A Review on Text- Independent Speaker Identification Using Gaussian Supervector SVM

Kauleshwar Prasad[1], Piyush Lotia[1] and M. R. Khan[2]

[1]SSCET Bhilai, India
[2]GEC Raipur, India
kauleshwarprasad@gmail.com, lotia_piyush@rediffmail.com,
mrkhan@cgdte.in

Abstract

Speech recognition is a challenging yet important speech technology. First, an introduction proposes components of typical automatic speaker recognition system .Two modes like enrollment mode and recognition is discussed. Then we discuss about Feature Extraction [8] .It is used to reduce the dimensionality of the input vector while maintaining the discriminating power of signal. After this Gaussian mixture modeling is discussed, which is the speaker modeling technique used in most systems. Vector Quantization Process is also discussed and then paper highlights on Supervectors. A few speaker modeling alternatives, namely, neural networks and support vector machines, are mentioned. Most recent technique to solve the Speaker Verification System is to combine GMM with SVM .So GMM Supervector [3] is also discussed. Here GMM supervector based SVM is applied to this field with spectral features. A GMM is trained for each utterance, and the corresponding GMM supervector is used as the input feature for SVM. Then, some applications of speaker verification are proposed, including on-site applications, remote applications, applications relative to structuring audio information, and games.

Keywords: *Feature extraction, vector quantization, supervectors, Gaussian Mixture Model (GMM), SVM*

1. Introduction

Figure 1 shows the components of an automatic speaker recognition system .The upper is the enrollment process while the lower panel illustrates the recognition process. The feature extraction module first transforms the raw signal into feature vectors in which speaker specific properties are emphasized and statistical redundancies suppressed. In the enrollment mode, a speaker model is trained using the feature vectors of the target speaker. In the recognition mode, the feature vectors extracted from the unknown person’s utterance are compared against the model(s) in the system database to give a similarity score. The decision module uses this similarity score to make the final decision.

Virtually all state-of-the-art speaker recognition systems use a set of background speakers or cohort speakers in one form or another to enhance the robustness and computational efficiency of the recognizer. In the enrollment phase, background speakers are used as the negative examples in the training of a discriminative model [14], or in training a universal background model from which the target speaker models are adapted [15]. In the recognition phase, background speakers are used in the normalization of the speaker match score [16, 17, 18, 19, 20, 21].

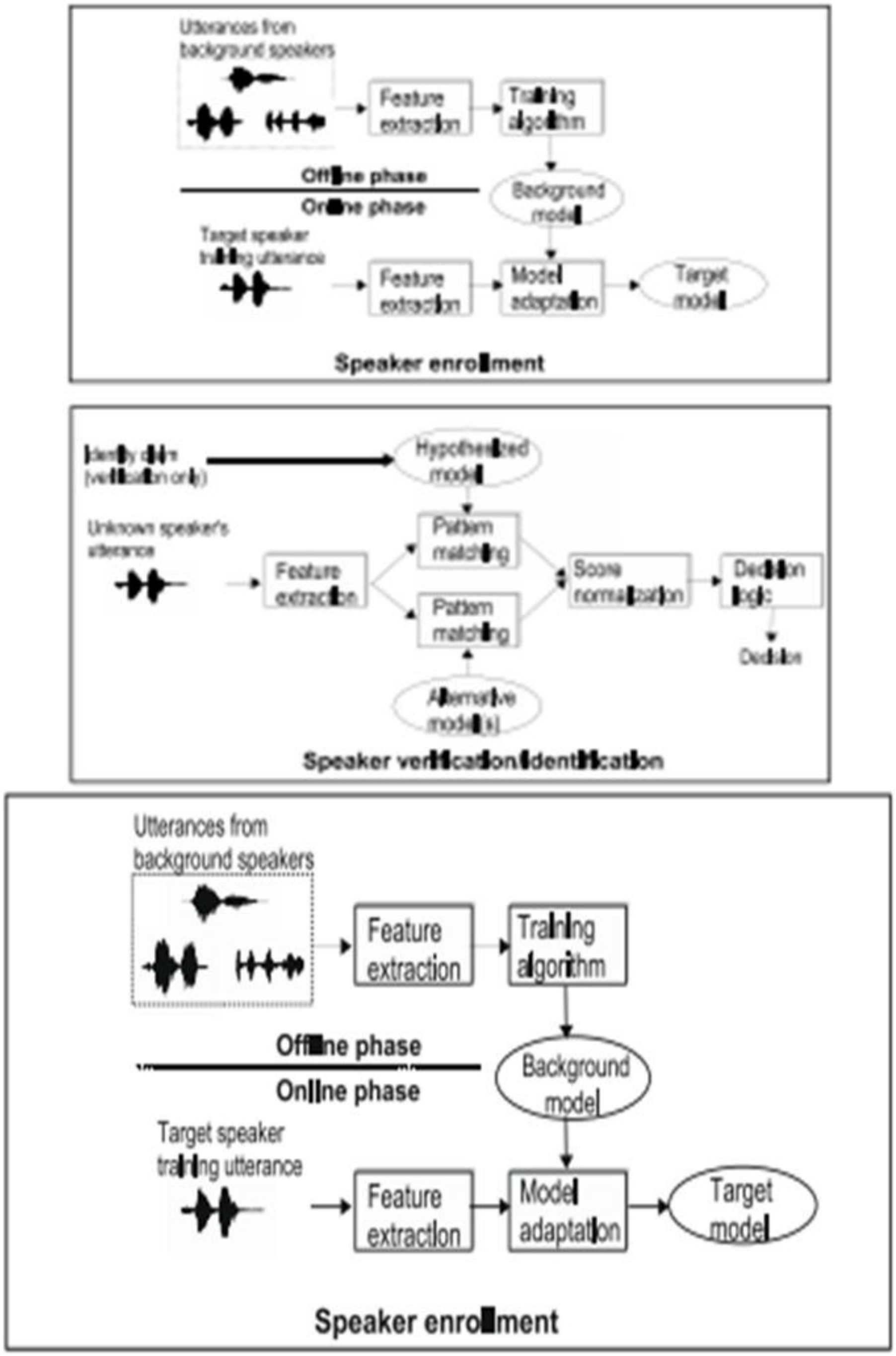

Figure 1: Components of a typical automatic speaker recognition system. In the enrollment mode, a speaker model is created with the aid of previously created background model; in recognition mode, both the hypothesized model and the background model are matched and background score is used in normalizing the raw score.

2. Feature Extraction

The purpose of this module is to convert the speech waveform to some type of parametric representation (at a considerably lower information rate). The speech signal is a slowly time varying signal (it is called quasi-stationary). When examined over a sufficiently short period of time (between 5 and 100 ms), its characteristics are fairly stationary. However, over long periods of time (on the order of 0.2s or more) the signal characteristics change to reflect the different speech sounds being spoken.

Therefore, short-time spectral analysis is the most common way to characterize the speech signal. A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), and others. MFCC is perhaps the best known and most popular.

3. Gaussian Mixture Model

For text independent speaker recognition, where there is no prior knowledge of what the speaker will say, the most successful likelihood function has been GMMs. In text dependent applications, where there is a strong prior knowledge of the spoken text, additional temporal knowledge can be incorporated by using hidden Markov models (HMMs) for the likelihood functions. To date, however, the use of more complicated likelihood functions, such as those based on HMMs, have shown no advantage over GMMs for text independent speaker detection tasks like in the NIST speaker recognition evaluations (SREs).For a *D*-dimensional feature vector *_x*, the mixture density used for the likelihood function is defined as follows:

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} w_i \, p_i(\vec{x}). \qquad (1)$$

The density is a weighted linear combination of *M* unimodal Gaussian densities $p_i(\vec{x})$ each parameterized by a *D*×1 mean Vector $\vec{\mu_i}$ and a *D* × *D* covariance matrix $\Sigma\, i$

$$p_i(\vec{x}) = \frac{1}{(2\Pi)^{D/2}|\Sigma_i|^{1/2}} e^{-(1/2)(\vec{x}-\vec{\mu_i})'\Sigma_i^{-1}(\vec{x}-\vec{\mu_i})} \qquad (2)$$

The mixture weights *wi* further satisfy the constraint

$$\sum_{i=1}^{M} w_i = 1 \qquad (3)$$

Collectively, the parameters of the density model are denoted as

$$\lambda = (w_i, \vec{\mu_i}, \Sigma_i),\ i = (1\ldots M) \qquad (4)$$

While the general model form supports full covariance matrices, that is, a covariance matrix with all its elements, typically only diagonal covariance matrices are used. This is done for three reasons. First, the density modeling of an *M*th-order full covariance GMM can equally well be achieved using a larger-order diagonal covariance GMM. Second, diagonal-matrix GMMs are more computationally efficient than full covariance GMMs for training since repeated inversions of a $D\times D$ matrix are not required. Third, empirically, it has been observed that diagonal-matrix GMMs outperform full-matrix GMMs.

Given a collection of training vectors, maximum likelihood model parameters are estimated using the iterative expectation-maximization (EM) algorithm. The EM algorithm iteratively refines the GMM parameters to monotonically increase the likelihood of the estimated model for the observed feature vectors, that is, for iterations *k* and *k*+1,

$$p\ (X|\ \lambda^{(k+1)}) \geq p(X|\lambda^{(k)}) \qquad (5)$$

Generally, five–ten iterations are sufficient for parameter convergence. The EM equations for training a GMM can be found in the literature. Under the assumption of independent feature vectors, the log likelihood of a model ë for a sequence of feature vectors

*X = {_x*1. . ._*xT}* is computed as follows:

$$\log p\ (X|\lambda) = 1/T\ \Sigma_t \log p\ (\vec{x}_t\ |\lambda) \qquad (6)$$

Note that the average log-likelihood value is used so as to normalize out duration effects from the log-likelihood value. Also, since the incorrect assumption of independence is underestimating the actual likelihood value with dependencies, scaling by *T* can be considered a rough compensation factor. The GMM can be viewed as a hybrid between parametric and nonparametric density models. Like a parametric model, it has structure and parameters that control the behavior of the density in known ways, but without constraints that the data must be of a specific distribution type, such as Gaussian or Laplacian.Like a nonparametric model, the GMM has many degrees of freedom to allow arbitrary density modeling, without undue computation and storage demands. It can also be thought of as a single-state HMM with a Gaussian mixture observation density, or an ergodic Gaussian observation HMM with fixed, equal transition probabilities.Here, the Gaussian components can be considered to be modeling the underlying broad phonetic sounds that characterize a person’s voice. A more detailed discussion of how GMMs apply to speaker modeling can be found elsewhere.

The advantages of using a GMM as the likelihood function are that it is computation nally inexpensive is based on a well-understood statistical model and, for text-independent tasks, is insensitive to the temporal aspects of the speech, modeling only the underlying distribution of acoustic observations from a speaker. The latter is also a disadvantage in that higher-levels of information about the speaker conveyed in the temporal speech signal are not used.

4. Vector Quantization and Supervectors

A Speaker recognition system must able to estimate probability distributions of the computed feature vectors. Storing every single vector that generate from the training mode is impossible, since these distributions are defined over a high-dimensional space. It is often easier to start by quantizing each feature vector to one of a relatively small number of template vectors, with a process called vector quantization. VQ is a process of taking a large set of feature vectors and producing a smaller set of measure vectors that represents the centroids of the distribution.

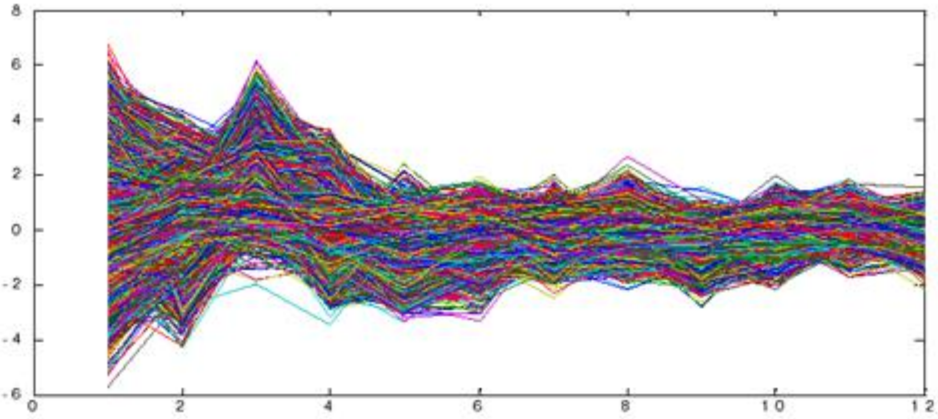

Figure 2. Vectors Generated from Training before VQ

The technique of VQ consists of extracting a small number of representative feature vectors as an efficient means of characterizing the speaker specific features .By means of VQ, storing every single vector that we generate from the training is impossible.

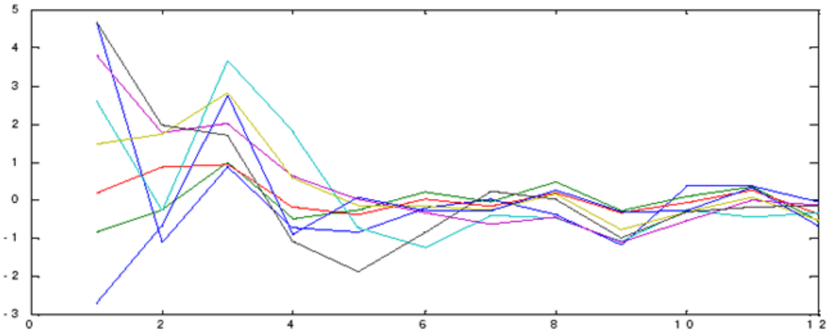

Figure 3. Representative Feature Vectors Resulted after VQ

By using these training data features are clustered to form a codebook for each speaker. In the recognition stage, the data from the tested speaker is compared to the codebook of each speaker and measure the difference. These differences are then use to make the recognition decision. In vector Quantization each utterances could be represented as a single vector .The average vectors would then be compared using a distance measure which is computationally very efficient but gives poor recognition accuracy. A robust way to present utterances using a single vector called supervector.These supervectors can be used as inputs to Support Vector Machine (SVM).

5. Support Vector Machine

SVMs are a principled technique to train classifiers that stems from statistical learning theory. Their root is the optimal hyperplane algorithm. They minimize a bound on the empirical error and the complexity of the classifier at the same time. Accordingly, they are capable of learning in sparse high-dimensional spaces with relatively few training examples.

Let $\{\mathbf{x}_i, y_i\}$, i= 1, 2, 3 ….N, denote *N* training examples where x*i* comprises an *M*-dimensional pattern and *yi* is its class label. Without any loss of generality we shall confine ourselves to the two-class pattern recognition problem.That is, $y_i \in \{-1, +1\}$. We agree that *yi* = +1 is assigned to positive examples, whereas *yi* = ¡1 is assigned to counter examples. The data to be classified by the SVM might be linearly separable in their original domain or not. If they are separable, then a simple linear SVM can be used for their classification. However, the power of SVMs is demonstrated better in the non separable case, when the data cannot be separated by a hyperplane in their original domain. In the latter case, we can project the data into a higher dimensional Hilbert space and attempt to linearly separate them in the higher dimensional space using kernel functions.

$$R^M \rightarrow \mathcal{H} \qquad (7)$$

Let $\emptyset$ denote a nonlinear map where $\mathcal{H}$ is a higher-dimensional Hilbert space. SVMs construct the optimal separating hyperplane in *H*. Therefore, their decision boundary is of the form:

$$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_I y_i K(x, x_i) + b \right) \qquad (8)$$

Where *K* (z1*;* z2) is a kernel function that defines the dot product between $\emptyset(z1)$ *and* $\emptyset$(z2) in *H*, and αi are the nonnegative Lagrange multipliers associated with the quadratic optimization problem that aims to maximize the distance between the two classes measured in *H* subject to the constraints

$$w^T \Phi(x_i) + b \geq 1 \qquad \text{for } y_i = +1$$

$$w^T \Phi(x_i) + b \leq 1 \qquad \text{for } y_i = -1$$

Frequently used kernel functions are:

1) The polynomial kernel:

   $K(x_i, x_j) = (m\, x_i^T x_j + n)^d$;

2) The Radial Basis Function (RBF) kernel:

   $K(x_i, x_j) = \exp(-\gamma \mid x_i - x_j \mid^2 \}$.

To enable the use of SVMs in visual speech recognition, when we model the speech as a temporal sequence of symbols corresponding to the different phones produced, we shall employ the SVMs as nodes in a Viterbi lattice. The nodes of such a Viterbi lattice are supposed to generate the posterior probabilities of the corresponding symbols to be emitted, and the standard SVMs do not provide such probabilities as output. Several solutions are proposed in the literature to map the SVM output to probabilities: the cosine decomposition proposed by Vapnik, the probabilistic approximation by applying the evidence framework to SVMs, the sigmoidal approximation by Platt. Here we adopt the solution proposed by Platt, since it is a simple solution which was already used in a similar application of SVMs to audio speech recognition. This solution shows that having a trained SVM, we can convert its output to probability by training the parameters of a sigmoidal mapping function:

$$P(y = +1|f(x)) = \frac{1}{1 + \exp(a_1 f(x) + a_2)} \text{ `} \quad (9)$$

where $\alpha 1$ and $\alpha 2$ are the parameters of the sigmoidal mapping to be derived for the trained SVM under consideration with $a1 < 0$. $P(y = +1\ j\ f(x))$ gives directly the posterior probability to be used in the Viterbi decoder. The parameters $a1$ and $a2$ are derived from the training set $\{ f(x_i),y_i\},I = 1,2,3,\ldots\ldots N$,using maximum likelihood estimation.

6. GMM Supervector

The UBM is trained using the background databases that are selected to reflect the alternative imposter speeches. The EM algorithm is used for the UBM training. The GMM probability density can be described as follows:

$$p(x) = \sum_{i=1}^{M} w_i f(x|m_i,\Sigma_i) \quad (10)$$

Where x is a -dimensional cepstral feature vector, and m_i, Σ_I, w_i, $(I = 1 \ldots M)$ are, respectively, the mean vector, the co-variance matrix, and the weight of the Gaussian component. F (.) denotes Gaussian density function, i.e.,

$$‘(x \mid m_i, \Sigma_i) = \frac{(2\prod)^{-D/2}}{|\Sigma_i|^{1/2}} \times \exp(-1/2(x - m_i)^T \Sigma_i^{-1}(x - m_i)) \quad (11)$$

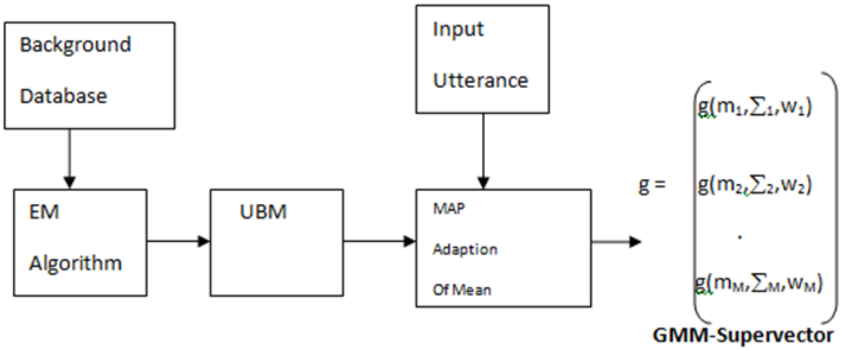

Figure 4. Process of Generating the GMM Supervector from an Utterance

g (m_i, $\sum_I$, w_i) is the function that represents the normalized mean aligned by covariance and weight.The UBM can be expressed by

$$u = \{w_i^{(u)}, m_i^{(u)}, \sum_i^{(u)} | \ I = 1,2,\ldots\ldots M\} \qquad (12)$$

The speaker GMM, λ , can be obtained by MAP adaptation, and it has the same form as follows:

$$\lambda = \{w_i^{(\lambda)}, \ m_i^{(\lambda)}, \sum_i^{(\lambda)} | \ I = 1,2,\ldots\ldots M\} \qquad (13)$$

The process of generating the GMM-supervector can be summarized in above figure. The GMM-supervector is formed by concatenating the normalized means of the Gaussian components.Suppose we have a Gaussian mixture model universal background model (GMM UBM),

$$g(x) = \sum_{i=1}^{N} \lambda_i \, N(x; m_i,\Sigma_i) \qquad (14)$$

where λ_i are the mixture weights, N() is a Guassian, and m_i and $\sum_i$ are the mean and covariance of the Guassians, respectively. We aasume diagonal covariances, $\sum$

Given a speaker utterance, GMM UBM training is performed by MAP adaptation of the means, mi. From this adapted model, we form a GMM supervector. The process is shown in Figure 5.

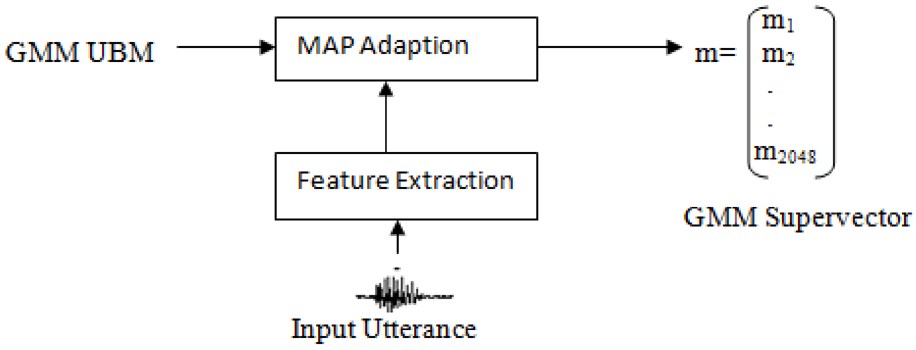

Figure 5. GMM Supervector Process

The GMM supervector can be thought of as a mapping between an utterance and a high-dimensional vector. This concept fits well with the idea of a SVM sequence kernel. The basic idea of a sequence kernel is to compare two speech utterances, utta and uttb, directly with a kernel, K(utta; uttb). The kernel can be written as K(utta; uttb) = b(utta) tb(uttb) because of the Mercer condition. The GMM supervector mapping is then part of the mapping of utta to b(utta).

7. GMM supervector based SVM Vss GMM

We use GMM KL divergence kernel in GMM Supervector based SVM system to compare it with standard GMM system [27]. In the standard GMM system , each emotion was modeled by a GMM trained with the corresponding emotional utterances .The GMMs were trained via EM algorithm, each of which consisted of 64 Gaussian components .A maximum likelihood Bayes classifier is used for decision .The accuracy of these two systems is shown in figure below :

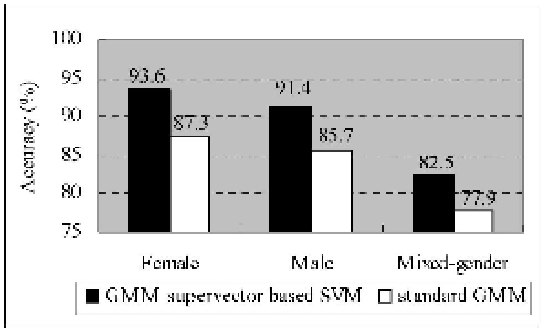

Figure 6: Accuracy of GMM Supervector based SVM using the GMM KL Divergence Kernel and Standard GMM

From above figure it can be seen that GMM supervector based SVM significantly outperforms the standard GMM for speech emotion recognition. More precisely compared with the standard GMM, the accuracy of GMM supervector based SVM is 6.3% higher for female subject. 5.7% higher for male subject and 4.6 % higher for mixed-gender subject.

The results also indicate that it is helpful to identify the speaker’s gender in the utterance first, and then perform the emotion recognition on a gender-dependent system. Since the gender-dependent emotion recognition system is preferred, we analyzed the confusion between different emotions in condition of separate-gender subject. Confusion matrices of the GMM supervector based SVM for female and male subjects are shown in Table1 and Table2, respectively.From the results, we can see that fear; happiness and anger are the most frequently confused emotions for both female and male subjects. This might be attributed to the similar arousal level [28] when speakers are in these three emotional states. It can be also found that sadness and fear are easily misclassified for male subject. The reason could be that the valence level of these two emotional states is close [28].

Table 1. Confusion Matrix of GMM Super Vector based SVM for Female Subject (A : anger , F : fear , H : happiness , N : neutral , S : sadness)

	Classified Emotion (%)				
Intended Emotion	A	F	H	N	S
A	**97.9**	0.7	1.4	0.0	0.0
V	8.0	**78.4**	9.6	0.0	4.0
H	1.7	7.5	**90.8**	0.0	0.0
N	0.0	0.0	0.0	**100.0**	0.0
S	0.0	0.7	0.0	0.7	**98.6**

Table 2. Confusion Matrix of GMM Super Vector based SVM for Male Subject (A : anger , F : fear , H : happiness , N : neutral , S : sadness)

	Classified Emotion (%)				
Intended Emotion	A	F	H	N	S
A	**96.5**	0.0	2.8	0.7	0.0
V	3.0	**91.1**	3.7	0.0	2.2
H	8.9	7.8	**80.0**	3.3	0.0
N	0.0	0.0	2.2	**96.3**	1.5
S	0.0	10.0	0.0	1.8	**88.2**

8. Applications of Speaker Verification

There are many applications to speaker verification. The applications cover almost all the areas where it is desirable to secure actions, transactions, or any type of interactions by identifying or authenticating the person making the transaction. We briefly review those various applications.

8.1 On-site Applications

On-site applications regroup all the applications where the user needs to be in front of the system to be authenticated. Typical examples are access control to some facilities (car, home, warehouse), to some objects (locksmith), or to a computer terminal. Currently, ID verification in such context is done by mean of a key, a badge ora password, or personal identification number (PIN).

8.2 Remote Applications

Remote applications regroup all the applications where the access to the system is made through a remote terminal, typically a telephone or a computer. The aim is to secure the access to reserved services (telecom network, databases, web sites, etc.) or to authenticate the user making a particular transaction (e-trade, banking transaction, etc.).

8.3 Information Structuring

Organizing the information in audio documents is a third type of applications where speaker recognition technology is involved. Typical examples of the applications are the automatic annotation of audio archives, speaker indexing of sound tracks, and speaker change detection for automatic subtitling. The need for such applications comes from the movie industry and from the media related industry recognition is a key technology for audio indexing.

8.4 Games

Finally, another application area, rarely explored so far, is games: child toys, video games, and so forth. Indeed, games evolve toward a better interactivity and the use of player profiles to make the game more personal.

9. Conclusions and Future Trends

In this paper, we have proposed a tutorial on text-independent speaker verification. After describing the training and test phases of a general speaker verification system, we detailed the cepstral analysis, which is the most commonly used approach for speech parameterization. Then, we explained how to build a speaker model based on a GMM approach. We propose to apply the GMM supervector based SVM with spectral features to speech emotion recognition. The GMM KL divergence kernel was shown to yield better performance than other commonly used kernels in the proposed system. The results suggest that the Gender information should be considered in speech emotion recognition, and demonstrate that the GMM supervector based SVM system significantly out performs standard GMM system. For the frequently confused emotional states, other type of features, such as prosodic and voice quality Features can be fused with our proposed method to enhance the emotion recognition performance in future work

References

[1] Reynolds DA, “Automatic speaker Recognition: Current Approaches and Future Trends”, MIT Lincoln Laboratory Lexington, MA USA dar@ll.mit.edu, ICASSP **(2001)**.

[2] Reynolds DA, “An overview on Automatic speaker Recognition Technology”, MIT Lincoln Laboratory Lexington, MA USA, dar@ll.mit.edu, IEEE, **(2002)**.

[3] Bimbot F, Bonastre JF, Fredouille C, Gravier G, Magrin-Chagnolleau I, Meignier S, Merlin T, Ortega-Garc ´ýa J, Petrovska-Delacr s ´etaz D, Reynolds DA, “A Tutorial on Text-Independent Speaker Verification “, EURASIP Journal on Applied Signal Processing, vol. 2004, no. 4, **(2004)**, pp. 430–451.

[4] Stolcke A, Shriberg E, Ferrer L, Kajarekar S, Sonmez K, Tur G, “Speech Technology and Research Laboratory”, SRI International, Menlo Park, CA, USA , speech recognition as feature extraction for speaker recognition , SAFE **(2007)** April 11-13,Washington ,DC, USA

[5] Campbell W, Sturim D, Reynolds DA, “Support Vector Machine using GMM Supervector for Speaker Verification”, IEEE Signal processing letters, vol. 13, no. 5, **(2006)** May.

[6] Deshak N, Chollet G, “support vector GMMs for speaker verification”, **(2006)** June.

[7] Schmidt M, Gish H, “Speaker identification via support vector classifiers”, in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP ’96), vol. 1, **(1996)** May, pp. 105–108, Atlanta, Ga, USA.

[8] Reynolds DA, Rose RC, “Robust text-independent speaker identification using Gaussian mixture speaker models”, IEEE Trans. Speech, and Audio Processing, **(1995)**.

[9] Reynolds DA, “Comparison of background methods Normalization for text-independent speaker verification”, in 5th European Conference on Speech Communication and Technology, **(1997)**.

[10] Reynolds DA, Quatieri TF, Dunn R, “Speaker verification using adapted Gaussian mixture models”, Digital Signal Processing, vol. 10, no. 1-3, **(2000)**, pp. 194.

[11] Kenny P, Dumouchel P, “Experiments in speaker verification using factor analysis likelihood ratios”, in Proc. Odyssey04, **(2004)**, pp. 219-226.

[12] Kenny P, Boulianne G, Dumouchel P, “Eigenvoice modeling with sparse training data”, IEEE Trans. Speech and Audio Processing, vol. 13, no. 3, **(2005)**, pp. 345-354.

[13] Campbell WM, “generalized linear discriminant sequence kernels for speaker recognition”, in Proceedings of the International Conference on Acoustics Speech and Signal Processing, **(2002)**, pp. 161-164.

[14] Campbell W, Campbell J, Reynolds D, Singer E, Torres-Carrasquillo P, “Support vector machines for speaker and language recognition”, Computer Speech and Language, vol. 20, no. 2-3, **(2006)** April, pp. 210–229.

[15] Reynolds D, Quatieri T, Dunn R, “Speaker verification using adapted Gaussian mixture models”, Digital Signal Processing, vol. 10, no. 1, **(2000)** January, pp. 19–41.

[16] Furui S, “Recent advances in speaker recognition”, Pattern Recognition Letters, vol. 18, no. 9, **(1997)** September, pp. 859–872.

[17] Higgins A, Bahler L, Porter J, “Speaker verification using randomized phrase prompting”, Digital Signal Processing , vol. 1, **(1991)** April, pp. 89–106.

[18] Li KP, Porter J, “Normalizations and selection of speech segments for speaker (New recognition scoring)”, In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1988) York, USA, **(1988)** April, pp. 595–598.

[19] Reynolds D, “Speaker identification and verification using Gaussian mixture speaker models”, Speech Communication, vol. 17, **(1995)** August, pp. 91–108.

[20] Reynolds D, Quatieri T, Dunn R, “Speaker verification using adapted Gaussian mixture models”, Digital Signal Processing, vol. 10, no. 1, **(2000)** January, pp. 19–41.

[21] Sivakumaran P, Fortuna J, Ariyaeeinia A, “Score normalization applied to open-set, text-independent speaker identification”, In Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland, **(2003)** September, pp. 2669–2672.

[22] Kinnunen T, Kilpeläinen T, Fränti P, “Comparison of Clustering Algorithms in Speaker Identification”, **(2000)**.

[23] Gill MK, “A Viable Technique: Speaker Recognition”, www.rimtengg.com/iscet/proceedings/pdf, **(2000)**.

[24] Réda A, Aoued B, “Artificial Neural Network & Mel-Frequency Cepstrum Coefficients-Based Speaker Recognition”, **(2005)** March 27-31, www.setit.rnu.tn/last_edition/setit20.

[25] Hasan R, Jamil M, Rabbani G, Rahman S, “Speaker Identification Using Mel Frequency Cepstral Coefficients”, ICECE, **(2004)**, pp. 565-568.

[26] Tychtl Z, Psutka J, “Speech Production Based on the Mel-Frequency Cepstral Coefficients”, www.informatik.unitrier.de/~ley/db/indices/a.../Tychtl:Zbynik.html, **(2000)**.

[27] Hu H, Xu MX, Wu W, “GMM Supervector based SVM with spectral features for speech emotion recognition”, Center for Speech Technology, Tsinghua National Lab for Information Science and Technology, Tsinghua University, Beijing, 100084, China, **(2000)**.