SUPPORT VECTOR MACHINE BASED SPEAKER IDENTIFICATION SYSTEMS USING GMM PARAMETERS

Vijendra Raj Apsingekar and Phillip L. De Leon

New Mexico State University
Klipsch School of Electrical and Computer Engineering
Las Cruces, New Mexico USA 88003
{vijendra, pdeleon}@nmsu.edu

ABSTRACT

Speaker identification is the task of determining which speaker characteristics from the speakers known to the system best matches the unknown voice sample. SI requires multiple decision alternatives and to implement SI system using SVM techniques requires multi-class SVM classifier. In this paper, speaker model clustering is implemented on a SVM based SI system. Here, instead of clustering the speakers, we build a SVM classifier which separates a group of speakers. Thus each hyperplane built using SVMs separates a group of speakers and this procedure is repeated in each sub-group until there is only one speaker in each group. Experiments performed on NIST-2002 speech corpus show an improvement in accuracy compared to the conventional multi-class SVM techniques.

Index Terms— Speaker recognition, Support Vector Machines, Kernel functions

1. INTRODUCTION

The objective of speaker *identification* (SI) is to determine which voice sample from a set of known voice samples best matches the characteristics of an unknown input voice sample [1]. The objective of SV is to verify the identity claim [2]. SI is a two-stage procedure consisting of training and testing. In the training stage shown in Fig. 1(a), speaker-dependent feature vectors, $\mathbf{x_m}$ are extracted from a training speech signal and a speaker model, $\boldsymbol{\lambda}_s$ is built. Of the various speaker modeling techniques, the Gaussian Mixture Model Universal Background Model (GMM-UBM) based approach and MAP adaptation of the speaker models has shown to be very successful in accurately identifying speakers from a large population and is presently state-of-the-art technique [3]. GMM-UBMs provide a probabilistic model of the distribution of feature vectors. A standard approach in estimating the parameters of the GMM-UBM (weights, mean vectors, and covariance matrices $\{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$) is to use the Expectation Maximization (EM) algorithm [2] and number of components in an GMM-UBM is $1024 - 2048$. In the testing stage shown in Fig. 1(b), features are extracted from a test signal (speaker unknown); features are compared and scored against all the S speaker models; and the most likely speaker identity, $\hat{s}$ is decided as

$$\hat{s} = \arg \max_{1 \le s \le S} \sum_{m=1}^{M'} \log p(\mathbf{x}_m^{\text{test}} | \lambda_s). \quad (1)$$

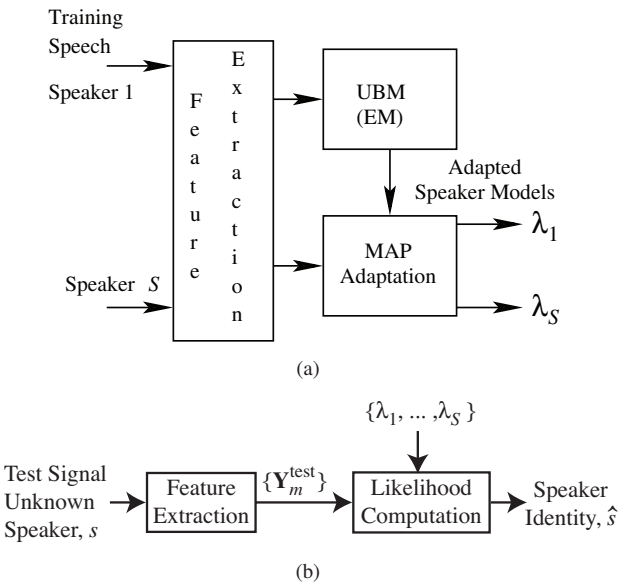

Fig. 1. (a) Training and (b) testing stages in SI

The SV is a binary decision strategy and decision alternatives in SI is equal to the number of speakers enrolled in the system. In effect, to determine the speaker identity we need a multi classifier that can classify the incoming test utterance to one among the possible S speakers.

Recently, support vector machines (SVMs) are used to obtain the scores between the incoming speech utterance and a model of the claimant in speaker verification (SV) applications [4]. SVMs are binary discriminant and linear classifiers

[4]. Given M training samples $\mathbf{X} = \{\mathbf{x}_1, \cdots \mathbf{x}_M\}$ with associated binary label $b_i = \{-1, 1\}$, the SVM will train a linear decision boundary of the form

$$\mathbf{\Psi}^T \mathbf{x} + d = 0 \tag{2}$$

where $\{\mathbf{\Psi}, d\}$ are the optimum decision boundary parameters [5].

SVM optimization function can be represented as

$$\begin{array}{lll} \min & \frac{1}{2}||\mathbf{\Psi}||_2^2 + C\sum_{i=1}^{M} \xi_i & \\ \text{w.r.t} & \mathbf{\Psi}, d, \boldsymbol{\xi} & \\ \text{s.t.} & b_i(\mathbf{\Psi}^T \mathbf{x}_i + d) \geq 1 - \xi_i & \forall i \\ & \xi_i \geq 0 & \forall i \end{array} \tag{3}$$

where ξ_i is the training error associated with $\mathbf{x}_i$, C is a constant that controls the tradeoff between maximizing the margin and reducing the empirical risk.

The optimization in (3) is a quadratic optimization problem and can be converted to a linear optimization form as

$$\begin{array}{ll} \max & \sum_{i=1}^{M} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{M} \alpha_i \alpha_j b_i b_j \mathbf{x}_i^T \mathbf{x}_j \\ \text{w.r.t} & \boldsymbol{\alpha} \\ \text{s.t.} & 0 \leq \alpha_i \leq C \end{array} \tag{4}$$

where α_i's are Lagrange multipliers associated with $\mathbf{x}_i$ Then the optimal weight vector can be written as [5]

$$\mathbf{\Psi} = \sum_{i=1}^{M} \alpha_i b_i \mathbf{x}_i. \tag{5}$$

The samples with nonzero and positive α_i are termed as support vectors and completely determine the position of the decision boundary. Using (5), given a test vector $\mathbf{y}^{\text{test}}$, decision function can be written as

$$f(x) = \sum_{i=1}^{M} \alpha_i b_i (\mathbf{x}_i)^T (\mathbf{y}^{\text{test}}) \tag{6}$$

There are few advantages of the form in (4). During training as well as testing, the main computation is the inner-product between the pairs of the data vectors. Moreover, computation does not depend on the dimensionality of the data vectors. Inner-product can be replaced with a *kernel* function $K(\mathbf{x}_i, \mathbf{x}_j)$.

However, SVMs cannot be directly applied to applications involving speech [4], as SVMs are linear classifiers and speech is nonlinear. Given a nonlinear mapping function $\phi : R^d \rightarrow F$, which maps the nonlinear input space to linear feature space, then SVMs can be implemented. Under the mapping function $\phi : R^d \rightarrow F$ the decision function of (6) can be written as

$$f(x) = \sum_{i=1}^{M} \alpha_i b_i \phi(\mathbf{x}_i)^T \phi(\mathbf{y}^{\text{test}}) + d. \tag{7}$$

If a kernel function $K(\mathbf{x}, \mathbf{y})$ which satisfies Mercer's condition is employed, such that $K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}_i)^T \phi(\mathbf{y}^{\text{test}})$ then the decision function in (7) can be written as [6]

$$f(x) = \sum_{i=1}^{M} \alpha_i b_i K(\mathbf{x}_i, \mathbf{y}^{\text{test}}) + d. \tag{8}$$

Researchers in recent years developed many kernel functions such as Fisher kernel [7], MLLR-kernel [8] and GMM-supervector kernel [9] and successfully used in SV applications.

2. SVMS FOR SPEAKER IDENTIFICATION

The use of SVMs in SI is very limited [6] as multiple hyperplanes need to be built for SI. There are two techniques for multi-class SVM problem: 1) One-separates-rest and 2) One-separates-one [10]. In One-separates-rest, one binary SVM classifier per speaker is built, separating data vectors of one speaker (with binary labels $b_i = 1$) from rest all other speakers' data vectors ($b_i = -1$). This technique is less efficient [6]. In One-separates-one, a total of $\frac{S(S-1)}{2}$ (S, the speaker population) binary SVM classifiers are formed, where each one is formed using data vectors from a pair of speakers.

Researchers in [6] developed cluster-based SVM for SI. Here k-means clustering is used to cluster the training data from each speaker and then the data vectors on the boundary of each cluster are determined and used as the support vectors [6]. Experiments were performed on 40 speakers of TCC-300 speech corpora and there was no loss in accuracy.

In [11], a hybrid GMM-SVM SI system was proposed. Here the testing is carried in two stages, first, a GMM based test stage is implements. Then top two speakers, ranked according to (1), are selected for second stage of testing using SVM techniques [11]. Experiments were performed using NTIMIT corpus, using GMM based SI system accuracies were 70.1% and using hybrid GMM-SVM system the accuracies were increased to 72.4% [11].

Hou and Wang in [12] proposed SI using probabilistic SVM with GMM adjustment. In [12], the probability of a class C given a test data vector is defined as $p(C|\mathbf{x}) = \frac{1}{1+\exp[f(\mathbf{x})g_\lambda]}$. Where $f(\mathbf{x})$ is defined as in (8) and g_λ is a function of log-likelihood score using speaker model λ_c. Experiments were performed using 20 speakers of NIST-2003 speech corpus and SI accuracies were reported as 98.47%, which is about 4% more than the conventional system.

From the above papers [6], [11] and [12], and the speaker population used in their the actual SVM experiments, the choice of present techniques is limiting the total number of speakers that can be used in the SVM based SI systems. Such systems are very inefficient for large population SI applications due to high computational complexity [6].

3. SVM SPEAKER CLUSTERING

In [1], [13], [14], [15], we have proposed GMM and GMM-UBM based speaker model clustering for efficient speaker recognition applications. In [1], we have vectorized each GMM of a speakers' training feature vectors as the weighted sum of all the mean vectors, called, weighted mean vector (WMV) of a GMM. Then used k-means algorithm to cluster the speaker models using these vectorized algorithms. Various distance measures to be used in k-means were proposed for efficient SI applications. During testing, only selected few cluster are searched to identify the speaker, thus resulting in faster ID times. A speed-up factor of $10\times$ with relatively no loss in accuracy compared to full search was achieved.

The idea of speaker model clustering is implemented on a SVM based SI system in this paper. Here, instead of clustering the speakers, we build a SVM classifier which separates a group of speakers. Thus each hyperplane built using SVMs separates a group of speakers and this procedure is repeated in each sub-group until there is only one speaker in each group (Algorithm is described in Algorithm 1).

Algorithm 1 SVM based Speaker Clustering for SI applications

1: Form a binary classifier between all pairs of speakers (enrolled in the system) taking two at a time
2: Select two speakers whose separating hyperplane has maximum margin and least training error
3: Then classify remaining speakers into two groups based on the classifier selected in step 2
4: Then repeat the procedure in each sub-group until there is only one speaker

The main advantage of this technique is that, the number of hyperplanes to be built, if each hyperplanes separates equal number of speakers is nearly $2\log_2 S$, where S is the number of speakers in the system. Fig. 2 compares the One-separates-one based multi-class SVM with the proposed method. In Fig. 2(a) each class numbered 1 to 4 and each hyperplane separating the classes numbered 1-2, 1-3 and so on. It can be observed that there are four classes and One-separates-one based SVM technique results in six hyper-plane separating all the possible classes. Using the proposed method in Fig. 2(b), it require only three hyperplanes separating four classes.

4. EXPERIMENTS AND RESULTS

Experiments have been performed on a set of randomly selected speakers from NIST-2002 speech corpus. Silence from the speech utterances is removed using an energy-based voice activity detector. Feature vectors composed of 13 MFCCs, extracted every 10 ms using a 25 ms hamming window, 13 delta-MFCCs are appended after CMS and RASTA applied on to MFCCs. Finally, Feature warping is applied on the feature

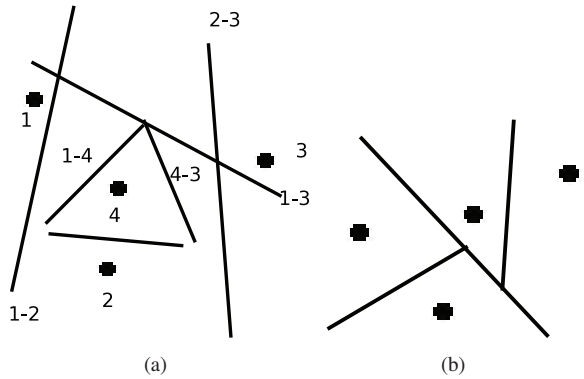

Fig. 2. Multiclass SVM classifier techniques for 4 classes (a) One-separates-one, requires $\frac{S(S-1)}{2} = 6$ hyperplanes to separate all the possible classes (b) Proposed Method of SVM based clusters, requires only 3 hyperplanes to separate 4 classes. Advantageous for higher number of classes

vectors. Our baseline system consists of a 1024 component GMM-UBM built using the training utterances of all the 330 speakers in one speaker detection cellular data from NIST-2002 corpus. Individual speaker models are MAP-adapted, only mean vectors, with a relevance factor of 16.

4.1. Data vectors for SVM classifier

To reduce the number of data vectors being used in the SVM classifier, the score $||\boldsymbol{\mu}_i^s - \boldsymbol{\mu}_i^{\mathrm{UBM}}||$ (where $\boldsymbol{\mu}_i^s$ is the ith mean vector of the speaker model s, and $\boldsymbol{\mu}_i^{\mathrm{UBM}}$ is the ith mean vector of the UBM) are ranked and the highest 32 components were used. Thus, only those Gaussian components in the MAP-adapted speaker models which are significantly different from that of the UBM are selected. Data vectors are formed by appending mean vectors with the weighted-covariance normalized mean vectors [13], $\mathbf{x}_i = [\boldsymbol{\mu}_i^T, w_i(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i)^T]^T$.

Principle component analysis (PCA) is applied on these data vectors to get uncorrelated feature vectors. A hybrid kernel, Gaussian radial basis kernel $K_G(\mathbf{x}_i, \mathbf{x}_j) = \exp(\frac{-||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2})$ with $\sigma = 60$, followed by GMM-supervector kernel were used in the SVM classifier. $\mathrm{SVM}^{\mathrm{light}}$ was used for training and testing the SVM techniques [10].

4.2. Results

Various sets of speaker population from NIST-2002 corpus were used, Table 1 shows the result of SI accuracy using various techniques. Proposed technique is compared against One-separating-one and One-separating-rest techniques. With a population of 15 speakers, One-separating-rest has an accuracy of 66.66%, GMM-UBM system has 100% accuracy.

Table 1. SVM based SI system, Proposed technique is compared against One-separating-one and One-separating-rest techniques. Accuracy of GMM-UBM based SI system is given in parenthesis.

Method	No. of Speakers	SI Accuracy (GMM-UBM based)
One-separating-Rest	15	66.66% (100%)
One-separating-One	40	96.00% (100%)
One-separating-One	64	92.00% (98.44%)
Proposed Method	64	97.00% (98.44%)

With speaker population of 40, One-separating-one has accuracy of 96%, GMM-UBM has 100% accuracy. When 64 speakers were used in the SI experiments, One-separating-one has achieved 92% accuracy and proposed technique achieved accuracy of 97% and GMM-UBM has accuracy of 98.44%. The proposed method increased the accuracy by 5%, outperforming the conventional SVM techniques. Using the proposed method, 62 speaker out-of 64 speakers are correctly identified and GMM-UBM based system identifies 63 speakers, there is only one extra speaker being misidentified compared to GMM-UBM system. Use of other hybrid kernels may increase the accuracy to that of the GMM-UBM based system.

5. CONCLUSION

In this paper, SVM based speaker cluster method based multiclass SVM have been proposed, in which a binary SVM is built to separate a group of speakers and this procedure is repeated in each sub-group until there is only one speaker in each group. This method has the fewer hyperplanes compared to conventional One-separating-one technique. SI experiments show that the proposed method can be used with speaker population of over 60 speakers and accuracy loss compared to GMM-UBM based system is relatively insignificant.

6. REFERENCES

[1] V. R. Apsingekar and P. DeLeon, "Speaker model clustering for efficient speaker identification in large population applications," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 4, pp. 848–853, May 2009.

[2] T. F. Quatieri, *Discrete-Time Speech Signal Processing Principles and Practice*. Prentice-Hall, Inc., 2002.

[3] H. Aronowitz and D. Burshtein, "Efficient speaker recognition using approximated cross entropy (ACE)," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 7, pp. 2033–2043, Sep. 2007.

[4] C. Longworth and M. J. F. Gales, "Combining derivative and parametric kernels for speaker verification," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 4, May. 2009.

[5] V. N. Vapnil, *Staristical Learning Theory*. Wiley Interscience, 1998.

[6] S. Sun, C. L. Tseng, Y. H. Chen, S. C. Chuang, and H. C. Fu, "Cluster-based support vector machines in text-independent speaker identification," in *Proc. Int. Joint Conf. Neural Networks*, 2003.

[7] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," *Proc. NIPS*, pp. 487–493, 1999.

[8] A. Stolcke, L. Ferrer, S. Kajarekar, and A. Venkataramam, "MLLR transformation as features in speaker recognition," *Interspeech*, pp. 2425–2428, 2005.

[9] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM-supervector kernel and NAP variability compensation," *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 2006.

[10] T. Joachims, *"Making large-Scale SVM Learning Practical," Advances in Kernel Methods–Support Vector Learning*, B. Scholkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT-Press, 1999.

[11] D. Mashao, "A hybrid gmm-svm speaker identification system," *IEEE AFRICON*, 2004.

[12] F. Hou and W. Bingxi, "Text-independent speaker recognition using probabilistic svm with gmm adjustment," *IEEE International Conference on Natural Language*, 2003.

[13] V. R. Apsingekar and P. DeLeon, "Proc. biometrics symposium," *Reducing Speaker Model Search Space in Speaker Identification*, 2007.

[14] ——, "Efficient speaker identification using speaker model clustering," *Proc. European Signal Processing Conf. (EUSIPCO)*, 2008.

[15] ——, "Efficient speaker identification using distributional speaker model clustering," *Asilomar Conf. on Signals, Systems and Computers*, 2008.