

ENHANCING GMM SCORES USING SVM “HINTS”

Shai Fine, Jiří Navrátil, Ramesh A. Gopinath

IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

{fshai, jiri, rameshg}@us.ibm.com

Abstract

This paper proposes a classification scheme that combines statistical models and support vector machines. It exploits the fact (observed in [1]) that GMM and SVM classifiers with roughly the same level of performance produce uncorrelated errors. We describe a novel scheme which employs an SVM classifier as an “advisor” to the GMM classifier in uncertain cases. The utility of the combined generative/discriminative approach is demonstrated on standard text-independent speaker verification and speaker identification tasks in matched and mismatched training and test conditions. Results indicate significant improvements in performance without much computational overhead.

1. INTRODUCTION

Gaussian mixture models (GMM) give state-of-the-art performance in text-independent speaker verification [2] and identification [3]. Well-designed GMM systems are robust to channel variations and achieve independency of text, topic and language.

Support Vector Machine (SVM) classifiers have recently generated [4] interest from speech community. SVMs are discriminative and can be used to train non-linear decision boundaries. As such SVM classifiers provide an attractive way to enhance classifiers based on generative models like GMM, HMM, etc.

A previous study [1] showed significant decorrelation between the errors of GMM and SVM classifiers trained on the same data - even when their performances were comparable. This paper develops on that work based on the following assumption: the GMM classifier performs well on most instances and fails once in a while, i.e. the GMM score is indecisive. In such instances an SVM classifier is called in to resolve the uncertainty. For this to work it suffices that the SVM classifier performs reasonably well in cases where the GMM is indecisive. By itself the SVM classifier does not need to perform as well as the GMM classifier on all instances.

Experimental results are reported on text-independent speaker verification (binary classifier) and identification (multi-class classifier) tasks.

2. The Baseline System

Here we describe just the baseline speaker verification system; for the identification system see [1]. Each class (speaker) is modeled using a diagonal GMM model enhanced with a speaker-dependent Maximum Likelihood Linear Transform (MLLT) [5]. MLLT-enhanced GMM models give the state-of-the-art performance in verification [6] and identification [3]. The model for the i th speaker is $\theta_i = \{\mathbf{T}, c_1^K, \mu_1^K, \Sigma_1^K, P_1^K\}_i$ where \mathbf{T} , μ_1^K , Σ_1^K , and P_1^K are respectively the MLLT matrix, means, covariances and priors for the K components of speaker i 's GMM. First the basic GMM parameters are trained by MAP adaptation of a speaker-independent universal background model (UBM) and then the MLLT matrix is estimated. The speaker identification, is then carried out as a maximum likelihood classification. In verification, the likelihood ratio between the speaker model and the UBM is calculated. The ratio serves as the discriminant measure for the threshold-based verification decision. Herein, the UBM plays an important role in normalizing the speaker model likelihood across different acoustic conditions, and stabilizes the value processed by the threshold.

3. The SVM System

An ensemble of binary SVM classifiers are built using Fisher features from the baseline GMM models. Recall that an SVM classifier builds the maximal margin *optimal hyperplane* that separates the two classes [4]:

$$f^* = \arg \max_f \min_i y_i f(x_i) \quad (1)$$

Here $f(x) = (x \cdot w) + b$ and $x, w \in \mathbb{R}^N$ and $b \in \mathbb{R}$, $y_i \in \{-1, 1\}$ are the labels corresponding to the training set $\{x_i\}$, and $sign(f(x))$ is the classification rule. The maximal margin classifier achieves robustness with respect to both the instances and the hypothesis space: small perturbations of either will not change the resulting classifier much. A regularization term, the norm of w , is introduced to make the optimization problem (1) is well-posed. The regularization term realizes complexity/capacity control hence implements Vapnik's *Structural Risk Minimization* principle [4]. As noted by Vapnik the optimization problem depends only on dot-products. This leads to the separation of the input space (where the data resides) from

the feature space where the dot-product is defined. In particular non-linear kernels could be used to map the input space into feature space and the resulting classifier is not much more complex computationally. To handle non-separable data Vapnik introduces *soft margin* SVM classifiers. Soft margin techniques handle outliers and mislabeled samples, by incorporating positive slack variables ξ_i in the SVM optimization problem and assigning an extra cost for the errors. This yield the following SVM (primal) optimization problem:

$$\begin{aligned} \min_{w,b,\xi} \quad & \|w\|^2 + C \sum_i \xi_i \\ \text{s.t.} \quad & y_i ((w \cdot x_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned} \quad (2)$$

(2) has a globally unique solution and is described solely using the Lagrange multipliers and dot-product values in the feature space, i.e. by kernel operations. Since the optimization problems we faced at the verification task were low ranked (large amount of data vs. low dimension feature space), we were able to converged to the optimal SVM solution, using a specially designed training algorithm which is highly efficient in storage requirement and computation load [7].

Finding an appropriate kernel function for a particular application can be difficult and remains largely an unresolved issue. One of the recent innovations in the field of kernel engineering has been made by Jaakkula and Hausler [8] who formed a link between generative and discriminative models: Generally speaking, generative models (such as GMM, HMM or graphical models) will focus on providing an efficient description of the data while discriminative models will strive for a better description of a decision boundary between the various classes. Denote $p(x|\theta)$ a generative model, where θ are its parameters, the mapping function is an analogous quantity to the model's sufficient statistics, known as the *Fisher score*:

$$U_\theta(x) = \nabla_\theta \log(p(x|\theta)) \quad (3)$$

Each component of U_x is a derivative of the log-likelihood score for the input vector x with respect to a particular parameter. The magnitude of the components specify the extent to which each parameter contributes to generating the input vector. The natural kernel for this mapping is the inner product between these feature vectors, possibly scaled by a positive definite matrix. It was shown [8] that subject to some mild assumptions, a kernel classifier employing the Fisher kernel would be at least as powerful as the original generative model, and in most cases will actually improve the discriminative power of the generative model.

3.1. An SVM System for Speaker Verification

Recall that the baseline system carries out the verification process as a sequence of likelihood ratios between coupled Gaussian pairs. We associated an SVM classifier with each such pair (who had sufficient amount of

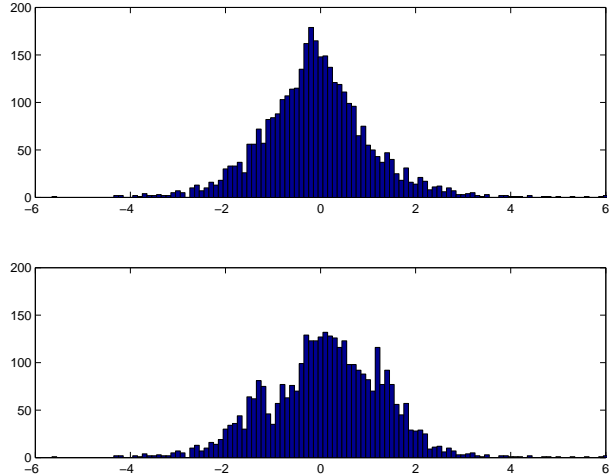


Figure 1: *The impact of SVM advisory on the likelihood ratio distribution: Before (upper) and after (lower) enhancement.*

data), using Fisher mapping (3) based on the GMMs of the UBM. Note that for a binary problem, this transformation implies asymmetric role for the positive and negative classes (since it depends on the generative model of only one of them). We chose to use a UBM based transformation (rather than the target speakers' GMM) to increase decorrelation with the baseline system. The training data for each binary optimization problem have been selected by hard clustering the original data sets (of the UBM and the target speaker) based on the individual Gaussians' scores. This resulted with very small target speaker clusters (negative sets) which forced us to consider training only about 30% of the possible binary classifiers. Even the optimization problems which we chose to solve were highly unbalanced, and we compensate by assigning different penalty terms (the C parameter at the objective function (2)) for the positive and negative training sets, based on their sizes.

At classification, we used the baseline target speaker's GMM to identify the associated coupled Gaussians per frame, and for those frame which we had an associated SVM binary classifier - we obtained a score. The final classification rule for a cell is a (un/weighted) voting across frames who provided SVM scores.

4. Enhancing The Baseline System

The baseline GMM classifier performs marginally better than the SVM classifier. This is probably because the SVM classifiers saw very little data and hence the classification is based on a small number of frames. Nevertheless, the errors produced by the systems are uncorrelated.

Instead of combining the scores from the GMM and SVM classifier we opted to use SVM advice on GMM confusions. More precisely, the GMM classifier scores

on a per-frame basis are examined and a list of frames where the GMM is indecisive is selected. The scores for just the selected frames are enhanced with scores from the SVM classifier. The SVM scores essentially nudge the GMM score towards the decision suggested by the SVM classifier. The magnitude of this change is proportional to the confidence of the SVM score.

Fig. 1 shows the impact of GMM score enhancement on the likelihood ratio distribution for one utterance: Notice how the unimodal distribution centered at the 0 (upper panel) is transformed into a bimodal distribution which its primary bump is slightly shifted to the right (lower panel). This marginal change was sufficient to resolve GMM confusion in our experiments. In other tasks one may need to adopt a more aggressive enhancement scheme to resolve confusions. The location and width of the “confusion window”, and the actual magnitude of the changes are application related parameters that should be tuned using held-out data.

5. EXPERIMENTS

5.1. Text-Independent Speaker Verification

5.1.1. Database

For this task we used the subset of the Switchboard telephone corpus used for NIST speaker recognition evaluations in 1996 and 1999. The 1996 data from Switchboard I was used to build the UBM. About 4.5 hours of speech was partitioned into 4 sets based on gender and handset type (electret and carbon button). Atomic GMMs with 256 components were trained and merged to give the 1024-component UBM model. The 1999 data (Switchboard II, Phase 3) was used for training and testing the targets. This corpus contains two 1-minute sessions of enrollment data for 539 speakers and provides for 37620 gender-matched verification trials. The test duration varies primarily between 15 to 45 seconds. The ratio of target vs. impostor trials among the 37k tests is roughly 1:10.

The front-end features were cepstral mean-removed 38-dimensional features (19-dim MFCC plus first derivatives) at a 10ms frame rate.

5.1.2. Evaluation Measure

The results were evaluated using Detection Error Trade-off (DET) curves [9] - a special case of ROC curves. On the DET curve two operating points are of interest - the Equal-Error Rate (EER) point and lowest detection cost point as defined by specific cost factors for false alarms, false misses normalized by expected relative frequencies of target and impostor tests. We used the detection cost function defined by NIST for the 1999 speaker recognition evaluation:

$$DCF = C_{fa} \Pr(fa|n) \Pr(n) + C_{miss} \Pr(miss|t) \Pr(t),$$

System	DCF 10^{-3}
Plain	52.2
MLLT	48.4
Enhanced GMM/SVM	47.0

Table 1: Verification detection costs

with $C_{fa} = 1$, $C_{miss} = 10$, $\Pr(n) = 0.99$, and $\Pr(t) = 0.01$, thus shifting the point of interest towards low false alarm rates. It turns out to be important which operating point is of primary interest as different techniques may bring different gains dependent on the region of the curve.

5.1.3. The Baseline MFCC System

The performance of the MFCC 1024-component GMM system in terms of the optimum detection cost function (DCF) for all 37k trials is shown in table 1.

The first row refers to MAP-adapted GMMs without MLLT. Consistent improvements are obtained by applying MLLT transforms to the common UBM and to the target models separately (“MLLT”).

5.1.4. The Enhanced GMM/SVM System

The SVM classifier used Fisher mapping to transform the 38-dim input vector to a 76-dim feature space. Only 30% (about 300) of possible binary classifiers were trained per enrolled speaker. The most expensive step in training was the clustering of the UBM (positive) training set using its GMM, while the actual binary classifiers training was relatively fast.

About 10% of the test set was used to determine the center and width of the “confusion window” and the magnitude of the perturbation. Results are reported on the full NIST99 eval data to maintain consistency with other experiments¹.

5.2. Text-Independent Speaker Identification

The system described in [1] was enhanced with MLLT to yield a baseline GMM classifier which achieves state-of-the-art performances on this database. GMM scores were further enhanced with SVM advice, using a modified scheme designed to handle the multi-class decisions.

5.2.1. Database

The Lincoln Lab Handset Database LLHDB [10] was used to train and test both system parts in text-independent mode. The database contains telephone-bandwidth speech from 52 speakers recorded over 4 types of carbon-button microphones - CB1 through CB4. Each speaker recorded

¹Comparing performances with the the baseline systems on 90% (without the held-out set) did not change the gain observed on the full training set.

System	Test Condition			
	CB1	C2	CB3	CB4
Baseline GMM CB1	5.4	7.5	48.8	23.9
Enhanced GMM/SVM CB1	3.8	6.7	46.9	24.1
Rel.red.% CB1	28.6	10.3	3.9	-0.8
Baseline GMM CB3	39.5	42.8	8.8	19.3
Enhanced GMM/SVM CB3	38.9	41.0	5.9	17.9
Rel.red.% CB3	1.5	4.1	32.6	7.0

Table 2: Identification error rates on the CB1- and CB3-trained systems for the four types of carbon button microphones (3-5 sec).

two long (30 sec) and ten short sentences (scripts from the TIMIT database) through each of the transducers. In all our experimental configurations, one long sentence of each speaker, namely the “rainbow” text, served for system training and the short utterances were used for testing, giving a total of about 2000 tests across the four microphone conditions.

5.2.2. The Baseline System

32-component GMM models were built for each speaker using MAP adaptation of an SI model. The SI model was built on a few hundred speakers taken from an internal telephone-quality speech database. Each speakers model is then enhanced with an MLLT transform.

Two separate systems were built: the first using training data from microphone CB1, the second on the CB3. The latter was chosen for comparison, based on the fact that the first system (CB1) performed worst on tests from this particular type CB3. Tests were carried out using all four types CB1 through CB4 on these two systems, thus having results for one matched and three mismatched condition rounds, for each of the two systems.

5.2.3. The Enhanced GMM/SVM System

Table 2 shows the identification rates for the two systems (CB1 and CB3) and tests across all conditions (CB1 through CB4) as described above. Obviously, the microphone mismatch in training and testing is the most influential factor in performance degradation, however also the particular type of microphone appears to play a role, which can be seen in the difference between matched tests for CB1 and CB3.

The utility of the advisory system is presented by the relative improvement in identification rate over the baseline system. In the matched (CB1/CB1 and CB3/CB3) and low mismatched (CB1/CB2 and CB3/CB4) tests the improvement ranges from 28-32% to 7-10% while in the strong mismatch the performances are in the same range as the baseline system.

6. Conclusion

Our experimental results show that through the use of the enhanced SVM/GMM scheme, significant improvements in both verification and identification tasks can be achieved. Measured on the 1999 NIST speaker recognition evaluation involving 37k mixed-microphone trials the enhanced system reduces the minimum detection cost from $48.4 \cdot 10^{-3}$ to $47 \cdot 10^{-3}$. On a separate database, the SVM reduces the identification error by up to 32%.

7. References

- [1] S. Fine, J. Navrátil, and R. A. Gopinath, “A hybrid gmm/svm approach to speaker identification,” in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.
- [2] D. Reynolds, R. Dunn, and J. McLaughlin, “The lincoln speaker recognition system: Nist eval2000,” in *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, 2000.
- [3] U. Chaudhari, J. Navrátil, S. Maes, and G. Ramaswamy, “Very large population text-independent speaker identification,” in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.
- [4] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.
- [5] R. Gopinath, “Maximum likelihood modeling with gaussian distributions for classification,” in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998.
- [6] J. Navrátil, U. V. Chaudhari, and G. N. Ramaswamy, “Speaker verification using target and background dependent linear transforms and multi-system fusion,” in *EuroSpeech*, 2001. Submitted.
- [7] S. Fine and K. Scheinberg, “Efficient svm training using low-rank kernel representation,” Tech. Rep. RC21911, IBM T. J. Watson Research Center, 2000.
- [8] T. S. Jaakkola and D. Haussler, “Exploiting generative models in discriminative classifiers,” in *Advances in Neural Information Processing Systems*, vol. 11, MIT Press, 1999.
- [9] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, “The det curve in assessment of detection task performance,” in *Proc. of the European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 1895–8, 1997.
- [10] D. Reynolds, “Htimit and llhdb: Speech corpora for the study of handset transducer effects,” in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1535–8, 1997.