Discriminative Kernel Classifiers in Speaker Recognition

Marcel Katz, Martin Schafföner, Edin Andelic, Sven E. Krüger, Andreas Wendemuth
IESK-Cognitive Systems, University of Magdeburg, Germany, marcel.katz@e-technik.uni-magdeburg.de

Abstract

The goal of automatic speaker recognition is to identify a speaker or to verify if a speaker is the person he claims to be. We present an overview of state-of-the-art speaker recognition systems which are usually based on speaker-dependent Gaussian Mixture Models (GMMs). In this paper we also describe different methods of integrating discriminative classifiers like the Support Vector Machine (SVM) into speaker recognition environments and show that it is possible to use the SVM methods directly on the frame-level for datasets with a small amount of speech data. On larger datasets a combination of generative and discriminative classifiers can be used. In speaker verification experiments the presented methods outperform the GMM baseline system on two datasets.

Introduction

In state of the art speaker verification systems Universal Background Models (UBM) and Maximum A-Posteriori (MAP) adapted target models are used.
During the last years several discriminative kernel classifiers like the Support Vector Machine (SVM) have shown a good performance on different classification tasks. Especially if the amount of data is limited, discriminative classifiers show a great ability of generalization and a better classification performance than GMMs, e.g., [4]. Since this is not feasible on larger speech datasets, it is possible to concatenate the GMM means of the MAP adapted speaker models and to classify these supervectors by SVMs.
One of the main problems in speaker recognition is the compensation of variabilities in different telephone transmission channels, e.g. cellular or landline, and telephone handsets (regular, handset). In [5] the SVM Nuisance Attribute Projection (NAP) is proposed to face this problems by projecting the features into a subspace which is more resistant to these channel effects.

GMM Based Speaker Verification

State of the art speaker verification systems are based on Gaussian Mixture Models. A GMM is the weighted sum of M Gaussian probability densities given by:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} c_i \, \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{1}$$

where c_i is the weight of the i'th component and $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is a multivariate Gaussian with mean vector $\boldsymbol{\mu}_i$ and the covariance matrix $\boldsymbol{\Sigma}_i$. In the UBM approach the mixture model is trained on a large amount of background data by standard methods like the Expectation Maximization (EM) algorithm. For each client of the system a speaker-dependent GMM is derived from the background model by adapting the parameters of the UBM using a Maximum A-Posteriori (MAP) approach [3]. The decision of detecting a client is based on the ratio between the summed log likelihoods of the specific speaker models and the background model. Defining the probability $P(\mathbf{X}|\lambda_k)$ as the probability of client C_k and $P(\mathbf{X}|\boldsymbol{\Omega})$ as the probability of the background model, each producing the sentence $\mathbf{X}$, the client is detected if the ratio is above a speaker-independent threshold δ:

$$\log \frac{P(\mathbf{X}|\lambda_k)}{P(\mathbf{X}|\boldsymbol{\Omega})} > \delta. \tag{2}$$

This results in two possible detection error probabilities: $P_{Miss|Target}$, the speaker is the claimed client but the resulting likelihood-ratio of equation (2) is lower than the threshold. $P_{FalseAlarm|NonTarget}$, the speaker is not the claimed one but the likelihood-ratio is higher than δ and the speaker is detected. For the performance measure of a speaker verification system the Decision Cost Function (DCF) is given. The DCF is defined as a weighted sum of these probabilities:

$$\begin{aligned} C_{det} &= C_{Miss} \times P_{Miss|Target} \times P_{Target} \\ &+ C_{FalseAlarm} \times P_{FalseAlarm|NonTarget} \times (P_{NonTarget}) \end{aligned} \tag{3}$$

with the predefined weights C_{Miss}, $C_{FalseAlarm}$ and prior probabilities P_{Target}, $P_{NonTarget} = 1 - P_{Target}$.

Support Vector Machines

Support Vector Machines (SVM) [1] are linear classifiers that can be generalized to non-linear classification problems by the so-called kernel trick. Instead of applying the linear methods directly to the input space $\mathbb{R}^d$, they are applied to a higher dimensional feature space $\mathcal{F}$ which is nonlinearly related to the input space via the mapping $\Phi : \mathbb{R}^d \to \mathcal{F}$. A kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ satisfying Mercer's conditions is used to compute the dot-product in $\mathbb{R}^d$. A possible kernel function is the Gaussian radial basis function (RBF) kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right). \tag{4}$$

The output of the SVM is a distance measure between a pattern and the decision boundary:

$$f(\mathbf{x}) = \sum_{i=1}^{N} \boldsymbol{\alpha}_i k(\mathbf{x}_i, \mathbf{x}) + b \tag{5}$$

For the posterior class probability we have to model the distributions $P(f|y = +1)$ and $P(f|y = -1)$ of $f(x)$ by the algorithm of Platt [2].

Experiments

For experiments we used the POLYCOST-BE1 dataset for the frame-based SVM system and the core task of the NIST 2006 speaker recognition evaluation for the supervector SVM system. The speech data were band-limited to the frequency range 300Hz-3400Hz. Energy-based speech detection was performed to discard frames containing low energy. Using a 20ms hamming window and a window shift of 10ms 13 mel-cepstral (MFCC) feature vectors were extracted. Additionally, the first and second order time differences of the MFCCs and the frame energy were computed and appended to the MFCCs resulting a 41 dimensional feature vector. Finally the feature vectors were normalized to fit a 0-mean and 1-variance distribution.

The GMM system

Gender dependent UBMs were trained on background data and consist of 32 mixture components for the POLYCOST corpus and 512 mixture components for the NIST SRE06 corpus. The speaker specific models were derived from the UBMs using a one step Maximum A-Posteriori adaptation [3]. Only the means of the mixtures were adapted with a relevance factor $\tau = 14$. During the detection test the top N-best encoded mixture components with respect to the UBM model were used for scoring. In our experiments we set $N = 10$.

The frame-based SVM systems

The experiments for the frame-based SVM systems were performed on the POLYCOST corpus. The SVM classifiers were trained in a one-vs-one approach using the RBF kernel of equation (4) and the non-probabilistic output of the SVM were transformed to a class probability. The equal error rate (EER) and the minimum DCF are given in table 1. As can be seen the SVM system outperforms the GMM baseline. The EER is reduced from 4.09% to 2.16% and the DCF from 0.034 to 0.019.

Table 1: Comparison of EER and DCF for the GMM and SVM system on the POLYCOST speaker verification task.

Classifier	EER (%)	DCF
GMM	4.09	0.034
SVM	2.16	0.019

The supervector SVM systems

The SVM system is based on the supervector approach presented by Campbell [5]. For all background and target speakers speaker specific GMMs were adapted from the UBM and all the resulting means of each GMM were concatenated to a single supervector. Using 512 mixture components and 41 acoustical features, this results in 20992 dimensional supervectors. To deal with channel and session variability in the GMM-space the supervectors were projected into a channel independent space using the Nuisance Attribute Projection (NAP). Finally the SVMs were trained with the transformed supervectors of the background and target speakers using a linear kernel. Figure 1 shows the detection error tradeoff (DET) curves of the three GMM and SVM systems. The best results were achieved by the supervector SVM/NAP system with an EER of 4.79% compared to 6.75% and 8.83% of the SVM and the GMM system respectively.

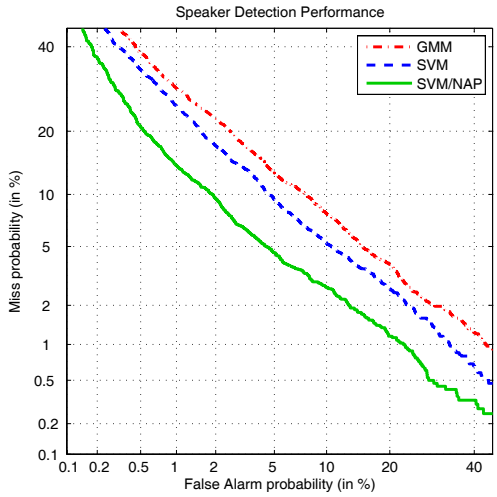

Figure 1: Performance comparison of GMM, SVM and NAP/SVM systems on the NIST 2006 Evaluation corpus

Conclusions

In this paper we presented an overview of state of the art speaker recognition systems. While it is possible to use kernel methods on the feature vectors directly for a small amount of data the supervector extension of the GMM system yielded excellent results on the NIST 2006 speaker recognition evaluation.

References

[1] Christopher Burges: "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery, pp. 121-167, 1998.

[2] John C. Platt: "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," Advances in Large-Margin Classifiers, pp. 61-74, 2000.

[3] Douglas A. Reynolds, T.F. Quatieri and R. Dunn: "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, pp. 19-41, 2000.

[4] Marcel Katz, S. E. Krüger, M. Schafföner, E. Andelic and A. Wendemuth: "Speaker Identification and Verification using Support Vector Machines and Sparse Kernel Logistic Regression," IWICPAS, Springer (Lecture Notes in Computer Science), 2006.

[5] William Campbell, D. Sturim, D. Reynolds and A. Solomonoff: "SVM Based Speaker Verification using a GMM SuperVector Kernel and NAP Variability Compensation," ICASSP, pp. 97-100, 2006.