DAILY SOUND RECOGNITION USING A COMBINATION OF GMM AND SVM FOR HOME AUTOMATION

M. A. Sehili[1,2]*, D. Istrate*[1]*, B. Dorizzi*[2]*, J. Boudy*[2]

[1]ESIGETEL, 1 Rue du Port de Valvins,77210 Avon, France
[2]Telecom SudParis, 9 Rue Charles Fourier, 91000 Evry, France

ABSTRACT

Most elderly people monitoring systems include the detection of abnormal situations, in particular distress situations, as one of their main goals. In order to reach this objective, many solutions end up combining several modalities such as video tracking, fall detection and sound recognition, so as to increase the reliability of the system. In this work we focus on daily sound recognition as it is one of the most promising modalities. We make a comparison of two standard methods used for speaker recognition and verification : Gaussian Mixture Models (GMM) and Support Vector Machines (SVM). Experimental results show the effectiveness of the combination of GMM and SVM in order to classify sound data sequences when compared to systems based on GMM.

Index Terms— Sound classification, Gaussian Mixture Models, Support Vector Machines

1. INTRODUCTION

The number of elderly people living alone in many countries has been steadily increasing for decades. In the last few years, many monitoring projects for elderly people have been proposed, each with its own set of solutions. The most common goals of such systems, regardless of the technologies involved, is to keep elderly people living in their homes as long as possible, to increase their safety and to help them to remain autonomous. Another motivation with regard to these systems is a better management of human and material resources.

Sound is one of the most promising modalities, not only because of the huge amount of information it carries about the elderly and their environment, but also because of the inexpensiveness, the high quality and the low intrusion level of the sensors. In the same time the sound recognition is a difficult task because of the very large number of types of sounds, because of the large variability of the same sound and because of the noise presence ; therefore the sound analysis will be coupled with other modalities. This makes them more acceptable to this population in comparison to other types of sensors like video cameras. Moreover, sound can be an ergonomic and natural communication solution between the subject and their environment via voice-command. Auditory Scene Analysis (ASA) aims at separating and recognizing the environmental sounds [1] in order to infer a real-life situation from one or many acoustic events. One common application is the separation of voice and non-voice signals to increase the reliability of an automatic speech recognition system. In the field of technological solutions for assisted living certain sound classes such as screaming or crying can reveal an abnormal situation. However, the great variation of sounds and the presence of noise may considerably affect the efficiency of this modality making it very challenging. This work is part of the Sweet-Home project whose goals are to enable audio-based interaction of elderly and frail people and to provide them a higher safety and security level through the detection of distress situations.

Sound recognition methods have taken advantage of the maturity of speaker recognition methods such as GMM, HMM and SVM. [2] [3]. Our significant contribution is to investigate in this paper the combination of GMMs which are a generative model, and the discrimination power of the kernel method SVM via the use of sequence discriminant kernel [4]. First results show that the combination of SVM and GMM increase the performance of the system.

The paper is organised as follows. In section 2 there is an outline of GMM-based systems and some of the most commonly used combinations of GMM and SVM systems and the details of the chosen method : GMM based on Supervectors Linear kernel. In section 3 we report the sound database used in our experiments, the sound features, the test protocol and the results obtained. In section 4 we report our conclusions and perspectives.

2. SOUND CLASSIFICATION APPROACHES

2.1. Gaussian Mixture Models

Gaussian mixture model classifiers [5] [6] have been used successfully for speaker recognition and verification and they have become a standard tool. This has encouraged their application in other sound classification tasks such as normal life

This work is a part of the Sweet-Home project founded by the French National Research Agency (Agence Nationale de la Recherche / ANR-09—VERS-011)

sounds [2] [7]. The basic idea of GMM classifiers applied to sound recognition is to consider a sound signal as a sequence of independent observations. The score of the whole sequence is then computed as the product of the vector's likelihood :

$$p(X|\lambda) = \prod_{t=1}^{T} p(\mathbf{x_t}|\lambda) \tag{1}$$

where X is a sequence of length T. $p(\mathbf{x_t}|\lambda)$ is the likelihood of vector $\mathbf{x_t}$ given a mixture model λ :

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{C} w_i \mathcal{N}(\mathbf{x}|\mathbf{m_i}, \mathbf{\Sigma_i}) \tag{2}$$

where $\mathbf{x}$ is a feature vector of dimension d, w_i are the weights of the C gaussians $\mathcal{N}(\mathbf{m_i}, \mathbf{\Sigma_i})$, $\mathbf{m_i}$ is the mean and $\mathbf{\Sigma_i}$ the covariance.

2.2. Support Vector Machines and Sequence Discriminant Kernels

Support Vector Machines [8] [9] is a discriminative method that has been used in many data classification applications in recent years. Its main advantage is its ability to classify data in a non-linear space thanks to kernel functions :

$$f(\mathbf{x}) = \sum_{i=1}^{N_{sv}} \alpha_i y_i K(\mathbf{x}, \mathbf{x_i}) + b \tag{3}$$

where $\mathbf{x_i}$ are the support vectors chosen from training data via an optimisation process and $y_i \in \{-1, +1\}$ are respectively their associated labels. $K(\mathbf{x}, \mathbf{x_i})$ is the kernel function and must fulfill some conditions :

$$K(\mathbf{x}, \mathbf{y}) = \mathbf{\Phi}(\mathbf{x})^t \mathbf{\Phi}(\mathbf{y}) \tag{4}$$

where $\mathbf{\Phi}$ is a mapping from the input space to a possible infinite-dimensional space.

The use of SVMs at frame level, as in the case of GMMs, shew its limitations for speaker recognition[10] and sound recognition[11] in terms of efficiency and training time as the number of frames increases. Sequence discrimination kernels [12] were proposed to overcome this problem by directly discriminating sequences of arbitrary length instead of operating the discrimination on the vectors which they compose. The Fisher kernel [4] is a score-space kernel that uses generative models to map a whole sequence into an unique high dimensional vector. The Fisher mapping is defined as follows :

$$\mathbf{\Psi}_{\text{Fisher}}(X) = \nabla_\theta \log P(X|\theta) \tag{5}$$

This sequence kernel was used for speech and speaker recognition[13][14] and speaker verification [15].

In [16], the general definition of a sequence kernel is :

$$K(X,Y) = \mathbf{\Phi}(X)^t \mathbf{R^{-1}} \mathbf{\Phi}(Y) = \left(\mathbf{R}^{-\frac{1}{2}} \mathbf{m}^X\right)^t \left(\mathbf{R}^{-\frac{1}{2}} \mathbf{m}^Y\right) \tag{6}$$

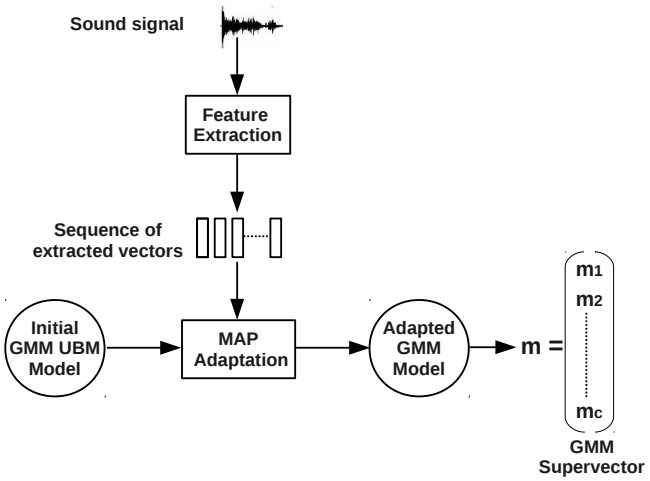

Fig. 1. GMM Supervector mapping process

where $\mathbf{\Phi}(X)$ is the high-dimensional vector resulting from the mapping of sequence X and $\mathbf{R^{-1}}$ is a diagonal normalization matrix. In [16], there is also a comparison made between two sequence kernels for speaker verification. The first kernel is the Generalized Linear Discriminant Sequence Kernel (GLDS) proposed initially for speaker and language recognition[17][18]. The second kernel is the GMM Supervector Linear kernel (GSL) [19] [16]. The GSL gave better performance and thus justify our choice.

2.3. GMM based on Supervector Linear Kernel

The map function of the GSL kernel $\mathbf{\Phi}_{\text{GLS}}$ uses a GMM Universal Background Model (UBM) of diagonal covariances to map a sequence of vectors X extracted from one signal into an unique high-dimensional vector $\mathbf{m}^X$ (8). The sequence is used to perform an adaptation of the mean vectors of the UBM via a MAP procedure [20]. In order to compute the kernel K as in equation (6), we define as $\mathbf{\Phi}_{\text{GSL}}(X)$ the supervector composed of the stacked means from the UBM components.

$$\mathbf{\Phi}_{\text{GSL}}(X) = \mathbf{m}^X \begin{bmatrix} \mathbf{m}_1^X \\ \mathbf{m}_2^X \\ \cdots \\ \mathbf{m}_C^X \end{bmatrix} \tag{7}$$

The normalization matrix $\mathbf{R^{-1}}$ is defined using the weights and the covariances of the UBM model :

$$diag(\mathbf{R}^{-\frac{1}{2}}) = \begin{bmatrix} \sqrt{w_1}\mathbf{\Sigma}_1^{-\frac{1}{2}} \\ \sqrt{w_2}\mathbf{\Sigma}_2^{-\frac{1}{2}} \\ \cdots \\ \sqrt{w_C}\mathbf{\Sigma}_C^{-\frac{1}{2}} \end{bmatrix} \tag{8}$$

Table 1. Sound dataset classes

Sound class	# of files	Total duration (sec.)
Breathing	50	106.44
Cough	62	181.69
Dishes	98	303.77
DoorClapping	114	62.70
DoorOpening	21	138.94
ElectricalShaver	62	420.33
FemaleCry	36	268.19
FemaleScream	70	216.83
GlassBreaking	101	99.52
HairDryer	40	224.86
HandsClapping	54	218.65
Keys	36	166.34
Laugh	49	272.65
MaleScream	87	202.11
Paper	63	330.66
Sneeze	32	51.67
Water	54	484.72
Yawn	20	95.87

3. EXPERIMENTS

3.1. Sound dataset

In order to compare GMMs and SVM using a GSL for daily sound recognition, we have used a database of 18 sound classes representing some human sounds and some human activity sounds. The sound files are 16KHz sampled wave files and were either recorded using a microphone or downloaded from Internet. The database can be acquired on request to the authors. Table 1 shows the sound classes and the corresponding number of files. We used 16 MFCC (Mel-Frequency Cepstral Coefficients) feature vectors extracted every 8 ms using a 16 ms Hamming window [21] [22].

The influence of the stationnary noise produced by a vacuum cleaner was also studied. The noise was added to useful signals in order to obtain an average Signal to Noise Ratio (SNR) of 5, 10 and 50 dB.

3.2. Test protocol

The sound database is divided into three equal parts. The first part is used to create the GMM models as well as the UBM model. The second part is used to test both systems and the third part, it is employed to generate the supervectors used for the SVM training.

The GMM models contain from 25 up to 50 components per class. For the SVM GSL system, we used an UBM of 512 and 1024 components which gives supervectors of 8192 and 16384 dimensions respectively. The SVM multiclasses classification is achieved using a one-to-one scheme (a combination of 2 classes SVMs). We used the Alize library [23] to create and update the Gaussian models and to generate the supervectors. For SVM, we used the libsvm library[24]. The Global Score is calculated as the average score of the individual classes scores in order to normalize with the number of files in each class.

Table 2. Comparative recognition performances for GMM and SVM GSL

Method	Good Recognition Ratio
GMM	0.69 $\pm$0.0028
SVM GSL	0.75 $\pm$0.0028

3.3. Results

Table 2 shows that the SVM GSL system outperforms the GMM system ; the improvement is about 8%. The best performances was obtained using an UBM of 1024 components. The Table 4 show the confusion matrix of the GSM GSL method. We can observe that breathing sounds are at most considered like door opening sounds which has 100% recognition rate and Dishes sounds are confused with Hands Clappings sounds, another class with a good recogniation rate (94%). The explanation can be based on the spectral contenance of the sound class in the spectrum of the other one.

The noise influence was studied for a stationnary noise which has long duration and in real condition can be simultaneous to a domotic sound. For the experiments the vacuum cleaner noise was choosen. We can observe in the Table 3 that SVM GSL system increasis the performances with about 18% for SNR between 5 and 10 dB.

4. CONCLUSION AND PERSPECTIVES

Advances in SVM sequence discriminant kernels have led to better performances compared to GMM systems in the fields of speaker recognition and verification in the recent years, and they have become a standard tool. In this work we have tested the GSL kernel which has shown good results over the classical GMM system for the daily sound recognition.

Another main point in perspective is the fact that several everyday sounds are actually very distinguishable from other

Table 3. Good Recognition Ratio in the case of noise presence

Method	Good Recognition Ratio for a SNR			
	Without Noise	5 dB	10 dB	50 dB
GMM	0.69	0.47	0.52	0.66
SVM GSL	0.75	0.55	0.62	0.72

sounds and thus could be recognized at a lower level using simpler features such as signal shape, which might lead to hierarchical classification schemes.

5. REFERENCES

[1] A. Bregman, "Auditory scene analysis," 1990, MIT Press, Cambridge.

[2] J.E. Rougui, D. Istrate, and W. Souidene, "Audio sound event detection for distress situations and context awareness," *31st Annual International Conference of the IEEE EMBS*, pp. 3501–3504, 2009.

[3] Andrey Temko and Climent Nadeu, "Classification of meeting-room acoustic events with support vector," in *Machines and Confusion-based Clustering", Proc. ICASSP'05*, 2005, pp. 505–508.

[4] Tommi Jaakkola and David Haussler, "Exploiting generative models in discriminative classifiers," in *In Advances in Neural Information Processing Systems 11*. 1998, pp. 487–493, MIT Press.

[5] Douglas A. Reynolds, "Gaussian mixture models," 2007.

[6] Douglas A. Reynolds and Richard C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, pp. 72–80, 1995.

[7] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *ICME 2005. IEEE International Conference on Multimedia and Expo*, 2005, pp. 1306–1309.

[8] Burges Christopher J. C., "A tutorial on support vector machines for pattern recognition," *Data Min. Knowl. Discov.*, pp. 121–167, 1998.

[9] B. Schölkopf and A. J. Smola, *Learning with Kernels*, MIT Press, 2002.

[10] M. Schmidt and H. Gish, "Speaker identification via support vector classifiers," in *Proceedings of the Acoustics, Speech, and Signal Processing, 1996. on Conference Proceedings., 1996 IEEE International Conference - Volume 01*, Washington, DC, USA, 1996, ICASSP '96, pp. 105–108, IEEE Computer Society.

[11] M. A. Sehili, D. Istrate, and J. Boudy, "Primary investigations of sound recognition for a domotic application using support vector machines," in *Annals of the University of Craiova, Series : Automation, Computers, electronics and Mathematics*, 2010, pp. 61–65.

[12] Andrey Temko, Enric Monte, and Climent Nadeu, "Comparison of sequence discriminant support vector machines for acoustic event classification," in *In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[13] S. Fine, J. Navratil, and R.A. Gopinath, "A hybrid gmm/svm approach to speaker identification," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, 2001, pp. 417–420.

[14] Quan Le and Samy Bengio, "Hybrid generative-discriminative models for speech and speaker recognition," Idiap-RR Idiap-RR-06-2002, IDIAP, 0 2002.

[15] Vincent Wan and Steve Renals, "Speaker verification using sequence discriminant support vector machines," 2005.

[16] B.G.B. Fauve, D.a Matrouf, N. Scheffer, and J.-F. Bonastre, "State-of-the-art performance in text-independent speaker verification through open-source software," in *IEEE Transactions on Audio, Speech, and Language Processing*, 2007, vol. 15, pp. 1960–1968.

[17] William M. Campbell, Joseph P. Campbell, Douglas A. Reynolds, Elliot Singer, and Pedro A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 210–229, 2006.

[18] William M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," 2002.

[19] William. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "Svm based speaker verification using a gmm supervector kernel and nap variability compensation," in *in Proceedings of ICASSP, 2006*, 2006, pp. 97–100.

[20] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted gaussian mixture models," in *Digital Signal Processing*, 2000, p. 2000.

[21] Beth Logan, "Mel frequency cepstral coefficients for music modeling," in *Proceedings of International Symposium on Music Information Retrieval*, 2000.

[22] Sigurdur Sigurdsson, Kaare Brandt Petersen, and Tue Lehn-Schiøler, "Mel frequency cepstral coefficients : An evaluation of robustness of mp3 encoded music," in *Proceedings of International Symposium on Music Information Retrieval*, 2006.

[23] J.-F. Bonastre, F. Wils, and S. Meignier, "Alize, a free toolkit for speaker recognition," in *ICASSP'05, IEEE*, Philadelphia, PA (USA), March, 22 2005.

[24] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM : A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27 :1–27 :27, 2011.

	Breathing	Cough	Dishes	DoorClapping	DoorOPening	ElectricalShaver	FemaleCry	FemaleScream	GlassBreaking	HairDryer	HandsClapping	Keys	Laugh	MaleScream	Paper	Sneeze	Water	Yawn
Breathing	**6/17**	1/17	0/17	0/17	6/17	0/17	0/17	0/17	0/17	0/17	0/17	2/17	0/17	0/17	2/17	0/17	0/17	0/17
Cough	0/21	**15/21**	0/21	0/21	0/21	0/21	0/21	0/21	0/21	0/21	0/21	0/21	4/21	0/21	2/21	0/21	0/21	0/21
Dishes	0/33	1/33	**12/33**	0/33	1/33	0/33	0/33	3/33	2/33	0/33	13/33	0/33	0/33	1/33	0/33	0/33	0/33	0/33
Door Clapping	0/38	0/38	0/38	**38/38**	0/38	0/38	0/38	0/38	0/38	0/38	0/38	0/38	0/38	0/38	0/38	0/38	0/38	0/38
Door Opening	0/7	0/7	0/7	0/7	**7/7**	0/7	0/7	0/7	0/7	0/7	0/7	0/7	0/7	0/7	0/7	0/7	0/7	0/7
Electrical Shaver	0/21	0/21	0/21	0/21	0/21	**21/21**	0/21	0/21	0/21	0/21	0/21	0/21	0/21	0/21	0/21	0/21	0/21	0/21
Female Cry	0/12	3/12	0/12	0/12	0/12	0/12	**8/12**	1/12	0/12	0/12	0/12	0/12	0/12	0/12	0/12	0/12	0/12	0/12
Female Scream	0/24	1/24	1/24	0/24	0/24	0/24	0/24	**20/24**	0/24	0/24	0/24	0/24	0/24	2/24	0/24	0/24	0/24	0/24
Glass Breaking	0/34	3/34	0/34	3/34	0/34	0/34	0/34	0/34	**27/34**	0/34	0/34	0/34	0/34	0/34	1/34	0/34	0/34	0/34
Hair Dryer	0/14	0/14	0/14	0/14	0/14	0/14	0/14	0/14	0/14	**14/14**	0/14	0/14	0/14	0/14	0/14	0/14	0/14	0/14
Hands Clapping	0/18	1/18	0/18	0/18	0/18	0/18	0/18	0/18	0/18	0/18	**17/18**	0/18	0/18	0/18	0/18	0/18	0/18	0/18
Keys	0/12	0/12	0/12	0/12	0/12	0/12	0/12	0/12	0/12	0/12	1/12	**11/12**	0/12	0/12	0/12	0/12	0/12	0/12
Laugh	0/17	10/17	1/17	1/17	0/17	0/17	0/17	0/17	0/17	0/17	0/17	0/17	**3/17**	1/17	0/17	1/17	0/17	0/17
Male Scream	0/29	2/29	0/29	0/29	0/29	0/29	0/29	0/29	0/29	0/29	0/29	0/29	0/29	**27/29**	0/29	0/29	0/29	0/29
Paper	0/21	1/21	0/21	0/21	0/21	2/21	0/21	0/21	1/21	0/21	0/21	0/21	0/21	0/21	**17/21**	0/21	0/21	0/21
Sneeze	0/11	0/11	0/11	0/11	0/11	0/11	0/11	0/11	0/11	0/11	0/11	0/11	0/11	0/11	0/11	**11/11**	0/11	0/11
Water	0/18	1/18	0/18	5/18	0/18	0/18	0/18	0/18	0/18	0/18	0/18	0/18	0/18	0/18	0/18	0/18	**12/18**	0/18
Yawn	3/7	0/7	0/7	0/7	0/7	0/7	0/7	0/7	1/7	0/7	0/7	0/7	1/7	0/7	0/7	0/7	0/7	**2/7**

Table 4. Confusion Matrix of SVM GSL