Speech-overlapped Acoustic Event Detection for Automotive Applications

Christian Müller[1], *Joan-Isaac Biel*[1], *Edward Kim*[2],*Daniel Rosario*[2]

[1]International Computer Science Institute, Berkeley, CA
[2]Electronics Research Lab, Volkswagen of America, Palo Alto, CA
{cmueller,biel}@icsi.berkeley.edu,{edward.kim,daniel.rosario}@vw.com

Abstract

We present two approaches on acoustic event detection for speech-enabled car applications: a generative GMM-UBM approach and a discriminative GMM-SVM supervector approach. The systems detect whether or not a certain acoustic event occurred while the built-in microphone of the car was active to record a spoken command, either before, while, or after the driver was speaking. These events can be *music playing*, *phone ringing*, a passenger different from the driver is *talking*, *laughing*, *or coughing*. The task is formally defined as a detection task along the lines of well established detection tasks such as speaker recognition or language recognition. Similarly, the evaluation procedure has been designed to resemble the respective official evaluation series performed by NIST (i.e. it was a blind 'one-shot' evaluation on a separately provided dataset). The performance of the system was calculated in terms of detection miss and false alarm probabilities ($C_{Miss} = C_{FA} = 1$, and $P_{Target} = 0.5$). The performance of the superior GMM-SVM system was 0.0345 for known test speakers and 0.1955 for novel test speakers. Frequency-filtered band energy coefficients (FFBE) outperformed MFCCS on that task. The results are promising and suggest further experiments on more data.

Index Terms: acoustic event detection, GMM-UBM, GMM-SVM supervector, Frequency-filtered band energy coefficients (FFBE), NIST-style evaluations.

1. Introduction

In speech recognition applications, the sounds corresponding to the words uttered are traditionally considered to be *signal* whereas everything else, might it be channel noise or the sound of a door shutting, is considered to be *background noise*. To not let the noise harm the accuracy of the recognition, it is usually either filtered out, or carefully provided in training in order let the recognizer learn how to deal with it.

However, the detection of *acoustic events* like a door being shut might very well be useful information for the downstream application. This is particularly the case with perceptually aware interfaces such as computer-based meeting-room assistants [1], medical tele-surveillance, or mobile robots working in diverse environments [2]. As a matter of fact, this problem is addressed by the field of auditory scene analysis (ASA) or (with a somewhat narrower scope) acoustic event detection (AED).

In this paper, we present a study on the AED for speech-enabled automotive applications. The following example illustrates the application scenario:

driver:	"Washington Mutual" [phone is ringing in the background]
without acoustic event detection	
system (recognition failed)	"Your command was not recognized. Please repeat."
with acoustic event detection	
system (recognition failed)	"Your command was not recognized because there was a phone ringing in the background. Please repeat."

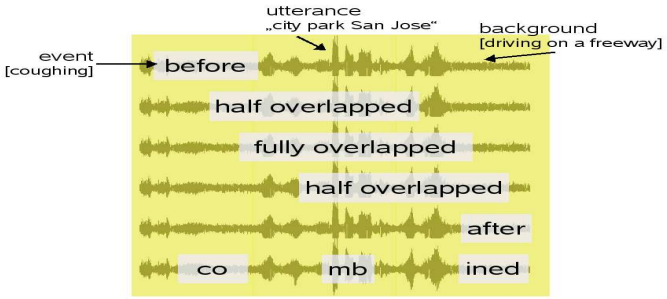

Figure 1: *Examples of occurrences of acoustic events to be detected in the given task.*

Particularly, our system is supposed to be able to detect the following acoustic events: music playing (MUSIC), phone ringing (PHONE), a passenger (p2) different from the driver/speaker is talking (TALKING), p2 is coughing (COUGHING), p2 is laughing (LAUGHING). In contrast to pure AED, however, the input does not only contain acoustic events but also the sounds of the words uttered. The events can occur at any time during the recording, either before, overlapped with, or after the utterance. Therefore, we refer to the task as *speech-overlapped* AED (SOAED). Also, the recording contains a varying amount of noise stemming from the engine, wheels, and wind resistance of the car in diversified driving conditions (see Figure 1).

Two issues should be pointed out in order to avoid confusion. First of all, the goal of the superordinate project is *also* to reduce word error rate in recognition (the above example might imply that this is not the case). Here, SOAED can for example help identifying appropriate acoustic models that facilitate a more robust recognition in that particular situation (see [3, 4]). However, in this paper we focus on the detect-and-explain scenario. Second, for many events that occur inside the car (such as music playing, phone ringing, and others that are not explicitly studied here) other sensors than speech exist that are likely to be more precise. However, it is clearly the goal of the sponsor to exploit those sensors that are actually available in today's line of production (such as the built-in microphone) as much as possible.

2. Task and evaluation procedure

The acoustic event detection task underlying this study has been defined along the lines of well established detection tasks such as speaker recognition [5] or language recognition [6]. Similarly, the evaluation procedure has been designed to resemble the respective official evaluation series performed by the National Institute of Standards and Technology (NIST)[1].

Accordingly, the task was the following [2]: Given a speech segment (s) and an acoustic event to be detected (*target event*, E_T), the task is to decide whether E_T was present in s (yes or no), based on an automated analysis of the data contains in s. The system performance was evaluated by presenting it with a set of trials. Each test segment was used for multiple trials, with one trial for each of following events: MUSIC, PHONE, TALKING, COUGHING, LAUGHING. The absence of all of these events (NO_EVENT) was explicitly included as a target event as well. Besides the decision as to whether the event of interest was actually present in the s, the output of the system contained a score indicating the system's confidence in its decision. More positive scores indicated greater confidence that the segment contains the target event.

The sponsor (the Volkswagen of America Electronics Research Lab, ERL) provided development and evaluation data as representative as possible for the application. Three months before the evaluation, the research site (International Computer Science Institute, ICSI) was provided with the development data only. At a pre-determined date, the blind evaluation data was provided to the research site for processing. Five days later, the system's output was submitted to the sponsor in the format NIST used in the latest language recognition evaluation. The sponsor then downloaded the scoring software from NIST's website, made the necessary modifications due to the changes in the labels (e.g. names of acoustic events instead of languages), and ran it on the submitted system output. The results were then disclosed to the research site along with the keys (truth) for further analysis.

The performance of the system was calculated in terms of detection miss and false alarm probabilities. Miss probability was computed separately for each target event and false alarm probability was computed separately for each target/non-target event pair. In addition, these probabilities were combined into a single number representing the cost performance of a system, according to the following cost model:

$C(E_T, E_N) = C_{Miss} * P_{Target} * P_{Miss}(E_T) + C_{FA} * (1 - P_{Target}) * P_{FA}(E_T, E_N)$

where E_T and E_N are the target and non-target events, and C_Miss, C_{FA}, and P_{Target} are application model parameters. Here, the application parameters are $C_{Miss} = C_{FA} = 1$, and $P_{Target} = 0.5$[3]

3. Data

The data consisted of single commands uttered while driving. For security reasons, the speaker was *not* driving but sitting on the passenger seat. Hence, the second passenger (p2) and third passenger (p3) were both sitting on the backseat. There were two driving conditions: CITY (between 25 and 40 mph) and HIGHWAY (between 35 and 70 mph). All data was collected utilizing a single type of car (Audi Q7). The recordings were made at 16KHz sing the built-in microphone located in the center console.

The total number of speakers in the database was 40. However, they were involved in the recording of different subsets of the corpus resulting in a number of 16 speakers per event. We are aware of the fact that this is a flaw in the procedure that should be fixed for future studies. In order to judge the speaker (in)dependency of the system, only data from 10 different speakers per target was selected for development, while the evaluation set contained all 16 speakers. Thus, it could be separated into a MATCHED SET with only speakers seen in training and a MISMATCHED SET with only novel speakers. Table 1 summarizes the amount of data available in the various datasets for development as well as evaluation (the information about the latter was not disclosed to the research site prior to the evaluation).

Table 1: *Amount of data available for development and evaluation. (*) not disclosed to the research site before evaluation.*

event	# of segments / ~hours					
	train		devtest		eval(*)	
music	1160	3	220	0.5	590	1.5
phone	1050	3.5	165	0.5	600	1.5
coughing	784	4	120	0.5	360	1.5
laughing	810	3	140	0.5	390	1.5
talking	790	2.5	150	0.5	420	1.5
no event	1400	4.5	150	0.5	560	2
background	4000	12	–	–	–	–

4. Approaches

Two different AED systems have been developed at the research site: A generative Gaussian Mixture Model – Universal Background Model (GMM-UBM) system and a discriminative Gaussian Mixture Model – Support Vector Machine Supervector (GMM-SVM) system.

4.1. GMM-UBM system

Generative classifiers such as GMMs and Hidden Markov Models (HMMs) have been widely used for acoustic event and classification tasks [1, 7]. HMMs have the advantage of capturing the temporal information of the sequence of speech frames. However, HMMs require a a larger amount of data to accurately train the models.

A GMM consists of a likelihood function based weighted linear combination of M Gaussian densities, each parameterized by a mean vector and a covariance matrix. For each target event, a separate GMM is trained. In testing, a likelihood ratio is computed as:

$\Lambda = \log p(X|\lambda_{hyp}) - \log p(X|\lambda_{\overline{hyp}})$

where $p(X|\lambda_{hyp})$ is the likelihood of the hypothesized event, and $p(X|\lambda_{\overline{hyp}})$ is the likelihood of the possible alternatives modeled by a universal background model (UBM). The UBM is trained on a pool of examples of all the possible events.

Due to its simplicity and effectiveness, the GMM-UBM approach has become one of the mainstay modeling techniques for *text-independent* speaker recognition. The reader is referred to [8] for a detailed description. The number of Mixtures in our system was 256. This value has been experimentally explored using the development test set.

[1] `http://www.nist.gov/speech/tests/sre/`, `http://www.nist.gov/speech/tests/lre/`

[2] `http://www.nist.gov/speech/tests/lre/2007/LRE07EvalPlan-v8b.pdf`

[3] Since no application oriented empirical values where available at evaluation time, we chose this rather general cost model. However, it is likely that false alarms will be given a higher cost value in practise.

4.2. GMM-SVM supervector system

Recently, a superior detection accuracy was reported using discriminative methods such as SVMs [1, 2]. However, this method is sensitive to variable segment lengths since input vectors for kernel evaluation are restricted to be of constant size. The most common approach to cope with that issue is to compute the mean and standard deviations of the feature trajectories, and to use these statistics as input features for the SVMs [1]. In other studies, this problem is solved by extracting features from a segment of a fixed length [9]. Again others performed a frame-by-frame SVM classification followed by a combination of the output scores [10]. Recently, the use of sequential kernels has shown significant improvements due to their ability of preserve the temporal information contained in the data [11].

The GMM-SVM approach combines the discriminative power of SVMs with the ability of GMMs to deal with variable length sequences. It was proposed in [12] as an alternative to the Fisher kernel, and was later on applied to speaker verification [13, 14]. The former uses a linear kernel derived from Kullback-Leiber distance:

$$K_{lin}(s_a, s_b) = \sum_{i=1}^{M} (\sqrt{w_i}\Sigma_i^{-\frac{1}{2}}\mu_i^a)(\sqrt{w_i}\Sigma_i^{-\frac{1}{2}}\mu_i^b)^t$$

where s_k is a GMM supervector obtained by pooling together all the Gaussian means μ_i^k of a means-only MAP-adapted GMM for the sequence k. Σ_i and w_i are the original weight and covariance of each Gaussian on the UBM model used for adaptation.

In our system, we trained a one-against-all SVM classifier for each class. Also the GMM supervectors were shifted by subtracting the original means of the UBM. This data centering method was proposed by [15]. It improved the results of the system on the development test set.

The optimal number of Gaussians was found to be 128 (0.022 avg cost). For the sake of comparison with other SVM-based systems proposed for AED, we as well created a SVM-only system, using the feature means computed on the whole audio segments. The results obtained on the development test set were exactly the same as the one obtained for the GMM-SVM using a single Gaussian (0.087 avg. cost). However, with a number of mixtures larger than one, the GMM-SVM approach is clearly superior.

5. Feature selection

We considered two different acoustic features as a possible representations for the audio signals: Mel-frequency cepstral coefficients (MFCC) and frequency-filtered band energy coefficients (FFBE). Whereas with MFCC, the logarithm of the filtered-bank energy (log FBE) are transformed using a discrete cosine transform in order to obtain a cepstral sequence, the decorrelation of the coefficients for the FFBE is achieved with a simple frequency filtering of the log FBE. It is computationally less expensive than the cepstral representation. Also, in [1] FFBE outperformed MFCC in the classification of acoustic events using both GMMs and SVMs classifiers.

For every audio segment, features were extracted using frames of 30ms with 10ms overlap using a Hamming window. For the MFCCs, 20 coefficients were extracted from 26 bands including the zero-th coefficient (Energy, E). For the FFBEs we used 20 coefficients obtained by filtering 20 bands with the usual second second-order filter $H(z) = z - z^{-1}$, which implies subtraction of the log FBEs of the two adjacent bands. Since FFBE already include energy information, no extra frame energy was added to these coefficients. In addition, first and second time derivatives (D, DD) for both types of features were computed.

Table 2: *Average costs obtained on the development test set for various sets of features.*

# set	type of features (size)	GMM-UBM	GMM-SVM
1	E+MFCC (20)	0.0652	0.0312
2	E+MFCC+D+DD (60)	0.0517	
3	**FFBE (20)**	**0.0548**	**0.0222**
4	FFBE+D+DD (60)	0.0457	–

Table 2 shows the average costs for different feature sets obtained on the development dataset. For both systems, FFBE features outperformed MFCCs. Using the derivatives improved the results with the GMM-UBM system. However, we decided that this improvement was not large enough to justify the introduction of the extra complexity. All results presented hereafter were obtained using feature set three.

6. Evaluation results

Table 3: *Average costs obtained on the evaluation set(s).*

	entire eval set	matched subset	mismatched subset
GMM-UBM	0.1458	0.0809	0.2341
GMM-SVM	0.1022	0.0345	0.1955

Table 3 shows the average costs obtained on the evaluation set. With 0.1458 for the GMM-UBM and 0.1022 for the GMM-SVM, the performances are significantly worse than the ones obtained on the development test set. However, as can be seen from the separation between the matched and mismatched cases (see section 3), this degradation is largely due to the presence of unknown speakers. We will get back to this point in the conclusion. Note, that a small degradation of results between development test and evaluation is to be expected, as numerous system parameters have been fine-tuned on the former set (including the decision threshold). This explains the difference of 0.0261 (GMM-UBM) and 0.0123 (GMM-SVM) between the results in Table 2 and Table 3 (matched subset).

Figure 2 (top) shows the decision-error-tradeoff curves of the GMM-SVM system for each target event obtained on the matched evaluation set. Evidently, COUGHING and LAUGHING stick out as the target classes for which the system exhibits the worst performance. As shown in Table 4, this is largely due to a high inter-confusion between those events. Generally, the events produced by the voices of other passengers (COUGHING, LAUGHING, TALKING) are rather likely to be confused with each other. This pattern is also to be seen in the DET curves obtained on the unmatched evaluation set (Figure 2, bottom) with the exception of the target event TALKING. We believe that the relatively good performance here is due to the presence of known speakers in the role of p2 or p3 (the second or third passenger talking with each other). However, we can not verify this hypothesis since this information is not available in the data.

7. Conclusions

To our knowledge, the task of speech overlapped acoustic event detection (SOAED) for car applications is novel. When planning this study, there has been a clear agreement between the

Table 4: *Confusion cost matrix between target classes (columns) and non-target classes (rows)*

	cough.	laugh.	talk.	music	no event	phone
cough.	–	0.1180	0.0456	0.0131	0.0227	0.0078
laugh.	0.1035	–	0.0535	0.0210	0.0091	0.0139
talk.	0.0408	0.0663	–	0.0107	0.0304	0.0078
music	0.0328	0.0509	0.0499	–	0.0309	0.0112
no event	0.0198	0.0443	0.0442	0.0117	–	0.0092
phone	0.0212	0.0466	0.0431	0.0204	0.0340	–
Avg Cost	0.0436	0.0652	0.0472	0.0154	0.0254	0.0100

sponsor (Volkswagen of America Electronics Research Lab, ERL) and the research site (International Computer Science Institute, ICSI) that, in order to be able to establish a first baseline, the variability in the data has to be restricted. Therefore, all data has been collected using only one type of car, for example. Since only 16 speakers have recorded in each of the conditions (from which only ten were used for training to leave speakers unseen for evaluation), it was especially questionable, whether the resulting models would be speaker independent. Nevertheless, with average costs of 0.2341 for the GMM-UBM respectively 0.1955 for the GMM-SVM, we obtained promising results even on the unmatched evaluation set that contained only novel speakers.

The results suggest that we need more speakers in order to make the systems speaker independent. We believe that with a comprehensive data collection from all forty speakers already in the pool, we could achieve cost values of below 0.1 in the unmatched condition. Besides that, the results of our experiments indicate, that GMM-SVM supervector approach is superior to traditional generative GMM-UBM approach for the task at hand. Also, Frequency-filtered band energy coefficients (FFBE) features outperformed MFCCs.

Acknowledgments The authors would like thank Oriol Vinyals for providing a GMM tool for train and test, and David Van-Leeuwen for the fruitful discussions about supervector kernels and feature normalization.

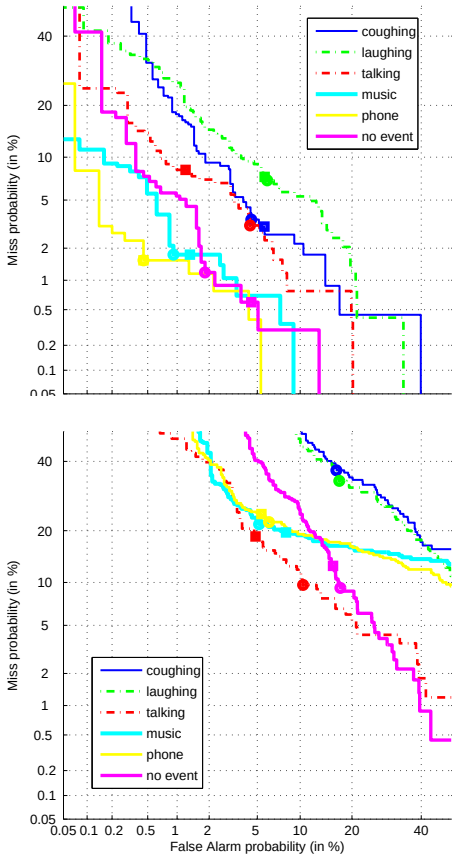

Figure 2: *Decision-Error Tradeoff (DET) curves of the GMM-SVM system for the matched (top) and unmatched (bottom) evaluation datasets. The box and circle symbols represent the actual and the optimal decision thresholds (minimum decision cost function), respectively.*

8. References

[1] A. Temko and C. Nadeu, "Classification of Acoustic Events using SVM-based Clustering Schemes," *Pattern Recognition*, vol. 39, pp. 682–694, 2006.

[2] S. Chu, S. Narayanan, C.C. Jay Kuo, and Maja J. Mataric, "Where am I? Scene Recognition for mobile robots using audio features," in *ICME 2006*, 2006.

[3] P. Ding, L. He, X. Yan, R. Zhao, and J. Hao, "Robust technologies towards automatic speech recognition in car environments," in *ICSP 2006*, 2006.

[4] A. Miguel, L. Buera, E. Lleida, A. Ortega, and O. Saz, "On-line feature and acoustic model space compensation for robust speech recognition in car environtment," in *2007 IEEE Intelligent Vehicles Symposium*, Istanbul, Turkey, June 2007.

[5] Alvin Martin, "Evaluations of Automatic Speaker Classification Systems," in *Speaker Classification*, Christian Müller, Ed., vol. 4343 of *Lecture Notes in AI*. Springer, Heidelberg, 2007.

[6] A. F. Martin and A. N. Le, "NIST 2007 Language Recognition Evaluation," in *Proceeedings of the Odyssey 2008 Workshop on Speaker and Language Recognition*, Stellenbosch, South Africa, 2008.

[7] Zhou, X. and Zhuang, X. and Liu, M. and Tang, H. and Hasegawa-Johnson, M. and Huang, T., "HMM-based Acoustic Event Detection with AdaBoost Feature Selection," in *CLEAR 07*, 2007.

[8] D. E. Sturim, W. M. Campbell, and D. A. Reynolds, "Classification Methods for Speaker Recognition," in *Speaker Classification*, Christian Müller, Ed., vol. 4343 of *Lecture Notes in AI*. Springer, Heidelberg, 2007.

[9] W. Huang, S. Lau, T. Tan, L. Li, and L. Wyse, "Audio Events Classification Using Hierarchical Structure," in *ICSP 2003*, December 2003, vol. 3.

[10] J-C. Wang, J-F. Wang, C-B. Lin, K-T. Jian, and W-H. Kuok, "Content-Based Audio Classification using support vector machines and independent component analysis," in *ICPR 06*, 2006.

[11] A. Temko, E. Monte, and C. Nadeu, "Comparison of Sequence discriminant support vector machines for acoustic event classification," in *ICASSP 06*, 2006.

[12] P.J. Moreno, P.P. Ho, and N. Vasconcelos, "A Generative Model Based Kernel for SVM Classication in Multime- dia Applications," in *NIPS*, 2003.

[13] W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff, "SVM Based Speaker Verication using a GMM Supervector Kernel and NAP Variability Compensation," in *IEEE ICASSP*, Toulouse, France, 2006, vol. 1, pp. 97–100.

[14] N. Dehak and G. Chollet, "Support Vector GMMs for Speaker Verication," in *IEEE Odyssey*, San Juan, Puerto Rico, 2006.

[15] Marina Meilă, "Data Centering in Feature Space," in *9th International Workshop on Artificial Intelligence and Statistics*, January 2003.