SPEAKER VERIFICATION BASED ON PHONETIC DECISION MAKING

Jesper Ø. Olsen

Center for PersonKommunikation, Aalborg University,
Fredrik Bajers Vej 7A-2, DK-9220 Aalborg Øst, Denmark

email: jo@cpk.auc.dk

ABSTRACT

Speaker verification based on phone modelling is examined in this paper. Phone modelling is attractive, because different phonemes have different levels of usefulness for speaker recognition, and because phone modelling essentially makes a speaker verification algorithm text independent. The speaker verification system used here is based on a two stage approach, where speech recognition (segmentation) is separated from the actual speaker modelling. Hidden Markov Models are employed in the initial stage, whereas Radial Basis Function networks are used in the second for modelling speaker identity. The system is evaluated on a large realistic telephone database.

1 INTRODUCTION

For speaker verification it is not necessary to explicitly model linguistic units such as phonemes or words, because a speakers identity does not depend on the linguistic content of a test utterance – at least it is undesirable to model speaker characteristics at this level. However, phone modelling is nevertheless advantageous because different phonemes carry different amounts and kinds of speaker information [1]. Each phoneme may be regarded as providing information about one aspect of the speaker (configuration of the articulators). Hence, to obtain a "complete" picture of the speaker, it is necessary to look at a broad range of different phonemes.

When basing a speaker verification system on phone modelling, it is characteristic that target speakers may have several good (dangerous) impostors for each phoneme under consideration, but that these impostors vary from phoneme to phoneme; ie. being a good impostor for one phoneme does not guarantee being a good impostor for a different phoneme. Target speakers, on the other hand, are able to validate their identity claims using, basically, any phoneme [2].

Speaker verification is a binary decision problem, and can therefore in the end be reduced to computing a score and verifying identity claims by determining whether or not the score is greater or less than a given threshold, t:

$$\text{Decide} \begin{cases} \text{accept} & \text{if score} > t \\ \text{reject} & \text{otherwise} \end{cases} \quad (1)$$

When computing this score, each phone segment in the speech signal makes a contribution (even when phones are not explicitly modelled). In a conventional text independent speaker verification algorithm, the contribution of the different phonemes to the overall score (eg. utterance likelihood) is unknown; the overall score depends on the particular frequency with which the phonemes are represented in the test utterance, and on the duration of each phone segment. This is clearly not optimal, since no regard is taken to the extent that local scores contributed by individual phone segments express speaker identity and the extent to which different phonemes express the same information about the speaker; eg. a nasal and a vowel presumably represent information which is largely complimentary whereas two back vowels, say, represent highly correlated information about the speaker.

2 METHOD

The algorithm described here has two parts: first phone segments are identified and the speaker identity modelled for each phoneme independently. The result of this is a number of local scores – one for each different phoneme in the test utterance – which subsequently must be combined in order to produce a global verification decision (equation 1).

2.1 Phone Modelling

The basic method for making phoneme dependent local verification decisions was introduced in [2] (see figure 1 below). Briefly, a two-stage approach is used where phone segments are identified in the first stage by means of forced Viterbi decoding of the test utterance using speaker independent Hidden Markov Models (HMMs). This is relatively easy to do, because the application considered here is text prompted speaker verification; hence the spoken text is already known, and only pronunciation ambiguities need to be resolved. Variable frame rate

coding is used for representing each phone segment by a fixed number of frames, which are concatenated to form a "phone" vector, $\vec{\phi}$, which then is subjected to a linear transformation of the speaker and phoneme dependent Fisher transform [2]:

$$\vec{\phi}' = \mathbf{A}^T \vec{\phi} \tag{2}$$

The Fisher transform is a discriminative transform, which requires a number of training impostors to be available when it is estimated: one impostor for each basis vector in the transform (the columns of **A**). Fisher's linear discriminant function is used for computing the basis vectors of the transform:

$$\vec{a}_i = \mathbf{U}_i^{-1}(\vec{\mu}_1 - \vec{\mu}_{2,i})^T \tag{3}$$

where $\vec{\mu}_1$ is the mean phone vector for the target speaker, $\vec{\mu}_{2,i}$ the mean phone vector for impostor speaker number i, and $\mathbf{U}_i$ the pooled phone vector covariance matrix for the target speaker and impostor speaker number i. The individual basis vectors, $\vec{a}_i$, in the transform are orthogonalised using the Gram-Schmidt process [3].

After being transformed, a phone vector, $\vec{\phi}'$, is passed as input to a phoneme dependent Radial Basis Function (RBF) network, which is used for computing the speaker probabilities: $P(I|\vec{\phi}')$ and $P(\neg I|\vec{\phi}')$, where I is the target speaker class, and $\neg I$ the impostor speaker class.

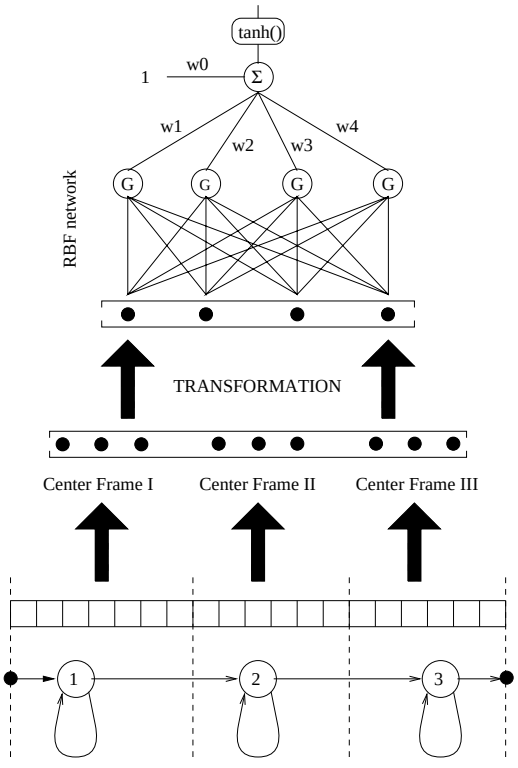

Figure 1: Model architecture: speaker independent HMMs are used for identifying phone segments; RBF networks for verifying the speaker identity.

The RBF networks compute the function $g_\Phi(\vec{\phi}')$:

$$g_\Phi(\vec{\phi}') = \tanh\left\{ S \sum_i w_i \exp\Big(C_i \frac{(\vec{\phi}' - \vec{\mu}_i)^2}{\vec{\sigma}_i^2} \Big) \right\} \tag{4}$$

where $\vec{\mu}_i$ and $\vec{\sigma}_i^2$ make up a codebook of centroids and corresponding variance vectors, S is the scaling factor of the activation function ($\tanh()$), C_i a set of basis function scales and finally w_i a set of basis function weights. These parameters are determined by a gradient descent based error minimisation algorithm. It can be shown [4] that the RBF networks approximate the Bayes optimal discriminant function

$$g_\Phi(\vec{\phi}') \approx P(I|\vec{\phi}') - P(\neg I|\vec{\phi}') \tag{5}$$

From this equation it is easy to compute the estimated probabilities for the target speaker, $P(I|\vec{\phi}')$ and the impostor speaker $P(\neg I|\vec{\phi}')$ given the phone vector, $\vec{\phi}'$. Two or more observations of a given phoneme, $\Phi^{(r)} = \vec{\phi}'_1, \ldots, \vec{\phi}'_r$, can be combined to obtain more reliable speaker probability estimate for that particular phoneme:

$$P(I|\Phi^{(r)}) = \frac{1}{1 + \prod_{i=1}^{r} \frac{g_\Phi(\neg I|\vec{\phi}'_i)}{g_\Phi(I|\vec{\phi}'_i)}} \tag{6}$$

In this way a speaker probability can be computed for each of the phonemes for which there are observations in the test utterance.

2.2 Decision Making

Each RBF network can be regarded as a speaker verification expert for a particular kind of phoneme. In order to produce a global verification decision, the local expert opinions must be combined into a global score, which can be used in connection with equation 1. A simple way of doing this is to use a principle of voting:

$$\text{score} = \sum_{\forall c=1}^{\#\Phi} \Big(P(I|\Phi_c^{(r)}) - P(\neg I|\Phi_c^{(r)}) \Big) \tag{7}$$

where the summation is over all phonemes represented in the test utterance (usually less than the total number of phonemes in the alphabet) and r is the number of observations of phoneme Φ_c.

2.2.1 Phoneme Correlations

The scoring procedure can be improved by weighting the votes of the individual experts differently. Different experts represent different information about the speaker, but this information is not necessarily orthogonal: different phonemes may, at least partly, represent the same speaker information.

The correlation between two variables, x and y, can be measured by the correlation coefficient, r:

$$r = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \tag{8}$$

where x_i and y_i are observations of respectively x and y, and where $\bar{x} = \frac{1}{N}\sum_i^N x_i$ and $\bar{y} = \frac{1}{N}\sum_i^N y_i$. Equation 8 can be used for defining a correlation coefficient for vector variables:

$$R_{\vec{x},\vec{y}} = \frac{1}{D}\sum_{d=1}^{D}|r_d(\vec{x},\vec{y})| \tag{9}$$

where $r_d(\vec{x},\vec{y})$ is the correlation between the d'th component of vector variables $\vec{x}$ and $\vec{y}$ (assumes that $\vec{x}$ and $\vec{y}$ have the same dimensionality).

The idea is now to use equation 9 for computing the correlation between different phonemes, and use the correlation information for giving higher weightings to expert opinions, which represent information that is relatively uncorrelated with the other experts. The correlations are computed between the phone vectors *after* the Fisher transform has been applied, ie. after information which is primarily not speaker discriminative has been discarded. A weighting scheme which satisfies the above requirements is:

$$m_{\Phi_c} = 1 + \sum_{i=1, i\neq c}^{\#\Phi}\ \sum_{j=i+1, j\neq c}^{\#\Phi} R_{\Phi_i,\Phi_j} \tag{10}$$

where m_{Φ_c} is the weight for the expert vote representing the c'th phoneme in the alphabet, that also is represented in the test utterance; $\#\Phi$ denotes the index of the last phoneme in the alphabet, which also is represented in the test utterance. Hence, the improved scoring rule is:

$$\text{score} = \sum_{\forall c=1}^{\#\Phi} m_{\Phi_c}\Big(P(I|\Phi_c^{(r)}) - P(\neg I|\Phi_c^{(r)})\Big) \tag{11}$$

3 SPEECH DATA

In this work, the Swedish Gandalf database [5] was used. The database contains speech recorded over the public telephone network; 58 speakers — 35 male + 23 female — made at least 23 telephone calls (sessions) over a one year period. In addition to this, the database contains an impostor part where 77 impostors — 49 male and 28 female — were recorded. The speech items from Gandalf that were used in these experiments consist of varied sentences, with on average seven words per sentence. The sentences prompted for in the test sessions, were not represented in the training sessions.

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. The total log energy in the frame normalised by the utterance energy was appended to each feature vector; feature vectors were extracted using a 25.6 ms Hamming window and a 10 ms frame period.

The HMM phone models used for segmenting the speech all had three emitting states; the variable frame rate coding procedure represented each phoneme by just three frames: one frame for each of the emitting states. Hence, after the first stage, the resulting feature (phone) vectors were 75 dimensional. The phone vectors were normalised to have norm one (in order to eliminate the signal gain), and subjected to a Fisher transformation [2], after which the dimensionality was reduced to 30. The Fisher transforms were estimated using the first (according to speaker ID) 30 training impostors of the same gender as the target speaker.

4 PHONE MODELS

The HMM phone models were context, speaker and gender independent. They were trained from the speech in target speaker sessions 1–2 + all the calls from 25 speakers who nominally belonged to Gandalf's target speaker set; these speakers were not, however, used as target speakers here, because they did not complete at least 23 recording sessions. Each HMM phone model had up to ten mixtures per state. For the purpose of creating a segmentation, the filter bank representation of the speech signal was, here, transformed into 12 MFCCs + normalised log energy + 13 delta + 13 acc. coefficients. Hence, different feature representations were used for speech and speaker recognition.

The RBF phone models were trained, for each target speaker, from up to 72 utterances recorded over a four month period (sessions 1–15); utterances where the spoken text differed from the text that was prompted for were removed. The test data was recorded in a number of calls (sessions 17–28) over the 6–11 months following the last training session. A number of different handsets were represented in the training calls (for each target speaker); all the target speaker test calls were from the so called favorite handset, which was used in approximately 50% of the training calls. Up to 30 RBF phone models were trained for each target speaker; a model was not trained if the number of training tokens from the target speaker was less than 10. The number of basis functions in each RBF model was adjusted to fit the number of training tokens: each model had $2(\lfloor N/20 \rfloor + 1)$ basis functions, where N is the number of training tokens from the target speaker. When training speaker models, the other target speakers were used as "training impostors" as was the above 25 speakers who were excluded from the test sets.

5 RESULTS & DISCUSSION

Using scoring rule 7 the target speaker acceptance (TA) and impostor speaker rejection (IR) error rates were respectively 3.0% (68/2233) and 3.0% (266/8786). Only impostors of the same gender as the target speaker were used, ie. male impostors were not used for female target speakers and vice versa. The value of the threshold (equation 1), was here fixed a posteori to the value $t = 0.1\#\Phi$, where again $\#\Phi$ is the number of different phonemes represented in a test utterance; the same threshold was used for all the speakers, and the results are therefore not "equal error rates".

Using scoring rule 11 the error rates were respectively 2.9% (65/2233) and 2.7% (241/8786); a small, but clear improvement over the case when phone correlations were not taken into account. In this case the threshold was fixed a posteori to $t = 0.45\#\Phi$.

An interesting question is whether all phonemes are useful for speaker modelling, or whether it is better to ignore some phonemes when doing speaker verification. Table 1 shows the error rates when decisions are based only on phonemes belonging to specific phoneme classes (scoring rule 7 was used).

Phoneme Class	**TA (%)**	**IR (%)**
Plosive (plc): /p/, /b/, /t/, /d/, /k/, /g/	11.4	9.7
Fricative (frc): /f/, /s/, /S/, /h/, /v/	12.5	11.5
Nasal (nas): /n/, /m/	12.8	11.4
Liquid (lqd): /l/, /r/, /j/	13.5	15.8
Unr. Front Vowel (ufv): /I/, /e/, /e:/, /a/, /a:/, /{:/	5.3	9.0
Central Vowel (cv): /@/	27.2	15.1
Rnd. Front Vowel (rfv): /Y/, /2/, /9/	85.0	0.3
Back Vowel (bv): /O/, /o:/, /u0/, /U/	26.0	11.4
plc+frc	7.4	6.9
plc+frc+nas	4.5	4.7
plc+frc+nas+lqd	3.5	4.2
plc+frc+nas+lqd+ufv	2.1	3.6
plc+frc+nas+lqd+ufv+cv	2.2	3.4
plc+frc+nas+lqd+ufv+cv+rfv	2.7	3.2
plc+frc+nas+lqd+ufv+cv+bv	2.6	3.2
plc+frc+nas+lqd+ufv+cv+rfv+bv	3.0	3.0

Table 1: Speaker verification error rates when decisions are based on phonemes from selected phoneme classes. Phoneme labels are in SAMPA notation [6].

As an evaluation of the usefulness of different phonemes, table 1 is not fair in the sense that the different phoneme classes have very different frequencies of occurrence in the test utterances, and simultaneously, the number of training tokens for each RBF phone model varied greatly, which allowed "complex" models to be trained for some phonemes (eg. /n/), but only simple models for others (eg. /9/). The table, however, shows that all phonemes are useful for speaker verification – also the "ill-reputed" fricative and plosive consonants: all phonemes can be used for reducing the global error rates.

In general phone models with low equal error rates can be trained even when only few (10–20) training tokens are available from a given target speaker, but in that case it is difficult to construct a model which will approximate closely the equal error rate on the test data (without actually "tuning" the model on the test data). The rounded front vowels (/Y/, /2/ and /9/) in table 1 is a good example of this.

6 CONCLUSIONS

Adjusting individual phone models to achieve the desired balance between the TA and IR error rates can be difficult if the number of training tokens is small, but fortunately, voting rules 7 and 11 are robust against this; it is unlikely that all models have the same bias, and individual models have only a relatively small influence on the overall classification decision; biases tend to be averaged away.

Taking the correlation between different phonemes into account is a useful and computationally cheap way of improving the error rates of a phone based speaker verification system. The phone models used in these experiments were context independent, and in general correlations between different phonemes were not very strong (typically 1–5%). To some extent this may be due to the correlation measure (equation 9) used here, which only considered "component to component" correlations within the phone vectors.

ACKNOWLEDGMENTS

I thank Håkan Melin and Telia Research AB for generously making the Gandalf database available to me for the experiments in this paper.

References

[1] J. Eatock and J.S. Mason, *A Quantitative Assessment of the Relative Speaker Discriminating Properties of Phonemes*, Proc. of ICASSP, 1994, 133–136, Vol. I

[2] J. Olsen, *A Two-Stage Procedure for Phone Based Speaker Verification*, Submitted to *Pattern Recognition Letters* (Elsevier), 1997

[3] T.M. Apostol, *Calculus*, Volume II Multi-Variable Calculus and Linear Algebra, with Applications to Differential Equations and Probability, Second Edition, Wiley International, 1969

[4] D.W. Ruck et al., *The Multilayer Perceptron as an Approximation to a Bayes Optimal Discriminant Function*, IEEE Trans. on Neural Networks, Vol. 1(4):296–298, December 1990

[5] H. Melin, *GANDALF - A Swedish Telephone Speaker Verification Database*, Proc. of ICSLP, 1996, pages 1954–1957, Volume III

[6] J. Wells et al., *Standard Computer-Compatible Transcription*, SAM STAGE REPORT Sen.3, ESPRIT PROJECT 2589 (SAM), MULTI-LINGUAL SPEECH INPUT/OUTPUT ASSESSMENT, METHODOLOGY AND STANDARDISATION, February 1992