DISENTANGLING SPEAKER AND CHANNEL EFFECTS IN SPEAKER VERIFICATION

Patrick Kenny and Pierre Dumouchel

Centre de recherche informatique de Montréal (CRIM)

{pkenny, pdumouch}@crim.ca

ABSTRACT

We show how a joint factor analysis of inter-speaker and intra-speaker variability in a training database which contains multiple recordings for each speaker can be used to construct likelihood ratio statistics for speaker verification which take account of intra-speaker variation and channel variation in a principled way. We report the results of experiments on the NIST 2001 cellular one speaker detection task carried out by applying this type of factor analysis to *Switchboard Cellular Part I*. The evaluation data for this task is contained in *Switchboard Cellular Part I* so these results cannot be taken at face value but they indicate that the factor analysis model can perform extremely well if it is perfectly estimated.

1. INTRODUCTION

Broadly speaking, state of the art methods in speech recognition and speaker recognition attempt to compensate for channel variation and intra-speaker variation — or channel variation for short, even if this is not strictly correct — by normalization techniques such as cepstral mean subtraction and feature warping and to model inter-speaker variation by adaptation techniques such as MLLR and EMAP (in the case of speech recognition) and classical MAP (in the case of speaker recognition). All of these model adaptation techniques conflate inter-speaker variation and channel variation so that they may be performing channel adaptation in some situations and speaker adaptation in others. Whether this is a bad thing in the case of speech recognition is not clear (nobody seems to have looked into the matter) but there is no question that it raises problems for speaker recognition. The challenge in the current NIST one speaker detection evaluations is to recognize a speaker given enrollment data extracted from a single recording and test data extracted from other recordings. Estimating a GMM from a speaker's enrollment data by classical MAP [1] produces a model which is adapted to the enrollment recording conditions as well as to the speaker. Using this GMM to recognize the speaker under different recording conditions is therefore problematic. It seems that collecting enrollment data from multiple sessions for each target speaker is the only way to deal with channel variation in speaker recognition using standard model adaptation methods. Online model adaptation techniques used in speech recognition such as MLLR and EMAP do not seem to be applicable to this problem precisely because they conflate channel variation and inter-speaker variation. So, although intra-speaker variation and channel variation are of critical importance for speaker recognition, the problem of how to model these effects (rather than trying to eliminate them in the front end) has hardly been studied.

In this paper we will indicate how, given a database comprising a large number of speakers in which each speaker is recorded under many different conditions, we can jointly model inter-speaker and channel variability by a probabilistic factor analysis. Our basic assumption is that speaker- and channel-dependent GMM supervectors are Gaussian distributed with most (but not all) of the variance in these supervectors being accounted for by a small number of hidden variables which we refer to as speaker and channel factors. The speaker factors and the channel factors play different roles in that, for a given speaker, the values of the speaker factors are assumed to be the same for all recordings of the speaker but the channel factors are assumed to vary from one recording to another. (Thus the channel factors may be capturing either channel variation or intra-speaker variation.) The prior on speaker-dependent supervectors used in eigenvoice MAP [2, 3] is a special case of the factor analysis model in which there are no channel factors and all of the variance in the speaker-dependent GMM supervectors is assumed to be accounted for by the speaker factors. (In this case, for a given speaker, the values of the speaker factors are the co-ordinates of the speaker's supervector relative to a suitable basis of the eigenspace.) The general factor analysis model is constructed by combining the prior for eigenvoice MAP with the priors for classical MAP and eigenchannel MAP [3].

We will report the results of speaker verification experiments on the NIST 2001 cellular one speaker detection task [4] using likelihood ratio statistics derived from factor analysis models trained on *Switchboard Cellular Part I*. The 2001 evaluation data was described in [4] as 'drawn from the *Switchboard-II Corpus, Phase 4*' but it turns out that it

is actually entirely contained in *Switchboard Cellular Part I*. Thus there is a flaw in our experimental design and it cannot be easily remedied since most of the speakers in *Switchboard Cellular Part I* serve as target speakers in the evaluation and there are no other cellular databases currently available through the LDC. Our results are extraordinarily good but they only indicate how well the factor analysis model can perform if it is perfectly estimated.

2. FACTOR ANALYSIS

We assume a fixed GMM structure containing a total of C mixture components. Let F be the dimension of the acoustic feature vectors.

2.1. Speaker and Channel factors

To begin with let us ignore the question of channel variation and assume that each speaker s can be modeled by a single supervector $\boldsymbol{M}(s)$ which is independent of channel effects. Classical MAP assumes that there is a diagonal matrix $\boldsymbol{d}$ such that, for a randomly chosen speaker s,

$$\boldsymbol{M}(s) = \boldsymbol{m} + \boldsymbol{d}\boldsymbol{z}(s) \tag{1}$$

where $\boldsymbol{m}$ is the speaker-independent supervector and $\boldsymbol{z}(s)$ is a hidden vector distributed according to the standard Gaussian density, $N(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$. Eigenvoice MAP assumes instead that there is a rectangular matrix $\boldsymbol{v}$ of low rank such that, for a randomly chosen speaker s,

$$\boldsymbol{M}(s) = \boldsymbol{m} + \boldsymbol{v}\boldsymbol{y}(s) \tag{2}$$

where $\boldsymbol{y}(s)$ is a hidden vector having a standard normal distribution. The strengths and weaknesses of classical MAP and eigenvoice MAP complement each other. (Eigenvoice MAP is preferable if small amounts of data are available for speaker adaptation and classical MAP if large amounts are available.) An obvious strategy to combine the two is to assume a decomposition of the form

$$\boldsymbol{M}(s) = \boldsymbol{m} + \boldsymbol{v}\boldsymbol{y}(s) + \boldsymbol{d}\boldsymbol{z}(s) \tag{3}$$

where $\boldsymbol{y}(s)$ and $\boldsymbol{z}(s)$ are assumed to be independent and have standard normal distributions. In this case it is no longer appropriate to speak of eigenvoices; rather $\boldsymbol{v}$ is a 'factor loading matrix' and the components of $\boldsymbol{y}(s)$ are 'speaker factors'.

Now let us consider channel effects. Suppose we are given recordings $h = 1, \ldots, H(s)$ of a speaker s. For each recording h, let $\boldsymbol{M}_h(s)$ denote the corresponding speaker- and channel-dependent supervector. We assume that the difference between $\boldsymbol{M}_h(s)$ and $\boldsymbol{M}(s)$ can be accounted for by a vector of channel factors $\boldsymbol{x}_h(s)$ having a standard normal distribution. That is, we assume that there is a rectangular matrix $\boldsymbol{u}$ of low rank (the loading matrix for the channel factors) such that

$$\left.\begin{array}{rcl} \boldsymbol{M}(s) & = & \boldsymbol{m} + \boldsymbol{v}\boldsymbol{y}(s) + \boldsymbol{d}\boldsymbol{z}(s) \\ \boldsymbol{M}_h(s) & = & \boldsymbol{M}(s) + \boldsymbol{u}\boldsymbol{x}_h(s) \end{array}\right\} \tag{4}$$

for each recording $h = 1, \ldots, H(s)$.

So if R_C is the number of channel factors and R_S the number of speaker factors, our factor analysis model is specified by a quintuple $\boldsymbol{\Lambda}$ of hyperparameters of the form $(\boldsymbol{m}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{d}, \boldsymbol{\Sigma})$ where $\boldsymbol{m}$ is $CF \times 1$, $\boldsymbol{u}$ is $CF \times R_C$, $\boldsymbol{v}$ is $CF \times R_S$ and $\boldsymbol{d}$ and $\boldsymbol{\Sigma}$ are $CF \times CF$ diagonal matrices. To explain the role of $\boldsymbol{\Sigma}$, fix a mixture component c and let Σ_c be the corresponding block of $\boldsymbol{\Sigma}$. For each speaker s and recording h, let $M_{hc}(s)$ denote the subvector of $\boldsymbol{M}_h(s)$ corresponding to the given mixture component. We assume that, for all speakers s and recordings h, observations drawn from mixture component c are distributed with mean $M_{hc}(s)$ and covariance matrix Σ_c.

2.2. The likelihood function

Suppose that we are given a set of hyperparameter estimates $\boldsymbol{\Lambda}$ and a set of recordings for a speaker s indexed by $h = 1, \ldots, H(s)$. For each recording h, assume that each frame has been aligned with a mixture component and let $\mathcal{X}_h(s)$ denote the collection of labeled frames for the recording. (We used a speaker-independent GMM or gender-dependent GMM to do the alignment in our experiments.) Let $\underline{\mathcal{X}}(s)$ be the vector obtained by concatenating the observable variables $\mathcal{X}_1(s), \ldots, \mathcal{X}_{H(s)}(s)$ and let $\underline{\boldsymbol{X}}(s)$ be the vector obtained by concatenating the unobservable variables $\boldsymbol{x}_1(s), \ldots, \boldsymbol{x}_{H(s)}(s), \boldsymbol{y}(s), \boldsymbol{z}(s)$. If $\underline{\boldsymbol{X}}(s)$ were given we could write down $\boldsymbol{M}_h(s)$ and calculate the (Gaussian) likelihood of $\mathcal{X}_h(s)$ for each recording h so the calculation of the joint likelihood of $\underline{\mathcal{X}}(s)$ would be straightforward. Since the values of the hidden variables are not given, calculating this joint likelihood requires evaluating the integral

$$\int P_{\boldsymbol{\Lambda}}(\underline{\mathcal{X}}(s)|\underline{\boldsymbol{X}}) N(\underline{\boldsymbol{X}}|\underline{\boldsymbol{0}}, \underline{\boldsymbol{I}}) d\underline{\boldsymbol{X}} \tag{5}$$

where $N(\underline{\boldsymbol{X}}|\underline{\boldsymbol{0}}, \underline{\boldsymbol{I}})$ is the standard Gaussian kernel

$$N(\boldsymbol{x}_1|\boldsymbol{0}, \boldsymbol{I}) \ldots N(\boldsymbol{x}_{H(s)}|\boldsymbol{0}, \boldsymbol{I}) N(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{I}) N(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I}).$$

We denote the value of this integral by $P_{\boldsymbol{\Lambda}}(\underline{\mathcal{X}}(s))$.

2.3. Estimating the hyperparameters

If we are given a training set in which each speaker is recorded in multiple sessions the hyperparameters $\boldsymbol{\Lambda}$ can be estimated by an EM algorithm which guarantees that the total likelihood of the training data increases from one iteration to the next. (The total likelihood of the training data

is $\prod_s P_{\mathbf{\Lambda}}(\underline{\mathcal{X}}(s))$ where s ranges over the speakers in the training set.) This estimation procedure can be derived by extending Proposition 3 in [2] to handle the hyperparameters $\boldsymbol{u}$ and $\boldsymbol{d}$ in addition to $\boldsymbol{v}$ and $\mathbf{\Sigma}$. We refer to this as the speaker-independent hyperparameter estimation procedure since it consists in fitting (4) to the entire collection of speakers in the training data rather than to an individual speaker.

In order to construct a likelihood ratio statistic for speaker verification, we also need a speaker-dependent estimation procedure. For this we assume that, for a given speaker s and recording h,

$$\left.\begin{array}{rcl} \boldsymbol{M}(s) & = & \boldsymbol{m}(s) + \boldsymbol{v}(s)\boldsymbol{y}(s) + \boldsymbol{d}(s)\boldsymbol{z}(s) \\ \boldsymbol{M}_h(s) & = & \boldsymbol{M}(s) + \boldsymbol{u}\boldsymbol{x}_h(s). \end{array}\right\} \quad (6)$$

That is, we make the hyperparameters $\boldsymbol{m}$, $\boldsymbol{v}$ and $\boldsymbol{d}$ speaker-dependent but we continue to treat $\boldsymbol{u}$ and $\mathbf{\Sigma}$ as speaker-indendent. Given enrollment data for the speaker s, we estimate the speaker-dependent hyperparameters $\boldsymbol{m}(s), \boldsymbol{v}(s)$ and $\boldsymbol{d}(s)$ by first using the speaker-independent hyperparameters and the enrollment data to calculate the posterior distribution of $\boldsymbol{M}(s)$ and then adjusting the speaker-dependent hyperparameters to fit this posterior. (More specifically, we find the prior of the form $\boldsymbol{m}(s)+\boldsymbol{v}(s)\boldsymbol{y}(s)+\boldsymbol{d}(s)\boldsymbol{z}(s)$ which is closest to the posterior in the sense that the Kullback-Leibler divergence is minimized. This idea is borrowed from [5].) Thus $\boldsymbol{m}(s)$ is an estimate of the speaker's supervector when channel effects are abstracted and $\boldsymbol{d}(s)$ and $\boldsymbol{v}(s)$ measure the uncertainty in this estimate.

Set $\mathbf{\Lambda}(s) = (\boldsymbol{m}(s), \boldsymbol{u}, \boldsymbol{v}(s), \boldsymbol{d}(s), \mathbf{\Sigma})$.

2.4. The likelihood ratio statistic

Given speaker-independent hyperparameters $\mathbf{\Lambda}$ and enrollment data for a speaker s, we estimate a set of speaker-dependent hyperparameters $\mathbf{\Lambda}(s)$. Given speech data $\mathcal{X}(t)$ uttered by a test speaker t, to test the hypothesis that $t = s$ against the hypothesis that $t \neq s$ we use the likelihood ratio

$$\frac{1}{T}\log\frac{P_{\mathbf{\Lambda}(s)}(\mathcal{X}(t))}{P_{\mathbf{\Lambda}}(\mathcal{X}(t))} \quad (7)$$

where T is the duration of the test utterance.

3. EXPERIMENTS

3.1. Signal processing

Speech data was sampled at 8 kHz and 12 liftered mel frequency cepstral coefficients and a log energy parameter were calculated at a frame rate of 10 ms. The acoustic feature vector consisted of these 13 parameters together with their first derivatives. Cepstral mean subtraction was not performed since the channel factors in the factor analysis model can account for convolutional noise. Similarly, the energy feature was not normalized. Except where otherwise indicated we did not use a silence detector on the enrollment and test data.

3.2. Fitting the model to *Switchboard Cellular Part I*

The *Switchboard Cellular Part I* corpus contains stereo recordings of 1306 conversations (each of 6 minutes duration) involving 254 speakers (129 females and 125 males). We limited ourselves to 10 conversation sides per speaker for computational reasons. We processed the data with an echo canceller and we used a silence detector to segment the data into conversation turns. Conversation turns were padded with silences to roughly match the speech/silence distribution in the NIST 2001 cellular data. The total amount of data used was 94 hours (including silences). We refer to this data set as the *training set* to distinguish it from the enrollment data for the target speakers provided by NIST.

We used 12 hours of training data (one conversation side per speaker) to estimate a speaker-independent GMM with 2K Gaussians by Baum-Welch training and 10 iterations of the speaker-independent hyperparameter estimation procedure to fit a factor analysis with 40 channel factors and 40 speaker factors. An indication of how the model fits the data can be obtained by noting that the trace of $\boldsymbol{u}\boldsymbol{u}^*$ gives a measure of the amount of variability in the training set which is attributable to channel effects and the trace of $\boldsymbol{d}^2 + \boldsymbol{v}\boldsymbol{v}^*$ is a measure of inter-speaker variability. The figures are

$$\begin{array}{rcl} \mathrm{tr}\left(\boldsymbol{d}^2\right) & = & 2388 \\ \mathrm{tr}\left(\boldsymbol{v}\boldsymbol{v}^*\right) & = & 2.35 \times 10^7 \\ \mathrm{tr}\left(\boldsymbol{u}\boldsymbol{u}^*\right) & = & 1.78 \times 10^7. \end{array}$$

The first thing to note here is that $\boldsymbol{d}$ is negligeable compared to $\boldsymbol{v}$. Our reason for introducing $\boldsymbol{d}$ was to remedy the rank deficiency problem in eigenvoice MAP but these figures suggest that this may not be a real problem after all, at least for the GMM configuration and training set under consideration here. Even more striking is the fact that the channel variability in the training set is almost as large as the speaker variability which raises the question of how speaker recognition is possible at all. We can offer a plausible answer to this question by pretending that the supervectors $\boldsymbol{M}_h(s)$ are observable. If we treat $\boldsymbol{d}$ as negligeable, then a supervector $\boldsymbol{M}_h(s)$ can be written in the form $\boldsymbol{m} + \boldsymbol{s} + \boldsymbol{c}$ where $\boldsymbol{s}$ (the speaker contribution) lies in the range of $\boldsymbol{v}$ and $\boldsymbol{c}$ (the channel contribution) lies in the range of $\boldsymbol{u}$. Since the range of $\boldsymbol{v}$ and the range of $\boldsymbol{u}$ are 40-dimensional subspaces of a very high dimensional space they (typically) only intersect at the origin. It follows that a decomposition of the form $\boldsymbol{m}+\boldsymbol{s}+\boldsymbol{c}$ is necessarily unique. (If $\boldsymbol{m}+\boldsymbol{s}+\boldsymbol{c} = \boldsymbol{m}+\boldsymbol{s}'+\boldsymbol{c}'$ then $\boldsymbol{s}-\boldsymbol{s}' = \boldsymbol{c}'-\boldsymbol{c}$. The left hand side here is in the range

of $\boldsymbol{v}$ and the right hand side is in the range of $\boldsymbol{u}$ so the common has to be $\mathbf{0}$.) Thus $\boldsymbol{s}$ — and hence the identity of the speaker — is uniquely determined by $\boldsymbol{M}_h(s)$.

3.3. Tests on NIST 2001 cellular data

By taking *Switchboard Cellular Part I* as our training set we inadvertently used all of the NIST 2001 cellular test data as well as additional enrollment data for the target speakers in estimating the speaker-independent hyperparameters. Although we used only the enrollment data provided by NIST for speaker-dependent hyperparameter estimation, the fact that the speaker-independent hyperparameters serve as the starting point for speaker-dependent hyperparameter estimation means that our experiment results on the NIST 2001 data cannot be taken at face value.

In the 2001 evaluation there were 174 target speakers (74 males and 100 females). For each target speaker 2 minutes of enrollment data extracted from a single conversation side were provided. Enrollment data was also provided for 38 male and 22 female development speakers which we used as T-Norm speakers. Test utterance durations ranged from 15 to 45 seconds (with some exceptions). There were 2038 distinct test utterances (850 male, 1188 female) with 10 imposter trials and 1 target trial for each utterance. The data consists of whole conversation turns (so that silences account for about 25% of the total) and was processed with an echo canceller (although there are plenty of residual echoes).

The results of a speaker verification experiment using the model described in Section 3.2 and the likelihood ratio statistic (7) are given in line 1 of Table 1. Lines 2–6 report the results obtained with several variants of the factor analysis model. All variants had the same number of speaker and channel factors. Comparing lines 1 and 2 shows that T-Norm is effective. Comparing lines 2 and 3 shows that using a silence detector leads a slight degradation in the DCF and a slight improvement in the EER. Comparing lines 2 and 5 shows that making the factor analysis hyperparameters gender-dependent rather than gender-independent gives a substantial improvement. Increasing the number of Gaussians from 2048 to 4096 does not appear to help in either the gender-dependent or gender-independent case (compare line 2 with line 4 and line 5 with line 6).

	Gaussians	SD	GD	TN	DCF	EER
1	2048				0.026	4.0%
2	2048			X	0.021	3.7%
3	2048	X		X	0.023	3.4%
4	4096			X	0.023	3.4%
5	2048		X	X	**0.016**	**2.5%**
6	4096		X	X	0.017	2.5%

Table 1. *Speaker verification experiments on the NIST 2001 cellular test set with 'perfect' estimates of the speaker-independent hyperparameters. SD = Silence Detection, GD = Gender Dependent, DCF = Detection Cost Function, EER = Equal Error Rate.*

4. DISCUSSION

Considering that state of the art speaker verification systems generally obtain DCF's of about 0.03 and EER's of about 8% on the NIST evaluation sets the results in Table 1 are probably too good to be true and may be attributable in large part to the flaw in our experimental design. Unfortunately the NIST evaluation sets and the *Switchboard* Corpora have been designed in such a way as to make it difficult to test the factor analysis model properly (that is, with disjoint training and test sets). For example, using the NIST 2002 or 2003 evaluation data for testing and *Switchboard Cellular Part I* for training would not be appropriate since the evaluation data consists principally of CDMA transmissions and there are essentially no CDMA transmissions in the training data. Again, if we used the NIST 2000 evaluation data for testing and *Switchboard II Phase 3* for training, the speaker populations would be mismatched (the test speakers would all be from the American Midwest and Northeast and the training speakers from the South). We would like to be able to report results on standard test sets but this constraint may prove to be more of a hindrance than a help.

5. REFERENCES

[1] 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.

[2] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Processing*, in press.

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

[4] (2001) The NIST year 2001 speaker recognition evaluation plan. [Online]. Available: http://www.nist.gov/speech/tests/spk/2001/doc/2001-spkrec-evalplan-v05.9.pdf

[5] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker adaptation using an eigenphone basis," *IEEE Trans. Speech Audio Processing*, in press.