THE CRIM SYSTEM FOR THE 2005 NIST SPEAKER RECOGNITION EVALUATION

Patrick Kenny

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

1. Introduction

This is a first draft of one of a series of articles [1, 2] devoted to exploring the use of corpus based methods in text-independent GMM-based speaker recognition. Our purpose is to explain how we built a speaker recognition system for the NIST 2005 speaker recognition evaluation using about 1000 hours of speech taken from seven publicly available corpora.

State of the art GMM methods are based on very simple models of speaker and channel variability: speaker variability is modeled by assuming that speaker-dependent GMM supervectors are normally distributed with a diagonal covariance matrix (the premise of classical MAP speaker adaptation [3, 4]) and channel effects are assumed to be discrete [5, 6]. Our view is that statistical learning methods applied to large corpora such as the six Switchboard databases should make it possible to develop more powerful models of these types of variability. Note that these corpora are all publicly available through the Linguistic Data Consortium (LDC); they contain tens of thousands of recordings of thousands of speakers; and the fact that most of them have not been phonetically transcribed is no obstacle to using them to develop models for text-independent speaker recognition although we are unaware of any other researchers who have attempted to exploit them for this purpose.

Our principal effort in this direction 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 [6] but it treats channel effects as continuous rather than discrete and it exploits correlations between Gaussians in modeling speaker variability. When trained and tested on Switchboard data we have found that this model gives encouraging results [2]. However it is much more mathematically and computationally demanding than the standard approach (namely GMM/UBM speaker adaptation together with feature mapping) and it seems to require a well balanced training set in which a majority of training speakers are recorded under a variety of channel conditions that is sufficiently broad to cover all of the channel variation that is likely to be encountered at recognition time.

These requirements turn out to be quite difficult to meet where the NIST 2004 and 2005 test sets [7, 8] are concerned. These test sets are taken from the Mixer corpus [9] and they manifest a far greater degree of variability than previous test sets (which were taken from the 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). Switchboard data is poorly suited to modeling this type of variability and the only data of Mixer type which is publicly available is the NIST 2004 evaluation data, a relatively small set (310 target speakers).

So in building a system for the 2005 evaluation we were confronted with the problem of how to use the 2004 data to best advantage in conjunction with the LDC corpora. We felt that a joint factor analysis of speaker and channel variability carried out on all of these corpora might not be a successful strategy because the shortage of speakers recorded over both landline and cellular channels would mislead the model into believing that some speakers are ‘landline speakers’ and others are ‘cellular speakers’. So, in order to avoid the need for a balanced training set in which speakers are typically recorded under a wide variety of channel conditions, we decided to divorce speaker modeling and channel modeling altogether and experiment with models of utterance and session variability which are simpler than the joint factor analysis model in [1].[1]

In the core condition of the evaluation an ‘utterance’ is a side of a 5 minute telephone conversation so we use the term ‘utterance variablity’ to refer to the variability of the population conversation sides. (It is usual to think of this type of variability as being attributable solely to speaker effects but if this really were the case the problem of speaker recognition would have been solved long ago.) Modeling this type of variability is important because it gives a probability distribution on GMM supervectors which can serve as a prior for estimating speaker GMM’s by MAP adaptation. In this article we combine

[1] In the light of the performance of the QUT_2 system, it seems that this may prove to have been an unfortunate decision.

the priors underlying classical MAP and eigenvoice MAP [11] for this purpose. This type of model is known in statistics as factor analysis and the earliest instance in the literature on speaker recognition is [12]. Our treatment is different from [12] in that we use a likelihood criterion similar to that of [11] to estimate the hyperparameters that specify the prior.

We use the term session variability to refer to the variability exhibited by a given speaker from one recording session to another. Our approach to this problem is based on a probabilistic principal components analysis of session variability which we introduced in [13]. In that paper we showed how to estimate a prior which could be used for MAP adaptation of a speaker GMM to the channel conditions in a test utterance *without* adapting it to the speaker in the utterance. The solution we proposed is formally almost identical to eigenvoice MAP so we dubbed it eigenchannel MAP. Eigenchannel MAP is a computationally expensive procedure (particularly if there is a large t-norm cohort) but it was used to good effect by Spescom DataVoice in the 2004 evaluation. In this article we use the same type of session variability model with a simplified decision criterion (similar to that which is used in [14, 10]) which handles large numbers of t-norm speakers at very little computational cost.

2. Models of Utterance and Session Variability

The models of utterance and session variability were developed 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. They are gender-dependent rather than gender-independent. We will describe the quantities of data that we used in the female case; the figures for the male case are similar.

2.1. 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 [15] using a 3 s sliding window. Delta coefficients are then calculated using a 5 frame window giving a 26-dimensional feature vector.

Where available, ASR transcripts containing time stamps (such as the ctm files provided in the evaluation) are used to suppress silence intervals. In other cases the ISIP voice activity detector is used [16].

First and second order Baum-Welch statistics are extracted from the non-silence portions of the speech signal using a standard universal background model. 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 2048 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 $F = 26$).

2.2. Factor analysis of utterance variability

We assume that if $\boldsymbol{M}$ is the $CF \times 1$ speaker- and channel-dependent supervector for a randomly chosen conversation side then

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{v}\boldsymbol{y} + \boldsymbol{d}\boldsymbol{z} \qquad (1)$$

where $\boldsymbol{m}$ is the speaker- and channel-independent supervector, $\boldsymbol{v}$ is a matrix of dimension $CF \times R$ where $R \ll CF$, $\boldsymbol{d}$ is a $CF \times CF$ diagonal matrix and $\boldsymbol{y}$ and $\boldsymbol{z}$ are random vectors having standard normal distributions. This is a factor analysis in the sense of [17] but since utterance variability conflates speaker and channel effects it is a much simpler model than the joint factor analysis of speaker and channel variability in [1].

In the terminology of [18], the elements of $\boldsymbol{y}$ are 'common factors' (because each of them serves to account for the variance in all of the elements of $\boldsymbol{M}$) and the elements of $\boldsymbol{z}$ are 'specific factors'. In the absence of the specific factors, (1) implies that all supervectors are contained in the linear span of $\boldsymbol{m}$ and the columns of $\boldsymbol{v}$. This is almost the same as the basic assumption of eigenvoice modeling. (Almost but not quite because in this case we are treating different utterances by a given speaker as being statistically independent. This why we use the term 'utterance variability' rather than 'speaker variability' in the title of this section but it is convenient to use eigenvoice terminology even though this is not strictly speaking correct.) 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 data to estimate $\boldsymbol{v}$ reliably.

The role of this model is to provide a prior distribution for MAP estimation of GMM supervectors for target speakers. This type of MAP estimation combines classical MAP [4] and eigenvoice MAP [11] 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.

In the case where $\boldsymbol{d} = \boldsymbol{0}$ and $\boldsymbol{m}$ is given (the UBM supervector is a natural choice), $\boldsymbol{v}$ can be estimated by the algorithm described in Proposition 3 of [11] which 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]. (But note that in order to model utterance variability rather than speaker variability, the algorithm has to be implemented in such a way that in situations where there are more than one utterance for a given training speaker, the Baum-Welch statistics are *not* pooled across utterances.)

The algorithms that we use to estimate $\boldsymbol{m}, \boldsymbol{v}$ and $\boldsymbol{d}$ in the general case are described in Theorems 4, 5 and 6 of [1] (take $\boldsymbol{u} = \boldsymbol{0}$ in the statement of each theorem). For the system we submitted, we trained a factor analysis model for each gender with $R = 25$. In the female case the training set consisted of 9291 conversation sides (393 hours of speech exclusive of silences). The eigenvalues corresponding to the non-zero eigenvectors of $\boldsymbol{v}\boldsymbol{v}^*$ (the eigenvoices) are shown in Fig. 1 where they are seen to 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

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

$$\mathrm{tr}\left(\boldsymbol{v}\boldsymbol{v}^*\right) = 730.48. \quad (3)$$

We note in passing that in our experience setting $\boldsymbol{v} = \boldsymbol{0}$ and estimating $\boldsymbol{d}$ by a maximum likelihood criterion does not give a better estimate than the empirical method in [4]. This is consistent with the observation in [4] that the effectiveness of relevance MAP is insensitive to the value of the relevance factor.

2.3. Principal components analysis of session variability

Our approach to speaker recognition is GMM-based in that we estimate a GMM supervector for each target speaker but we use these supervectors for making speaker verification decisions in a non-traditional way. 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 we will briefly sketch the model-adaptation techniques that have been developed to tackle the problem of channel compensation in GMM-based speaker recognition.

2.3.1. Background

The most widely used method of GMM adaptation is feature mapping [6]. 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 (i) that 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

$$\boldsymbol{M} = \boldsymbol{S} + \boldsymbol{C} \quad (4)$$

where $\boldsymbol{C}$ depends only on the channel effects in the recording and (ii) that channel effects can be treated as discrete and identified in a pre-processing step (so that there is one channel supervector $\boldsymbol{C}$ for carbon-button handsets, another for GSM cellular transmissions and so forth). Given an enrollment utterance for a target speaker, the speaker supervector $\boldsymbol{S}$ can be estimated by subtracting the channel supervector for the enrollment utterance from the data; adding the 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 factor analysis model also takes (4) as its starting point but it treats channel supervectors as continuous rather than discrete and does away with the need for channel detection in a pre-processing step. Disentangling speaker and channel effects in (4) is more difficult but still manageable provided that a large training database is available in which speakers are recorded under a variety of channel conditions. The ideal situation is that the recordings for a typical training speaker are sufficiently numerous and diverse that channel effects can be averaged out as this enables reliable inferences concerning

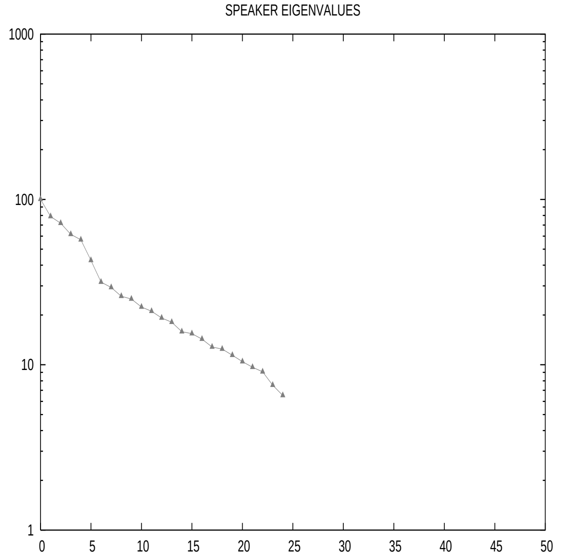

Figure 1: *Eigenvalues obtained by fitting a factor analysis model of utterance variability with 25 eigenvoices to female data. Female data. Compare with Fig. 4.*

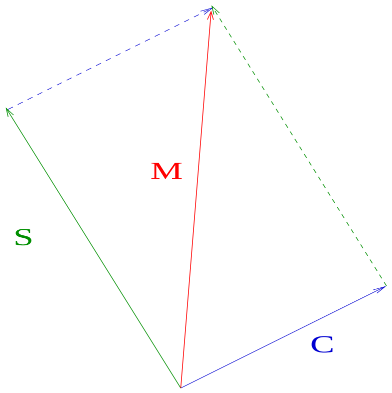

Figure 2: *Feature mapping and joint factor analysis of speaker and channel variability are based on a decomposition of the form $\boldsymbol{M} = \boldsymbol{S} + \boldsymbol{C}$. $\boldsymbol{M}$ is the speaker- and channel-dependent supervector for a given recording, $\boldsymbol{S}$ depends only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

the speaker supervector $\boldsymbol{S}$ in (4) to be made. However, for the purposes of developing a system to be tested on a dataset such as the NIST 2004 or NIST 2005 test sets this turns out to be a difficult requirement to meet because these test sets were specifically designed to evaluate the robustness of speaker recognition systems to gross channel mismatches and there are no publicly available speech corpora that are really suitable for modeling these channel effects. 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 database were recorded more than once. So it is hard to get hold of speakers who have been recorded under a variety of channel conditions. There is little hope of averaging out channel effects under these conditions so despite the investment we have made in developing the factor analysis model we decided not to use this approach in the NIST 2005 evaluation.

The assumptions underlying speaker model synthesis [5] are slightly different from those in feature mapping [6]. As in feature mapping, channel effects are assumed to be discrete and channel detection is performed in a preprocessing step. 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 assumptions (and not on the speaker) as in Fig. 3.

The eigenchannel model in [13] 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 [20].) Suppose we are given an enrollment utterance for a speaker and we used it to estimate a supervector $\boldsymbol{M}$ (using, say, classical MAP or factor analysis MAP). Given a test utterance we can synthesize a supervector $\boldsymbol{M}'$ for the same speaker under the test channel conditions by a type of MAP adaptation (dubbed eigenchannel MAP in [13]) which operates in a similar way to eigenvoice MAP. Similar types of model synthesis were used in the systems submitted by Spescom DataVoice in the 2004 evaluation and also in [21].

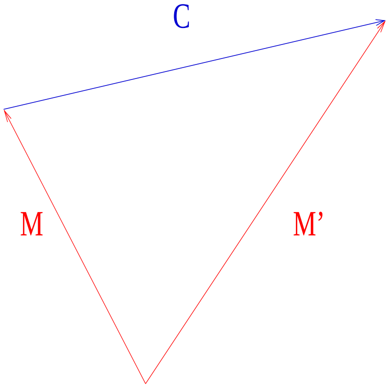

Figure 3: *In speaker model synthesis and eigenchannel modeling it is assumed that if $\boldsymbol{M}$ and $\boldsymbol{M}'$ are the speaker- and channel-dependent supervectors for two recordings of a given speaker and $\boldsymbol{C} = \boldsymbol{M}' - \boldsymbol{M}$ then $\boldsymbol{C}$ depends only on channel effects.*

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 speaker recognition [22]. It is interesting to note that although the approach in SVM 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 variabity, the key algorithm in 'nuisance attribute projection' is also formulated as an eigenvalue problem.

2.3.2. Eigenchannel estimation

Suppose that we have a pair of conversation sides for a given speaker. Let $\boldsymbol{M}$ and $\boldsymbol{M}'$ denote the corresponding speaker- and channel-dependent supervectors. The assumption in eigenchannel modeling is that

$$\boldsymbol{M}' = \boldsymbol{M} + \boldsymbol{u}\boldsymbol{x} \qquad (5)$$

where $\boldsymbol{u}$ is a rectangular matrix of low rank and $\boldsymbol{x}$ has a standard normal distribution. In other words, the assumption is that all of the channel compensation supervectors in Fig. 4 can be expressed as linear combinations of the columns of $\boldsymbol{u}$; in the terminology of [13], the non zero eigenvectors of $\boldsymbol{u}\boldsymbol{u}^*$ are the eigenchannels.

Given a training set consisting of a suitably large collection of pairs of utterances by different speakers, the algorithms used to estimate the factor analysis model in Section 2.2 can easily be modified to estimate $\boldsymbol{u}$. The idea here is that we can use one of the utterances in each pair (we chose the longer of the two) to calculate a point estimate $\hat{\boldsymbol{M}}$ of the supervector $\boldsymbol{M}$ appearing in the right hand side of (5), namely the MAP estimate of $\boldsymbol{M}$ calculated using the prior (1) and the Baum-Welch statistics extracted from the utterance. If $\boldsymbol{M}$ is replaced by $\hat{\boldsymbol{M}}$ in (5) and $\boldsymbol{d}$ is set to zero in (1), then (1) and (5) have the same form so $\boldsymbol{u}$ can be estimated by the same method as $\boldsymbol{v}$. However, the problem of estimating $\boldsymbol{u}$ does differ from that of estimating $\boldsymbol{v}$ in one important respect, namely that it is much easier to gather a large training set of utterance *pairs* so that even with a relatively large number of eigenchannels $\boldsymbol{u}$ can probably be estimated more reliably than $\boldsymbol{v}$.

For the system that we submitted, we fitted an eigenchannel model of rank 50 for each gender. In the female case the training set consisted of 27,399 utterance pairs. (The training set for the factor analysis model in Section 2.2 was obtained by choosing the longer utterance in each of these pairs.) Fig. 4 shows the eigenvalues sorted in decreasing order. 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. It is interesting to compare Figs. 1 and 4: utterance and session variability have essentially the same magnitude. (Session variability can be quantified as $\operatorname{tr}(\boldsymbol{u}\boldsymbol{u}^*)$ which turns out to be 741.85; utterance variability can be quantified as $\operatorname{tr}\left(\boldsymbol{d}^2 + \boldsymbol{v}\boldsymbol{v}^*\right)$ whose value is given by (2) and (3)).

Note that there is no difficulty in modifying the estimation algorithms to accommodate a residual term in the model (5) analogous to the term $\boldsymbol{d}\boldsymbol{z}$ in (1) but our experience has been that this hurts performance. This is to be expected because including such a term would result in a covariance matrix for channel supervectors which is of full rank and this would imply that any speaker could be made to sound like any other by varying the channel conditions. This seems unreasonable to us or, at any rate, we hope that it is not the case since it would seem to make a complete solution to the speaker recognition problem impossible in principle.

The eigenchannel approach to channel compensation is weaker than the factor analysis model in that it only compensates for channel effects in test utterances whereas the factor analysis model handles enrollment utterances as well. (The same remark applies to speaker model synthesis *vis-à-vis* feature mapping.) Eigenchannel modeling is simpler mathematically than joint factor analysis and it does not require as well balanced a training set; all that is required is that for any given pair of channel conditions there should be at least some training speakers recorded under both conditions. Speaker and channel effects do not have to be disentangled in enrolling a speaker but by the same token the variance of the channel supervectors in Fig. 3 will be twice as great as the variance of the channel supervectors in Fig. 2 and this is obviously undesirable.

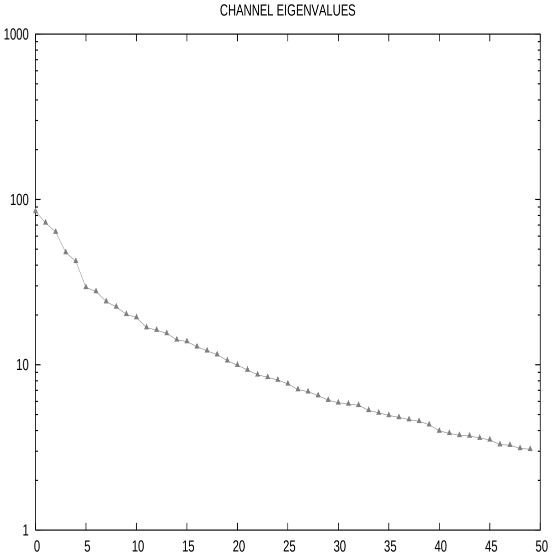

Figure 4: *Eigenvalues corresponding to 50 eigenchannels. Female data.*

3. Building a Speaker Verification System

In this section we explain how we use the models of utterance and session variability that we have just described to construct a speaker verification system. We will describe how we estimate a model for each target speaker, how we evaluate the likelihood of a test utterance using a target speaker model and how we normalize likelihoods calculated in this way so that a common decision threshold can be used in all speaker verification trials.

3.1. Enrolling a target speaker

Given Baum-Welch statistics extracted from an enrollment utterance, we use the prior distribution (1) to calculate the posterior distribution of the speaker- and channel-dependent supervector $\boldsymbol{M}$. We denote the posterior mean by $E[\boldsymbol{M}]$ and the diagonal of the posterior covariance matrix by $\operatorname{Cov}(\boldsymbol{M}, \boldsymbol{M})$.

In the case $\boldsymbol{d} = \boldsymbol{0}$, the calculation is described in

Proposition 1 of [11]; a modification is needed to handle the general case (see Section III 3 of [1]).

3.2. The likelihood function

In our early experiments with eigenchannel MAP [13] we proceeded in the same way as speaker model synthesis, synthesizing a new model for each speaker whenever a new test utterance was encountered and evaluating that model with the standard GMM likelihood function. 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.

On the other hand there is a natural likelihood function which serves as the objective function for estimating eigenchannels (similar to that for eigenvoices [11]) and although it is not related to the GMM likelihood function it can serve as a basis for constructing a decision criterion for speaker verification. Since it is generally a good idea to use the same objective function in training and testing no matter what the task is, this is the decision criterion that we decided to use. This approach can accommodate large numbers of t-norm speakers at little computational cost.

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 $\mathcal{X}$ and let $\boldsymbol{M}$ and $\boldsymbol{M}'$ be the speaker- and channel-dependent supervectors for the enrollment and test utterances respectively.

If we assume to begin with that $\boldsymbol{M}$ is known the likelihood of $\mathcal{X}$ under the alternative hypothesis — let us denote it by $P(\mathcal{X}|\boldsymbol{M})$ — can be calculated by the methods in [11]. By (5) there is a random vector $\boldsymbol{x}$ such that

$$\boldsymbol{M}' \quad = \quad \boldsymbol{M} + \boldsymbol{u}\boldsymbol{x}. \tag{6}$$

If $\boldsymbol{x}$ was known, we would know the supervector $\boldsymbol{M}'$ so it would be straightforward matter to calculate the conditional (Gaussian) likelihood of the test utterance, $P(\mathcal{X}|\boldsymbol{M},\boldsymbol{x})$, using the Baum-Welch statistics extracted from the utterance (Lemma 1 in [11]). So, since $\boldsymbol{x}$ is assumed to have a standard normal distribution, $P(\mathcal{X}|\boldsymbol{M})$ is given by

$$P(\mathcal{X}|\boldsymbol{M}) \quad = \quad \int P(\mathcal{X}|\boldsymbol{M},\boldsymbol{x})N(\boldsymbol{x}|\boldsymbol{0},\boldsymbol{I})d\boldsymbol{x} \tag{7}$$

where $N(\cdot|\boldsymbol{0},\boldsymbol{I})$ is the standard Gaussian kernel. Proposition 2 in [11] 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

$$\begin{aligned} F_c &= \sum_t X_t & (8) \\ S_c &= \operatorname{diag}\left(\sum_t X_t X_t^*\right) & (9) \end{aligned}$$

where the sum extends over all observations X_t aligned with the given mixture component, and $\operatorname{diag}()$ sets off-diagonal entries to 0. (As we have written them these are Viterbi statistics but we use Baum-Welch statistics in practice.) 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 $c = 1, \ldots, C$).[2] We denote the first and second order moments of $\mathcal{X}$ around $\boldsymbol{M}$ by $\boldsymbol{F_M}$ and $\boldsymbol{S_M}$ so that

$$\begin{aligned} \boldsymbol{F_M} &= \boldsymbol{F} - \boldsymbol{N}\boldsymbol{M} \\ \boldsymbol{S_M} &= \boldsymbol{S} - 2\operatorname{diag}(\boldsymbol{F}\boldsymbol{M}^*) + \operatorname{diag}(\boldsymbol{N}\boldsymbol{M}\boldsymbol{M}^*). \end{aligned} \tag{10}$$

Finally, let

$$\boldsymbol{l} \quad = \quad \boldsymbol{I} + \boldsymbol{u}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{u}, \tag{11}$$

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

$$\boldsymbol{l} \quad = \quad \boldsymbol{l}^{1/2}\boldsymbol{l}^{1/2*} \tag{12}$$

(that is, the Cholesky decomposition of $\boldsymbol{l}$). Then the likelihood function is given by

$$\begin{aligned} \log P(\mathcal{X}|\boldsymbol{M}) \quad = \quad & \sum_{c=1}^{C} N_c \log \frac{1}{(2\pi)^{F/2}|\Sigma_c|^{1/2}} \\ & - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S_M}\right) - \frac{1}{2}\log|\boldsymbol{l}| \\ & + \frac{1}{2}\|\boldsymbol{l}^{-1/2}\boldsymbol{u}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{F_M}\|^2 \end{aligned} \tag{13}$$

provided that $\boldsymbol{M}$ is known. In practice $\boldsymbol{M}$ has to be estimated from the enrollment data for the hypothesized

[2] It is convenient to use $\boldsymbol{S}$ to stand for 'second order statistics' rather than for 'speaker' as we did in (4).

speaker so we replace $\boldsymbol{F_M}$ and $\boldsymbol{S_M}$ by their posterior expectations, $E[\boldsymbol{F_M}]$ and $E[\boldsymbol{S_M}]$, which are given by

$$\begin{aligned} E[\boldsymbol{F_M}] &= \boldsymbol{F} - \boldsymbol{N} E[\boldsymbol{M}] \\ E[\boldsymbol{S_M}] &= \boldsymbol{S} - 2\,\mathrm{diag}(\boldsymbol{F}\boldsymbol{M}^*) \\ &\quad + \mathrm{diag}\big(\boldsymbol{N}(E[\boldsymbol{M}]\,E[\boldsymbol{M}^*] \\ &\quad + \mathrm{Cov}(\boldsymbol{M}, \boldsymbol{M}))\big) \qquad (14) \end{aligned}$$

(in accordance with the notation introduced in Section 3.1). The term $\mathrm{Cov}(\boldsymbol{M}, \boldsymbol{M})$ will be non-negligble if the amount of enrollment data for the hypothesized speaker is small. Because the term $\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S_M}\right)$ enters into (13) with a negative sign, its effect is to diminish the value of the likelihood function by an amount which is proportional to the uncertainty in the point estimate $E[\boldsymbol{M}]$ of $\boldsymbol{M}$. Thus it provides a mechanism for penalizing hypothesized speakers with small amounts of enrollment data.

The most interesting thing to note about (13) is that the likelihood function depends on the hypothesized speaker only through the computations in (14) and the cost of these computations is negligeable (since $E[\boldsymbol{M}]$ and $\mathrm{Cov}(\boldsymbol{M}, \boldsymbol{M})$ 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).

The likelihood function in [10] differs from ours in just two respects: it does not take account of the term $\mathrm{Cov}(\boldsymbol{M}, \boldsymbol{M})$ and it uses the MAP estimate of $\boldsymbol{x}$ instead of integrating with respect to $\boldsymbol{x}$ as in (8). The value of the integral is actually very closely related to the MAP estimate of $\boldsymbol{x}$ (see Proposition 2 in [1]) so that the approach in [10] enjoys the same computational advantages as ours.

3.3. Likelihood normalization

In our first experiments 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 if the number of eigenvoices is set to 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. 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

$$\begin{aligned} E[\boldsymbol{M}] &= \boldsymbol{m} \\ \mathrm{Cov}(\boldsymbol{M}, \boldsymbol{M}) &= \boldsymbol{0}. \qquad (15) \end{aligned}$$

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

$$\begin{aligned} E[\boldsymbol{M}] &= \boldsymbol{m} \\ \mathrm{Cov}(\boldsymbol{M}, \boldsymbol{M}) &= \mathrm{diag}\left(\boldsymbol{d}^2 + \boldsymbol{v}\boldsymbol{v}^*\right). \qquad (16) \end{aligned}$$

We will refer these two versions of z-norm as 'z-norm without uncertainty' and 'z-norm with uncertainty' respectively.

4. Experiments

The system we submitted for the evaluation used a UBM with 2048 Gaussians, 25 eigenvoices, 50 eigenchannels and t-norm score normalization. Taking all trials of the core condition as the test bed [8], it resulted in an equal error rate (EER) of 11.7% and a DCF of 0.042. These results were not as good as we expected so we conducted a series of experiments after the evaluation to see how our system might be improved.

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 re-ran 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.

Our first series of experiments we designed to evaluate the effect of modifying the configurations of the utterance and session models. The results are summarized in Table 1 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 1) 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).

The experiments reported in Table 1 were conducted using only t-norm score normalization. We tested the other types of normalization strategies described in Section 3.3 on two model configurations: 5 eigenvoices and 25 eigenchannels (the best configuration according to Table 1) and 0 eigenvoices and 25 eigenchannels (the configuration most similar to [10]). The results are summarized in Tables 2 and 3. For each configuration the best results are obtained with zt-norm, confirming the results

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

Table 1: *Effect of different utterance and session model configurations. All trials, core condition. T-norm score normalization.*

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. Curiously, zt-norm seems to be much less effective in the case of 5 eigenvoices than in the case of 0 eigenvoices. Thus it turns out that our best result, 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 1 might suggest.

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

Table 2: *Effect of different types of score normalization. All trials, core condition. 5 eigenvoices, 25 eigenchannels.* z_1 *indicates z-norm with uncertainty,* z_2 *z-norm without uncertainty.*

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

Table 3: *Effect of different types of score normalization. All trials, core condition. 0 eigenvoices, 25 eigenchannels.* z_1 *indicates z-norm with uncertainty,* z_2 *z-norm without uncertainty.*

5. Discussion

The approach to the problem of speaker verification described in this article is very similar to that of [10] although our results are not as good as those obtained by the QUT_2 system in the evaluation. The differences in the way we make verification decisions are very minor (we have opted to integrate over hidden variables such as speaker and channel factors rather than use point estimates of them) but our method of estimating eigenchannels is different to that of [10] (which more closely resembles the factor analysis model). We will have to investigate this question before we can draw any conclusions from our results.

6. Acknowledgements

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

7. References

[1] P. Kenny, “Factor analysis of speaker and channel variability,” in preparation.

[2] ——, “Speaker and session variability in GMM-based speaker verification,” in preparation.

[3] J.-L. Gauvain and C.-H. Lee, “Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains,” *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 291–298, Apr. 1994.

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

[5] R. Teunen, B. Shahshahani, and L. Heck, “A model-based transformational approach to robust speaker recognition,” in *Proc. ICSLP*, Beijing, China, Oct. 2000.

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

[7] (2004) The NIST year 2004 speaker recognition evaluation plan. [Online]. Available: http://www.nist.gov/speech/tests/spk/2004

[8] (2005) The NIST year 2005 speaker recognition evaluation plan. [Online]. Available: http://www.itl.nist.gov/iad/894.01/tests/spk/2005

[9] J. Campbell *et al.*, “The MMSR bilingual and cross-channel corpora for speaker recognition research and evaluation,” in *Proc. Odyssey 2004*, Toledo, Spain, June 2004.

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

[11] P. Kenny, G. Boulianne, and P. Dumouchel, “Eigenvoice modeling with sparse training data,” *IEEE Trans. Speech Audio Processing*, vol. 13, no. 3, May 2005.

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

[13] P. Kenny, M. Mihoubi, and P. Dumouchel, "New MAP estimators for speaker recognition," in *Proc. Eurospeech*, Geneva, Switzerland, Sept. 2003.

[14] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," in *Proc. ICASSP 2005*, Philadelphia, PA, Mar. 2005.

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

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

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

[18] 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.

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

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

[21] 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.

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