IMPROVEMENTS IN FACTOR ANALYSIS BASED SPEAKER VERIFICATION

Patrick Kenny, Gilles Boulianne, Pierre Ouellet and Pierre Dumouchel

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

ABSTRACT

We present the results of speaker verification experiments conducted on the NIST 2005 evaluation data using a factor analysis of speaker and session variability in 6 telephone speech corpora distributed by the Linguistic Data Consortium. We demonstrate the effectiveness of zt-norm score normalization and a new decision criterion for speaker recognition which can handle large numbers of t-norm speakers and large numbers of speaker factors at little computational cost. The best result we obtained was a detection cost of 0.016 on the core condition (all trials) of the evaluation.

1. INTRODUCTION

We extend the work presented in [1] by greatly increasing the size and diversity of the training set used to estimate factor analysis models, by introducing a new decision criterion for speaker verification which steers a middle course between the 'exact' and 'simplified' decision rules presented in [1] and by using new methods of score normalization. A more detailed treatment of the material in this paper can be found in [2] and the references cited there.

Whereas we used just 2 of the Switchboard Corpora for the experiments reported in [1], in the present article we use 5 Switchboard Corpora together with the NIST 2004 evaluation data for training factor analysis models.

When using the factor analysis model for speaker verification, the problem of enrolling a target speaker consists in estimating a supervector for the speaker in such a way as to suppress the channel effects in the enrollment utterance(s). It has always been our experience that it is advantageous to take account of the uncertainty entailed in estimating this speaker supervector which arises from the fact that the amount of enrollment data is limited and that channel effects can never be perfectly suppressed. Ignoring this uncertainty altogether led to the 'simplified' decision rule in [1] whose performance was found to be unsatisfactory when compared with the 'exact' decision rule but the exact decision rule is very computationally expensive. In this paper we present a new decision rule which handles this type of uncertainty in a computationally inexpensive way by shifting the computational burden to the enrollment phase. The new decision rule can handle large numbers of t-norm speakers and large numbers of speaker factors at very little computational cost and it has enabled us to obtain some excellent results on the 2005 set by using a very large number of speaker factors.

2. JOINT FACTOR ANALYSIS

Joint factor analysis is a model of speaker and session variability in Gaussian mixture models which are widely used in text-independent speaker recognition.

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 $\boldsymbol{m}$ denote the universal background supervector, that is, the supervector obtained by concatenating the mean vectors in a UBM with C mixtuer components. If $\boldsymbol{s}$ is the supervector for a randomly chosen speaker (that is, the supervector obtained by concatenating the mean vectors in a speaker dependent GMM having the same topology as the UBM) then we assume that $\boldsymbol{s}$ is distributed according to

$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{v}\boldsymbol{y} + \boldsymbol{d}\boldsymbol{z} \tag{1}$$

where $\boldsymbol{d}$ is diagonal, $\boldsymbol{v}$ is a rectangular matrix of low rank and $\boldsymbol{y}$ and $\boldsymbol{z}$ are independent random vectors having standard normal distributions. The components of $\boldsymbol{y}$ are *common speaker factors* and the components of $\boldsymbol{z}$ are *special speaker factors*. If $\boldsymbol{M}$ is the channel dependent supervector corresponding to a particular recording of the speaker we assume that

$$\boldsymbol{M} = \boldsymbol{s} + \boldsymbol{u}\boldsymbol{x} \tag{2}$$

where $\boldsymbol{u}$ is a rectangular matrix of low rank. The components of $\boldsymbol{x}$ are *channel factors*. We assume that $\boldsymbol{x}$ is also normally distributed (the question of whether this is an appropriate assumption is addressed in [3]). Finally, for each mixture component c there is a diagonal covariance matrix Σ_c whose role is to model the variability which is not captured by the speaker variability model (1) or the channel variability model (2). We denote by $\boldsymbol{\Sigma}$ the $CF \times CF$ supercovariance matrix whose diagonal is the concatenation of these covariance matrices.

For the experiments reported in this paper we trained two gender dependent factor analysis models using the following data bases: 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. As in [1], the acoustic features that we used were Gaussianized cepstral features and their first derivatives. For each gender the hyperparameters $\boldsymbol{m}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{d}$ and $\boldsymbol{\Sigma}$ were estimated using the simplified training algorithm in [1] subject to one modification, namely that we skipped the 'adaptation to the target speaker population' step mentioned in Section 3 of that paper in order to adhere strictly to the NIST evaluation protocol.

3. SPEAKER VERIFICATION

We now explain how we use the joint factor analysis model to construct a speaker verification system. We have to describe how

1-4244-0469-X/06/$20.00 ©2006 IEEE

we estimate a GMM supervector for each target speaker, how we evaluate the likelihood of a test utterance using a target speaker GMM and how we normalize likelihoods calculated in this way so that a common decision threshold can be used in all speaker verification trials.

3.1. Enrolling a target speaker

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

As we mentioned in the introduction any estimate of a target speaker's supervector is necessarily imprecise and taking account of this type of uncertainty can result in decision rules for speaker verification which are very computationally expensive. In this paper we propose a new way of addressing this question which is no more computationally expensive than the simplified scoring method in [1]. This entails calculating not only the posterior expectation of the target speaker's supervector $\boldsymbol{s}$ which, following the notation in [4], we denote by $E\left[\boldsymbol{s}\right]$ at enrollment time but also the diagonal of posterior covariance matrix of $\boldsymbol{s}$ which we denote by $\operatorname{Cov}\left(\boldsymbol{s}, \boldsymbol{s}\right)$. (Although we calculated $\operatorname{Cov}\left(\boldsymbol{s}, \boldsymbol{s}\right)$ exactly for the experiments reported here it has been our experience that the approximation

$$\operatorname{Cov}\left(\boldsymbol{s}, \boldsymbol{s}\right) \simeq \operatorname{diag}\left(\boldsymbol{v} \operatorname{Cov}\left(\boldsymbol{y}, \boldsymbol{y}\right) \boldsymbol{v}^*\right) + \boldsymbol{d} \operatorname{Cov}\left(\boldsymbol{z}, \boldsymbol{z}\right) \boldsymbol{d}. \qquad (3)$$

which ignores the cross correlations between $\boldsymbol{y}$ and $\boldsymbol{z}$ works quite well in practice. This is easy to implement (the calculation of the posterior covariances in (3) is explained in Section III of [4]) and it gives exact results in the two cases which are of greatest interest, namely $\boldsymbol{d} = \boldsymbol{0}$ and $\boldsymbol{v} = \boldsymbol{0}$.

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

3.2. The likelihood function

Suppose we are given a target speaker and a test utterance and that we wish to test the null hypothesis that the utterance speaker is different from the target speaker against the alternative hypothesis that the two speakers are the same. Denote the speaker supervector for the target speaker by $\boldsymbol{s}$ and denote the test utterance by $\mathcal{X}$. If we assume to begin with that $\boldsymbol{s}$ is known the likelihood of $\mathcal{X}$ under the alternative hypothesis — let us denote it by$P(\mathcal{X}|\boldsymbol{s})$ — can be calculated by the methods in [5]. By (2) there is a random vector $\boldsymbol{x}$ such that the speaker- and channel-dependent supervector for the test utterance is

$$\boldsymbol{s} + \boldsymbol{u}\boldsymbol{x}. \qquad (4)$$

If $\boldsymbol{x}$ was known, we would know the value of this supervector so it would be straightforward matter to calculate the conditional (Gaussian) likelihood of the test utterance, $P(\mathcal{X}|\boldsymbol{s}, \boldsymbol{x})$, using the Baum-Welch statistics extracted from the utterance (Lemma 1 in [5]). So, since $\boldsymbol{x}$ is assumed to have a standard normal distribution, $P(\mathcal{X}|\boldsymbol{s})$ is given by

$$P(\mathcal{X}|\boldsymbol{s}) = \int P(\mathcal{X}|\boldsymbol{s}, \boldsymbol{x}) N(\boldsymbol{x}|\boldsymbol{0}, \boldsymbol{I}) d\boldsymbol{x} \qquad (5)$$

where $N(\cdot|\boldsymbol{0}, \boldsymbol{I})$ is the standard Gaussian kernel. Proposition 2 in [5] 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 N_c be the total number of observation vectors in $\mathcal{X}$ for the given mixture component and set

$$F_c = \sum_t X_t \qquad (6)$$

$$S_c = \operatorname{diag}\left(\sum_t X_t X_t^*\right) \qquad (7)$$

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. We use gender-dependent UBM's to perform the alignments.) 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 concatenatingF_c (for $c = 1, \ldots, C$). Similarly, let $\boldsymbol{S}$ be the $CF \times CF$ diagonal matrix whose diagonal blocks are S_c (for $c = 1, \ldots, C$). We denote the first and second order moments of $\mathcal{X}$ around $\boldsymbol{s}$ by $\boldsymbol{F_s}$ and $\boldsymbol{S_s}$ so that

$$\begin{aligned} \boldsymbol{F_s} &= \boldsymbol{F} - \boldsymbol{N}\boldsymbol{s} \\ \boldsymbol{S_s} &= \boldsymbol{S} - 2\operatorname{diag}\left(\boldsymbol{F}\boldsymbol{s}^*\right) + \operatorname{diag}\left(\boldsymbol{N}\boldsymbol{s}\boldsymbol{s}^*\right). \end{aligned} \qquad (8)$$

Finally, let

$$\boldsymbol{l} = \boldsymbol{I} + \boldsymbol{u}^* \boldsymbol{\Sigma}^{-1} \boldsymbol{N} \boldsymbol{u}, \qquad (9)$$

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

$$\boldsymbol{l} = \boldsymbol{l}^{1/2} \boldsymbol{l}^{1/2*} \qquad (10)$$

(that is, the Cholesky decomposition of $\boldsymbol{l}$). Then applying some algebraic manipulations to the formula given in the statement of Proposition 3 in [4] leads to the following expression for the likelihood function:

$$\begin{aligned} \log P(\mathcal{X}|\boldsymbol{s}) &= \sum_{c=1}^{C} N_c \log \frac{1}{(2\pi)^{F/2} |\Sigma_c|^{1/2}} \\ &\quad - \frac{1}{2} \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1} \boldsymbol{S_s}\right) - \frac{1}{2} \log |\boldsymbol{l}| \\ &\quad + \frac{1}{2} \|\boldsymbol{l}^{-1/2} \boldsymbol{u}^* \boldsymbol{\Sigma}^{-1} \boldsymbol{F_s}\|^2 \end{aligned} \qquad (11)$$

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

$$\begin{aligned} E\left[\boldsymbol{F_s}\right] &= \boldsymbol{F} - \boldsymbol{N}E\left[\boldsymbol{s}\right] \\ E\left[\boldsymbol{S_s}\right] &= \boldsymbol{S} - 2\,\mathrm{diag}\left(\boldsymbol{F}E\left[\boldsymbol{s}^*\right]\right) \\ &\quad + \mathrm{diag}\big(\boldsymbol{N}(E\left[\boldsymbol{s}\right]E\left[\boldsymbol{s}^*\right] \\ &\quad + \mathrm{Cov}\left(\boldsymbol{s},\boldsymbol{s}\right))\big) \qquad (12) \end{aligned}$$

(in accordance with the notation introduced in Section 3.1). Because the term $\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S_s}\right)$ enters into (11) with a negative sign, the effect of including the term $\mathrm{Cov}\left(\boldsymbol{s},\boldsymbol{s}\right)$ in (12) is to diminish the value of the likelihood function by an amount which is inversely proportional to the amount of the speaker's enrollment data. (In order to ensure that the same criterion is used in training and testing we incorporate a similar modification to the Baum-Welch statistics in training the factor analysis model.)

The most interesting thing to note about (11) is that the likelihood function depends on the hypothesized speaker only through the computations in (8) and the cost of these computations is negligible (since $E\left[\boldsymbol{s}\right]$ and $\mathrm{Cov}\left(\boldsymbol{s},\boldsymbol{s}\right)$ are calculated at enrollment time). The principal computation is the evaluation of $\boldsymbol{l}^{-1/2}$ (the value of the determinant $|\boldsymbol{l}|$ is a by-product) and this only needs to be done once (independently of the number of speakers hypothesized and the number of t-norm speakers). Note also that the number of common speaker factors has a major effect on the computational burden of evaluating the terms $E\left[\boldsymbol{s}\right]$ and $\mathrm{Cov}\left(\boldsymbol{s},\boldsymbol{s}\right)$ in (12) but these terms are only evaluated at enrollment time. The calculation in (12) is completely insensitive to the number of common speaker factors; this is another major advantage over the decision criterion used in [6].

3.3. Score normalization

In our first experiments with the factor analysis model we used only t-norm for score normalization but we learned from [7] that zt-norm (that is, z-norm followed by t-norm and not the other way round) could be very effective at least in the case where $\boldsymbol{v} = \boldsymbol{0}$. Unlike t-norm, z-norm requires a way of evaluating the likelihood of a test utterance under the assumption that the actual speaker is somebody other than the hypothesized speaker — the 'unknown speaker' as it were. For the unknown speaker it is natural to take $E\left[\boldsymbol{s}\right] = \boldsymbol{m}$; as for $\mathrm{Cov}\left(\boldsymbol{s},\boldsymbol{s}\right)$ we investigated two possibilities:

$$\mathrm{Cov}\left(\boldsymbol{s},\boldsymbol{s}\right) = \mathrm{diag}\left(\boldsymbol{v}\boldsymbol{v}^* + \boldsymbol{d}^2\right) \qquad (13)$$

$$\mathrm{Cov}\left(\boldsymbol{s},\boldsymbol{s}\right) = \frac{1}{N}\sum_{n=1}^{N}\mathrm{Cov}\left(\boldsymbol{s}_n,\boldsymbol{s}_n\right) \qquad (14)$$

where the sum on the right hand side of (14) extends over the set of t-norm speakers and N is the number of t-norm speakers. We will refer these two types of score normalization as z-norm and z′-norm respectively. We used 120 t-norm speakers for each gender and 120 z-norm utterances (20 from each of the databases that we used for development). Our experience has been that z'-norm is much more effective than z-norm if common speaker factors are included in the speaker variability model (1).

4. EXPERIMENTS

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

4.1. No common speaker factors

Our first experiments were carried out with $\boldsymbol{v} = \boldsymbol{0}$. (This is the configuration of the factor analysis model which most resembles the traditional GMM/UBM approach.) We carried out the trials in both the forward direction (that is, with the training and test utterance designations given by NIST) and in the reverse direction. The two strategies give essentially the same overall results but averaging the results gives a small improvement in DCF.

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

Table 1. *Joint factor analysis with no common speaker factors and 25 channel factors. F = forward, R = reverse.*

Our best result with this type of configuration of the joint factor analysis model on the NIST 2005 test set, namely an EER of 6.2% and a DCF of 0.019, was obtained by 50 channel factors in the same way as the result in the last line of Table 1. (Increasing the number of channel factors from 50 to 100 gave essentially the same results.)

4.2. 300 common speaker factors

We now turn to the opposite extreme where the number of common speaker factors is very large, namely 300. In this situation, the speaker variability model (1) behaves like a pure eigenvoice model (i.e. $\boldsymbol{d} \simeq \boldsymbol{0}$). The effects of various types of score normalization are shown in Table 2.

The general trend is that, just as we found in the case where the number of common speaker factors was 0, zt-norm is more effective than z-norm and z-norm is more effective than t-norm but in this situation the z′ flavor of z-norm is the more effective. Note that the best result in Table 2 (EER = 5.2%, DCF = 0.017) is considerably better than the best result we obtained with no common speaker factors (EER = 6.6%, DCF = 0.021).

4.3. Varying the number of common speaker factors

So far we have only considered the two extreme cases where the the number of common speaker factors is zero or very large. Results obtained with different numbers of common speaker factors and 50 channel factors are reported in Table 3. Adding a small number of speaker factors (1 or 5) is seen to hurt performance particularly as measured by the the DCF.

Our reason for developing the speaker variability model (1) was to try to take advantage of the complementary strengths of

normalization	trial type	EER	DCF
t-norm	F	9.0%	0.034
t-norm	R	8.9%	0.033
z-norm	F	9.2%	0.033
z-norm	R	7.5%	0.030
z$'$-norm	F	6.8%	0.026
z$'$-norm	R	6.2%	0.023
zt-norm	F	7.4%	0.023
zt-norm	R	6.7%	0.022
zt-norm	F+R	6.6%	0.020
z$'$t-norm	F	5.4%	0.018
z$'$t-norm	R	**5.2%**	**0.017**
z$'$t-norm	F+R	5.3%	0.017

Table 2. *Joint factor analysis with 300 common speaker factors and 50 channel factors.*

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

Table 3. *Joint factor analysis with varying numbers common speaker factors and 50 channel factors. Forward scoring only. z$'$t score normalization*

classical MAP and eigenvoice MAP in estimating target speaker models from limited amounts of data. But the results in Table 3 show that the best performance is obtained in the two extreme cases where $\boldsymbol{d} = \mathbf{0}$ and $\boldsymbol{v} = \mathbf{0}$ and this suggests that fusion at the score level may be the best strategy for achieving this goal. It turns out that a linear fusion of four systems, namely forward and reverse scoring with 0 common speaker factors and 300 common speaker factors does indeed give a slightly improved value of the detection cost function (0.016 versus 0.017) when compared with the best result in Table 2 but there is a slight degradation in the equal error rate (5.4% versus 5.2%). Thus more sophisticated fusion techniques such as logistic regression or a multilayer perceptron may be worth investigating.

5. DISCUSSION

The NIST 2005 test set presents an interesting challenge for the joint factor analysis model because there is reason to doubt that the model can be properly trained for this task using currently available telephone speech corpora which do not reflect the variety of channel conditions encountered the in evaluation data.

Indeed it is something of a challenge even to find speakers who have been recorded over both landline and cellular transmission channels. As a rule, each speaker in the Switchboard collections was recorded over either landline channels or cellular channels but not both. Only a small fraction of the speakers in the Fisher English database were recorded more than once (and furthermore the speaker identifications in this database are not reliable). So if a joint factor analysis model is trained on the union of the telephone speech corpora currently available through the Linguistic Data Consortium, the model could be misled into believing that some speakers are 'landline speakers' and others are 'cellular speakers'. Fortunately this effect does not appear to be too serious.

The success of our approach is due in large part to the effectiveness of the zt-norm technique [7] and to the new scoring procedure described in Section 3.2 which enabled us to turn around a large number of experiments very quickly because of the efficiency with which it handles t-norm speakers. A remarkable feature of this scoring procedure is that its computational cost is independent of the number of common speaker factors in the factor analysis model. This enabled us to experiment with large numbers of common speaker factors and obtain some excellent results.

In the the extreme case where the number of common speaker factors is very large (e.g. 300), the factor analysis model of speaker variability behaves essentially like an eigenvoice model ($\boldsymbol{d} \simeq \mathbf{0}$). It may be that the reason why this model performs so well is that it implicitly models long term features. (Eigenvoice methods take account of the correlations between the various Gaussians in a speaker model.)

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

6. REFERENCES

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

[2] ——, "Joint factor analysis versus eigenchannels in speaker recognition," submitted to *IEEE Trans. Audio Speech and Language Processing*. [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

[3] P. Kenny, P. O. G. Boulianne, and P. Dumouchel, "The geometry of the channel space in GMM-based speaker recognition," in *Proc. IEEE Odyssey 2006*, San Juan, Puerto Rico, June 2006 (submitted).

[4] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and alogorithms." [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

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

[6] 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*. [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

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