Activity Recognition Using a Spectral Entropy Signature

Jessica Beltrán-Márquez
CICESE, MEXICO
jbeltran@cicese.mx

ABSTRACT

Context identification is one of the key challenges in Ubicomp. An application example is providing contextual information to caregivers of person with dementia to identify assistance needs. Environmental audio provides significant and representative information of the context and the challenge is to automatically identify audio cues coming from overlapping sound sources without sophisticated microphone arrangements. My thesis proposes a succinct representation of the audio, based on the spectral entropy of the signal, and we show experimentally its robustness to source overlap and noise. This would permit ubiquitous applications that perform sound-based activity identification directly in mobile phones.

Author Keywords

Activity Recognition, Context Awareness, Auditory Scene Analysis, Ambient Assisted Living.

ACM Classification Keywords

H.4.m Information Systems Applications: Miscellaneous.

General Terms

Algorithms, Design, Experimentation, Performance, Theory.

PROBLEM STATEMENT

Caring for a Person with Dementia (PwD) is a demanding task often performed by a close relative facing stress and burnout providing assistance 24/7. Ubiquitous computing applications can be designed to assist PwD and their caregivers. Providing caregivers with automatic information of the activities performed by the PwD would improve the accuracy of pharmacological interventions, track illness and disease progression, and also opportunistically identify potential risks or unusual activities of the PwD[16].

While microphones and video cameras are the most ubiquitous sensors[5], microphones have the advantage of capturing information in all directions and are robust to change in position and orientation, thus making data collection less intrusive. Audio also provides significant and highly representative information of the context and its capture require only a microphone, which is already available in any mobile phone [11].

Sound-based ubiquitous activity identification is possible because several activities produce characteristic sounds which facilitate their identification using audio analysis. For the case of elders, audio can be used to detect activities which may be interesting for their caregivers, like showering, walking on the street or preparing aliments among others. Audio also permits to detect if the users present disease symptoms like coughing or sneezing. Another example is identifying if a person is having a conversation in order to know her socialization habits.

As a downside, audio cannot be used with activities that do not produce characteristic sounds, and also noise is always present in real situations thus making audio difficult to analyze. In addition, sound may have multiple interpretations and need disambiguation. Fortunately, audio can be complemented with other sensors to design robust ubiquitous activity detection systems.

The literature for activity detection using audio, attempt to classify a small set of categories in relatively limited and predefined contexts of interest. Also, the profile of the users have not yet been used to improve the accuracy of the methods.

The objective of my thesis is a method to identify sounds unprocessed as a way to contribute with sound-based activity detection research. We aim at tackling the problem in mobile environments under realistic circumstances, hence we use monophonic audio as the one available in standard cellular phones. We also aim at identifying activities executed by different people. Additionally we have the goal of encountering a compact audio representation that permits indexing directly on the phone. This would help address privacy concerns because it would make possible completing activity identification in mobile phones without needing to send audio information to servers.

The rest of the paper is organized as follows, the next section shows an overview of previous work. Then we discuss the methodology proposed. We also discuss the research products obtained so far and finally, we outline the conclusions and the work that remains to complete the research.

UbiComp '12, Sep 5-Sep 8, 2012, Pittsburgh, USA.

PREVIOUS WORK

Audio analysis for activity identification usually follows the next process. Signal acquisition through microphones with a proper selection of a sampling frequency and bit resolution. Afterwards comes a preprocessing step including basic operations like filtering or smoothing. Then follows an audio feature extraction step with the purpose of having a concise representation of the auditory signals. Finally, a classifier is used to obtain the activity categories of the input audio. A large number of audio features have been used in audio related applications, being the Mel Frequency Cepstrum Coefficients (MFCC) the most popular feature. The procedure to calculate the MFCC is described in [13]. In the classification front we find a great diversity also, the most used methods are Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM).

We briefly discuss some examples of studies aiming at inferring activities using auditory information. A study uses features in the frequency domain and GMM classifying among the following activities: brush, wash, shave, electric brushing, electric shaving and other activities [9]. Another study uses two microphones to identify the activities: hammering, sawing, drilling, grinding and filing, to achieve this, they use temporal and frequency domain features and a K-NN classifier [14]. A recent approach uses MFCC plus time and frequency features with Bayes and HMM classifiers to identify the activities: walking, driving cars, riding elevators and riding bus [8]. This latter approach results interesting for us because they propose and experiment with an application for mobile phones. In [4] a vector of 20 time and frequency features are used to train a HMM to identify the semantic scenes: Restaurant, street, lecture, conversation and other, the identification of scenes provide hints of the activity been performed.

Audio analysis can also be used to identify minor health distress of a person. For example, there is a study with the propose of detecting coughs, this is achieved generating eigenvectors by a Principal Component Analysis (PCA) on the audio spectrogram and using a Random Forest (RF) classifier [6]. In another approach the Short Time Fourier Transform with Support Vector Machine are used to detect an elder falling, the sensors used are a microphone located on the foot and also use accelerometers [2]. In other example, daily audio patterns like the amount of human speech are analyzed in order to identify social and mental well being, in this work are used features to detect the structure of voiced speech and a HMM [12].

A recent paper addresses [3] the identification of mixture overlapped signals problem using an unsupervised non-negative matrix factorization (NMF) to produce separated tracks. Then MFCC and HMM are used on each separated track. Another work aims to detect acoustic events in meeting rooms, they experimented with artificially overlapped acoustic data using log filter-bank energies, along with the first and the second time derivatives with other common temporal and frequency features [15].

We find less work aiming at compact audio representation. There are Information Retrieval (IR) techniques that can be used to achieve this, for example the model of Bag of Words (BoW) as is used in [7] to represent audio for an application of content-based video copy detection .

In this investigation we propose analyzing audio with a modification of the Multi Band Spectral Entropy Signature (MBSES). The MBSES is a signature formed with the entropy of the signals in 24 Barks sub-bands which has been used in music information retrieval proving robustness to noise, equalization and loudness [1].

METHODOLOGY

The methodology proposed is divided in three main phases:

Understanding the relation between auditory information and the activities than can be estimated. In this stage we analyze which sounds are relevant to estimate activities.

Design and experimentation. This is a cyclic stage that consist in designing and experimenting. The algorithm comprises sound feature extraction and classification techniques. With the analysis of the auditory information we will use the most salient features to represent the activities. We will use a database to train and evaluate the robustness of our proposed algorithms. We plan to have three iterations and the expected results are: An initial approach, results improvement and idiosyncrasy consideration, and finally we will compare the results of audio analysis with those obtained with the addition of other contextual information such as location or movement.

In situ experiments with elders and caregivers. We will perform preliminary experiments using a personal computer, when we have satisfactory classification results we are going to implement the system in a smartphone in order to conduct in situ experiments with PwD patients and caregivers currently participating in a study conducted by our lab which requires them to use a smartphone.

RESEARCH CARRIED SO FAR

We have used a binary signature formed with the entropy of the signals in 24 Barks sub-bands which we preliminary validate with an experiment described below. We use a database of synthetic mixed sounds to prove the robustness in noisy conditions. We then engineered another experiment where we generated a vocabulary of representative sequences to represent audio using a model like the bag-of words from classic information retrieval. We then used standard classifiers from the software WEKA [10] from the Universy of Waikato to evaluate our proposed representation. WEKA is a collection of machine learning algorithms for data mining tasks implemented in java.

Feature extraction

To calculate the Multiband Spectral Entropy Signature (MBSES) we followed the procedure described in [1]. The MBSES calculation produce an entropygram which gives the amount of information along the time for every critical band

in the Bark scale. A procedure to obtain a procedure to obtain the binary MBSES to eliminate a peculiar shift is also described in [1]. In the binary MBSES, every frame of the sound is a string of 24 bit symbols. The size of the string depends on the duration of the audio. This also allows to use Hamming distance to compare signatures.

As a contribution of our research, we modified the the MBSES in order to obtain a signature more robust to noise. We computed the MBSES as usual, and then compute a discrete cosine transform of the 24 entropy values in each frame. We call this signature the CMBSES and also obtained a corresponding binary version.

First approach

We conducted an experiment using four different features in order to evaluate the performance of our proposed signature with mixed sounds. To achieve this we collected nine audio segments representing various sound sources from the collaborative database http://www.freesound.org and with 44100 Hz of sampling frequency and 16 bits depth (baby crying, keys, siren sound, bird singing, tooth brushing, music with voice, music without voice, male voice, and female voice).

All sounds were cut to have duration of three seconds. We created a database by mixing the nine original sounds. First we formed the dataset "mixture A"; this was obtained by mixing all the combinations of pairs of sounds with a 0dB Signal to Noise Ratio (SNR). After the previous procedure, we made four mixtures between the combination of two of the nine original sounds and all the elements from "mixture A". We avoided repetitions of sounds in a single mixture. These mixtures were obtained with four different SNR values (3.4dB, 5dB, 10dB and 20dB) where the nine original sounds were taken as the signal and the elements of mixture A as the noise. With this procedure, four data sets of 252 elements were obtained for each SNR value. In each data set every original sound contributes as a signal or as a noise in 84 mixtures.

Experiments and results

The goal of the experiment was identifying the occurrences of a given sound in the database, for this reason, the nine original sounds were compared with each SNR dataset by using a corresponding distance. The features used were: MFCC, MBSES, binary MBSES and binary CMBSES. The corresponding distance to compare two sounds are the Euclidean distance for MFCC, a shifted euclidean distance for the MBSES and the hamming distance for the binary MBSES and binary CMBSES. To establish a ground truth we appeal 48 subjects between 21-30 years old with headphones to hear the 21 mixtures of sounds from a web page that we created for the experiment (can be checked in http://sound.natix.org).

Table 1 shows the recall obtained with each feature. Due to space restriction we only show results with the 3.4dB mixtures. We searched every sound in all the mixtures to obtain the results. It is also important to mention that humans heard several times every mixture. On average they heard 2.52 times the 3.4dB mixtures.

	3.4 Db Mix				
	(a)	(b)	(c)	(d)	(e)
(i)	32.14	8.33	96.43	**100**	96.42
(ii)	51.19	64.29	96.43	**100**	96.42
(iii)	53.57	96.43	**97.62**	85.71	95.23
(iv)	2.38	78.57	76.19	82.14	**90.47**
(v)	**100**	**100**	**100**	**100**	97.61
(vi)	35.71	88.10	86.90	**100**	92.85
(vii)	**100**	98.81	98.81	**100**	94.04
(viii)	40.48	23.81	**100**	**100**	97.61
(ix)	95.24	58.33	**100**	**100**	97.61
Avg.	56.75	68.52	94.71	**96.43**	95.37

Table 1. The results of our experiment. Rows are different sounds as Baby Crying (i), Bird singing (ii), Keys (iii), Siren sound (iv), Tooth brushing (v), Music with voice (vi), Instrumental music (vii), Male voice (viii), Female voice (ix). We tested the following features (all with the same classifier) MFCC (a), MBSES (b), Binary MBSES (c), and Binary CMBSES (d) . The (e) column is the average for the 48 volunteers. Both tables correspond to the different signal to noise ratio in the mixture.

This experiment provides evidence of the efficacy of this approach because it is apparent that our signature performs better than MFCC features and matches the performances of the volunteers. As we mentioned, in the database we used synthetic mixed sounds with a static duration of three seconds. To improve our sound representation we made additional processing which is discussed in the next section.

Improvement to audio representation

We use a vocabulary to represent audio similarly to the 'Bag of Words' model from information retrieval area. We represent each sound with a fixed-dimensional vector, where each coordinate represents the frequency of the vocabulary elements in the sound. The vocabulary is formed with *base-sequences* described below.

As mentioned in the feature extraction section, the binary signatures MBSES or CMBSES produce a string for every frame. We define a *sequence* as the strings of m adjacent frames, where m is the size of the sequence. For example, sequences of size one are formed with one frame, sequences of size two are formed with two adjacent frames. In general, a sequence of size m is formed with m adjacent frames. A *base-sequence* is a sequence that represents several similar sequences. There are different methods to select representative sequences to create a vocabulary, for example clustering algorithms like K-means or based on neural networks.

A sound is represented as a fixed-dimensional vector, where each coordinate represents the frequency of base-sequences encountered in the sound. If we generate the vocabulary selecting k_m base-sequences for every m size, the sound vector size is $\sum_{m=1}^{M} k_m$ where M is the maximum sequence size. We experimented using CMBSES binary signature with 95% of overlapp using sequences of size from one to fifteen frames.

To test our vector representation, we conducted an experiment using a database with 112 sounds examples. We created the database by recording with a smartphone seven classes

of audio (bouncing ball, toot brushing, cricket, washing hands, crying baby, keys, typing). The recording was performed with a 44100 sampling frequency and 16 bits of depth having duration from three to five seconds. The sounds were recorded from four different subjects and each subject contributed with four examples. It is important to mention that sounds do not have the same duration, and all classes have examples from four subjects that generated the sounds in their own way. We used the sequential minimal optimization algorithm (SMO) from WEKA for training a support vector machine classifier with the 112 sound vectors with 10 fold cross validation for testing the classifier.

CONCLUSION AND FUTURE WORK

Our results have a 99% overall precision, where 6 of 7 classes have precision of 100%. As mentioned earlier, our proposed signature is robust to noisy environments which make us believe that we can successfully use the signature to identify sound events generated from different persons in real situations. Additionally, the final sound representation is compact and permit a fast search in a database.

As future work we will experiment with more classifiers and test the indexing capabilities of the vector representations. We still need to test more techniques to find base-sequences in a database. We need to evaluate the classifiers in real time environmental audio streams and we also need to test the performance of our proposed sound classification technique implemented in a mobile phone.

Our contribution to activity estimation represent an important step because we are concerned with real aspects of the auditory analysis problem like the noisy mixed condition and in finding a compact representation of the sound.

ACKNOWLEDGEMENTS

This work is partially supported by a grant from CONACYT trough a scholarship provided to the author.

REFERENCES

1. A. Camarena-Ibarrola, E. Chávez, and E. S. Tellez. Robust radio broadcast monitoring using a multi-band spectral entropy signature. In *CIARP 2009*, LNCS, pages 587–594, Berlin, Heidelberg, 2009. Springer-Verlag.

2. C. Doukas and I. Maglogiannis. Advanced Patient or Elder Fall Detection based on Movement and Sound Data. *The Proceedings of the Second ICST International Conference on Pervasive Computing Technologies for Healthcare 2008*, pages 6–10, 2008.

3. T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen. Sound event detection in multi-source environments using source separation. pages 36–40. Workshop on machine listening in Multisource Environments, 2011.

4. N. Kern, B. Schiele, and A. Schmidt. Recognizing context for annotating a live life recording. *Personal Ubiquitous Comput.*, 11:251–263, April 2007.

5. N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell. A survey of mobile phone sensing. *Comm. Mag.*, 48:140–150, September 2010.

6. E. C. Larson, T. Lee, S. Liu, M. Rosenfeld, and S. N. Patel. Accurate and privacy preserving cough sensing using a low-cost microphone. In *Proceedings of the 13th international conference on Ubiquitous computing*, UbiComp '11, pages 375–384, New York, NY, USA, 2011. ACM.

7. Y. Liu, W.-L. Zhao, C.-W. Ngo, C.-S. Xu, and H.-Q. Lu. Coherent bag-of audio words model for efficient large-scale video copy detection. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '10, pages 89–96, New York, NY, USA, 2010. ACM.

8. H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell. Soundsense: scalable sound sensing for people-centric applications on mobile phones. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*, MobiSys '09, pages 165–178, New York, NY, USA, 2009. ACM.

9. C.-h. Min, N. F. Ince, and A. H. Tewfik. *Early Morning Activity Detection Using Acoustics and Wearable Wireless Sensors*. Number Eusipco. 2008.

10. T. U. of Waikato. Weka. http://www.cs.waikato.ac.nz/ml/weka/. Version 3.6.

11. I. Potamitis and T. Ganchev. Generalized recognition of sound events: Approaches and applications. pages 41–79. 2008.

12. M. Rabbi, S. Ali, T. Choudhury, and E. Berke. Passive and in-situ assessment of mental and physical well-being using mobile sensors. In *Proceedings of the 13th international conference on Ubiquitous computing*, UbiComp '11, pages 385–394, New York, NY, USA, 2011. ACM.

13. T. L.-s. S. Sigurdsson, K. B. Petersen. Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music. In *ISMIR*, pages 286–289, 2006.

14. M. Stager, P. Lukowicz, and G. Troster. Implementation and evaluation of a low-power sound-based user activity recognition system. In *Proceedings of the Eighth International Symposium on Wearable Computers*, ISWC '04, pages 138–141, Washington, DC, USA, 2004. IEEE Computer Society.

15. A. Temko and C. Nadeu. Acoustic event detection in meeting-room environments. *Pattern Recognition Letters*, 30(14):1281 – 1288, 2009.

16. D. H. Wilson. *Assistive Intelligent Environments for Automatic Helath Monitoring*. PhD thesis, Carnegie Mellon University, 2005.