LONG-TIME SPAN ACOUSTIC ACTIVITY ANALYSIS FROM FAR-FIELD SENSORS IN SMART HOMES

Jing Huang, Xiaodan Zhuang, Vit Libal, Gerasimos Potamianos***

IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, U.S.A.

Emails: jghg@us.ibm.com; xzhuang2@uiuc.edu; libalvit@us.ibm.com; gpotam@ieee.org

ABSTRACT

Smart homes for the aging population have recently started attracting the attention of the research community. One of the problems of interest is this of monitoring the activities of daily living (ADLs) of the elderly, in order to help identify critical problems, aiming to improve their protection and general well-being. In this paper, we report on our initial attempts to recognize such activities, based on input from networks of far-field microphones distributed inside the home. We propose two approaches to the problem: The first models the entire activity, which typically covers long time spans, with a single statistical model, for example a hidden Markov model (HMM), a Gaussian mixture model (GMM), or GMM supervectors in conjunction with support vector machines (SVMs). The second is a two-step approach: It first performs acoustic event detection (AED) to locate distinctive events, characteristic of the ADLs, and it is subsequently followed by a post-processing stage that employs activity-specific language models (LMs) to classify the output sequences of detected events into ADLs. Experiments are reported on a corpus containing a small number of acted ADLs, collected as part of the Netcarity Integrated Project inside a two-room smart home. Our results show that SVM GMM supervector modeling improves six-class ADL classification accuracy to 76%, compared to 56% achieved by the GMMs, while also outperforming HMMs by 8% absolute. Preliminary results from LM scoring of acoustic event sequences are comparable to those from GMMs on a three-class ADL classification task.

Index Terms— Acoustic scene analysis, acoustic event detection, smart homes, activities of daily living

1. INTRODUCTION

Smart homes equipped with multiple audio and visual sensors have recently started to attract the attention of the research community. One of the scenarios of interest is this of ambient assisted living for the elderly, where smart homes are envisaged to help them improve their well-being, independence, safety, and health [1]. Automatic audio and visual analysis of the home environment based on camera and microphone input can help identify critical health situations, e.g., a person falling, and monitor the activities of daily living (ADLs) of the elderly, in order to issue reminders, identify deviations from the daily routine, and keep distant loved ones in the loop. In this work, we study the feasibility of achieving these goals by using only unobtrusive far-field microphones in the smart home. We are particularly interested in investigating whether characteristic ADLs can be accurately classified by acoustic scene analysis, based on such far-field audio sensors.

Most of the recent literature work on acoustic scene analysis focuses on the problems of acoustic event classification (AEC) and acoustic event detection (AED) of short-time span events [2, 3]. Much of these efforts in the context of smart spaces has been carried out as part of the CHIL project [4]. Various approaches have been adopted for these tasks, for example support vector machines (SVMs), left-to-right or fully-connected hidden Markov models (HMMs) for statistic modeling, and a number of acoustic front ends, for example general speech/audio features or features derived by data-driven approaches [2, 3].

In contrast to the above efforts, in this paper we are interested in applying acoustic scene analysis to classify longer-time span activities taking place in the home environment. In particular, we focus on a small number of ADLs, for example making coffee, watching TV, cleaning, ironing, reading, or eating/drinking. To further simplify the problem somewhat, we assume that the ADL temporal boundaries are a-priori known. Nevertheless, the problem remains extremely challenging for a variety of reasons: First, most ADLs are of long durations (at least a couple of minutes) and of complex structure. Among the same type of ADLs, there exists large variation of acoustic events happening within, such as different cleaning noise of vacuuming the floor or wiping the table for example. In addition, some activities, such as reading and ironing, lack distinctive acoustic footprints. In addition, there is large variation of background noise, produced for example by another person on the phone or a TV set turned on. Activities may also overlap / be interspersed with each other, such as cleaning while answering the phone. Finally, far-field sensors typically provide data of very low signal-to-noise ratio (SNR)

* X. Zhuang is with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. Work has been performed during his Summer 2008 internship at the IBM Thomas J. Watson Research Center.

** G. Potamianos is currently with the Institute of Informatics & Telecommunications (IIT), National Centre of Scientifi c Research "Demokritos", Athens, Greece.

that are further degraded by additional background noise or activity overlaps.

In this paper, we propose two main modeling approaches to the ADL classification problem that differ significantly in their philosophy. The first considers ADLs in their entirety, employing Gaussian mixture models, hidden Markov models, or support vector machines (SVMs) built on GMM supervectors, to model them holistically, thus avoiding to explicitly characterize individual events contained within ADLs. In contrast, the second approach constitutes a hierarchical, two-stage process that first performs acoustic event detection to segment the long-time span activity into a sequence of distinctive acoustic events, on which, subsequently, activity-specific language models are employed to classify the overall activity. The latter approach is motivated by the fact that most activities contain characteristic individual events, for example that of kettle whistling in a coffee making activity. Of course, detecting such events in time is also a very difficult problem, as it also involves their temporal localization [3, 5]. We apply both approaches to a corpus of ADLs, collected by the FBK Research Center in Trento, Italy, as part of the Netcarity Integrated Project [1]. The database contains a small number of simple ADLs, recorded in a two-room smart home by a number of actors.

The rest of the paper is organized as follows: Section 2 describes the holistic approach and specific statistical models used, whereas Section 3 presents the hierarchical approach to ADL classification. Section 4 describes the Netcarity ADL data corpus, followed by experiments and results. The paper finally concludes with a summary in Section 5.

2. HOLISTIC ACTIVITY CLASSIFICATION

To model the audio signal of whole activities, the long-time span audio is represented as a sequence of feature vectors, extracted from the 25 ms Hamming windows with a 15 ms overlap with each other. In this work, we employ 13-dimensional perceptual linear prediction (PLP) coefficients as features, with "utterance-level" cepstral mean subtraction applied for their normalization. We then adopt two general statistical methods to model the distribution of these feature vectors.

The first method leverages on a continuous-density HMM with left-to-right topology for each activity. A special case of this constitutes the GMM, basically an HMM with only one emitting state. All activity-specific HMMs are employed within the framework of maximum likelihood classification.

The second technique, referred to as the SVM-GMM-supervector method, approximates the joint distribution of the feature vectors in each long-time span audio segment with a GMM adapted from a universal background model using the maximum-a-posteriori (MAP) approach. A high-dimensional GMM supervector is constructed from the normalized means of the adapted GMM. Kernels constructed on these supervectors are used in a support vector machine (SVM) for activity classification. Details of this method are given in [5, 6].

3. HIERARCHICAL ACTIVITY CLASSIFICATION

Activities of daily life may be distinguishable from each other in two ways. First, each activity may contain distinctive characteristic events, such as kettle whistling during a coffee making activity, or phone rings in the phone answering activity. Second, the temporal structure of event sequences might reveal different activities. For example, it is more likely to detect a water pouring event during an eating activity, than during phone answering. Similarly, it is more plausible that kettle whistling occurs after water pouring during the coffee making activity than while eating.

Motivated by the above facts, we first train an HMM-based acoustic event detector. The acoustic events are a-priori chosen to reflect characteristics of the activities, as motivated by the above discussion. A background model is also included for better detection performance. Each acoustic event (including background) is modeled by a ten-state HMM with left-to-right topology. An activity-independent bigram language model is trained using all training acoustic event sequences, and it is used in Viterbi decoding for acoustic event detection.

Then, at a second stage, and for each activity, an activity-dependent language model is built using training acoustic event sequences of each particular activity. These language models are used to rescore the acoustic event detection output, resulting in one score for each activity for the long-span audio signal. The one with the highest score is chosen as the hypothesized activity.

4. EXPERIMENTS AND RESULTS

4.1. Data Resources

A corpus containing activities of daily life (ADLs) has been collected as part of the Netcarity project by partner site FBK at Trento, Italy. An apartment has been used for this purpose, with two of its rooms (living room and kitchen) specially equipped with video cameras and microphones (see Fig. 1). Three T-shaped omni-directional microphone arrays were placed in each of the two rooms, with four microphones per array, giving a total of 24 audio channels. Each channel provides 16 bit/sample audio data at a 48 kHz sampling rate. In addition, three cameras were used to capture the video signal. In this work, the latter have only been used to give a visual reference supporting the audio data annotation.

The corpus has been organized into multiple sessions, each about 1.2 hours long in duration. Each session contains one main subject / actor and an additional one that creates "interference". The main subjects perform activities within a prescribed set of twelve, while the interfering subjects act randomly and perform other activities in order to yield a complex but realistic acoustic environment. Twenty such sessions have been used in this work, of which 16 have been employed for training and four for testing purposes.

Six basic activities are chosen for this work for ADL classification to experiment with various statistical models

(a) Living room

(b) Kitchen

Fig. 1. Typical images of the two apartment rooms used for the collection of the ADL corpus by Netcarity partner site FBK.

(GMMs, HMMs, and SVM-GMM-supervector modeling), in conjunction with the holistic approach. These six classes, believed to be acoustically distinguishable, are: eating-drinking, reading, ironing, cleaning, phone answering, and TV watching. Notice that most of the activities are recorded with the TV set on – making the problem even more challenging. Furthermore, in the corpus, so-called "parallel" activities exist, such as phoning & cleaning, eating & TV watching. We merge the former into phone answering and the latter into the eating-drinking class.

These six activities yield 576 training samples in the database, totaling 13.5 hours in duration; among them, there exist 128 instances for each of eating/drinking (EAT), reading (RDG) and TV watching (TVW), and 64 instances for each of ironing (IRN), phone answering (PHN), and cleaning (CLN). Each activity ranges from approximately one to five minutes in duration. The test set contains 144 activity samples, totaling 3.5 hours; among these, there are 32 instances for each of EAT, RDG and TVW, and 16 instances for each of IRN, PHN and CLN.

To experiment with the hierarchical approach, we limit ourselves to only three activities, namely phone answering,

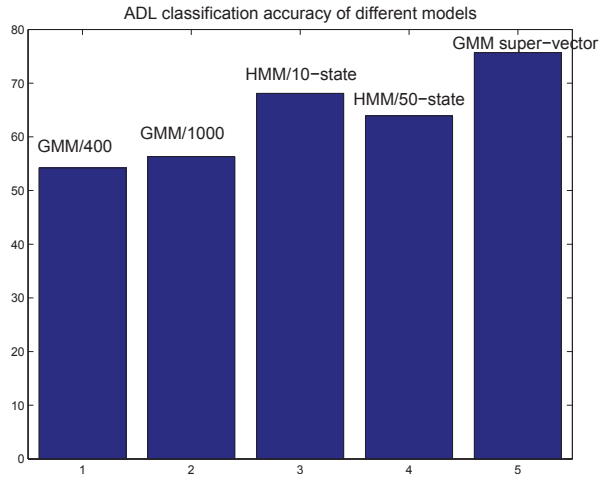

Fig. 2. Accuracy on the six-ADL classifi cation task on the Netcarity test set employing fi ve statistical models under the holistic approach.

eating/drinking, and coffee making (COF). This is due to limited time available to manually annotate ADLs with low-level acoustic event information. Such acoustic events were chosen to be a "wrap rustling" noise during eating, "laughing", "greeting", "speech", "phone ringing", "ending of a phone conversation", "kettle whistling", "hesitation", "boiling", and "water pouring". Non-labeled ADL segments were annotated as "background".

The three activities considered in conjunction with the hierarchical approach occur 335 times in training, totaling 6 hours in duration. Among them are 192 instances of EAT, 128 of PHN and 15 of COF. Furthermore, there exist 84 testing activities, totaling 2.5 hours, among which are 48 of EAT, 32 of PHN, and 4 of COF. Concerning lower-level acoustic events, there exist 3605 and 802 non-background events annotated within the training and testing activities, respectively.

4.2. Experimental Results

Results on the six-activity ADL classification task on the Netcarity ADL test dataset are depicted in Fig. 2. There, performance of a number of statistical models in conjunction with the holistic approach is given, measured by the overall classification accuracy. GMMs with 256 Gaussians per activity class, GMMs with 1000 Gaussians per class, 10-state HMMs for each class with 256 Gaussians per state, 50-state HMMs for each ADL with 256 Gaussians per state, and the SVM-GMM supervector approach with 256 Gaussians.

The results in Fig. 2 demonstrate that increasing the number of Gaussians in GMMs from 256 per class to 1000 per class does not help much, improving performance somewhat, from 54.2% to 56.3%. On the other hand, HMMs outperform GMMs significantly, with the 10-state HMM achieving an accuracy of 68.1% and the 50-state HMM 63.9%. The SVM-GMM-supervector approach turns out to be the best, reaching a 75.7% classification accuracy, which represents more than 30% relative improvement over the GMMs.

	COF	EAT	PHN
COF	3	1	0
EAT	9	35	5
PHN	4	1	27

Table 1. Confusion matrix of GMM-based classification of three ADL classes.

To investigate performance of the hierarchical approach, we first consider the acoustic event classification (AEC) and detection (AED) problems [2, 3]. On the AEC task, the classification accuracy is 57.5% on eleven acoustic event labels (lower level events, characteristic of the three ADLs of interest). The AED performance, measured by the AED-ACC metric [3], is 34.4%. These results are comparable to the ones reported in recent international evaluations of far-field AEC/AED technology in smart spaces [3].

We next compare classification accuracy on three ADLs using the hierarchical approach (LM rescoring of AED outputs) vs. the use of two holistic statistical models (GMMs and 10-state HMMs). Using a unigram LM to rescore AED outputs results in 70.2% classification accuracy on three ADLs, which is similar to the 76.2% achieved by the GMM, but significantly worse than the 10-state HMM performance of 88.1%. Interestingly, using a bigram LM for rescoring actually hurts performance, reducing accuracy from 70.2% to 34.5%. The reason for this degradation is likely due to the fact that the training acoustic events are dominated by the "background" label, coupled with the fact that too few data for reliable bigram estimation exist, a problem especially acute for the coffee making (COF) ADL.

Tables 1, 2, and 3 depict the confusion matrices of GMMs, HMMs (holistic approaches), and AED with unigram scoring (hierarchical method). Obviously, the HMM improves upon the GMM mostly on the EAT class, and AED with unigram scoring performs the worst on the COF class, due to the lack of sufficient training data.

5. SUMMARY

In this paper we studied the problem of long-time span acoustic activity classification from far-field sensors in smart homes, aiming at analysis of activities of daily living (ADLs). We proposed two approaches to the problem: The first one is holistic, modeling the entire activity with GMMs, HMMs, or the recently introduced approach of GMM supervectors. Among those, the latter performed the best, reaching a 76% classification accuracy on six ADL classes on the Netcarity database. The second approach to the ADL classification problem is of a hierarchical nature, first detecting lower-level characteristic events of ADLs via an acoustic event detection stage, the output of which is rescored by activity-specific language models of these events. Preliminary results of this approach, applied on the problem of classifying three ADLs, demonstrate that this approach is at par with the GMM holistic method, but lags behind better statistical models, such as HMMs. Improving both AED stage and language models would most certainly result in improvements to the hierarchical approach. Both are areas of interest for future work. Furthermore, we would like to reduce the labor-intensive stage of manual annotation of events, necessary in this approach. This can be replaced by a partially unsupervised method based on iterative sound clustering and segmentation.

	COF	EAT	PHN
COF	2	2	0
EAT	2	45	1
PHN	3	2	27

Table 2. Confusion matrix of HMM-based classification.

	COF	EAT	PHN
COF	0	0	4
EAT	4	36	8
PHN	1	8	23

Table 3. Confusion matrix of the hierarchical ADL classification approach on three ADLs, when employing unigram rescoring.

6. ACKNOWLEDGEMENTS

This work has been partially supported by the European Commission as part of Integrated Project Netcarity. The authors wish to thank the following colleagues at Netcarity partner site FBK, in Trento, Italy, for the design and collection of the corpus: Fabio Pianesi, Massimo Zancanaro, Paul Chippendale, Nadia Mana, Alessandro Cappelletti, Stefano Messelodi, and Francesco Tobia. In addition, IBM colleague Larry Sansone has annotated acoustic events in part of the corpus and Stanley Chen has provided insightful algorithmic suggestions.

7. REFERENCES

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

[2] 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.

[3] 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.

[4] *CHIL – Computers in the Human Interaction Loop*. [Online] `http://chil.server.de`

[5] X. Zhuang, J. Huang, G. Potamianos, and M. Hasegawa-Johnson, "Acoustic fall detection using Gaussian mixture models and GMM supervectors," In *Proc. Int. Conf. Acoustics Speech Signal Process.*, Taipei, Taiwan, 2009.

[6] 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.