $\mathrm{Cov}(\boldsymbol{s},\boldsymbol{s})$). Both of these can be calculated using the methods in [1].[2]

It has been our experience that it in the case where $\boldsymbol{v}=\boldsymbol{0}$, $\mathrm{Cov}(\boldsymbol{s},\boldsymbol{s})$ is invariably quite large (typically about 75% of the total speaker variability $\boldsymbol{d}^2$). On the other hand, in the case of a pure eigenvoice model ($\boldsymbol{d}=\boldsymbol{0}$), this uncertainty is quite small (since enrolling a target speaker entails estimating only as many free parameters as there are eigenvoices). In general, for any configuration of the joint factor analysis model, $\mathrm{Cov}(\boldsymbol{s},\boldsymbol{s})$ will be largest for target speakers with the least amount of training data. As we shall see, incorporating this term into the scoring mechanism for speaker recognition provides a way for penalizing hypothesized speakers with small amounts of enrollment data.

In using the eigenchannel model for speaker recognition, enrolling a target speaker is much simpler. We use the prior distribution (6) to calculate the posterior distribution of the speaker- and channel-dependent supervector for the enrollment utterance, $\boldsymbol{M}$. In the case $\boldsymbol{d}=\boldsymbol{0}$, the calculation is described in Proposition 1 of [9] and its corollaries; a modification is needed to handle the general case. (Take $\boldsymbol{u}=\boldsymbol{0}$ in Section III-C of [1].)

B. The likelihood function

Again we consider the factor analysis model first. Suppose we are given a target speaker and a test utterance and that

[2]Although we calculated $\mathrm{Cov}(\boldsymbol{s},\boldsymbol{s})$ exactly for the experiments reported here, it has been our experience that the approximation

$$\mathrm{Cov}(\boldsymbol{s},\boldsymbol{s}) \simeq \mathrm{diag}\left(\boldsymbol{v}\,\mathrm{Cov}(\boldsymbol{y},\boldsymbol{y})\,\boldsymbol{v}^*\right) + \boldsymbol{d}\,\mathrm{Cov}(\boldsymbol{z},\boldsymbol{z})\,\boldsymbol{d} \tag{11}$$

which ignores the cross correlations between $\boldsymbol{y}$ and $\boldsymbol{z}$ works quite well in practice. This is easy to implement and it gives exact results in the two cases which are of greatest interest, namely $\boldsymbol{d}=\boldsymbol{0}$ and $\boldsymbol{v}=\boldsymbol{0}$.

we wish to test the null hypothesis that the utterance speaker is different from the target speaker against the alternative hypothesis that the two speakers are the same. Denote the speaker supervector for the target speaker by $\boldsymbol{s}$ and denote the test utterance by $\mathcal{X}$.

If we assume to begin with that $\boldsymbol{s}$ is known the likelihood of $\mathcal{X}$ under the alternative hypothesis — let us denote it by $P(\mathcal{X}|\boldsymbol{s})$ — can be calculated by the methods in [9]. By (4) there is a random vector $\boldsymbol{x}$ such that the speaker- and channel-dependent supervector for the test utterance is

$$\boldsymbol{s} + \boldsymbol{u}\boldsymbol{x}. \tag{12}$$

If $\boldsymbol{x}$ was known, we would know the value of this supervector so it would be straightforward matter to calculate the conditional (Gaussian) likelihood of the test utterance, $P(\mathcal{X}|\boldsymbol{s},\boldsymbol{x})$, using the Baum-Welch statistics extracted from the utterance (Lemma 1 in [9]). So, since $\boldsymbol{x}$ is assumed to have a standard normal distribution, $P(\mathcal{X}|\boldsymbol{s})$ is given by

$$P(\mathcal{X}|\boldsymbol{s}) \;=\; \int P(\mathcal{X}|\boldsymbol{s},\boldsymbol{x})N(\boldsymbol{x}|\boldsymbol{0},\boldsymbol{I})d\boldsymbol{x} \tag{13}$$

where $N(\cdot|\boldsymbol{0},\boldsymbol{I})$ is the standard Gaussian kernel. Proposition 2 in [9] explains how to derive a closed form expression for this type of integral so we will simply state the result here in a form which is appropriate for t-norm score normalization.

First some notation. For each mixture component c, let Σ_c be the corresponding $F \times F$ covariance matrix; we take this to be diagonal and assume that it is speaker- and channel independent. Let $\boldsymbol{\Sigma}$ be the $CF \times CF$ covariance matrix whose diagonal blocks are Σ_c ($c = 1,\ldots,C$). Let N_c be the total number of observation vectors in $\mathcal{X}$ for the given mixture component and set

$$F_c \;=\; \sum_t X_t \tag{14}$$

$$S_c \;=\; \mathrm{diag}\left(\sum_t X_t X_t^*\right) \tag{15}$$

where the sum extends over all observations X_t aligned with the given mixture component, and $\mathrm{diag}\,()$ sets off-diagonal entries to 0. (As we have written them these are Viterbi statistics but we use Baum-Welch statistics in practice. We use gender dependent UBM's to extract these statistics in all cases.) Let $\boldsymbol{N}$ be the $CF \times CF$ diagonal matrix whose diagonal blocks are $N_c I$ (for $c = 1,\ldots,C$) where I is the $F \times F$ identity matrix. Let $\boldsymbol{F}$ be the $CF \times 1$ vector obtained by concatenating F_c (for $c = 1,\ldots,C$). Similarly, let $\boldsymbol{S}$ be the $CF \times CF$ diagonal matrix whose diagonal blocks are S_c (for $c = 1,\ldots,C$). We denote the first and second order moments of $\mathcal{X}$ around $\boldsymbol{s}$ by $\boldsymbol{F_s}$ and $\boldsymbol{S_s}$ so that

$$\begin{aligned} \boldsymbol{F_s} &= \boldsymbol{F} - \boldsymbol{N}\boldsymbol{s} \\ \boldsymbol{S_s} &= \boldsymbol{S} - 2\,\mathrm{diag}\,(\boldsymbol{F}\boldsymbol{s}^*) + \mathrm{diag}\,(\boldsymbol{N}\boldsymbol{s}\boldsymbol{s}^*)\,. \end{aligned} \tag{16}$$

Finally, let

$$\boldsymbol{l} \;=\; \boldsymbol{I} + \boldsymbol{u}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{u}, \tag{17}$$

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

$$\boldsymbol{l} \;=\; \boldsymbol{l}^{1/2}\boldsymbol{l}^{1/2*} \tag{18}$$

(that is, the Cholesky decomposition of $\boldsymbol{l}$). Then applying some algebraic manipulations to the formula given in the statement of Proposition 2 in [1] leads to the following expression for the likelihood function:

$$\begin{aligned} \log P(\mathcal{X}|\boldsymbol{s}) \;=\; & \sum_{c=1}^{C} N_c \log \frac{1}{(2\pi)^{F/2}|\Sigma_c|^{1/2}} \\ & - \frac{1}{2}\,\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S_s}\right) - \frac{1}{2}\log|\boldsymbol{l}| \\ & + \frac{1}{2}\|\boldsymbol{l}^{-1/2}\boldsymbol{u}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{F_s}\|^2 \end{aligned} \tag{19}$$

provided that $\boldsymbol{s}$ is known. In practice $\boldsymbol{s}$ has to be estimated from the enrollment data for the hypothesized speaker so we replace $\boldsymbol{F_s}$ and $\boldsymbol{S_s}$ by their posterior expectations, $E\,[\boldsymbol{F_s}]$ and $E\,[\boldsymbol{S_s}]$, which are given by

$$\begin{aligned} E\,[\boldsymbol{F_s}] \;=\; & \boldsymbol{F} - \boldsymbol{N}E\,[\boldsymbol{s}] \\ E\,[\boldsymbol{S_s}] \;=\; & \boldsymbol{S} - 2\,\mathrm{diag}\,(\boldsymbol{F}E\,[\boldsymbol{s}^*]) \\ & + \mathrm{diag}\big(\boldsymbol{N}(E\,[\boldsymbol{s}]\,E\,[\boldsymbol{s}^*] \\ & + \mathrm{Cov}\,(\boldsymbol{s},\boldsymbol{s}))\big) \end{aligned} \tag{20}$$

(in accordance with the notation introduced in Section V-A). Because the term $\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S_s}\right)$ enters into (19) with a negative sign, the effect of including the term $\mathrm{Cov}\,(\boldsymbol{s},\boldsymbol{s})$ in (20) is to diminish the value of the likelihood function by an amount which is inversely proportional to the amount of the speaker's enrollment data. (In order to ensure that the same criterion is used in training and testing we incorporate a similar modification to the Baum-Welch statistics in estimating the factor analysis model.)

The most interesting thing to note about (19) is that the likelihood function depends on the hypothesized speaker only through the computations in (20) and the cost of these computations is negligible (since $E\,[\boldsymbol{s}]$ and $\mathrm{Cov}\,(\boldsymbol{s},\boldsymbol{s})$ are calculated at enrollment time). The principal computation is the evaluation of $\boldsymbol{l}^{-1/2}$ (the value of the determinant $|\boldsymbol{l}|$ is a by-product) and this only needs to be done once (independently of the number of speakers hypothesized and the number of t-norm speakers). Note also that the number of common speaker factors has a major effect on the computational burden of evaluating the terms $E\,[\boldsymbol{s}]$ and $\mathrm{Cov}\,(\boldsymbol{s},\boldsymbol{s})$ in (20) but these terms are only evaluated at enrollment time. The calculation in (19) is completely insensitive to the number of common speaker factors; this is another major advantage over the decision criterion used in [4].

In the case of the eigenchannel model, we formulate the hypothesis test in a slightly different way but the calculations are formally identical. Suppose we are given two utterances and we wish to test the null hypothesis that they were uttered by different speakers against the alternative hypothesis that they were both uttered by the same speaker. We designate one of the utterances (the longer of the two in our implementation) as the enrollment utterance and the other as the test utterance. Denote the test utterance by $\mathcal{X}$, let $\boldsymbol{M}$ be the speaker- and channel-dependent supervector for the enrollment utterance

and let $E[\boldsymbol{M}]$ and $\mathrm{Cov}(\boldsymbol{M},\boldsymbol{M})$ be the corresponding posterior mean and covariance matrix (calculated in the enrollment phase). Then the likelihood of $\mathcal{X}$ under the alternative hypothesis can be calculated in exactly the same way as $P(\mathcal{X}|\boldsymbol{s})$ in (19) by replacing $\boldsymbol{s}$ by $\boldsymbol{M}$ throughout.

C. Score normalization

In our first experiments in the present article, as in [4], [3], we used only t-norm for score normalization but we learned from [10] that zt-norm (that is, z-norm followed by t-norm and not the other way round) could be very effective for the type of model under consideration at least in the case where $\boldsymbol{v} = \boldsymbol{0}$. Unlike t-norm, z-norm requires a way of evaluating the likelihood of a test utterance under the assumption that the actual speaker is somebody other than the hypothesized speaker — the 'unknown speaker' as it were. (This is the likelihood which appears in the denominator of the log likelihood ratio used in Neyman-Pearson style hypothesis tests. These denominators are redundant if t-norm is used for score normalization.)

We consider only the case of the joint factor analysis model. (The modifications needed to handle the eigenchannel model are self-evident.) The solution proposed in [10] is to take the speaker in the center of the acoustic space as the unknown speaker. That is the likelihood of a test utterance for the unknown speaker is evaluated in the same way as for a target speaker by taking

$$\begin{aligned} E[\boldsymbol{s}] &= \boldsymbol{m} \\ \mathrm{Cov}(\boldsymbol{s},\boldsymbol{s}) &= \boldsymbol{0} \end{aligned} \tag{21}$$

where $\boldsymbol{m}$ is the speaker- and channel-independent supervector in (5). However, since our likelihood function takes account of the uncertainty in the point estimate of a target speaker's supervector produced by the enrollment procedure, it is more natural for us to take the speaker for whom no enrollment data is available as the unknown speaker. This is tantamount to setting

$$\begin{aligned} E[\boldsymbol{s}] &= \boldsymbol{m} \\ \mathrm{Cov}(\boldsymbol{s},\boldsymbol{s}) &= \mathrm{diag}\left(\boldsymbol{v}\boldsymbol{v}^* + \boldsymbol{d}^2\right) \end{aligned} \tag{22}$$

where $\boldsymbol{v}$ and $\boldsymbol{d}$ are as in (5). Our experience has been that this is an effective way of evaluating the denominator of the likelihood ratio in the case where $\boldsymbol{v} = \boldsymbol{0}$ but we also found it necessary to experiment with the following variant:

$$\begin{aligned} E[\boldsymbol{s}] &= \boldsymbol{m} \\ \mathrm{Cov}(\boldsymbol{s},\boldsymbol{s}) &= \frac{1}{N}\sum_{n=1}^{N}\mathrm{Cov}(\boldsymbol{s}_n,\boldsymbol{s}_n) \end{aligned} \tag{23}$$

where the sum on the right hand side extends over the set of t-norm speakers and N is the number of t-norm speakers. We will refer these three versions of z-norm as z_1-norm (or 'z-norm without uncertainty'), z_2-norm and z_3-norm respectively. Our experience has been that z_1 is not really appropriate for the scoring procedure that we are using and that z_3-norm is more effective than z_2-norm if common speaker factors are included in the speaker variability model (5).

For most of our experiments we used 120 t-norm speakers for each gender and 120 z-norm utterances (20 from each of the databases that we used for development). In our experience there is little to be gained by increasing the number of t-norm speakers or z-norm utterances beyond this number.

VI. Experiments

All of the results we report are on the core condition of the NIST 2005 evaluation. We used all of the trials in this condition rather than the 'common' subset [5]. (In all there were 2771 target trials and 28, 472 non-target trials.) We report both equal error rates (EER) and the minimum values of the NIST detection cost function (DCF).

In all of experiments we used UBM's with 2048 Gaussians. Our first experiment was conducted with an eigenchannel model having 25 eigenvoices (EV), 50 eigenchannels (EC) and t-norm score normalization. This resulted in a EER of of 11.7% and a DCF of 0.042 which was disappointing considering the size of the databases that we used for development. This led us to experiment with several variants of the eigenchannel model but eventually we abandoned it in favor of the joint factor analysis model.

Before describing these experiments we make a brief remark about silence detection. Unlike most participants in the evaluation we used the time stamps provided by NIST to suppress silences in the enrollment and test utterances. This gave us about 25% more speech data to work with than a conventional silence detector. To evaluate the effect of this decision we reran our system using the ISIP silence detector and found that we obtained poorer results (an EER of 12.3% and a DCF of 0.045). Thus we did not use the silence detector in our subsequent experiments.

A. Eigenchannels

Our first series of experiments was designed to evaluate the effect of modifying the configuration of the eigenchannel model. The results are summarized in Table I which shows that our best results were obtained with a configuration of 5 eigenvoices and 25 eigenchannels. It is apparent that care is needed to avoid over fitting the utterance and session models in spite the large amounts of training data that we used. The benefit of adding eigenvoice MAP to classical MAP is not great (compare the last two lines of Table I) but this is perhaps not surprising since eigenvoice methods were developed to deal with situations where very little data is available for model adaptation (far less data than a whole conversation side).

TABLE I

Results obtained on all trials of the core condition of the NIST 2005 evaluation using various configurations of the eigenchannel model. T-norm score normalization.

EV	EC	EER	DCF
25	50	11.7%	0.042
5	50	11.7%	0.036
5	**25**	**10.2%**	**0.036**
0	25	11.7%	0.038

The experiments reported in Table I were conducted using only t-norm score normalization. We tested the other types of normalization strategies described in Section V-C on two model configurations: 5 eigenvoices and 25 eigenchannels (the best configuration according to Table I) and 0 eigenvoices and 25 eigenchannels (the configuration most similar to [10]). The results are summarized in Tables II and III. For each configuration the best results are obtained with zt-norm, confirming the results in [10]. In each case z-norm with uncertainty gives better results than z-norm without uncertainty in implementing zt-norm, as one might expect. However zt-norm is substantially more effective in the case where there are 0 eigenvoices than in the case of 5 eigenvoices. Thus it turns out that our best result with eigenchannel modeling, namely an EER of 8.7% and a DCF of 0.029, is obtained without using any eigenvoices contrary to what the results in Table I might suggest.

TABLE II

Eigenchannel model with 5 eigenvoices and 25 eigenchannels. Effect of different types of score normalization.

normalization	EER	DCF
t-norm	10.2%	0.036
z_1-norm	12.9%	0.056
z_2-norm	13.8%	0.047
z_1t-norm	10.9%	0.040
z_2t-norm	**9.5**%	**0.034**

TABLE III

Eigenchannel model with 0 eigenvoices and 25 eigenchannels. Effect of different types of score normalization. Compare with Table II.

normalization	EER	DCF
t-norm	11.7%	0.038
z_1-norm	12.1%	0.055
z_2-norm	9.9%	0.034
z_1t-norm	9.5%	0.034
z_2t-norm	**8.7**%	**0.029**

B. Joint factor analysis with no common speaker factors

Since we found that with the eigenchannel model zt-norm was much more effective with no eigenvoices than with 5 eigenvoices we set the number of common speaker factors to be 0 for our first experiments with the joint factor analysis model. The results are substantially better than the results we obtained with the eigenchannel model and they are reported in Table IV. With the eigenchannel model it was natural to treat the longer of the two utterances in a trial as the training utterance and the shorter as the test utterance but with the joint factor analysis model this distinction does not seem to be a natural one. So we carried out the trials in both the forward direction (that is, with the training and test utterance designations given by NIST) and in the reverse direction. The two strategies give essentially the same overall results but averaging the results gives a small improvement in DCF.

Most of the free parameters in this configuration of the joint factor analysis model are devoted to modeling session variability rather than speaker variability. Since session variability should be to a large extent gender independent we were interested to see if gender independent joint factor analysis model with this configuration could use the training data at our disposal to greater advantage. (Recall that one of our principal concerns with the factor analysis model was the paucity of Mixer type training data.) The results of these experiments are reported in Table V where a slight degradation in performance is apparent. (In performing score normalization in these experiments we used the same gender-dependent t-norm speaker sets and z-norm utterance sets that we used for the experiments presented in Table IV.)

TABLE IV

Joint factor analysis with no common speaker factors and 25 channel factors. F = forward, R = reverse. Compare with Table III.

normalization	trial type	EER	DCF
t-norm	F	11.5%	0.035
z_2-norm	F	7.8%	0.027
z_2t-norm	F	6.9%	0.022
z_2t-norm	R	6.4%	0.023
z_2t-norm	F + R	**6.6**%	**0.021**

TABLE V

Gender independent *joint factor analysis with no common speaker factors and 25 channel factors. F = forward, R = reverse. Compare with Table IV.*

normalization	trial type	EER	DCF
t-norm	F	10.9%	0.040
z_2-norm	F	8.2%	0.030
z_2t-norm	F	7.4%	0.025
z_2t-norm	R	6.8%	0.026
z_2t-norm	F + R	6.9%	0.024

So we concluded that even in the situation where the number of common speaker factors is set to 0, gender dependent modeling is the best strategy. Our best result with this type of configuration of the joint factor analysis model on the NIST 2005 test set, namely an EER of 6.2% and a DCF of 0.019, was obtained by gender dependent joint factor analysis modeling with no common speaker factors and 50 channel factors in the same way as the result in the last line of Table IV. The corresponding DET curve is shown in Fig. 5. (Increasing the number of channel factors from 50 to 100 gave essentially the same results.)

C. Joint factor analysis with 300 common speaker factors

So far we have only considered a special case of the joint factor analysis model, namely the case when the number of common speaker factors is zero. We now turn to the opposite extreme where the number of common factors is very large, namely 300. In this situation, the speaker variability model (5) behaves like a pure eigenvoice model (i.e. $\boldsymbol{d} \simeq \mathbf{0}$). The effects of various types of score normalization are shown in Table VI.

The general trend is that, just as we found in the case where the number of common speaker factors was 0, zt-norm is more

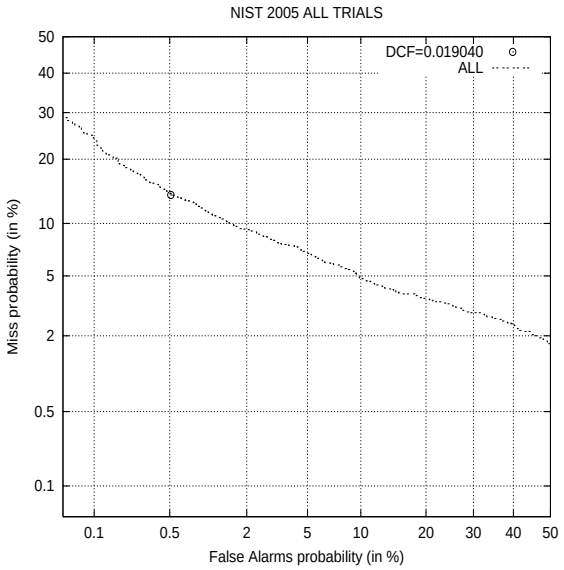

Fig. 5. *DET curve showing results on the NIST 2005 test set, all trials, core condition. Gender dependent joint factor analysis with no common speaker factors and 50 channel factors. Trials evaluated by averaging the results obtained in both forward and reverse directions.* z_2t *score normalization. EER = 6.2%, DCF = 0.019.*

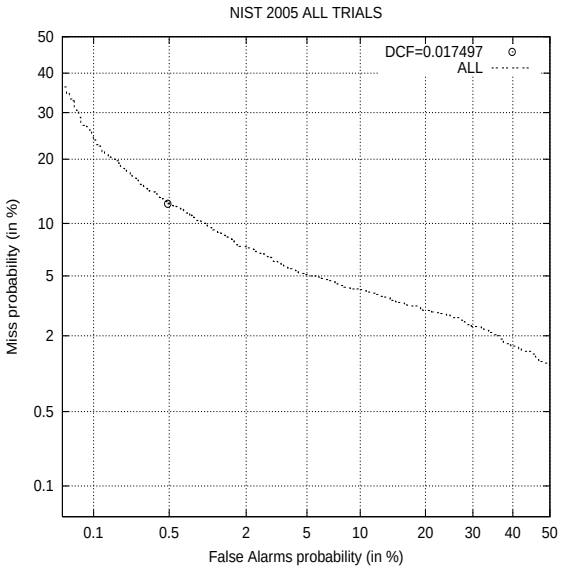

Fig. 6. *DET curve showing results on the NIST 2005 test set, all trials, core condition. Gender dependent joint factor analysis with 300 common speaker factors and 50 channel factors. Trials evaluated in the reverse direction.* z_3t *score normalization. EER = 5.2%, DCF = 0.017. Compare with Fig. 5.*

TABLE VI

Joint factor analysis with 300 common speaker factors and 50 channel factors.

normalization	trial type	EER	DCF
t-norm	F	9.0%	0.034
t-norm	R	8.9%	0.033
z_2-norm	F	9.2%	0.033
z_2-norm	R	7.5%	0.030
z_3-norm	F	6.8%	0.026
z_3-norm	R	6.2%	0.023
z_2t-norm	F	7.4%	0.023
z_2t-norm	R	6.7%	0.022
z_2t-norm	F+R	6.6%	0.020
z_3t-norm	F	5.4%	0.018
z_3t-norm	R	**5.2**%	**0.017**
z_3t-norm	F+R	5.3%	0.017

effective than z-norm and z-norm is more effective than t-norm but in this situation the z_3 flavor of z-norm is more effective than z_2. Note that the best result in Table VI (EER = 5.2%, DCF = 0.017) is considerably better than the best result we obtained with no common speaker factors (EER = 6.2%, DCF = 0.019 in Fig. 5). The DET curve corresponding to the best result is shown in Fig. 6.

D. Varying the number of common speaker factors

So far we have only considered the two extreme cases where the number of common speaker factors is zero or very large. Results obtained with different numbers of common speaker factors and 50 channel factors are reported in Table VII. Adding a small number of speaker factors (1 or 5) is seen to hurt performance particularly as measured by the the DCF.

Our reason for developing the speaker variability model (5) was to try to take advantage of the complementary strengths

TABLE VII

Joint factor analysis with varying numbers of common speaker factors and 50 channel factors. Forward scoring only. z_3t *score normalization.*

Common Speaker Factors	EER	DCF
0	6.8%	0.021
1	7.1%	0.029
5	7.3%	0.036
20	6.9%	0.029
100	5.8%	0.020
300	5.4%	0.018

of classical MAP and eigenvoice MAP in estimating target speaker models from limited amounts of data. But the results in Table VII show that the best performance is obtained in the two extreme cases where $\boldsymbol{d} = \mathbf{0}$ and $\boldsymbol{v} = \mathbf{0}$ and this suggests that fusion at the score level may be the best strategy for achieving this goal. It turns out that a linear fusion of four systems, namely forward and reverse scoring with 0 common speaker factors and 300 common speaker factors does indeed give a slightly improved value of the detection cost function (0.016 versus 0.017) when compared with the result in Fig. 6 but there is a slight degradation in the equal error rate (5.4% versus 5.2%). Thus more sophisticated fusion techniques such as logistic regression or a multilayer perceptron may be worth investigating.

VII. Discussion

The NIST 2005 test set presents an interesting challenge for the joint factor analysis model because there is reason to doubt that the model can be properly trained for this task using currently available telephone speech corpora. The results we have presented far surpass our initial expectations which were so pessimistic that we thought that the more primitive eigenchannel model stood a better chance of working.

The success of our approach is due in large part to the effectiveness of the zt-norm technique [10] and to the new scoring procedure described in Section V-B which enabled us to turn around a large number of experiments very quickly because of the efficiency with which it handles t-norm speakers. A remarkable feature of this scoring procedure is that its computational cost is independent of the number of common speaker factors in the factor analysis model. This enabled us to experiment with large numbers of common speaker factors and obtain some excellent results. In the the extreme case where the number of common speaker factors is very large (e.g. 300), the factor analysis model of speaker variability behaves essentially like an eigenvoice model ($\boldsymbol{d} \simeq \mathbf{0}$). It may be that the reason why this model performs so well is that it implicitly models long term features. (Eigenvoice methods take account of the correlations between the various Gaussians in a speaker GMM.)

It is rather surprising that it was possible to train this configuration of the factor analysis model with a training set which consisted of only a few hundred speakers (500 in the male case and 700 in the female case). It is also interesting to note that since the number of free parameters that have to be estimated in order to enroll a target speaker with an eigenvoice model is far less than with classical MAP, it may be that the methods presented here will prove to be effective with smaller amounts of enrollment data than have traditionally been provided in the NIST evaluations.

VIII. Acknowledgments

The authors would like to thank Robbie Vogt and especially Niko Brümmer for numerous stimulating and fruitful discussions.

References

[1] 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/

[2] D. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. ICASSP*, Hong Kong, China, Apr. 2003.

[3] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," in *Proc. ICASSP 2005*, Philadelphia, PA, Mar. 2005. [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

[4] ——, "Speaker and session variability in GMM-based speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, May 2007.

[5] (2005) The NIST year 2005 speaker recognition evaluation plan. [Online]. Available: http://www.itl.nist.gov/iad/894.01/tests/spk/2005

[6] P. Kenny, M. Mihoubi, and P. Dumouchel, "New MAP estimators for speaker recognition," in *Proc. Eurospeech*, Geneva, Switzerland, Sept. 2003. [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

[7] R. Teunen, B. Shahshahani, and L. Heck, "A model-based transformational approach to robust speaker recognition," in *Proc. ICSLP*, Beijing, China, Oct. 2000, pp. 495–498.

[8] M. Gales, "Acoustic factorisation," in *Proc. ASRU 2001*, Trento, Italy, Dec. 2001.

[9] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 3, May 2005. [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

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

[11] H. Aronowitz, D. Burshtein, and A. Amir, "A session-GMM generative model using test utterance Gaussian mixture modeling for speaker verification," in *Proc. ICASSP 2005*, Philadelphia, PA, Mar. 2005.

[12] A. Solomonoff, W. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. ICASSP 2005*, Philadelphia, PA, Mar. 2005.

[13] D. Rubin and D. Thayer, "EM algorithms for ML factor analysis," *Psychometrika*, vol. 47, pp. 69–76, 1982.

[14] J. A. Bilmes, "Graphical models and automatic speech recognition," in *Mathematical Foundations of Speech and Language Processing*, M. Johnson, S. P. Khudanpur, M. Ostendorf, and R. Rosenfeld, Eds. New York, NY: Springer-Verlag, 2004, pp. 191–246.

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

[16] S. Lucey and T. Chen, "Improved speaker verification through probabilistic subspace adaptation," in *Proc. Eurospeech*, Geneva, Switzerland, Sept. 2003, pp. 2021–2024.

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

[18] Institute for Signal and Information Processing, Mississippi State University. [Online]. Available: http://www.isip.msstate.edu/projects/speech/software/legacy/index.html

[19] M. Tipping and C. Bishop, "Mixtures of probabilistic principal component analysers," *Neural Computation*, vol. 11, pp. 435–474, 1999.

Patrick Kenny received the BA degree in Mathematics from Trinity College, Dublin and the MSc and PhD degrees, also in Mathematics, from McGill University. He was a professor of Electrical Engineering at INRS-Télécommunications in Montreal from 1990 to 1995 when he started up a company (Spoken Word Technologies) to spin off INRS's speech recognition technology. He joined CRIM in 1998 where he now holds the position of principal research scientist. His current research interests are concentrated on Bayesian speaker- and channel-adaptation for speech and speaker recognition.

Gilles Boulianne received the B.Sc. degree in Unified Engineering from Universite du Quebec a Chicoutimi and the M.Sc. degree in Telecommunications from INRS-Telecommunications, Montreal. He worked on speech analysis and articulatory speech modeling at UQAM Linguistics Department until 1990, then on large vocabulary speech recognition at INRS and Spoken Word Technologies until 1998. He has been since with the Computer Research Institute of Montreal Speech Recognition Team. His research interests include finite state transducer approaches and practical applications of large vocabulary speech recognition such as live closed-captioning and content indexation.

Pierre Ouellet obtained the BSc degree in Computer Science from McGill University in 1994. He joined the Ecole de Technologie Supérieure in Montreal in 1997 to work on speaker identification in the context of dialogs in noisy environments. Since 1998, he has been working in the CRIM Speech Recognition team, where he contributes to ASR software development. His interests are software implementation issues and the application of adaptation techniques.

Pierre Dumouchel , B.Eng. (Universit McGill), M.Sc., Ph.D. (INRS-Tlcommunications). Pierre is actually Scientific Vice-President at CRIM and full professor at the cole de technologie suprieure (TS) of Universit du Qubec. Pierre was the vice-president Research and Development of CRIM from 1999 to 2004. Before he assumed the role of Principal Researcher of the CRIMs Automatic Speech Recognition team and was a scientific columnist at Radio-Canada, the French Canadian National Radio. He has more than 20 years of expertise in Speech Recognition Research, eight years in managing a research team and three years in managing the Research and Development unit of CRIM. His research has resulted in many technology transfers to such companies as Nortel, Locus Dialog, Canadian National Defence, Le Groupe TVA, as well as many SME, as such as Ryshco Media. His research interests are in search by transduction and automatic adaptation to new environment. He favoured applications of speech recognition for the hard-of-hearing and audio-visual film indexation.

APPENDIX

In this appendix we summarize the calculation of the posterior distribution of the hidden variables of the joint factor analysis model and we outline the simplified training procedures that we used to estimate the hyperparameters $\boldsymbol{m}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{d}$ and $\boldsymbol{\Sigma}$ in our experiments.

a) Posterior calculations in a joint factor analysis model: As we explained in Section V-A, when we use the factor analysis model to build a speaker verification system, enrolling a target speaker consists in calculating the posterior distribution of the speaker's supervector $\boldsymbol{s}$ which is related to the hidden variables in the factor analysis model by (5). Here we summarize the calculations needed to evaluate the joint posterior distribution of the hidden variables in a joint factor analysis model given a single enrollment utterance for the target speaker. (The case where multiple enrollment utterances are provided, as in the NIST extended data tasks, is treated in Section III of [1].) If

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix}$$

then the posterior distribution of $\boldsymbol{X}$ is Gaussian of the same form as the posterior distribution described in Proposition 1 of [9]. Specifically, if $\boldsymbol{V}$ and $\boldsymbol{L}$ are the matrices defined by

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{u} & \boldsymbol{v} & \boldsymbol{d} \end{pmatrix} \quad (24)$$

$$\boldsymbol{L} = \boldsymbol{I} + \boldsymbol{V}^* \boldsymbol{\Sigma}^{-1} \boldsymbol{N} \boldsymbol{V}. \quad (25)$$

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

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

$$\begin{pmatrix} \boldsymbol{a} & \boldsymbol{b} & \boldsymbol{c} \\ \boldsymbol{b}^* & \boldsymbol{I} + \boldsymbol{v}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{v} & \boldsymbol{v}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{d} \\ \boldsymbol{c}^* & \boldsymbol{d}\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{v} & \boldsymbol{I} + \boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{d}^2 \end{pmatrix} \quad (26)$$

where

$$\begin{aligned} \boldsymbol{a} &= \boldsymbol{I} + \boldsymbol{u}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{u} \\ \boldsymbol{b} &= \boldsymbol{u}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{v} \\ \boldsymbol{c} &= \boldsymbol{u}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{d}. \end{aligned}$$

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

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

where

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

with

$$\begin{aligned} \boldsymbol{\alpha} &= \begin{pmatrix} \boldsymbol{a} & \boldsymbol{b} \\ \boldsymbol{b}^* & \boldsymbol{I} + \boldsymbol{v}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{v} \end{pmatrix} \\ \boldsymbol{\beta} &= \begin{pmatrix} \boldsymbol{c} \\ \boldsymbol{v}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{d} \end{pmatrix} \\ \text{and } \boldsymbol{\gamma} &= \boldsymbol{I} + \boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{d}^2. \end{aligned}$$

Of course, since the dimensions of $\boldsymbol{L}$ are enormous (namely $(CF + R_C + R_S) \times (CF + R_C + R_S)$ where R_C is the rank of $\boldsymbol{u}$ and R_S is the rank of $\boldsymbol{v}$), care has to be taken to evaluate only those entries of $\boldsymbol{L}^{-1}$ which are actually needed.

b) Training a pure speaker factor analysis model: We consider a speaker factor analysis model of the form

$$\boldsymbol{s}(s) = \boldsymbol{m} + \boldsymbol{v}\boldsymbol{y}(s) + \boldsymbol{d}\boldsymbol{z}(s) \quad (27)$$

where s is a randomly chosen speaker. (This is a model of speaker variability alone rather than a joint model of speaker and channel variability. Since it is necessary to distinguish between training speakers in what follows we need to indicate the dependence on s explicitly.) We use $\boldsymbol{\Lambda}$ to denote a generic hyperparameter set $(\boldsymbol{m}, \boldsymbol{v}, \boldsymbol{d}, \boldsymbol{\Sigma})$.

Given an initial hyperparameter set $\boldsymbol{\Lambda}_0 = (\boldsymbol{m}_0, \boldsymbol{v}_0, \boldsymbol{d}_0, \boldsymbol{\Sigma}_0)$, we summarize two algorithms, known as maximum likelihood and minimum divergence estimation [1], for producing a better hyperparameter set $\boldsymbol{\Lambda}$ ('better' in a sense that can be made precise using the likelihood function defined in [1]). Throughout, we will use the notation $E[\cdot]$ to refer to posterior expectations calculated with the initial parameter set $\boldsymbol{\Lambda}_0$ as outlined above. (We only use the case where $\boldsymbol{u} = \boldsymbol{0}$ since we are dealing with a pure speaker factor analysis rather than a joint factor analysis.)

In order to describe the maximum likelihood estimation procedure it will be convenient to eliminate $\boldsymbol{m}$ by writing

$$\begin{aligned} \boldsymbol{Y}(s) &= \begin{pmatrix} \boldsymbol{y}(s) \\ 1 \end{pmatrix} \\ \text{and } \boldsymbol{V} &= \begin{pmatrix} \boldsymbol{v} & \boldsymbol{m} \end{pmatrix} \end{aligned}$$

so that (27) can be written in the form $\boldsymbol{s}(s) = \boldsymbol{V}\boldsymbol{Y}(s) + \boldsymbol{d}\boldsymbol{z}(s)$.

The following statistics are accumulated over the training set:

$$N_c = \sum_s N_c(s) \tag{28}$$

$$\mathfrak{A}_c = \sum_s N_c(s) E\left[\boldsymbol{Y}(s)\boldsymbol{Y}^*(s)\right] \tag{29}$$

$$\mathfrak{B} = \sum_s \boldsymbol{N}(s) E\left[\boldsymbol{z}(s)\boldsymbol{Y}^*(s)\right] \tag{30}$$

$$\mathfrak{C} = \sum_s \boldsymbol{F}(s) E\left[\boldsymbol{Y}^*(s)\right] \tag{31}$$

$$\mathfrak{a} = \sum_s \operatorname{diag}\left(\boldsymbol{N}(s) E\left[\boldsymbol{z}(s)\boldsymbol{z}^*(s)\right]\right) \tag{32}$$

$$\mathfrak{b} = \sum_s \operatorname{diag}\left(\boldsymbol{F}(s) E\left[\boldsymbol{z}^*(s)\right]\right). \tag{33}$$

In (28) and (29), c ranges over all mixture components. The new hyperparameter set $(\boldsymbol{m}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{d}, \boldsymbol{\Sigma})$ is defined by

(i) For each mixture component $c = 1, \ldots, C$ and for each $f = 1, \ldots, F$, set $i = (c-1)F + f$ and let V_i denote the ith row of $\boldsymbol{V}$ and d_i the ith entry of $\boldsymbol{d}$. Then V_i and d_i are defined by the equation

$$\begin{pmatrix} V_i & d_i \end{pmatrix} \begin{pmatrix} \mathfrak{A}_c & \mathfrak{B}_i^* \\ \mathfrak{B}_i & \mathfrak{a}_i \end{pmatrix} = \begin{pmatrix} \mathfrak{C}_i & \mathfrak{b}_i \end{pmatrix} \tag{34}$$

where $\mathfrak{B}_i$ is the ith row of $\mathfrak{B}$, $\mathfrak{a}_i$ is the ith entry of $\mathfrak{a}$, $\mathfrak{C}_i$ is the ith row of $\mathfrak{C}$ and $\mathfrak{b}_i$ is the ith entry of $\mathfrak{b}$.

(ii) Let $\mathfrak{M}$ be the diagonal $CF \times CF$ matrix given by

$$\mathfrak{M} = \operatorname{diag}\left(\mathfrak{C}\boldsymbol{V}^* + \mathfrak{b}\boldsymbol{d}\right).$$

Then

$$\boldsymbol{\Sigma} = \boldsymbol{N}^{-1}\left(\sum_s \boldsymbol{S}(s) - \mathfrak{M}\right) \tag{35}$$

where $\boldsymbol{N}$ is the $CF \times CF$ diagonal matrix whose diagonal blocks are $N_1 I, \ldots, N_C I$.

For the minimum divergence estimation procedure, the new hyperparameter set $\boldsymbol{\Lambda}$ is given by

$$\boldsymbol{m} = \boldsymbol{m}_0 + \boldsymbol{v}_0\boldsymbol{\mu}_{\boldsymbol{y}} + \boldsymbol{d}_0\boldsymbol{\mu}_{\boldsymbol{z}} \tag{36}$$

$$\boldsymbol{v} = \boldsymbol{v}_0\boldsymbol{K}_{\boldsymbol{yy}}^{1/2} \tag{37}$$

$$\boldsymbol{d} = \boldsymbol{d}_0\boldsymbol{K}_{\boldsymbol{zz}}^{1/2} \tag{38}$$

$$\begin{aligned} \boldsymbol{\Sigma} = \boldsymbol{N}^{-1}\sum_s \Big(& \boldsymbol{S}(s) \\ & - 2\operatorname{diag}\left(\boldsymbol{F}(s) E\left[\boldsymbol{s}^*(s)\right]\right) \\ & + \operatorname{diag}\left(E\left[\boldsymbol{s}(s)\boldsymbol{s}^*(s)\right]\boldsymbol{N}(s)\right) \Big) \end{aligned} \tag{39}$$

where

$$\boldsymbol{\mu}_{\boldsymbol{y}} = \frac{1}{S}\sum_s E\left[\boldsymbol{y}(s)\right]$$

$$\boldsymbol{\mu}_{\boldsymbol{z}} = \frac{1}{S}\sum_s E\left[\boldsymbol{z}(s)\right]$$

$$\boldsymbol{K}_{\boldsymbol{yy}} = \frac{1}{S}\sum_s E\left[\boldsymbol{y}(s)\boldsymbol{y}^*(s)\right] - \boldsymbol{\mu}_{\boldsymbol{y}}\boldsymbol{\mu}_{\boldsymbol{y}}^*$$

$$\boldsymbol{K}_{\boldsymbol{zz}} = \operatorname{diag}\left(\frac{1}{S}\sum_s E\left[\boldsymbol{z}(s)\boldsymbol{z}^*(s)\right] - \boldsymbol{\mu}_{\boldsymbol{z}}\boldsymbol{\mu}_{\boldsymbol{z}}^*\right),$$

S is the number of training speakers, the sums extend over all speakers in the training set and the square root sign in (37) and (38) indicates Cholesky decomposition.

c) Training a channel model: The matrix $\boldsymbol{u}$ in the channel model (3), (4) can be estimated by applying the maximum likelihood and minimum divergence training algorithms with $\boldsymbol{d} = \boldsymbol{0}$, $\boldsymbol{m} = \boldsymbol{0}$ and $\boldsymbol{v}$ replaced by $\boldsymbol{u}$ by using a modified set of first and second order Baum-Welch statistics as the input. For each recording of a speaker, the modification consists of centralizing the statistics extracted from the recording by applying the transformations (20) so as to remove speaker effects.

View publication stats