

The Geometry of the Channel Space in GMM-Based Speaker Recognition

Patrick Kenny, Gilles Boulianne, Pierre Ouellet and Pierre Dumouchel

Centre de recherche informatique de Montréal (CRIM)

{Patrick.Kenny, Gilles.Boulianne, Pierre.Ouellet, Pierre.Dumouchel}@crim.ca

Abstract

We describe an extension of the joint factor analysis model of speaker and channel variability in which channel supervectors are modeled by mixtures of low-rank Gaussians rather than by a unimodal Gaussian. This version of the joint factor analysis model includes data-driven feature mapping and the standard joint factor analysis models as limiting cases and it enables us to explore a range of possibilities between these two extremes. Our experimental results indicate that unimodal models of relatively high rank perform better than mixture models of lower rank and they confirm the appropriateness of the unimodal assumption in the standard joint factor analysis model.

1. Introduction

Two approaches to the problem of modeling session variability (and channel variability in particular) in GMM-based speaker recognition that have turned out to be very successful are feature mapping [1, 2] and joint factor analysis [3, 4, 5, 6, 7].¹ Both approaches are based on a simple additive model of speaker and channel effects and they differ principally in the assumptions made concerning how speaker and, especially, channel variability should be modeled. The assumption in joint factor analysis is that channel effects can be modeled by a normal distribution concentrated on a low dimensional subspace of the GMM supervector space whereas feature mapping assumes a discrete distribution concentrated on a finite set of points in the supervector space.

Our aim in this paper is to describe an extension to the joint factor analysis model in which channel supervectors are modeled by mixtures of low-rank Gaussian distributions rather than by a single Gaussian as in our previous work. This provides a framework which embraces both the basic assumptions in feature mapping (the rank 0 case) and the basic assumptions in standard factor analysis modeling (the case where there is just one mixture component) as special cases and enables us to experiment with a range of possibilities between these two extremes.

This paper is a follow-up to [7] which is an abridged version of [4]. The principal focus of the paper is to show

how the unimodal algorithms in [7] can be extended to the multimodal (mixture) case. The experimental results that we will produce indicate that unimodal models of relatively high rank perform better than mixture models of lower rank and they confirm the appropriateness of the unimodal assumption in the standard joint factor analysis model.

2. Session Variability: Unimodal or Multimodal?

Joint factor analysis is a model of speaker and session variability in Gaussian mixture models which are commonly used in text-independent speaker recognition. In this section, we begin by recapitulating how the standard version of joint factor analysis models these two types of variability using unimodal Gaussian distributions in supervector space and then explain how we relaxed the unimodal assumption in the case of session variability for the experiments reported later in the paper.

We assume a fixed GMM structure containing a total of C mixture components each modeled by a diagonal Gaussian. Let F be the dimension of the acoustic feature vectors. (We took $C = 2048$ and $F = 26$ throughout.) Let \mathbf{m} denote the universal background supervector.

The joint factor analysis model treats speaker variability by assuming that if \mathbf{s} is the supervector for a randomly chosen speaker then \mathbf{s} is distributed according to

$$\mathbf{s} = \mathbf{m} + \mathbf{v}\mathbf{y} + \mathbf{d}\mathbf{z} \quad (1)$$

where \mathbf{d} is diagonal, \mathbf{v} is a rectangular matrix of low rank and \mathbf{y} and \mathbf{z} are independent random vectors having standard normal distributions. Including the term $\mathbf{v}\mathbf{y}$ in (1) enabled us to obtain our best results on the core condition of the NIST 2005 test set [4, 7] but ignoring it results in only a minor degradation in performance. We used the 2005 test set for the experiments reported here and we assumed simply that

$$\mathbf{s} = \mathbf{m} + \mathbf{d}\mathbf{z}. \quad (2)$$

This is the model of speaker variability which is implicit in the standard GMM/UBM approach (where \mathbf{d} is estimated empirically by assuming that it depends on a single free parameter known as the relevance factor [8, 9]).

¹See also <http://www.crim.ca/perso/patrick.kenny>

If M is the channel dependent supervector corresponding to a particular recording of the speaker then the standard joint factor analysis model assumes that

$$M = s + ux \quad (3)$$

where u is a rectangular matrix of low rank and x has a standard normal distribution. We refer to the rank of u as the *channel rank* and to the components of x as channel factors.

The standard model is completely specified by additionally defining, for each mixture component c , a diagonal covariance matrix Σ_c whose role is to model the variability which is not captured by the speaker variability model (2) or the channel variability model (3). We denote by Σ the $CF \times CF$ supercovariance matrix whose diagonal is the concatenation of these covariance matrices.

In previous articles we have shown how to estimate the hyperparameters m, d, u and Σ that define the standard joint factor analysis model using large speech corpora in which speakers are recorded in multiple sessions; how the joint factor analysis model can be used to build speaker verification systems; and that this approach is particularly effective if it is used in conjunction with feature warping [10]. The purpose of the present paper is to see if any improvements in performance can be gained by relaxing the channel modeling assumptions (3).

As we explained in [4], the basic assumption in joint factor analysis is the same as in feature mapping [1], namely that a speaker and channel dependent supervector can be decomposed into a sum of two components one of which depends only on the speaker and the other only on the channel (or recording session). The main difference is that channel effects are not discretized in factor analysis so that, in particular, there is no need for handset detection. (The data-driven version of feature mapping [2] also does away with this requirement.) Instead, channel supervectors are assumed to vary continuously in a low-dimensional subspace of the supervector space, namely the range of uu^* .

There is very strong evidence in favor of this type of assumption (see Fig. 2 for example) but the additional assumption that channel supervectors are normally distributed in the channel space is rather dubious because it is reasonable to suppose that various types of handset or headset and various channel effects should produce multiple modes in the distribution of the channel supervectors. Given the success of Gaussian mixtures in speech modeling generally, a Gaussian mixture seems to be a more natural candidate than a unimodal Gaussian for modeling the distribution of channel supervectors.

Of course, using a mixture model instead of a unimodal model leads to a fragmentation of the training data (different training utterances will be used to estimate different mixture components in the channel supervector

distribution) but this will not be problem in practice if the LDC corpora are used for training since these contain tens of thousands of recordings. In this connection it is also interesting to note that a matrix v in (2) of rank 300 can be robustly estimated with as few as 500 speakers [4, 7]. Thus there should be no difficulty in principle in training mixture models having large numbers of mixture components and a high channel rank.

If we take the number of mixture components to be P , then the model we are considering in this paper can be formally specified by a set of parameters of the form $\{(\pi_p, o_p, u_p, \Sigma_p) : p = 1, \dots, P\}$ where, for $p = 1, \dots, P$,

1. π_p is a mixture weight
2. o_p is a $CF \times 1$ vector
3. u_p is a matrix of dimension $CF \times R$ (R for channel rank).
4. Σ_p is a $CF \times CF$ diagonal covariance matrix.

In place of (3), the basic assumption is that session effects in an utterance produced by a speaker whose GMM supervector is s are accounted for by first sampling a mixture component from the weight distribution $\{\pi_p : p = 1, \dots, P\}$. If mixture component p is chosen, then the speaker and channel dependent supervector for the utterance is given by

$$M = s + o_p + u_p x \quad (4)$$

where x is a hidden $R \times 1$ vector having a standard normal distribution.

For each $p = 1, \dots, P$, the covariance matrix Σ_p plays the same role with respect to mixture component p as the covariance matrix Σ in the unimodal version of the model.

The offset vectors o_p define the modes of the channel supervector distribution. (This term does not appear in (3) because we assume that the mean of the channel supervector distribution is $\mathbf{0}$ in the unimodal case.) Of course, in the case where $P = 1$ and $o_1 = \mathbf{0}$, the mixture model reduces to the standard unimodal joint factor analysis model.

In the case where the channel rank R is 0 (so that the term $u_p x$ in (4) disappears) then our basic assumption is essentially the same as in the data-driven version of feature mapping [2].

Our experience with the unimodal version of the factor analysis model is that the dimension of the channel space is of the order of 100 (see Table 2 for example). Our main interest in pursuing the current line of research was to see if we could obtain a better fit to the distribution of channel supervectors by treating the channel space as a curved submanifold of dimension R for some $R < 100$

and modeling it by the union of the R -dimensional linear manifolds

$$\mathbf{o}_p + \text{range}(\mathbf{u}_p \mathbf{u}_p^*) \quad (p = 1, \dots, P),$$

as illustrated in Fig. 1.

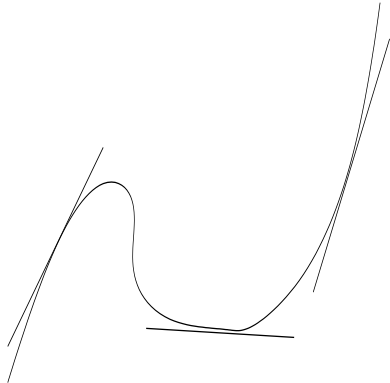


Figure 1: An R -dimensional curved submanifold of the supervector space can be approximated by a collection of linear submanifolds of the same dimension.

3. Implementing the Mixture Model

In Section 3 of [7] we outlined the algorithms that we are currently using to construct a speaker verification system from a standard (unimodal) joint factor analysis model: namely how we estimate a GMM supervector for each target speaker and how we evaluate the likelihood of a test utterance using a target speaker GMM. It is a straightforward matter to extend these algorithms to a mixture model if the hyperparameters which specify the model (namely $\{(\pi_p, \mathbf{o}_p, \mathbf{u}_p, \mathbf{\Sigma}_p) : p = 1, \dots, P\}$) are given.

Suppose we have a test utterance \mathcal{X} and a GMM supervector \mathbf{s} for a target speaker and we wish to calculate the likelihood of \mathcal{X} given \mathbf{s} . For each mixture component p , let $P(\mathcal{X}|\mathbf{s}, p)$ denote the likelihood of the test utterance calculated with the speaker's GMM on the assumption that the mixture component p is responsible for the session effects in the utterance. Calculating $P(\mathcal{X}|\mathbf{s}, p)$ is just a matter of replacing \mathbf{u} by \mathbf{u}_p , $\mathbf{\Sigma}$ by $\mathbf{\Sigma}_p$ and \mathbf{s} by $\mathbf{s} + \mathbf{o}_p$ in equations (5) – (12) of [7]. The total likelihood of the utterance \mathcal{X} is then given by

$$\sum_{p=1}^P \pi_p P(\mathcal{X}|\mathbf{s}, p). \quad (5)$$

(We used the maximum in place of the sum in our implementation.)

Similarly, to enroll a target speaker using an enrollment utterance \mathcal{X} , we first identify the mixture component p that best accounts for the utterance in the sense that $P(\mathcal{X}|\mathbf{s}, p)$ is maximal and then, using this mixture

component, we estimate a GMM supervector for the target speaker by means of the same algorithm that we use in the unimodal case.

Note that for a given channel rank R , the computational and memory requirements of these calculations are P times more expensive than in the unimodal case (where P is the number of mixture components) and, as in [4, 7], the marginal cost of handling large numbers of t-norm speakers is very small.

As for training, we use the same type of EM-type algorithms (maximum likelihood and minimum divergence estimation) as in [4, 7], starting from initial random estimates of the hyperparameters and refining them on successive iterations. (We first trained unimodal models, then models with 2 components, 4 components and so on by splitting each of the mixture components in two prior to each training run.)

For a given number of mixture components, the simplest approach to training is to assume a deterministic assignment of training utterances to mixture components on each iteration, just as we do in enrolling target speakers. We then use the utterances associated with each mixture component p to estimate the hyperparameters $\pi_p, \mathbf{o}_p, \mathbf{u}_p$ and $\mathbf{\Sigma}_p$ just as in the unimodal case. This is analogous to Viterbi-style training of HMM's or GMM's. A minor complication that needs to be dealt with is that, as we mentioned in Section 2, the offset vectors \mathbf{o}_p in (4) have no analog in the unimodal model. These can be made to disappear by a standard trick: if

$$\mathbf{X} = \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix}$$

and $\mathbf{U}_p = \begin{pmatrix} \mathbf{u}_p & \mathbf{o}_p \end{pmatrix}$

then (4) can be written in the form

$$\mathbf{M} = \mathbf{s} + \mathbf{U}_p \mathbf{X}, \quad (6)$$

just as in the unimodal case.

A more sophisticated approach to training (the one which we actually implemented) entails calculating on each training iteration, for each training utterance and each mixture component p , the posterior probability that the utterance is generated by the given mixture component. (These posteriors are derived from the likelihoods $\pi_p P(\mathcal{X}|\mathbf{s}, p)$ which appear in (5) by normalizing them to sum to 1.) By using the posteriors to weight the statistics extracted from the various training utterances we derive training algorithms which are analogous to Baum-Welch-style training of HMM's or GMM's.

We will spare the reader the details and simply point out that our standard unimodal training algorithms are versions of Probabilistic Principal Components analysis [11] designed to work in situations where the supervectors are unobservable (in the sense that speaker and channel dependent GMM supervectors cannot be estimated

from small amounts of data by maximum likelihood estimation). As such, these algorithms can be extended from the unimodal case to the multimodal case in much the same way as ordinary Probabilistic Principal Components analysis [12].

4. Experiments

For the experiments reported in this paper we trained two gender dependent factor analysis models using the same databases as in [4, 7]: the LDC releases of Switchboard II, Phases 1, 2 and 3; Switchboard Cellular, Parts 1 and 2; and the NIST 2004 evaluation data. Where possible we selected only those speakers for which 6 or more different number conversation sides were available. The female training set consisted of 612 speakers and 6764 conversation sides; the male training set consisted of 463 speakers and 5254 conversation sides.

We used the NIST 2005 evaluation data for testing (core condition, all trials). We used the NIST designation of enrollment and test utterances for each trial. (Modest improvements in performance can be obtained by reversing the roles of the enrollment and test utterances and combining the results but we did not take advantage of this in our experiments.) As in [6], the acoustic features that we used were Gaussianized cepstral features and their first derivatives. As we mentioned in Section 2, we did not use any common speaker factors in the sense of [7] in our models (that is, we suppressed the term $v\mathbf{y}$ in (1)). We used zt-norm for score normalization [13, 4, 7].

Our first series of experiments was designed to find the optimal dimension of the channel space in the unimodal case. The minimum value of the NIST detection cost function and the equal error rate for each of the dimensions we tested are reported in Table 1. The performance is seen to be fairly stable across a wide range of dimensions. The reason for this is apparent from Figure 2 which shows the eigenvalues of the matrix $\mathbf{u}\mathbf{u}^*$ sorted in decreasing order in the case where the channel rank is 200. Since the scale on the y -axis is logarithmic it is clear that most of the session variability is captured by 50–100 leading eigenvectors.

Table 1: *The effect of varying the dimension of the channel space in the unimodal case. Core condition, all trials, NIST 2005 evaluation data. DCF = detection cost, EER = equal error rate.*

Channel Rank	DCF	EER
25	0.022	6.9%
50	0.020	6.6%
100	0.019	6.8%
200	0.022	7.2%

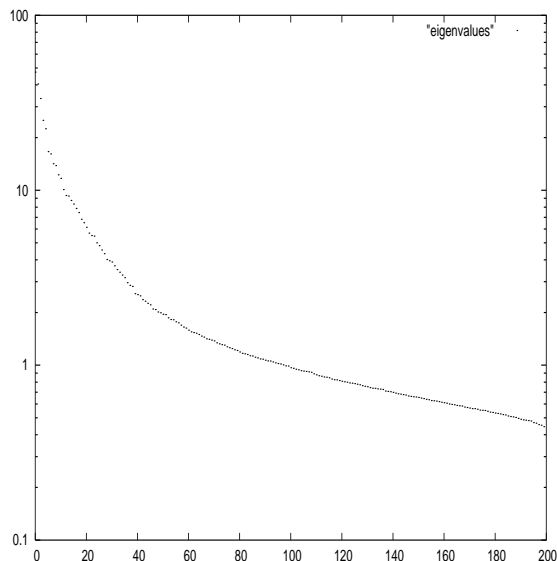


Figure 2: *Eigenvalues of $\mathbf{u}\mathbf{u}^*$ in the unimodal case with channel rank 200. Female training data.*

In testing the mixture model we choose a series of configurations with essentially the same number of free parameters by keeping the value of the product $R \times P$ fixed. (Recall that R is the channel rank and P is the number of mixture components.) We used a larger number of free parameters than in the case of a single mixture component because the results in Table 1 indicated that increasing the channel rank beyond 200 in the unimodal case would not be helpful and, as we explained in Section 2, there is no shortage of data in practice for training mixture models.

The results are reported in Table 2. There is a clear trend: if the channel rank is high (e.g. 40) then performance is similar to the unimodal case but performance degrades as P increases and R decreases. The reason for this behavior is easy to find. In the case where $P = 64$ and $R = 5$ (for instance) the eigenvalues of the matrices $\mathbf{u}_p\mathbf{u}_p^*$ ($p = 1, \dots, 64$) are much bigger than 0 (when compared with the tail of the graph in Fig. 2). This indicates that a much larger channel rank would be needed to capture session variability in a mixture model having a large number of mixture components (so that the computational cost of implementing such a model would be prohibitive).

5. Conclusion

Our experimental results indicate that a mixture model of high channel rank performs better than a mixture model of lower rank when the total number of free parameters in the model is held fixed and, in particular, a unimodal model of sufficiently high rank is capable of achieving better performance than mixture models. Thus our results

Table 2: Trading off the number of mixture components and channel rank in the mixture case. Core condition, all trials, NIST 2005 evaluation data. DCF = detection cost, EER = equal error rate.

Mixture Components	Channel Rank	DCF	EER
8	40	0.022	6.7%
16	20	0.023	6.9%
32	10	0.028	8.1%
64	5	0.034	9.7%

confirm the appropriateness of the unimodal assumption in the standard version of the joint factor analysis model. Using a large number of mixture components to model discrete channel effects such as those attributable to different types of microphone and codec is clearly helpful (just as in feature mapping) but our results suggest that methods which are capable of modeling continuous effects as well (such as those attributable to electrical and acoustic background noise) are more powerful.

It is perhaps a bit surprising that unimodal models should outperform mixture models in our experiments since there is no shortage of training data in the LDC corpora and such a result seems to contradict some obvious facts about channels. It seems quite likely that feature warping is at least partly responsible for this phenomenon, firstly because short-term Gaussianization probably reinforces the unimodal Gaussian assumptions in the standard joint factor analysis model and secondly, because feature warping is a very powerful channel normalization technique in its own right.

Of course there is a trivial sense in which mixture models *have* to perform better than unimodal models if they are properly configured but eigenvalue inspections indicate that a high channel rank is needed to achieve optimal results in the multimodal case just as in the unimodal case. Handling a large number of mixture components of high rank is very computationally expensive so that unimodal modeling is to be recommended in practice at least on the NIST core test sets (which involve only telephone speech). Other test beds, such as the NIST auxiliary microphone tasks, remain to be explored.

6. References

- [1] Douglas Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. ICASSP*, Hong Kong, China, Apr. 2003.
- [2] M. Mason, R. Vogt, B. Baker, and S. Sridharan, "Data-driven clustering for blind feature mapping in speaker verification," in *Proc. Eurospeech*, Lisbon, Portugal, Sept. 2005, pp. 3109–3112.
- [3] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," submitted to *IEEE Trans. Audio Speech and Language Processing*.
- [4] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," to appear in *IEEE Trans. Audio Speech and Language Processing*.
- [5] P. Kenny and P. Dumouchel, "Disentangling speaker and channel effects in speaker verification," in *Proc. ICASSP*, Montreal, Canada, May 2004.
- [6] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," in *Proc. ICASSP 2005*, Philadelphia, PA, Mar. 2005.
- [7] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Improvements in factor analysis based speaker verification," in *Proc. ICASSP 2006*, Toulouse, France, May 2006.
- [8] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.
- [9] 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.
- [10] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey*, Crete, Greece, June 2001.
- [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] M. Tipping and C. Bishop, "Mixtures of probabilistic principal component analysers," *Neural Computation*, vol. 11, pp. 435–474, 1999.
- [13] R. Vogt, B. Baker, and S. Sridharan, "Modeling session variability in text-independent speaker verification," in *Proc. Eurospeech*, Lisbon, Portugal, Sept. 2005.