New MAP Estimators for Speaker Recognition

P. Kenny, M. Mihoubi and P. Dumouchel

Centre de recherche informatique de Montréal (CRIM)
{pkenny, mmihoubi, pdumouch}@crim.ca

Abstract

We report the results of some experiments which demonstrate that eigenvoice MAP and eigenphone MAP are at least as effective as classical MAP for discriminative speaker modeling on *SWITCHBOARD* data. We show how eigenvoice MAP can be modified to yield a new model-based channel compensation technique which we call eigenchannel MAP. When compared with multi-channel training, eigenchannel MAP was found to reduce speaker identification errors by 50%.

1. Introduction

Research on speaker identification and speaker verification (as in most of the NIST evaluations [1]) is primarily focused on the problem of how to estimate HMMs or GMMs capable of discriminating between speakers using small amounts of data (typically on the order of one minute per speaker recorded in just one or two sessions).

MAP estimation has proved to be a powerful modeling technique for this problem [2, 3, 4]. Its appeal is that it provides a principled way of interpolating between an unreliably estimated speaker model and a reliably estimated speaker-independent or universal background model. However classical MAP estimation (that is, in the form in which it was originally presented in [5]) was developed as a method of acoustic phonetic modeling for speech recognition rather than for speaker recognition. The classical MAP estimator is constructed in such a way that, in situations where there are few observations for a given speaker and Gaussian, it falls back to speaker-independent estimates. This type of behavior is reasonable for speech recognition modeling but its suitability for discriminating between speakers is open to question. This suggests that other types of MAP estimator with more subtle fall back behavior might be worth considering. For example, structured MAP estimation [6] has been shown to be more effective in speaker verification than classical MAP estimation [7, 8].

Our primary purpose in this paper is to investigate the effectiveness of the eigenvoice MAP estimator introduced in [9] for speaker identification. At the same time we will investigate a dual estimator which we refer to as eigenphone MAP [10]. (These estimators are dual in the sense that eigenvoice MAP exploits correlations between GMM or HMM mixture components on the assumption that speakers are statistically independent whereas eigenphone MAP exploits correlations between speakers on the assumption that mixture components are statistically independent.) Speaker modeling using eigenvoices (without the MAP framework) has already proved to be effective in speaker identification and verification [11] but no comparable studies have been carried out on eigenphones. However the eigenphone approach is arguably more natural for discriminative speaker modeling because it is based on an explicit model for *inter*-speaker variability (namely an inter-speaker correlation matrix). We will report experimental results demonstrating that eigenphone MAP and eigenvoice MAP are both more effective than classical MAP for text-independent speaker identification using GMMs on *SWITCHBOARD* data.

We will also show how a slight change in perspective enables the eigenvoice methodology to be applied effectively to the problem of modeling *intra*-speaker variability. (That is, the variability exhibited by a speaker from one recording session to another. This is perhaps the most difficult problem in speaker recognition [1]. It is well known that training speaker models on data in which each speaker is recorded multiple sessions is an effective way of dealing with this problem [12, 13, 14] but this is not always feasible.) The intra-speaker variability in *SWITCHBOARD* (our test bed) seems to be principally attributable to microphone and channel effects so the problem we are addressing here can be stated as follows: can model-based channel compensation help to discriminate between target speakers and imposters? Given a test speaker and channel, we need to be able to adapt speaker models to the channel *without* adapting them to the test speaker if channel compensation is to be successful. MAP model adaptation can be made to respect this type of constraint if it is based on a suitable prior distribution for channel compensations.

The MAP estimator we propose for this purpose adapts speaker models to a given channel in essentially the same way as eigenvoice MAP adapts a speaker-independent model to a given speaker. Accordingly we dub our approach eigenchannel MAP. Suppose we are given a test utterance for speech recognition using a speaker independent HMM. Even if the test utterance comes from a previously unseen speaker, eigenvoice MAP can be brought to bear by using the test utterance itself to carry out unsupervised speaker adaptation prior to making the recognition decision. Similarly, suppose we are given a test utterance for speaker identification using a set of speaker GMMs. For each hypothesized speaker, eigenchannel MAP uses a prior distribution on channel compensations to adapt[1] the speaker GMM to the test utterance and the speaker identification descision is based on likelihood evalaution with these adapted models. Like cepstral mean subtraction and RASTA processing this approach to channel compensation is blind (no knowledge of microphone or channel characteristics is assumed and no attempt is made to detect these in the signal) but since it is model-based rather than feature-based it is far more flexible: the adaptation mechanism differs from one speaker to another and, for a given speaker GMM, it differs from one mixture component to another. (It also has to be pointed out that it is far more computationally expensive.)

Just as eigenvoice modeling can account for most *inter*-speaker variability with a relatively small number of eigen-

[1] In the case of GMMs there is no distinction between supervised and unsupervised adatpation since the phonetic transcription of the test utterance is irrelevant.

voices, eigenchannel modeling can account for most *intra*-speaker variability with a relatively small number of eigenchannels. In order to estimate the eigenchannels we need speaker models for a large collection of speakers and a training set comprising several recordings of each of these speakers. These speakers should be representative of the target speaker population (i.e. the subjects for speaker identification) but they need not be part of it and, although every possible channel/microphone combination should be well represented in the training set, it is not necessary that each training speaker be recorded under every condition. The *SWITCHBOARD* databases are well suited to eigenchannel modeling since they comprise hundreds of speakers with an average of 10 recordings per speaker. Whereas the number of eigenvoices that can be estimated from a given training set is bounded by the number of training speakers (which is generally insufficient), the number of eigenchannels that can be estimated is bounded only by the number of conversation sides (which is probably more than enough).

2. Speaker Models

Suppose we are given a population of S speakers and we wish to construct a GMM or a HMM having C mixture components for each speaker. For each $c = 1, \ldots, C$ let $\mu(c)$ be the speaker independent mean vector associated with the mixture component c and, for each speaker s, let $\mu_s(c)$ denote the corresponding speaker-dependent mean vector. The MAP approach to speaker modeling assumes that for each mixture component c and speaker s, there is an unobservable offset vector O_{sc} such that

$$\mu_s(c) = \mu(c) + O_{sc} \tag{1}$$

and that the prior distribution of the matrix (O_{sc}) is known. Point estimates of the speaker-dependent mean vector can be obtained by calculating the mode of the posterior distribution (that is, the distribution obtained by conditioning on the training data) of the matrix (O_{sc}). This strategy can be extended to adapt the HMM variances as well.

Various assumptions can be made concerning the form of the prior (without giving rise to intractable posteriors):

1. Classical MAP assumes that the entries of the matrix (O_{sc}) are statistically independent [2].
2. Eigenphone modeling assumes that the column vectors of the matrix (O_{sc}) are independent and identically distributed [10].
3. Eigenvoice modeling assumes that the row vectors of the matrix (O_{sc}) are independent and identically distributed [9].

We refer the reader to [10, 9] for descriptions of the eigenvoice MAP and eigenphone MAP estimation procedures. We implemented eigenvoice modeling using a single stream (that is, without making any statistical independence assumption concerning the acoustic features) but for technical reasons we had to treat the acoustic features as being statistically independent for eigenphone modeling.

3. The channel model

In much the same way as MAP estimation can be used to adapt a speaker independent (or universal background) model to a given speaker, it can also be used to adapt a speaker model to a given channel. In order to be of practical use, the prior on channel compensations ought to be chosen in such a way as to permit adaptation at recognition time to channels that have not previously been seen. Note that eigenvoice modeling lends itself easily to adapting previously unseen speakers at recognition time because it uses a prior in which speakers are statistically independent and identically distributed. So for purposes of channel compensation it is natural to use a prior in which the channel compensations for all speakers and channels are independent and identically distributed.

To be more specific, let $\mu_{sh}(c)$ denote the mean vector corresponding to a speaker s, a recording h and a mixture component c and let o_{shc} be an (unobservable) vector such that

$$\mu_{sh}(c) = \mu_s(c) + o_{shc}. \tag{2}$$

Let $\boldsymbol{o}_{sh}$ be the supervector obtained by concatenating $o_{shc} \quad (c = 1, \ldots, C)$. Since the variability from one recording to another is presumably primarily attributable to channel effects, we refer to such a supervector as a channel compensation supervector. We assume the channel compensation supervectors $\{\boldsymbol{o}_{sh}\}$ (where s ranges over all speakers and h ranges over all recordings of speaker s) are independent and identically distributed with mean $\mathbf{0}$ and a covariance matrix $\mathbf{C}$. If we are given a training set in which a variety of recording conditions are represented together with speaker models for each training speaker, the methods in [9] can easily be modified to estimate the principal eigenvectors of $\mathbf{C}$ (the eigenchannels). Note that we can constrain channel compensation supervectors to lie in a low dimensional vector space by specifying the number of eigenchannels to be estimated. (We used 50 eigenchannels and we made no statistical independence assumptions concerning the acoustic features for the experiments reported below.) Once $\mathbf{C}$ has been estimated the MAP estimation procedure described in [9] can be used to adapt a given speaker GMM to a previously unseen channel (using some adaptation data recorded over the channel).

Channel compensation will not be helpful in discriminating between a target speaker and imposters if its effect is to adapt imposter models to the test speaker rather than to the test channel. To see why our procedure can be expected to perform channel adaptation rather than speaker adaptation, observe that since the prior distribution on channel compensations is concentrated on the range of $\mathbf{C}$, the same is true of the posterior distribution. Thus the MAP estimator takes values in the range of $\mathbf{C}$. An elementary argument shows that since $\mathbf{C}$ is the covariance matrix of the channel compensation supervectors encountered in training, its range is spanned by these supervectors. Thus the only channel compensations permitted by our model are those which lie in the linear span of channel compensations needed to fit the various speaker models to the channels observed in training.

4. Experiments

4.1. Database

We used a subset of the *SWITCHBOARD-1* training set for our experiments. For each speaker, we designated the longest conversation side as the primary conversation side (or primary channel), the second longest conversation side as the secondary conversation side (or secondary channel) and so on. We used the primary channels to train speaker models, the secondary and tertiary channels for testing them and the remaining data for channel modeling. More specifically:

1. We trained speaker models using primary channel data for 319 speakers (138 females and 181 males). These

speakers were selected by the requirement that the duration of the primary conversation side should be at least one minute and in this case the entire primary conversation side was used for training. The average amount of data per speaker was almost 3 minutes. (For eigenphone and eigenvoice modeling it is advantageous to use as many training speakers as possible so we did not restrict ourselves to speakers for which secondary channel data was available.)

2. For testing we took 30 seconds of data from each of 535 secondary and tertiary conversation sides representing 289 speakers.

3. For training the channel model we took 30 seconds of data from each of 1330 conversation sides representing 214 speakers. Since speaker models are needed train the channel model, we restricted ourselves to conversation sides in which the speaker was one of 319 training speakers. (The channel model could be trained using data from an entirely different set of speakers but we would have to use a much larger database to explore this possibility.)

The durations cited here are not quite exact because we extracted whole turns (without silence detection or truncation) from the conversation sides. Silences were not suppressed because it is very likely that they contain information which is useful for channel modeling.

For signal processing we used a 10 ms frame rate and a 26 dimensional feature vector (log energy, cepstral coefficients 1–12 and their first derivatives). Except where otherwise indicated we performed cepstral mean subtraction on a turn-by-turn basis.

4.2. Test Protocol

For each speaker identification trial we randomly choose a test conversation side and 10 imposters of the same sex as the target speaker. Each of the experiments reported below consisted of 2000 trials. Note that since we only had 535 test conversation sides to work with these trials are not fully randomized so statements made below concerning confidence intervals and statistical significance have to interpreted loosely.

Limiting the number of imposters per trial was necessary because of the the computational cost of the channel modeling experiments (which require that compensation be performed for each of the imposters in a trial as well as for the target speaker). But note that if the speaker identification error rate with 10 imposters can be reliably estimated then it easy to calculate approximate error rates with 20 imposters, 30 imposters etc.

4.3. Speaker Modeling

In order to compare the effectiveness of classical MAP, eigenvoice MAP and eigenphone MAP for text-independent speaker identification we used each of these methods to estimate speaker GMMs with 256 components from the primary channel training data. We made no attempt to deal with microphone and channel mismatches in testing. The results are given in Table 1.

Eigenvoice and eigenphone modeling are both seen to give better results than classical MAP. Eigenphone MAP was implemented using an inter-speaker correlation matrix of full rank in each feature dimension (i.e. with 319 eigenphones per feature). The results for eigenvoice modeling with 300 eigenvoices are essentially the same as for eigenphone modeling.

	M	MV
Classical MAP	16.15%	
Eigenphones	14.55%	14.80%
300 Eigenvoices	14.75%	14.65%
100 Eigenvoices	16.45%	17.70%

Table 1: *Speaker identification error rates obtained by training on the primary channel data. Confidence limits are* $\pm 1.5\%$. M *indicates mean adaptation,* MV *indicates mean and variance adaptation.*

The experiment with 100 eigenvoices is reported to underscore the importance of using a large number of eigenvoices for discriminative speaker modeling. (The result in [4] to the effect that classical MAP is significantly better than eigenvoice modeling for speaker verification were obtained using only 70 eigenvoices. Note however that this result pertains to speaker verification and not to speaker identification.)

4.4. Channel Modeling

The purpose of these experiments was to evaluate the effectiveness of eigenchannel MAP adaptation.

4.4.1. Pilot Experiments

For the pilot experiments choose eigenphone modeling (adapting variances as well as mean vectors) with cepstral mean subtraction as the speaker modeling technique.

We estimated two sets of speaker models, one using the primary channel data alone and the other using an extended training set obtained by adding the multi-channel data reserved for channel modeling to the primary channel data. As expected, using multi-channel data reduces the speaker identification error rate substantially, from 14.80% to 8.80%.

Next we estimated the 50 principal eigenvectors of the correlation matrix $\mathbf{C}$ from the data reserved for channel modeling in conjunction with the speaker models trained on the primary channel data. For each speaker-identification trial we used eigenchannel MAP estimation to adapt the primary channel speaker models to the test data. Under these conditions we were able to halve the speaker identification error rate, going from 8.80% to 4.40%. The difference between these error rates is highly statistically significant ($p < 0.00005$).

In order to see if the computational burden of eigenchannel MAP could be reduced somewhat, we tried using only the first 10 seconds of test data in each trial for channel compensation. Under these conditions we obtained an error rate of 4.30% (compared with 4.40% when all of the test data is used for adaptation).

When we replicated this experiment using the extended training set to estimate the speaker models (instead of using just the primary channel data) and also to estimate the channel model (instead of using just the data reserved for channel modeling) we obtained higher error rate, namely 5.50%. This suggests that if multi-channel data is used to estimate speaker models then channel compensation ought to be integrated into the estimation procedure. More research seems to be needed here.

4.4.2. Variants

We carried out some additional experiments with the channel model to evaluate the usefulness of cepstral mean subtraction in this context and to compare eigenvoice modeling with eigenphone modeling. All of these experiments were carried out using only the primary channel data to estimate speaker models.

	CMS	no CMS
Eigenphones	4.40%	4.90%
Eigenvoices	5.95%	5.30%

Table 2: *Speaker identification error rates using channel compensation. CMS indicates cepstral mean subtraction. Confidence limits are* ±*1.0%*

The results of these experiments are summarized in Table 2. Eigenphones are seen to perform better than eigenvoices. Since eigenvoices and eigenchannels both model *intra*-speaker dependencies it is not surprising that eigenvoice MAP and eigenchannel MAP should interact differently with eigenchannel MAP.

The results with cepstral mean subtraction are paradoxical. One would not expect cepstral mean subtraction to be useful in the context of the channel compensation. In the case of eigenvoice modeling this is exactly what we found but the opposite behavior is apparent in the eigenphone case. The reason for this seems to be that (because of its similarity to the eigenchannel model) eigenvoice modeling has a built in immunity to channel distortions in the primary channel training data which eigenphone modeling does not. This is another indication of the need to carry out channel compensation in training.

5. Discussion

Our results show that eigenchannel MAP is a very effective way of dealing with intra-speaker variability in speaker identification on *SWITCHBOARD* data. We are unaware of any work in this direction by other authors although our approach is closely analogous to the deformable modeling methods used in handwriting recognition [15] (which also incorporate MAP estimation into the likelihood evaluations) and it seems very likely that suitably chosen priors could enable other types of MAP model adaptation to be applied to the channel compensation problem. (In particular, MAPLR would seem to be a natural candidate.)

In tackling this type of problem it is natural to concentrate on speaker identification rather than speaker verification to begin with. More experimental work will be needed to see whether channel modeling can be used effectively with the universal background model or cohort models needed for speaker verification. A key question here seems to be whether, given a recording of a test speaker, it is feasible to adapt a universal background model to the test channel without adapting it to the test speaker.

A hard problem raised by our work is how to integrate speaker modeling and channel modeling into a coherent framework so that both models can be trained on a common data set comprising multiple channels for each speaker (rather than using only primary channel data to estimate the speaker models). This problem would have to solved in order to use channel modeling in conjunction with complex HMMs so that the effectiveness of channel modeling in speech recognition on *SWITCHBOARD* could be investigated.

6. Conclusion

Classical MAP estimation is currently the most widely used speaker modeling technique for speaker recognition. Eigenvoice MAP and eigenphone MAP are generalizations which (like classical MAP) were developed originally for speech recognition tasks. We found that when applied to speaker identification on *SWITCHBOARD* data, both of these approaches to speaker modeling decreased error rates by about 10% when compared with classical MAP.

We also introduced a blind model-based channel compensation technique called eigenchannel MAP and found that it decreased error rates by about 50% when compared with multisession training. In the context of eigenchannel MAP, eigenphone MAP outperformed eigenvoice MAP by about 15%.

7. References

[1] A. Martin, M. Przybocki, G. Doddington, and D. Reynolds, “The NIST speaker recognition evaluation: Overview, methodology, systems, results, perspectives (1998),” *Speech Communication*, vol. 31, pp. 225–254, 2000.

[2] J.-L. Gauvain, L. Lamel, and B. Prouts, “Experiments with speaker verification over the telephone,” in *Proc. Eurospeech*, Madrid, Spain, Sept. 1995.

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

[4] J. Mariéthoz and S. Bengio, “A comparison of adaptation methods for speaker verification,” in *Proc. ICSLP*, Denver, Colorado, Sept. 2002.

[5] 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.

[6] K. Shinoda and C.-H. Lee, “A structural bayes approach to speaker adaptation,” *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 276–287, 2001.

[7] B. Xiang and T. Berger, “Structural Gaussian mixture models for efficient text-independent speaker verification,” in *Proc. ICSLP*, Denver, Colorado, Sept. 2002.

[8] M. Liu, E. Chang, and B.-Q. Dai, “Hierarchical Gaussian mixture model for speaker verification,” in *Proc. ICSLP*, Denver, Colorado, Sept. 2002.

[9] P. Kenny, G. Boulianne, and P. Dumouchel, “Maximum likelihood estimation of eigenvoices and residual variances for large vocabulary speech recognition tasks,” in *Proc. ICSLP*, Denver, Colorado, Sept. 2002.

[10] ——, “Bayesian adaptation revisited,” in *Proc. ISCA ITRW*, Paris, France, Sept. 2000.

[11] O. Thyes, R. Kuhn, *et al.*, “Speaker identification and verification using eigenvoices,” in *Proc. ICSLP*, Beijing, China, Oct. 2000.

[12] L. Lamel and J.-L. Gauvain, “Speaker verification over the telephone,” in *Proc. RLA2C*, Avignon, France, Apr. 1998.

[13] L. Heck and N. Mirghafori, “Unsupervised on-line adaptation in speaker verification: Confidence-based updates and improved parameter estimation,” in *Proc. ISCA ITRW*, Sophia Antipolis, France, Aug. 2001.

[14] K. Farrell, “Speaker verification with data fusion and model adaptation,” in *Proc. ICSLP*, Denver, Colorado, Sept. 2002.

[15] A. Jain, A. Zhong, and S. Lakshamanan, “Object matching using deformable templates,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, pp. 267–278, Mar. 1996.