ACOUSTIC FALL DETECTION USING GAUSSIAN MIXTURE MODELS AND GMM SUPERVECTORS

Xiaodan Zhuang[1]*, *Jing Huang*[2], *Gerasimos Potamianos*[2]**, *Mark Hasegawa-Johnson*[1]

[1] Dept. of ECE, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA
[2] IBM T.J. Watson Research Center, Yorktown Heights, New York, USA
Emails: xzhuang2@uiuc.edu; jghg@us.ibm.com; gpotam@ieee.org; jhasegaw@uiuc.edu

ABSTRACT

We present a system that detects human falls in the home environment, distinguishing them from competing noise, by using only the audio signal from a single far-field microphone. The proposed system models each fall or noise segment by means of a Gaussian mixture model (GMM) supervector, whose Euclidean distance measures the pairwise difference between audio segments. A support vector machine built on a kernel between GMM supervectors is employed to classify audio segments into falls and various types of noise. Experiments on a dataset of human falls, collected as part of the Netcarity project, show that the method improves fall classification F-score to 67% from 59% of a baseline GMM classifier. The approach also effectively addresses the more difficult fall detection problem, where audio segment boundaries are unknown. Specifically, we employ it to reclassify confusable segments produced by a dynamic programming scheme based on traditional GMMs. Such post-processing improves a fall detection accuracy metric by 5% relative.

Index Terms— fall detection, Gaussian mixture model, GMM supervector, support vector machine

1. INTRODUCTION

Assistance to dependent people, particularly to the elderly living alone at home, has been attracting increasing attention in today's aging societies [1]. Specifically, one public health problem of interest relevant to the elderly is this of falls that often go undetected and may result in injury [2]. Reliable and speedy detection of such potentially devastating events by automatic monitoring of the home is expected to be of benefit to both elderly and caregivers.

Not surprisingly, the research community has started to explore automatic fall detection based on input from a variety of sensors in specially equipped smart home environments. Wearable accelerometers, cameras, and microphones have been the most commonly used such devices, giving rise to a number of initial approaches for detecting falls that range from single-sensory to multi-sensory and multimodal algorithms [3]–[6]. In our work, among these sensors, we are interested in far-field microphones due to their unobtrusiveness, easy deployment, and in general lower cost and data stream bandwidth compared to cameras.

Automatic detection of human falls based on far-field audio is of course a non-trivial problem: First, falls are inconsistent phenomena; for example, ten fall types are identified in [3]. In addition, the acoustic signature of falls is affected by human characteristics and the impact surface. Furthermore, falls happening in realistic environments are easily confusable with daily noise, such as dropping objects, moving chairs, closing doors, and walking steps, and may overlap with a large variety of background noise.

Initial efforts to the problem of far-field acoustic fall detection are reported in [4]–[6]. These however mostly suffer from at least one of two shortcomings: They in general employ simple classification algorithms, thus achieving relatively low performance, or investigate somewhat simple experimental setups lacking in data variability with respect to the factors reported above. In this work, we attempt to address both these issues: On the algorithmic front, motivated by progress in speaker identification [7], we propose using a support vector machine (SVM) built on Gaussian mixture model (GMM) supervectors to distinguish falls from other competing noise. In our proposed approach, a universal background model learns the shared acoustic feature space for falls and other noise, and the supervectors extracted from the GMMs adapted using each audio segment serve as robust summary of the acoustic signal. On the experimental front, we report results on a relatively large database of human falls, collected as part of the Netcarity Integrated Project [1]. This set contains desirable variability of falls and other noise activity expected in realistic home environments [6].

The rest of the paper is organized as follows: Section 2 introduces the general framework for fall classification and detection. Section 3 discusses the GMM supervectors for audio segments. Section 4 derives the distance between the GMM supervectors and the corresponding GMM supervector kernel used in an SVM. Experiments on the Netcarity fall dataset are presented in Section 5, followed by a summary in Section 6.

2. PROBLEM OVERVIEW

We are interested in *detecting* falls acoustically, pinpointing their temporal occurrence and effectively distinguishing them from other possible confusable or background noise. For this purpose, we employ a dynamic programming algorithm that is based on standard GMMs of falls and of a number of noise types, resulting to a segmentation of the audio input. These segments can then be subsequently reclassified – as done in this work – by a more complex scheme. In a simpler version of the above problem, the segments are known a-priori, in which case the problem reduces to that of *classification* alone. The two problems are instances of the general acoustic event classification and detection paradigm [8, 9].

To better distinguish falls from competing noise, we choose to model falls and nine classes of noise in the home environment. These classes, depicted in Table 1, are the result of a labor-intensive annotation effort on the Netcarity fall database (see Section 5.1), taking

* Work performed during a Summer 2008 internship at IBM Research.
** Currently with the Institute of Informatics & Telecommunications (IIT), National Centre of Scientific Research "Demokritos", Athens, Greece.

Table 1. Sound classes for fall classification and detection.

FA	sound resulting from the subject falling
ST	noise when the subject sits down on the chair, possibly leading to a bit of chair movement
CL	noise of clapping hands
GU	noise when the subject gets up from the floor
MP	noise of moving, placing, or catching an object
DO	noise of dropping an object on the floor
DN	noise of opening/closing doors
WK	noise of walking steps
MO	other noise, including speech and non-speech human voices, telephone rings and other acoustically salient noise
BG	background noise, usually not perceptually salient

three considerations into account: Each noise class should appear a sufficient number of instants in the training data; should be relatively distinguishable from the others; and should help in the ultimate goal of discriminating falls from noise.

Central to this effort is the approach to audio segment modeling. Each audio segment is first represented by a sequence of feature vectors, extracted from evenly sampled and partially overlapping time-domain Hamming windows. In particular, in this work, 12 perceptual linear predictive (PLP) coefficients of the windowed signal are extracted (over 25 ms) and are augmented with the overall energy to give rise to 13-dimensional feature vectors. This process is repeated every 10 ms, and it is followed by "utterance-level" cepstral mean subtraction, applied for feature normalization.

Two approaches are then used to model the distribution of these feature vectors, as schematically depicted in Figs. 1 and 2. The first follows the traditional GMM paradigm that approximates the joint distribution of all feature vectors in *each event class* with a GMM. For a test audio segment, a maximum likelihood classifier is used to obtain the hypothesized event class. We propose to use a second approach to model audio segments, referred to as the SVM-GMM-supervector method, approximating the joint distribution of all feature vectors in *each audio segment* with a GMM, from which a GMM supervector is constructed as a summary of the segment. The pairwise Euclidean distances between these supervectors characterize the difference between the audio segments. Kernels derived from these distances are used in an SVM for classification.

3. GAUSSIAN MIXTURE MODELS AND GMM SUPERVECTORS

A GMM approximates the distribution of the observed features with a Gaussian mixture density function $g(z) = \sum_{k=1}^{K} w_k \mathcal{N}(z; \mu_k, \Sigma_k)$, where w_k, μ_k, and Σ_k denote the weight, mean, and covariance matrix of the k^{th} Gaussian component, and K is the total number of such components. Covariance matrices Σ_k are restricted to be diagonal for computational efficiency. Maximum likelihood parameters of a GMM can be obtained by using the well-known expectation-maximization (EM) algorithm.

3.1. UBM-MAP

Instead of separately estimating parameters for each GMM, we can also train GMMs by adapting from a "global" GMM, known as the universal background model (UBM). The potential merits of adapting GMMs from a UBM are two-fold: First, the parameters may be robustly estimated with a relatively small amount of training data. Second, there is correspondence between Gaussian components in different GMMs when these models are adapted from the same UBM.

More specifically, we obtain the GMMs by adapting the mean vectors of the global GMM using the maximum-a-posteriori (MAP) criterion. The mixture weights and covariance matrices are retained for simplicity and robustness of parameter estimation. MAP adaptation of GMMs can be implemented by applying the EM algorithm. In the E-step, we compute $Pr(k|z_i)$, namely the posterior probability of the unimodal Gaussian component k given observed feature vector z_i, as

$$Pr(k|z_i) = \frac{w_k \mathcal{N}(z_i; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} w_j \mathcal{N}(z_i; \mu_j, \Sigma_j)}, \quad \text{for } i = 1, ..., H, \qquad (1)$$

where $w_k, \mu_k, \Sigma_k, k \in \{1, \ldots, K\}$ are the UBM parameters, and H denotes the total number of observed feature vectors. This step uses the UBM to assign each feature vector to the unimodal Gaussian components probabilistically. This mechanism establishes correspondence between the components of adapted GMMs, because the component parameters, i.e., the means, are estimated from statistics obtained involving the same UBM. In the M-step, the mean of each Gaussian component is updated as

$$\hat{\mu}_k = \frac{N_k}{N_k + \tau} E_k(Z) + \frac{\tau}{N_k + \tau} \mu_k ,$$

where

$$E_k(Z) = \frac{1}{N_k} \sum_{i=1}^{H} Pr(k|z_i)\, z_i ,$$

N_k is the occupation likelihood of the observed data on the k^{th} Gaussian component ($N_k = \sum_{i=1}^{H} Pr(k|z_i)$), and τ represents the weight placed on the prior knowledge, i.e., the UBM means, compared to the observed data. In this work, τ is adjusted empirically according to the amount of available training data.

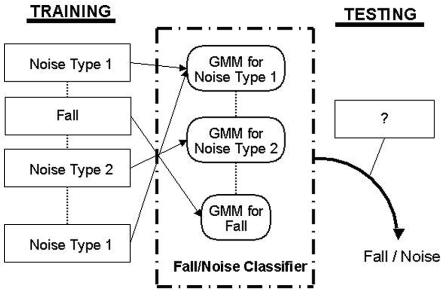

Fig. 1. The GMM approach to fall/noise modeling.

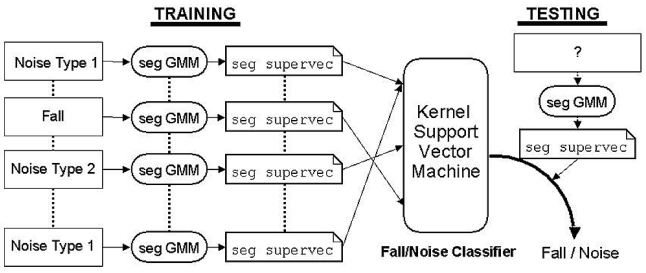

Fig. 2. SVM-GMM-supervector approach to fall/noise modeling.

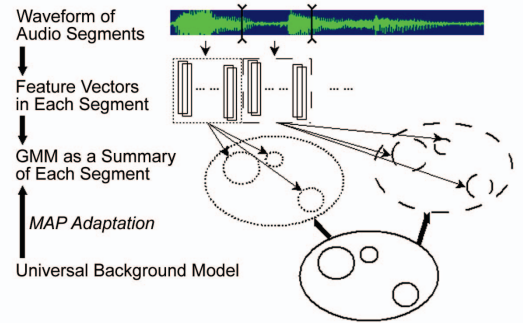

Fig. 3. GMMs (depicted as ovals) summarize audio segments using multiple unimodal Gaussians (illustrated as circles).

3.2. Summarizing Audio Segments

Feature vectors extracted from an audio segment may carry a lot of noise. We use a GMM adapted from the UBM to capture the inner structure of the ensemble of feature vectors in each audio segment, as shown in Fig. 3. According to (1), the feature vectors are assigned to different unimodal Gaussian components probabilistically based on the UBM. We concatenate the adapted means of all the unimodal Gaussian components as a vector in a high dimensional space defined by the UBM, each dimension roughly corresponding to one dimension in the mean vector of one particular Gaussian component in the UBM. This high-dimensional vector, called a GMM supervector, serves as a summary of the audio segment.

4. GMM SUPERVECTOR SPACE

4.1. Approximating Kullback-Leibler Divergence

As detailed in Section 3.2, we can summarize audio segments with supervectors constructed from GMMs adapted from the UBM. We denote two such *segment* GMMs as g_a and g_b. A natural similarity measure between these two GMMs is the Kullback-Leibler divergence,

$$D(g_a||g_b) = \int_z g_a(z) \, \log \frac{g_a(z)}{g_b(z)} \, dz \; .$$

The Kullback-Leibler divergence does not satisfy the conditions for a metric function, but there exists an upper bound using the log-sum inequality,

$$D(g_a||g_b) \leq \sum_{k=1}^{K} w_k \, D(\mathcal{N}(z;\mu_k^a,\Sigma_k)||\mathcal{N}(z;\mu_k^b,\Sigma_k)) \, ,$$

where μ_k^a and μ_k^b denote the adapted means of the k^{th} component from the segment GMMs g_a and g_b , respectively. Since the covariance matrices are shared across all adapted GMMs and the UBM, the right hand side is equal to

$$d(a,b)^2 = \frac{1}{2} \sum_{k=1}^{K} w_k (\mu_k^a - \mu_k^b)^T \Sigma_k^{-1} (\mu_k^a - \mu_k^b) \; .$$

We can consider $d(a,b)$ as the Euclidean distance between the normalized GMM supervectors in a high-dimensional feature space,

$$d(a,b) = \|\phi(Z_a) - \phi(Z_b)\|_2 \; , \tag{2}$$

where

$$\phi(a) = [\, \sqrt{\frac{w_1}{2}} \Sigma_1^{-\frac{1}{2}} \mu_1^a \, ; \cdots ; \sqrt{\frac{w_K}{2}} \Sigma_K^{-\frac{1}{2}} \mu_K^a \,] \; . \tag{3}$$

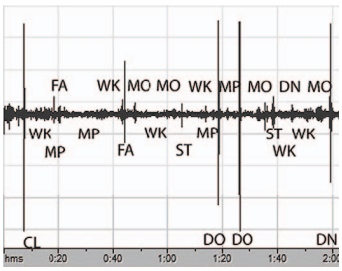

Fig. 4. Snapshot of the labeled Netcarity fall dataset (see also Table 1). Segment boundaries are omitted for simplicity.

4.2. Kernel for SVM

We use the GMM supervectors in an SVM for fall/noise classification. Since there are multiple types of noise, we tackle the problem as multi-class classification, implemented as binary classification problems via the one-vs-one method using LibSVM [10]. The distance defined in (2) can be evaluated using kernel functions, as

$$d(a,b) = \sqrt{K(a,a) - 2K(a,b) + K(b,b)} \; . \tag{4}$$

It is straightforward that kernel function $K(a,b) = \phi(a) \bullet \phi(b)$ satisfies (4), where $\phi(a)$ and $\phi(b)$ are defined as in (3).

5. EXPERIMENTS

5.1. The Netcarity Fall Database

Our experiments are carried out on the acoustic fall data collected as part of European Integrated Project Netcarity [1, 6]. The dataset is about 7 hours long, consisting of 32 sessions that involve 13 different actors as subjects that may fall or perform other activities, as well as additional subjects that produce noise in the background, simulating relatively well an environment that elderly people may encounter at home. Fig. 4 provides a snapshot of an acoustic signal from this database, manually annotated. Note that we map the labels in the Netcarity dataset to the ten classes detailed in Table 1 as the ground truth. For our experiments, we split the dataset into 20 training, 7 testing, and 5 held-out sessions, the latter for tuning system parameters. Note that the subjects in the training and held-out sessions do not overlap with the test subjects.

5.2. Experimental Results

Our first experiment aims at the *classification* of audio segments, whose boundaries are provided by the manual database annotation. Both the GMM baseline and the proposed SVM-GMM-supervector approach are employed for this purpose, implemented using 512 Gaussian components for each GMM. To compare their performance, we report results based on two metrics: Classification accuracy on all ten classes of Table 1, reflecting the overall performance of the classifiers, and the F-score of the fall segments, reflecting the capability to distinguish falls from all other noise. Results of this experiment are illustrated in Fig. 5. It is clear that both metrics improve significantly for the proposed approach over the GMM baseline. In particular, the F-score improves by about 12% relative.

The second experiment focuses on the *detection* of falls over entire database sessions. To measure performance we use the acoustic event detection accuracy metric (AED-ACC), defined in [9] as the harmonic mean between precision and recall. In its calculation, correctness is defined as the temporal center of either a hypothesized

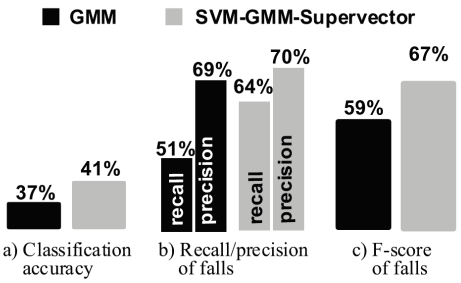

Fig. 5. Fall classification results on the Netcarity test dataset.

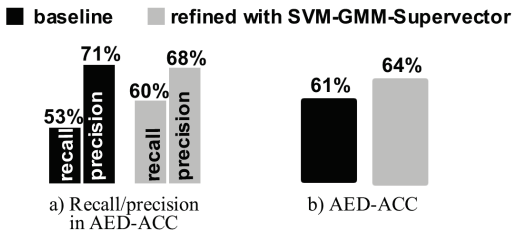

Fig. 6. Fall detection results on the Netcarity test dataset.

fall segment or a reference fall segment located within the span of the other [8, 9]. In our experiment, we further require that all proposed fall segments do not exceed a duration of 5 seconds, so that the system output can be used for timely response to falls. Any fall segments that exceed 5 seconds are removed from the output before scoring. As already mentioned earlier, to perform detection we employ a dynamic programming algorithm with GMM audio segment modeling. The output of this process can be further refined by applying the SVM-GMM-supervector approach to reclassify the resulting audio segments. Here, we limit this post-processing to segments with perceptually confusable labels. These are chosen to be falls (FA), dropping objects (DO), getting up (GU), and walking (WK). Results of this experiment are illustrated in Fig. 6. It is clear that the proposed reclassification approach using SVM-GMM supervectors improves performance, resulting in a relative 5% improvement in the AED-ACC metric.

6. SUMMARY

In this paper, we presented a classification and detection system of human falls based on input from a single far-field microphone. We proposed modeling each fall or noise segment using a GMM supervector and employing an SVM built on a GMM supervector kernel to classify audio segments into falls and various types of noise. We reported experiments on an appropriate dataset, containing falls performed by multiple subjects interspersed with other characteristic activities and noise expected in realistic home environments. Our experiments demonstrated that the proposed fall/noise modeling boosts classification performance, compared to a standard event class GMM classifier. The proposed approach also effectively improved fall detection accuracy, when applied as a post-processing stage to reclassify confusable labels at the output of dynamic programming using the GMM classifier.

Recent work in speaker verification applications has shown further improvement using new classifiers based on GMM supervectors, compared to approaches similar to the SVM-GMM-supervector method presented in this paper [11]. This suggests the possibility of further improvements in fall detection based on GMM supervectors.

7. ACKNOWLEDGEMENTS

This work was partially funded by the European Commission, as part of Integrated Project Netcarity. We would like to thank the authors of [6] from three Netcarity partner sites in Italy, at the University of Pavia, CNR-IMM, and FBK-irst, for the design and collection of the human fall dataset. In addition, IBM colleagues Vit Libal and Larry Sansone have assisted with database organization and annotation.

8. REFERENCES

[1] *Netcarity – Ambient Technology to Support Older People at Home*. [Online] `http://www.netcarity.org`

[2] S. Sadigh, A. Reimers, R. Andersson, and L. Laflamme, "Falls and fall-related injuries among the elderly: A survey of residential-care facilities in a Swedish municipality," *J. Community Health*, 29(2): 129–140, 2004.

[3] N. Noury, A. Fleury, P. Rumeau, A.K. Bourke, G.Ó. Laighin, V. Rialle, and J.E. Lundy, "Fall detection – principles and methods," In *Proc. Int. Conf. of the IEEE Engineering in Medicine and Biology Soc.*, Lyon, France, pp. 1663–1666, 2007.

[4] B.U. Töreyin, Y. Dedeoğlu, and A.E. Çetin, "HMM based falling person detection using both audio and video," In *Computer Vision in Human-Computer Interaction (HCI/ICCV 2005)*, N. Sebe, M.S. Lew, and T.S. Huang (Eds.), Springer-Verlag, LNCS 3766, pp. 211–220, 2005.

[5] A. Fleury, M. Vacher, H. Glasson, J.-F. Serignat, and N. Noury, "Data fusion in health smart home: Preliminary individual evaluation of two families of sensors," In *Proc. Int. Conf. of the Int. Soc. for Gerontechnology*, Pisa, Italy, 2008.

[6] M. Grassi, A. Lombardi, G. Rescio, P. Malcovati, A. Leone, G. Diraco, C. Distante, P. Siciliano, M. Malfatti, L. Gonzo, V. Libal, J. Huang, and G. Potamianos, "A hardware-software framework for high-reliability people fall detection," In *Proc. IEEE Conf. on Sensors*, Lecce, Italy, pp. 1328–1331, 2008.

[7] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process. Letters*, 13(5): 308–311, 2006.

[8] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "CLEAR evaluation of acoustic event detection and classification systems," In *Multimodal Technologies for Perception of Humans (CLEAR 2006)*, R. Stiefelhagen and J. Garofolo (Eds.), Springer-Verlag, LNCS 4122, pp. 311–322, 2007.

[9] R. Stiefelhagen, K. Bernardin, R. Bowers, R. Travis Rose, M. Michel, and J. Garofolo, "The CLEAR 2007 evaluation," In *Multimodal Technologies for Perception of Humans (CLEAR 2007 and RT 2007)*, R. Stiefelhagen, R. Bowers, and J. Fiscus (Eds.), Springer-Verlag, LNCS 4625, pp. 3–34, 2008.

[10] C.-C. Chang and C.-J. Lin, *LIBSVM: A Library for Support Vector Machines*. [Online] `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

[11] R. Dehak, N. Dehak, P. Kenny, and P. Dumouchel, "Linear and non linear kernel GMM supervector machines for speaker verification," In *Proc. Interspeech*, Antwerp, Belgium, pp. 302–305, 2007.