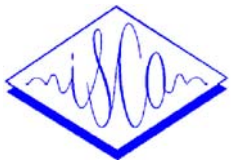

ODYSSEY04 - The Speaker and Language Recognition Workshop
Toledo, Spain
May 31 - June 3, 2004

Fusing Discriminative and Generative Methods for Speaker Recognition: Experiments on Switchboard and NFI/TNO Field Data*

W. M. Campbell, D. A. Reynolds, J. P. Campbell

MIT Lincoln Laboratory
Lexington, MA USA
{wcampbell,dar,jpc}@ll.mit.edu

Abstract

Discriminatively trained support vector machines have recently been introduced as a novel approach to speaker recognition. Support vector machines (SVMs) have a distinctly different modeling strategy in the speaker recognition problem. The standard Gaussian mixture model (GMM) approach focuses on modeling the probability density of the speaker and the background (a generative approach). In contrast, the SVM models the boundary between the classes. Another interesting aspect of the SVM is that it does not directly produce probabilistic scores. This poses a challenge for combining results with a GMM. We therefore propose strategies for fusing the two approaches. We show that the SVM and GMM are complementary technologies. Recent evaluations by NIST (telephone data) and NFI/TNO (forensic data) give a unique opportunity to test the robustness and viability of fusing GMM and SVM methods. We show that fusion produces a system which can have relative error rates 23% lower than individual systems.

1. Introduction

The Gaussian mixture model (GMM) is a dominating technology in the area of speaker recognition [1]. Its success stems from many factors including a probabilistic framework, training methods scalable to large data sets, and high-accuracy recognition. The GMM approach is generative. I.e., a GMM is used to model the probability density of the observed feature vectors produced by target speakers and by a universal background speaker set (a universal background model).

Outside of the speech literature, many alternatives to the GMM have emerged for classification problems. These methods include rule-based systems, multi-layer perceptrons, support vector machines, Bayesian nets, radial basis functions, etc. A drawback of many of these approaches is that they lack positive attributes in the three areas of strength of the GMM mentioned above. Rather than trying to challenge the GMM with yet another classifier, we take a slightly different tact and consider discriminatively trained classifiers as complementary to the GMM approach. That is, we want to use classifiers that are (computationally) reasonable to train and test for speech applications and explore methods to fuse them with a GMM. We expect that the resulting system will combine decision boundaries in a synergistic manner yielding reduced error rates.

* This work was sponsored by the United States Government Technical Support Working Group under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

An exciting area of recent focus has been the application of support vector machines (SVMs) in many different fields. In several applications, SVMs have excited much fervor because of their superior performance. SVMs have been applied to speaker recognition in several instances [2, 3, 4]. Because of the dramatically different training philosophy, SVMs seem a likely candidate for testing fusion ideas.

We focus on the approach used in [4] for SVM-based speaker recognition. We explore several methods for fusing SVMs with GMMs. Attacking the fusion problem involves some unique challenges because of the non-probabilistic nature of the SVM output. Thus our methods involve novel solutions to the problem.

The outline of the paper is as follows. In Section 2, we discuss the standard NIST evaluation corpus used for part of our experiments. In Section 3, we discuss a unique training/testing corpus from the NFI/TNO used as a forensic evaluation. In Section 4, we discuss our GMM system. In Section 5, we describe our support vector system. Section 6 discusses methods for applying T-norm to a discriminative system. Section 7 discusses fusion of discriminative and GMM systems. Section 8 details experiments which showcase our methods and their performance.

2. The NIST 2003 Speaker Recognition Evaluation

The NIST 2003 speaker recognition evaluation (SRE) included multiple tasks for both one and two speaker detection. For the purposes of this paper, we focus on the one speaker detection task from limited data.

The data in the one-speaker limited-data detection task was taken from the second release of the cellular Switchboard corpus of the Linguistic Data Consortium. Training data was nominally 2 minutes of speech from a target speaker excerpted from a single conversation. The training corpus contained 356 target speakers. Each test segment contained a single speaker. The primary task was detection of the speaker from a segment of length 15 to 45 seconds. The test set had 2,215 true trials and 25,945 false trials (impostor attempts). For evaluation, NIST used the decision cost function

$$\begin{aligned} C_{\text{det}} = & C_{\text{miss}} P(\text{miss}|\text{target}) P(\text{target}) + \\ & C_{\text{FA}} P(\text{FA}|\text{nontarget}) P(\text{nontarget}) \end{aligned} \quad (1)$$

as well as reporting standard measures such as equal error rate (EER). In (1), $C_{\text{miss}} = 10$, $C_{\text{FA}} = 1$ and $P(\text{target}) = 0.01$. More details on the evaluation may be found in [5].

3. The NFI/TNO Forensic Speaker Recognition Evaluation

The Netherlands Forensic Institute (NFI) and the Netherlands organization for applied research (TNO) jointly organized an evaluation of current speaker recognition systems on real police investigations in the fall of 2003. The goal of the evaluation was to determine the state of the art in text independent speaker recognition and the possibility of using these systems for police investigations.

An interesting aspect of the NFI/TNO evaluation was the data variability. Many factors were uncontrolled–telephone line quality, SNR, speaking time, language, speaking style, etc. In addition, no development data was provided, so that systems were applied "out of the box." This evaluation setup provided an opportunity to test the SVM system, the GMM system, and the fusion setup with no prior tuning.

Speech for training target models was either a 30s, 60s, or 120s segment and was potentially from multiple conversations. For testing, the speech segments were approximately of length 7s, 15s, and 30s collected from a single conversation. Tests from all conversation lengths were pooled resulting in 521 true trials and 9,676 false trials. Note that all speakers in the corpus were male. A decision cost function was used as in (1) with $C_{\mathrm{FA}} = 10 * C_{\mathrm{miss}}$ and $P(\mathrm{target}) = 0.5$. Our results correspond to the primary task (experiment 1) in the NFI/TNO data. For more information, we refer to the evaluation plan [6].

4. GMM

The basic system used is a likelihood ratio detector with target and alternative probability distributions modeled by Gaussian mixture models (GMMs). A universal background model (UBM) GMM is used as the alternative hypothesis model and target models are derived using Bayesian adaptation (also known as Maximum A-Posteriori (MAP) training) [1]. The techniques of feature mapping and T-norm were also used.

4.1. Feature extraction

A 19-dimensional mel-cepstral vector is extracted from the speech signal every 10ms using a 20ms window. The mel-cepstral vector is computed using a simulated triangular filterbank on the DFT spectrum. Bandlimiting is then performed by only retaining the filterbank outputs from the frequency range 300Hz-3138Hz. Cepstral vectors are processed with RASTA filtering to mitigate linear channel bias effects. Delta-cepstral coefficients are then computed over a ± 2 frame span and appended to the cepstra vector producing a 38 dimensional feature vector. Lastly, the feature vector stream is processed through an adaptive, energy-based speech detector to discard low-energy vectors.

4.2. Feature mapping

All features are passed through a feature mapping technique to help remove channel effects [7]. Briefly the feature mapper works as follows. A channel independent root model is trained using all available channel specific data. Next, channel specific models are derived by using MAP adaptation of root parameters with channel specific data. For an input utterance, the most likely channel specific model is first identified then each feature vector in the utterance is shifted and scaled using the top-1 scoring mixture parameters in the root and channel-specific models to map the feature vector to the channel-independent feature space. Ten channel models derived from Switchboard landline and cellular corpora were used.

4.3. T-norm

T-norm is a technique where scores from a collection of fixed non-target models are used to normalize a target model score for a test file [8]. The target model score normalization is accomplished by subtracting the mean and dividing by the standard deviation of the non-target model scores per test file. For the NFI/TNO evaluation data, we used a fixed set of 100 male models from the NIST 2001 SRE trained with 2 minutes of data from a single utterance and performed gender-dependent T-norming. For the NIST 2003 Evaluation data, we used gender dependent T-norming with speakers taken from the Switchboard 2 part 1 corpus (100 per gender).

5. SVM

The LPCC SVM system uses a novel sequence kernel [4]. The sequence kernel compares entire utterances using a generalized linear discriminant. We describe the particular aspects of the classifier–the front end, training, and scoring below.

5.1. Front end processing

Feature extraction was performed using a 30ms window with a rate of 100 frames/second. A Hamming window was applied and then 12 LP coefficients were extracted. 18 cepstral coefficients (LPCC's) were derived from 12 LP coefficients. Both cepstral mean subtraction and variance normalization (on a per coefficient basis) were performed. Deltas were extracted from the 18 LPCC's. The final feature vector was dimension 36 (18 LPCC's and deltas).

5.2. Training

The SVM used a Generalized Linear Discriminant Sequence kernel (GLDS) with an expansion into feature space using a monomial basis. All monomials up to degree 3 were used, resulting in a feature space expansion of dimension 9139. We used a diagonal approximation to the kernel inner product matrix. A "background" for the SVM consisted of a set of speakers taken from a database not used in the train/test set. The SRE 01 evaluation was used as a background. SVM training was performed as a two-class problem where all of the speakers in the background had SVM target -1 and the current speaker under training had SVM target +1. For each conversation in the background and for the current speaker under training, an average feature expansion was created. SVM training was then performed using the GLDS kernel implemented using SVMTorch.

5.3. Scoring

For each utterance the standard front end was used. An average feature expansion was then calculated. Scores for each target speaker were an inner product between the speaker model and the average expansion. A T-Norm score was also computed using 100 males and 100 females from the NIST SRE '01 task; we describe this in detail next.

6. T-norm for Discriminative Classifiers

We explored two methods for T-norm. The first, shown in Figure 1 is based upon a background corpus and a *separate* corpus for the T-norm speakers. For each T-norm speaker, we train a

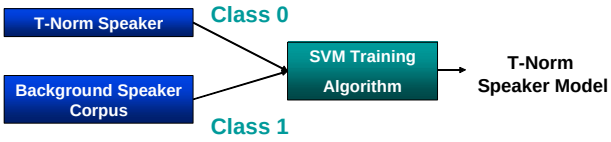

Figure 1: First method for T-norm

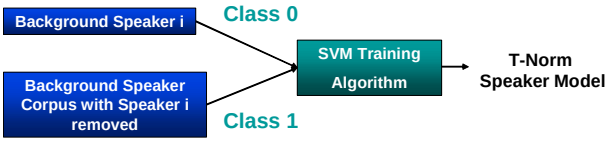

Figure 2: Leave-one-out method for T-norm

model using the T-norm speaker as class 0 and the background speaker set as class 1. This process is repeated for all T-norm speakers. An advantage of this approach is that we can construct an SVM speaker recognition system and then just cycle the T-norm speakers through the system as if they were target speakers. A disadvantage is that it requires another corpus to for T-norm speakers.

A second approach to constructing T-norm speakers is shown in Figure 2. The basic idea is to leave a speaker out of the background and then train a T-norm model using the extracted data and the modified background; we call this approach a leave-one-out T-norm method. Note that we are required to take the T-norm speaker out of the background because of the discriminative nature of the SVM; the SVM training would be impaired by having the exact same data labeled in and out of class. An advantage of this approach to constructing T-norm speakers is that it does not require an additional corpus. A disadvantage is that the training process is more complex.

7. Fusing Discriminative and Generative Classifiers

Given the palette of approaches discussed in Sections 4 and 5, we next set out to examine fusion of the different systems to see if they are indeed providing complementary information to improve speaker recognition accuracy.

7.1. Linear combination approach

For this fusion approach, we apply T-norm to the SVM and GMM output scores. The T-norm non-target speakers are *not* constrained to be the same for both systems. The scores are then added using a linear combination; i.e., if s_{GMM} and s_{SVM} are the T-normed scores, we make a speaker recognition decision based upon the score, s_{fuse} given by

$$s_{\mathrm{fuse}} = \alpha s_{\mathrm{SVM}} + (1-\alpha) s_{\mathrm{GMM}}. \qquad (2)$$

The rationale for using this approach is straightforward–T-norm provides a way of normalizing the variance of the score. After this variance normalization, the scores are on comparable scales and linear fusion is natural way of combination.

7.2. Perceptron approach

Another approach for fusion is to use a neural network. Following our previous work [9], we chose a single-layer perceptron to fuse the scores from the various systems. This perceptron has a layer consisting of inputs for each system (and a bias term), no hidden layer, and an output layer using sigmoids on the target and nontarget nodes. The perceptron is similar to a linear discriminant with a sigmoidal output squashing function. The priors can be adjusted in LNKnet [10] to minimize detection cost [9].

For both the NIST SRE and the NFI/TNO evaluations, we trained a fusion system using male cellular data from NIST SRE 2002. This trained system was then applied to the GMM and SVM systems' T-normed scores from each evaluation.

8. Experiments

Our first set of experiments was to determine which T-norm approach was the most appropriate for the SVM system. Experiments were performed using the development set from the NIST 2003 evaluation; a reduced feature set of only 12 LPCCs plus deltas was used for quick evaluation. Figure 3 shows the results of the base system and the two T-norm approaches outlined in Section 6. The figure shows that the T-norm method 2 (leave-one-out) has the best accuracy; it improves performance at low probability of false alarm. The figure also indicates that the T-norm method 1 degrades accuracy somewhat.

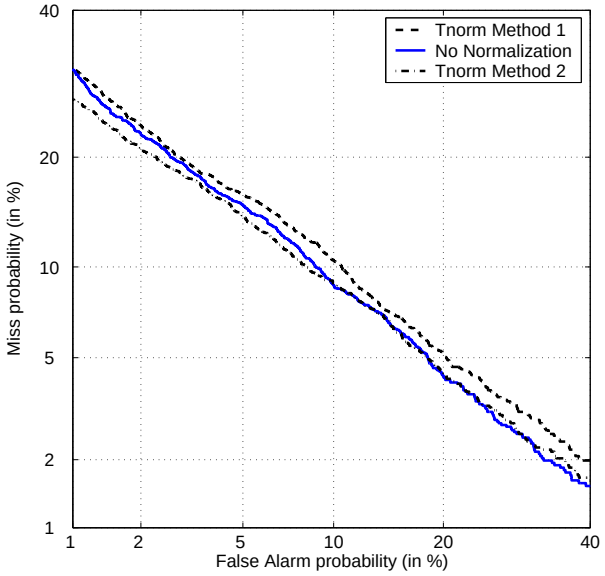

Figure 3: Comparison of T-norm approaches on the NIST 2003 Evaluation Corpus

Next we performed experiments on the 2003 NIST SRE evaluation data described in Section 2; we note for clarity that these experiments were performed using an improved system over the actual NIST 2003 evaluation. We implemented T-norm based fusion using $\alpha = 0.5$ in (2), see Table 1 and Figure 4. We used leave-one-out T-norm and a background set from the NIST SRE 2001 evaluation corpus for the SVM. The GMM system used the setup described in Section 4. Both the figure and the table show that the SVM and GMM fuse in a complementary way reducing error rates substantially.

We then performed experiments on the NFI/TNO evaluation data, see Figure 5 and Table 2. T-norm based fusion was implemented using equal weighting of the SVM and GMM. Again we see that the EER drops and the individual systems

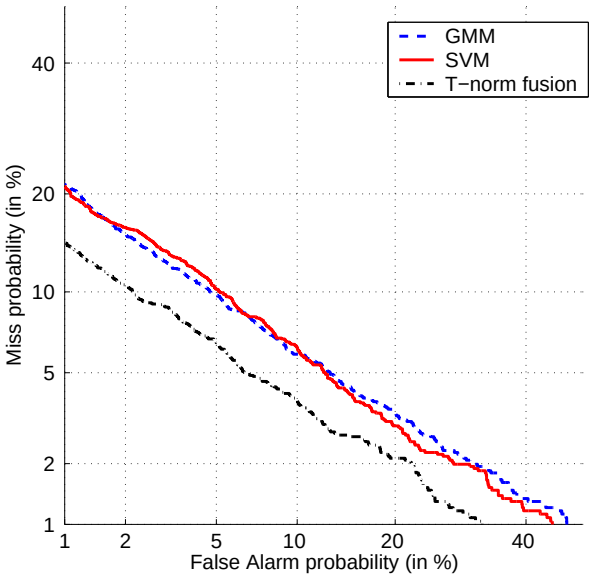

Figure 4: NIST 2003 1sp limited data results (perceptron fusion is indistinguishable from T-norm fusion and is not shown)

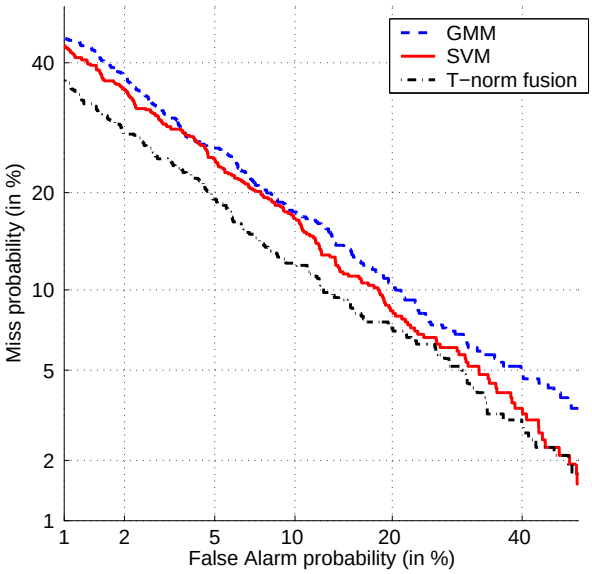

Figure 5: NFI/TNO results (perceptron fusion is indistinguishable from T-norm fusion and is not shown)

Table 1: Comparison of EER and minDCF for different systems on the 2003 NIST SRE 1sp limited data evaluation

System	EER	minDCF
GMM	7.47%	0.0306
SVM	7.72%	0.0303
T-norm Fusion	5.73%	0.0237
Perceptron Fusion	5.73%	0.0237

are complementary. This result is particularly significant since the system for the NFI/TNO evaluation was performed "blind"; i.e., no development data was available for tuning fusion.

Table 2: Comparison of EER and minDCF for different systems on the NFI/TNO evaluation

System	EER	minDCF
GMM	14.01%	0.267
SVM	13.05%	0.261
T-norm Fusion	11.32%	0.228
Perceptron Fusion	11.32%	0.229

9. Conclusion

We have explored the fusion of two speaker recognition systems based on GMMs and SVMs. Two fusion methods were tried based upon linear combination and perceptron fusion of T-normed scores. We found the linear combination and perceptron fusions of T-normed scores each produced nearly identical results. In order to use T-norm for the SVM, a method was proposed to implement T-norm speaker training. Fusion of the GMM and SVM systems yielded excellent results on typical and forensic-style telephone data showing that the two systems could be combined in a robust and complementary manner.

10. References

[1] Douglas A. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[2] Shai Fine, Jiří Navrátil, and Ramesh A. Gopinath, "A hybrid GMM/SVM approach to speaker recognition," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2001.

[3] Vincent Wan and William M. Campbell, "Support vector machines for verification and identification," in *Neural Networks for Signal Processing X, Proceedings of the 2000 IEEE Signal Processing Workshop*, 2000, pp. 775–784.

[4] W. M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, 2002, pp. 161–164.

[5] M. Przybocki and A. Martin, "The NIST year 2003 speaker recognition evaluation plan," http://www.nist.gov/speech/tests/spk/2003/index.htm, 2003.

[6] David A. van Leeuwen and Jos S. Bouten, "The NFI/TNO forensic speaker recognition evaluation plan," available at http://speech.tm.tno.nl/aso/, 2003.

[7] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, 2003, vol. 2, pp. II–53–56.

[8] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[9] J. P. Campbell, D. A. Reynolds, and R. B. Dunn, "Fusing high- and low-level features for speaker recognition," in *Proceedings of Eurospeech*, 2003, pp. 2665–2668.

[10] R. Lippmann and et. al., "LNKnet," Available at http://www.ll.mit.edu/IST/lnknet/.