SPEAKER VERIFICATION AS A TARGET-NONTARGET TRIAL TASK

Valiantsina Hubeika
Doctoral Programme (1), FIT BUT
E-mail: ihubeika@fit.vutbr.cz

Supervised by: Jan Černocký
E-mail: cernocky@fit.vutbr.cz

ABSTRACT

This paper deals with a speaker detection task. In this work, a different interpretation of the problem is introduced than the one used so far. In the standard approach, each speaker is modeled by their own model and the task is to decide whether the test speech segment was generated by the given model or not. In this work, only two models are used: one represents the target trials and the other represents nontarget trials, where the trial is represented by two speech segments, both from the same speaker, and two from different speakers, respectively. As the input features, fixed-length low-dimensional vectors derived from speaker factors generated by Joint Factor Analysis are used. Gaussian Mixture Models framework is used to model the feature distribution. The achieved results are compared to the state of the art systems.

1 INTRODUCTION

Given two speech segments, we have to decide:

- $H1$: The two speech segments come from the same speaker.
- $H2$: The two speech segments come from two different speakers.

In the standard approach of speaker verification [1], each speaker in the training set is represented by their own model. In the testing phase, given the test feature vector and the speaker model, the task is to decide whether the features were generated by the model depending on the output score. The score is usually represented by means of the (log)likelihood ratio between the tested model and the background model (or a set of impostor models).

In this work, only two models are trained. Given two speech segments, map them into one feature vector which is denoted as a trial feature vector. It results in having two sets of trial features: target trial features and nontarget trial features. Distribution of each set is then represented by one model. Given an input test trail feature vector, the task is to decide whether it was generated by the target trial model or the nontarget trial model.

When foloowing the standard approach, the detection system contains a variable number of models for a different task. In case a new speaker is inrolled in the system, a new model has to be trained. Whereas in the approach used in this work, the number of models is always 2.

The outline of the paper is as follows: Section 2 describes the approach; Section 3 presents the results achieved and Section 4 concludes the paper.

2 APPROACH

This work follows a supervector-based approach. Having two speech segments from one trial:

1. For each input speech segment, divide it into frames and extract a low-dimensional feature vector for every frame. This results to a separate variable-length sequence of feature vectors for each of the two speech segments.
2. Map each of the feature-vector sequences to a fixed-size supervector. Thus, each supervector represents a speech segment.
3. According to the trial list, defining the pairs of the segments that have to be tested against each other, derive new fixed-size vectors called target and nontarget trial supervectors.
4. Train two models: a target trial model and a nontarget trial model.
5. Given an input trial supervector, decide between $H1$ or $H2$ (test which of the two models generated the input trial supervector).

2.1 LOW-DIMENSIONAL FIXED-LENGTH SPEAKER VECTORS

The features are derived from the speaker factors generated using Joint Factor Analysis (JFA) [5] of a speaker and channel variability.

Assuming that if **s** is the speaker supervector for a randomly chosen speaker then

$$\mathbf{s} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{U}\mathbf{x} + \mathbf{D}\mathbf{z} \tag{1}$$

where **m** is the speaker- and channel-independent supervector, **V** is a rectangular matrix of low rank, **D** is a diagonal matrix and **y** (common or speaker factors) and **z** (specific factors) are random vectors having standard normal distribution. In practice, the speaker factors account for most of the variance in the data. In this work, the speaker factor vectors, **y**, are 300 dimensional vectors. Speaker factors represent speaker model but it is used as a feature vector to the classifier.

2.2 LOW-DIMENSIONAL FIXED-LENGTH TRIAL VECTORS

Let $\mathbf{y}_a$ and $\mathbf{y}_b$ be speaker factors of dimension $n = 300$ representing speech segments A and B, respectively. The trial vector could be represented as a concatenation of $\mathbf{y}_a$ and $\mathbf{y}_b$ but it would lead to a correlated distribution as it is presented in figure 3.1 (left). Therefor the trial vector is defined according to the following definition which provides a simple decorelation of the features:

$$\mathbf{Y} = [y_{a1} + y_{b1}, y_{a2} + y_{b2}, ..., y_{an} + y_{bn}, y_{a1} - y_{b1}, y_{a2} - y_{b2}, ..., y_{an} - y_{bn}] \tag{2}$$

The dimensionality of the trial vector is thus 600. The plot of the distribution of the 1^{st} component depending on the 301^{th} component (the most informative components corresponding to eigenvoice with largest variability) of the trial vector is presented in figure 3.1 (right).

2.3 DATA

As the training data, the Mixer 2004 and 2005 databases [7] (containing 2760 target trials and 24849 nontarget trials) are used. Additionally, SwitchBoard data (100000 target and 100000 nontarget randomly selected trials) are used to estimate transformation matrices (see below). Mixer 2006 (containing 2021 target trials and 26922 nontarget trials) is used as the test data. Both, the training and test data are composed of recordings with telephone conversational speech of approximately 2.5 min. Only female recordings are used.

2.3.1 TRIAL LISTS

In the experiments, the trail lists were used as defined by NIST [6]. The amount of target and nontarget trials is biased, as, obviously, it is much easier to define a nontarget trial. Cross-language trials and Nnglish only trials are used.

2.3.2 HETEROSCEDASTIC LINEAR DISCRIMINANT ANALYSIS (HLDA)

HLDA provides a linear transformation that can decorrelate the features and reduce the dimensionality while preserving the discriminative power of the features. The theory of HLDA is described in [2]. The HLDA transformation matrix is calculated using both Mixer and Switchboard data.

3 GAUSSIAN MODEL (GM)

The distribution of the features was modeled by the Gaussian distribution. The uni-modal Gaussian probability density, $p(x|\lambda)$, each represented by a $D \times 1$ vector, μ, and $D \times D$ covariance matrix, Σ:

$$p(\mathbf{x}|\lambda) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp[-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)] \tag{3}$$

Diagonal covariance matrix is used instead of the full covariance matrix.

3.1 MAXIMUM LIKELIHOOD (ML) PARAMETER ESTIMATION

Maximum Likelihood Estimation of the GMM parameters can be efficiently applied when a large amount of data is available to estimate model parameters. Given the training data $\mathbf{x}$, ML estimate, θ_{ML}, is defined as:

$$\theta_{ML} = \arg\max_{\theta} p(\mathbf{x}|\theta) \tag{4}$$

The maximum likelihood estimates of μ and Σ are just simple averages, that is:

$$\mu = \frac{1}{T}\sum_{t=1}^{T} x_t \ , \qquad \Sigma = \frac{1}{T}\sum_{t=1}^{T} (x_t-\mu)(x_t-\mu)' \tag{5}$$

3.2 MAXIMUM MUTUAL INFORMATION (MMI) PARAMETER ESTIMATION

Unlike ML training which aims to maximize the overall likelihood of training data given the class labels, the MMI objective function is to maximize the posterior probability of correctly

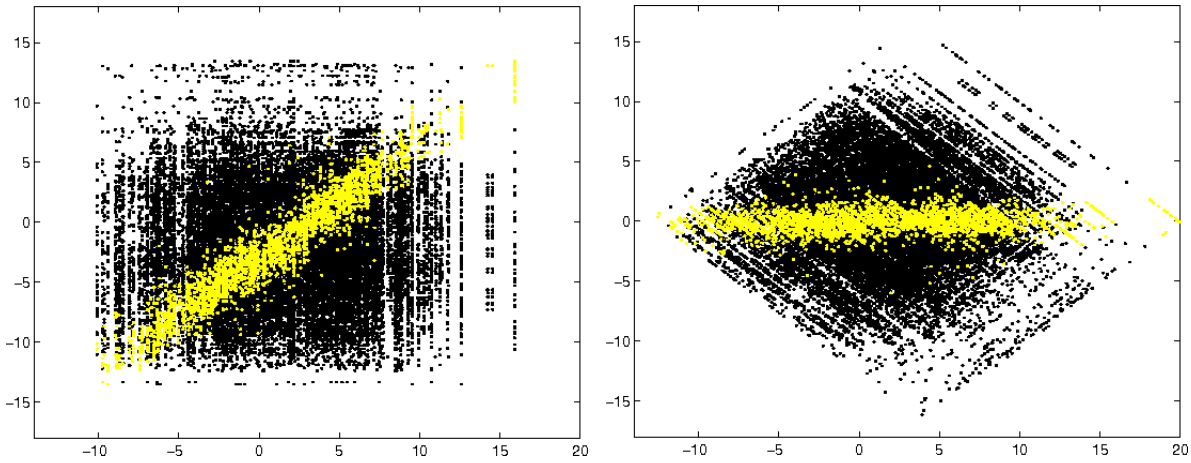

Figure 1: The left picture presents distribution of the first (most informative) speaker factors of two speech segments tested against each other. The black color represents the distribution when two segments come from different speakers and the yellow color denotes that the segments come from the same speaker. The right-hand picture presents the distribution of the 1^{st} and 301^{th} (most informative) feature components in the trail supervector. The black color denotes the distribution of the nontarget trials and the yellow color denotes the distribution of the target trials.

recognizing all training segments:

$$\theta_{MMI} = \arg\max_{\theta} \sum_{r=1}^{R} \log \frac{p_\theta(\mathbf{x}_r|s_r)^{K_r} P(S_r)}{\sum_{\forall s} p_\theta(\mathbf{x}_r|s)^{K_r} P(s)} \tag{6}$$

where $p_\theta(\mathbf{x}_r|s_r)^{K_r} P(S_r)$ is likelihood of $r-th$ training supervector, x_r, given the correct labeling of the supervector, s_r, and model parameters. R is the number of training supervectors and the denominator represents the overall probability density, $p(x_r)$. Definition of the mean and variance re-estimation formulae can be found in [4]. In this work, only Gaussian means are MMI re-estimated.

4 RESULTS

The results are presented in terms of equal error rate (EER) [1], a standard metric in speaker verification. As it can be seen from the table 4, using a simple ML approach gives already good results. When the ML GMM system is trained on the HLDA decorrelated features (without dimensionality reduction yet) the error decreases by 17 % relative to the baseline (see table 4, GMM-HLDA600to600 system). When additionally to decorrelation, dimensionality of features is reduced, the decrease in error is almost 30 % relative against the baseline. Several experiments were run, trying to find the optimal dimensionality of the feature vector. The best results were achieved when the reduction was done from 600 to 300 (see table 4, GMM-HLDA600to300 system). MMI parameter re-estimation, on the other hand, did not meet the expectations. All the ML GMM systems (without and with HLDA) were discriminatively re-trained but in all the cases the error increased comparing to the counterpart ML systems (see table 4, {GMM Baseline|GMM-HLDA600to600|GMM-HLDA600to300}-MMI).

	DET1	DET3
GMM Baseline	4.21	3.09
GMM-HLDA600to600	4.55	2.56
GMM-HLDA600to300	3.65	2.20
GMM-MMI	4.40	3.57
GMM-HLDA600to600-MMI	6.14	4.50
GMM-HLDA600to300-MMI	3.91	2.29

Table 1: Results of the GMM system represented in terms of equal error rate (EER). DET1 denotes cross-language trials and DET3 denoted English only trials.

5 CONCLUSION

This work introduced a simple speaker detection system. The novelty of the approach is that it views the verification task as a two class, target trial and nontarget trial, problem in contrast to the standard approach where each speaker was modeled individually by their own model. Results show that modeling classes using uni-modal Gaussian on HLDA transformed features in ML fashion brings the best results. Unfortunately, a supposed believed benefit of MMI parameter re-estimation was not achieved.

REFERENCES

[1] D. A. Reynolds. Speaker verication using adapted Gaussian Mixture Models, Digital Signal Processing 10 (2000), pages 19-41, 2000.

[2] N. Kumar, Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition. Ph.D. dissertation, Johns Hopkins Univ., Baltimore, MD, 1997.

[3] L. Burget, P. Matějka, and J. Černocký, Discriminative training techniques for acoustic language identification. In Proceedings of ICASSP 2006, pages 209-212, 2006.

[4] P. Matějka, L. Burget, P. Schwarz, and J. Černocký. Brno university of technology system for nist 2005 language recognition evaluation. In IEEE Odyssey: The Speaker and Language Recognition Workshop, pages 57-64, 2006.

[5] P. Kenny, G. Boulianne, P. Ouellet, P. Durmouchel, Joint Factor Analysis versus Eigenchannels in Speaker Recognition, Transcriptions on Audio, Speech, and Language Processing, Volume: 15, Issue: 4, pages 1435-1447, 2007

[6] NIST Speaker Recognition Evaluation Web-site, [Online], available: http://www.nist.gov/speech/tests/sre

[7] Ch. Cieri, J. P. Campbell, H. Nakasone, D. Miller, K. Walker, The Mixer Corpus of Multilingual, Multichannel Speaker Recognition Data, Proc. 4 th International Conference on Language Resources and Evaluation, [online], available: http://papers.ldc.upenn.edu/LREC2004/LREC2004_Mixer_Paper.pdf