SPEAKER RECOGNITION BASED ON DISCRIMINATIVE PROJECTION MODELS

Jesper Ø. Olsen

Center for PersonKommunikation, Aalborg University,
Fredrik Bajers Vej 7A-6, DK-9220 Aalborg Øst, Denmark

email: jo@cpk.auc.dk, http://www.kom.auc.dk/∼jo

ABSTRACT

A new discriminant speaker model is introduced in this paper. The model is text dependent and relies on characterising speakers in terms of the angular distance between "projection vectors", which allow good discrimination between individual speakers. The projection models require only little enrolment data to be available per target speaker, but at the same time require a set of "cohort speakers" to be available for which a relatively large amount of training speech is available per cohort speaker. The projection model technique is evaluated on the Gandalf database and compared to conventional Gaussian Mixture Models (GMMs). It is found that the projection models require less storage per target speaker, while at the same time achieving lower error rates, particularly when applied for speaker identification and recognition under mismatched conditions.

1. INTRODUCTION

Speaker recognition is the generic term used for two related classification problems: *speaker verification*, where the task is to verify a claimed speaker identity (2-way classification) and *speaker identification* where the task is to identify a speaker from a group of known speakers (N-way classification). An automatic speaker recognition system can be characterised in terms of

- The classification error rate.
- The amount of enrolment speech needed.
- The number of free parameters.

The classification error rate is generally important for speaker recognition to be useful in applications and a competitive alternative to other identification/verification techniques. The enrolment speech is important in commercial applications, where speakers (customers) may be unwilling to spend time on a long and tedious enrolment procedure. The number of free parameters or the complexity of the models is important for computational reasons, and because it is a limitation for recognition systems with many users, e.g. in a nation wide speaker verification system, storage of speaker models and rapid access to specific models can be a major problem.

In order to minimise the classification error rate, classification decisions should ideally be based on the a posteriori class probabilities given the observed speech signal $\mathbf{X} = \vec{x}_1, \ldots, \vec{x}_T$. For speaker identification this means we should identify the speaker for which $P(\lambda_i|\mathbf{X})$ is greatest:

$$\text{Decide: speaker } i, \text{ where } P(\lambda_i|\mathbf{X}) \geq P(\lambda_j|\mathbf{X}), \forall j \neq i \quad (1)$$

where the λ_i's represent the individual speaker models in the system. For speaker verification we should accept the identity claim if the probability of the target speaker is greater than the probability of an impostor speaker:

$$\text{Decide} \begin{cases} \text{accept} & \text{if } g(\mathbf{X}) = \frac{P(\lambda_{\text{tar}}|\mathbf{X})}{P(\lambda_{\text{imp}}|\mathbf{X})} > 1 \\ \text{reject} & \text{otherwise} \end{cases} \quad (2)$$

The a posterior class probabilities may in practise be estimated directly using *discriminative* models (e.g. neural networks), which model the decision surfaces of the classification problem, or they may be estimated indirectly using *non-discriminative* models, which separately model the Probability Density Functions (pdf's) characterising the individual classes. In the last case, the a posterior class probabilities are then computed from the pdf estimates using Bayes rule:

$$P(\lambda_i|\mathbf{X}) = \frac{P(\lambda_i)p(\mathbf{X}|\lambda_i)}{p(\mathbf{X})} \quad (3)$$

Discriminative models have the advantage that the modelling is focused on feature variations which are particularly relevant for the classification problem. Non-discriminative models, however, are typically easier to train, when the number of different classes is large, because the modelling problem is decomposed into a number of simpler subproblems, namely the construction of approximations for the pdf's, $p(\mathbf{X}|\lambda_i)$, characterising the individual classes. When non-discriminative models are used, speaker identification decisions can in practise be based directly on the class conditional likelihoods, $p(\mathbf{X}|\lambda_i)$, since $p(\mathbf{X})$ is common to all speakers. For speaker verification it is here necessary to specifically construct a "speaker model" for the impostor class. This process is generally referred to as *cohort normalisation* or *score normalisation* [1, 2]:

$$g(\mathbf{X}) \approx \frac{p(\mathbf{X}|\lambda_{\text{tar}})}{\frac{1}{C}\sum_{c=1}^{C} p(\mathbf{X}|\lambda_{\text{coh},c})} \quad (4)$$

where the cohort set $\{\lambda_{\text{coh},c}|1 \leq c \leq C\}$ consists of C speaker models, which have been selected to represent the impostor class.

The remainder of this paper is organised as follows. Section 2 gives a brief introduction to Gaussian Mixture Models (GMMs), which is one of the most popular non-discriminative speaker models used in state-of-the-art automatic speaker recognition systems; GMMs are used for creating baseline results in this study. Section 3 describes an alternative, discriminative speaker model: the *projection model*. Section 4 summarises the speech database used for the experimental part of the paper and finally section 5 reports the results of a set of speaker verification and identification experiments, where the projection models were evaluated.

2. GAUSSIAN MIXTURE MODELS

A popular modelling technique for building speaker models is to use Speaker Dependent (SD) GMMs for modelling the likelihood of the speech signal [3]:

$$p(\mathbf{X}|\lambda_i) = \prod_{t=1}^{T} p(\vec{x}_t|\lambda_i) \tag{5}$$

where

$$p(\vec{x}|\lambda_i) = \sum_{m=1}^{M} c_m \mathcal{N}(\vec{x}; \vec{\mu}_m, \mathbf{U}_m) \tag{6}$$

where M is the number of mixtures and

$$\mathcal{N}(\vec{x}; \vec{\mu}_m, \mathbf{U}_m) = \frac{1}{(2\pi)^{D/2}|\mathbf{U}_m|^{0.5}} e^{-0.5(\vec{x}-\vec{\mu}_m)^{\mathrm{TRP}}\mathbf{U}_m^{-1}(\vec{x}-\vec{\mu}_m)} \tag{7}$$

A GMM does not necessarily assume that the features are normally distributed: provided a GMM has a sufficient number of mixtures, it can be shown that it can approximate any distribution arbitrarily well. For text independent modelling this property is vital, because a normal distribution does not characterise the feature variations well and many mixtures (e.g. 32–256) are needed. For text (e.g. phoneme) dependent models the Gaussian assumption is better and GMMs with fewer mixtures can be employed. In general text dependent speaker modelling is more accurate than text independent modelling, because the acoustic overlap between the speakers is smaller and discrimination therefore easier.

3. PROJECTION MODELS

Constraining assumptions about the feature distributions characterising observations from different classes can significantly simplify the construction of a classifier. For phoneme dependent speaker models a normal assumption is not bad, but can be made to fit even better by using segmental features: feature vectors are extracted which represent phoneme segments rather than individual speech frames [4, 5]. This can, for instance, be done by first explicitly identifying the phoneme segments using a speech recogniser and then time sampling each phoneme segment in order to obtain a fixed dimensional representation.

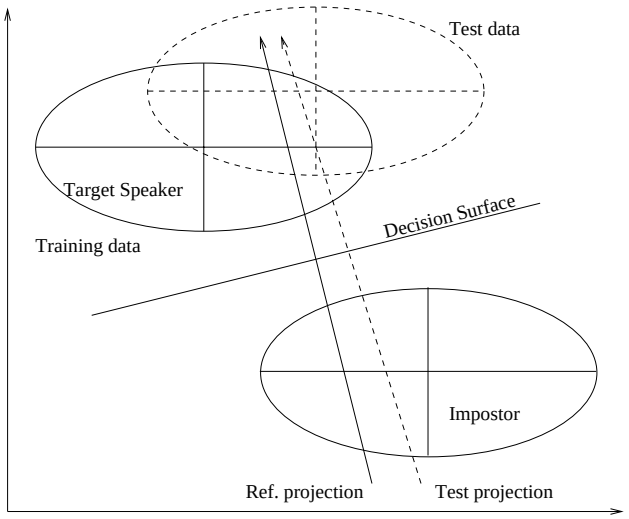

Figure 1: A reference and a test projection component.

If we have a binary classification problem (speaker verification) and assume that the two classes are characterised by normal distributions with different means, but the same covariance matrix:

$$p(\vec{x}|\lambda_{\text{tar}}) = \mathcal{N}(\vec{x}; \vec{\mu}_{\text{tar}}, \mathbf{U}) \tag{8}$$

$$p(\vec{x}|\lambda_{\text{imp}}) = \mathcal{N}(\vec{x}; \vec{\mu}_{\text{imp}}, \mathbf{U}) \tag{9}$$

use of equation 2 leads to Fisher's linear discriminant function:

$$\text{Decide} \begin{cases} \text{target} & \text{if } \vec{a}(\vec{x}-\vec{b}) > \ln\left(\frac{P(\lambda_{\text{imp}})}{P(\lambda_{\text{tar}})}\right) \\ \text{impostor} & \text{otherwise} \end{cases} \tag{10}$$

where

$$\vec{a} = \mathbf{U}^{-1}(\vec{\mu}_{\text{tar}} - \vec{\mu}_{\text{imp}}) \tag{11}$$

and

$$\vec{b} = 0.5(\vec{\mu}_{\text{tar}} + \vec{\mu}_{\text{imp}}) \tag{12}$$

However, even though speaker verification is a binary classification problem, equation 10 can not be used directly, because the impostor class can not be modelled well by a single normal distribution. If a normal distribution characterises the feature distribution from individual speakers, equation 10 can, however, be used for discriminating well between pairs of individual speakers.

When equation 10 is used, a hyper plane is used as the *decision surface* separating the two classes. The hyper plane is characterised by the normal vector $\vec{a}$ (equation 11). The class identity is determined by projecting the test sample onto $\vec{a}$ in order to find out on which side of the decision surface the sample is positioned (see figure 1); the normal vector $\vec{a}$ will here be referred to as a *projection vector*. In [4, 6] a number of such projection vectors were computed from a set of available "cohort speakers" and used for constructing a linear speaker dependent speech transform (the Fisher transform). The Fisher transform has been shown to be an effective way of preprocessing the feature vectors before actually trying to determine the class identities [6].

The projection vectors characterise a target speaker, and rather than using them for constructing a speech transform, they can in themselves be used as a speaker model, which here will be referred to as a *projection model*. In order to estimate a projection model, one needs a set of cohort speakers to be available, and for each cohort speaker a relatively large amount of speech should be available, so that for each projection component, robust estimates of the covariance matrix, $\mathbf{U}$, and the cohort speaker mean, $\vec{\mu}_{\text{imp}}$, can be obtained. Ideally the covariance matrix for a given component, should be estimated from both the target speaker enrolment speech and the speech from the relevant cohort speaker, but in practise the available enrolment speech from the target speaker is severely limited and will not make much impact. Hence, it is suggested here that only the target speaker mean vector, $\vec{\mu}_{\text{tar}}$ be estimated from the target speaker enrolment data.

A reference projection model is estimated from the enrolment speech and in the test situation, a "test projection model" is estimated for the same set of cohort speakers as for the reference projection model. The test samples are here used directly as estimates of $\vec{\mu}_{\text{tar}}$. The distance between two projection components

can be measured in terms of the angle between the reference projection, $\vec{a}_{\text{ref}}$, and the test projection, $\vec{a}_{\text{test}}$:

$$\Theta(\vec{a}_{\text{ref}}, \vec{a}_{\text{test}}) = \text{acos}\left(\frac{\vec{a}_{\text{ref}} \cdot \vec{a}_{\text{test}}}{|\vec{a}_{\text{ref}}||\vec{a}_{\text{test}}|}\right) \quad (13)$$

An angle of 0° indicates a perfect match ($\vec{a}_{\text{ref}}$ and $\vec{a}_{\text{test}}$ are parallel and in the same direction), whereas an angle of 180° indicates a maximally poor match ($\vec{a}_{\text{ref}}$ and $\vec{a}_{\text{test}}$ are parallel but point in opposite directions). The distance between a reference projection model, $A_{\text{ref}} = \vec{a}_{\text{ref},1}, \ldots, \vec{a}_{\text{ref},M}$, and a test projection model, $A_{\text{test}} = \vec{a}_{\text{test},1}, \ldots, \vec{a}_{\text{test},M}$, is computed as the average sum of the angles between the M individual components in the two projection models:

$$\Theta(A_{\text{ref}}, A_{\text{test}}) = \frac{1}{M}\sum_{i=1}^{M}\Theta(\vec{a}_{\text{ref},i}, \vec{a}_{\text{test},i}) \quad (14)$$

The projection models are easy to estimate and require relatively little storage per target speaker, because no matter how many components are included in the model, only $\vec{\mu}_{\text{tar}}$ needs to be stored: $\vec{\mu}_{\text{imp}}$ and $\mathbf{U}$ are part of the "cohort library", which is shared between all target speakers.

4. SPEECH DATA

In this work, the Swedish Gandalf database [7] was used. The database contains speech recorded over the public telephone network; The target speaker set consists of 58 speakers (23 female + 35 male) recorded over a one year period, and the impostor set consists of 77 speakers (28 female + 49 male). The speech items from Gandalf that were used in these experiments consist of digit strings. For enrolment purposes, the target speakers were prompted for 25 5-digit utterances in a single session (session 1). Utterances that contained speaker or technical recording errors were removed from the test and training sets. On average this meant that 12.1 training tokens were available per digit for each target speaker (see table 1). The test trials (sessions 2–28) were based on 4-digit utterances: each speaker verification decision was based on one such utterance. The test utterances consist of two parts: the *favorite* part, where the target speakers used the same telephone handset as in the enrolment call, and the *non-favorite* part, where a number of different telephone handsets were used: all different from the favorite handset.

The speech data was parameterised as the logarithmic energy outputs of a filter bank with 24 triangular filters spaced linearly along the logarithmic mel scale; each filter overlapped 50% with each of its two neighbours. Feature vectors were extracted using a 25.6 ms Hamming window and a 10 ms frame period. Phoneme segments were identified by forced Viterbi decoding using SI HMMs. For the projection models, the phoneme segments were represented by extracting three feature vectors from each phoneme segment (one feature vector per emitting state in the HMM phoneme models). The three 24 dimensional vectors were concatenated to form one long 72 dimensional *phoneme vector*. In order to eliminate the signal gain, the phoneme vectors were normalised to have norm one [5].

word	count	transcription
noll	12.5	/n O l/
ett	11.6	/e t/
två	11.6	/t v o:/
tre	11.6	/t r e:/
fyra	13.6	/f Y r a/
fem	11.6	/f e m/
sex	12.6	/s e k s/
sju	11.5	/S u0/
åtta	12.5	/O t a/
nio	11.5	/n I U/

Table 1: Word transcriptions (SAMPA) and frequencies in the enrolment data.

5. EXPERIMENTS & RESULTS

A number of experiments were conducted in order to evaluate the usefulness of the projection models for both speaker verification and speaker identification. In each experiment, a total of 30 speaker models were trained for each target speaker corresponding to the 30 within word triphones needed to transcribe the digits (see table 1). Speaker models were constructed for different numbers of training tokens per digit: 1, 2, 5 and all tokens (on average 12.1 tokens/digit).

In order to compare the results to alternative speaker modelling techniques, a set of baseline experiments were conducted using GMMs as speaker models [3]. As for the projection models, 30 triphone models were trained for each target speaker and in the evaluation the same speech segmentations were used. However, the GMMs were based on a *frame level* evaluation rather than on a *phoneme segment* evaluation: all the speech frames in a phoneme segment were used. The cosine transform was used for converting the filter bank coefficients into cepstral coefficients; all coefficients were retained ($c_1, \ldots, c_{24}$). The GMMs where experimentally optimised with regard to the number of components to include in the models. The covariance matrices were diagonal and were tied between the mixtures within one GMM; a minimum variance floor was imposed by means of a gender dependent estimate of the variance for the particular phoneme.

5.1. Speaker Verification

Table 2 summarises the evaluation of the projection models when used for speaker verification. The table shows the EERs as a function of the number of components (#cmp) in the model; the "count" column indicates the number of training tokens per digit that were used for estimating $\vec{\mu}_{\text{tar}}$. The evaluation was based on 9413 same sex impostor speaker trials and respectively 5115 favorite and 1769 non-favorite target speaker trials. The EERs were computed a posteori using a single speaker independent decision threshold for balancing the false acceptance and false rejection error rates. Further speaker verification results are reported for this test set in [6].

Table 3 summarises the baseline speaker verification experiments, where GMMs where used as speaker models. The number of mixtures in each phoneme dependent GMM was respectively 2 (#count=1), 4 (#count=2), 8 (#count=5) and 8 (#count$\approx$12.1).

For both the GMM and the projection model experiments, the telephone handset is seen to play a major role: the favorite EERs are 2–4 times lower than the non-favorite EERs. For both test sets the projection models need approximately 20 components in order to characterise the target speakers well. For the GMMs, at

Projection models; Favorite handset				
#cmp	**count = 1**	**count = 2**	**count = 5**	**count ≈ 12.1**
1	25.7%	22.4%	17.6%	15.8%
2	22.3%	18.1%	13.3%	11.6%
5	16.7%	11.8%	8.2%	6.7%
10	13.1%	9.1%	5.6%	4.7%
20	12.5%	8.3%	5.2%	4.1%
40	11.7%	7.9%	4.9%	4.0%

Projection models; Non-favorite handset				
#cmp	**count = 1**	**count = 2**	**count = 5**	**count ≈ 12.1**
1	33.9%	30.3%	27.3%	26.3%
2	30.7%	26.9%	24.4%	23.0%
5	24.6%	20.6%	18.0%	16.8%
10	21.4%	18.0%	15.6%	14.2%
20	21.4%	18.0%	15.3%	13.6%
40	20.7%	17.8%	14.6%	13.2%

Table 2: Speaker verification EERs using projection models.

GMMs; Favorite handset				
#cohorts	**count = 1**	**count = 2**	**count = 5**	**count ≈ 12.1**
0	19.0%	10.6%	7.2%	6.6%
1	19.2%	14.6%	9.4%	9.3%
2	17.5%	10.8%	6.2%	5.7%
5	17.8%	9.8%	5.4%	4.7%
10	17.7%	9.4%	5.3%	4.5%
20	17.7%	9.4%	5.7%	4.7%
40	18.0%	9.2%	5.7%	4.5%

GMMs; Non-favorite handset				
#cohorts	**count = 1**	**count = 2**	**count = 5**	**count ≈ 12.1**
0	30.5%	25.9%	23.9%	23.5%
1	30.0%	26.9%	24.3%	24.1%
2	28.5%	24.3%	21.3%	21.1%
5	27.7%	22.6%	19.7%	19.4%
10	27.9%	22.2%	19.9%	18.7%
20	28.3%	22.4%	20.1%	19.3%
40	28.3%	22.3%	19.7%	18.8%

Table 3: Speaker verification EERs using GMMs.

least 5 cohort speakers are needed for score normalisation. The score normalisation, however, is seen to be only really effective when the speaker models actually characterise the target speakers well, i.e. when a relatively large amount of enrolment speech is available.

For the favorite handset the projection model EERs are 10–15% lower than the corresponding GMM EERs – for the non-favorite handset the EER reduction is 20–30%: the GMMs are somewhat more sensitive to the change of handset.

With regard to complexity, the projection models require that 72 coefficients ($\vec{\mu}_{\text{tar}}$) be stored per projection model. In addition to this a 72×72 matrix and a 72 dimensional vector needs to be stored per component, but as mentioned previously, these data are shared between all the target speakers in the system as are the cohort speaker models for the GMMs. The GMMs generally require more storage than the projection models: the GMMs with eight mixture components used for the **count=5** and **count≈12.1** experiments require three times as many parameters as the projection models (i.e. $(8+1)\cdot 24 = 216$ coefficients per model), and the GMMs with two mixture components used for the **count=1** experiments require the same amount of storage as the projection models.

5.2. Speaker Identification

Table 4 summarises the evaluation of the projection models when used for speaker identification. The set of speakers to be identified consisted of the 58 speakers used as target speakers in the verification experiments. Table 5 summarises the corresponding identification error rates when GMMs are used. The telephone handset again plays a major role and the GMMs in particular are sensitive to the handset change. Compared to the GMM error rates, the projection model error rates are reduced by 10–20% for the favorite handset and 30–40% for the non-favorite handset.

Projection models; Favorite handset				
#cmp	**count = 1**	**count = 2**	**count = 5**	**count ≈ 12.1**
1	73.2%	63.7%	56.3%	50.4%
2	62.7%	52.1%	40.9%	34.4%
5	46.0%	33.4%	21.2%	16.8%
10	36.9%	26.1%	13.5%	10.0%
20	32.2%	22.1%	11.5%	7.8%
40	27.7%	16.4%	8.0%	5.2%

Projection models; Non-favorite handset				
#cmp	**count = 1**	**count = 2**	**count = 5**	**count ≈ 12.1**
1	83.9%	81.4%	77.8%	74.7%
2	78.7%	73.4%	66.9%	62.9%
5	65.2%	57.4%	47.7%	44.5%
10	59.1%	50.5%	39.5%	36.1%
20	55.6%	47.8%	37.6%	33.3%
40	50.3%	39.6%	29.2%	25.0%

Table 4: Speaker identification error rates using projection models.

GMMs				
handset	**count = 1**	**count = 2**	**count = 5**	**count ≈ 12.1**
favorite	39.1%	18.8%	8.9%	6.5%
non-favorite	70.5%	56.0%	48.7%	44.4%

Table 5: Speaker identification error rates using GMMs.

6. CONCLUSIONS

In this paper, projection models have been introduced and evaluated for speaker verification and speaker identification. The projection models make strong parametric assumptions about the feature distributions characterising different speakers, and these constraints limits the complexity of the models. Compared to baseline experiments using GMMs, the projection models were found to offer slightly improved performance under matched conditions and significantly improved performance under mismatched conditions, while at the same time generally requirering fewer parameters.

7. REFERENCES

[1] K.P. Li and J.E. Porter, *Normalisations and selection of speech segments for speaker recognition scoring*, Proc. of ICASSP, 1988, pp. 595–598

[2] A.L. Higgins and L. Bahler and J.E. Porter, *Speaker verification using randomized phrase prompting*, Digital Signal Processing, Vol. 1, pp. 89–106 1991,

[3] D.A. Reynolds, *Speaker identification and verification using Gaussian mixture speaker models*, Speech Communication 17 (1–2):91–108, 1995

[4] J. Olsen, *A two-stage procedure for phone based speaker verification*, Pattern Recognition Letters (Elsevier) 18:889–897, 1997

[5] J. Olsen, *Phoneme based speaker recognition*, PhD Thesis, Aalborg University, 1997.

[6] J. Olsen, *Speaker Verification with Ensemble Classificers Based on Linear Speech Transforms*, Proc. of ICSLP, 1998

[7] H. Melin, *GANDALF - A Swedish telephone speaker verification database*, Proc. of ICSLP, 1996, pp. 1954–1957