Auditory Context Recognition Using SVMs

Mikko Perttunen[1], Max Van Kleek[2], Ora Lassila[3], Jukka Riekki[1]

[1]*Department of Electrical and Information Engineering, 90014 University of Oulu, Finland*
{first.last}@ee.oulu.fi

[2]*MIT CSAIL 32 Vassar St. Cambridge, MA, 02139, USA*
emax@csail.mit.edu

[3]*Nokia Research Center Cambridge Cambridge, MA 02142, USA*
ora.lassila@nokia.com

Abstract

We study auditory context recognition for context-aware mobile computing systems. Auditory contexts are recordings of a mixture of sounds, or ambient audio, from mobile users' everyday environments. For training a classifier, a set of recordings from different environments are segmented and labeled. The segments are windowed into overlapping frames for feature extraction. While previous work in auditory context recognition has often treated the problem as a sequence classification task and used HMM-based classifiers to recognize a sequence of consecutive MFCCs of frames, we compute averaged Mel-spectrum over the segments and train a SVM-based classifier. Our scheme outperforms an already reported HMM-based scheme. This result is achieved using the same dataset. We also show that often the feature sets used by previous work are affected by attenuation, limiting their applicability in practice. Furthermore, we study the impact of segment duration on recognition accuracy.

1. Introduction

Context-aware systems adapt to the context of users, where context comprises of information related to the current situation of the user [1]. Commonly used sensing systems include indoor and outdoor positioning systems, accelerometers, and video analysis. Perhaps surprisingly, a less-studied source is the auditory environment of our daily activities.

Auditory scenes consisting of a mixture of sounds from everyday objects is a natural source of context information for context-aware computing. Most humans can quite naturally listen to audio of a scene, and deduce certain characteristics about the setting – people in the scene, whether it is outdoors or indoors, the types of other objects in the scene and their relative positions [2]. However, definitively classifying multiple locations exclusively from audio taken from those locations is difficult even for humans; Eronen et al [2] demonstrated that humans required on average 14 seconds of audio, and achieved only 69% accuracy when given a 28-scene identification task. Nonetheless, the identification of scenes and locations by computers could have significant use in context aware computing, since it requires no centralized infrastructure, and no additional hardware besides microphones, which are already pervasively available in most portable devices. Peltonen et al. call this automated classification of auditory contexts, computational auditory scene recognition (CASR) [3].

Following the earliest work in CASR by Sawhney [4], many researchers have reported experiments using variations in classifiers, feature sets and datasets. For example, Ma et al. classified 12 auditory contexts using a hidden Markov model (HMM) based classifier [5]. They achieved 96% accuracy by using 9-state left-to-right HMMs, with one Gaussian mixture component per state, using MFCCs features and their first and second-order deltas, with a log energy term. Similarly, Eronen et al. developed a HMM-based classifier for 28 auditory contexts, using a different dataset that they collected [6]. To select the most suitable features for the task, they tried 11 different feature sets using a Gaussian mixture model and 1-nearest neighbor classifiers. They also studied the effect of audio segment length on recognition performance, showing a steady increase in recognition accuracy until 20s, and a plateau of 72% accuracy at 60s. In [7], Lu et al. apply support vector machines (SVMs) for classifying among four classes: non-pure speech, pure speech, background sound, demonstrating 80% to 96% accuracy from 0.1 to 1 second duration audio segments, respectively.

Like our approach presented in this paper, they derive a single set of MFCC based features (means and variances) for each segment, instead of treating the problem as sequence classification task using HMMs.

Unfortunately, with all of these different experiments reporting varying degrees of success, run independently using different data sets, classifiers and features, it is difficult to compare and definitively identify the best set of methods to use, or to say with any confidence how likely the results are to generalize to new scenes and audio capture devices. We feel that the field of CASR needs more work in achieving consistent, comparable results using common methodology, which can then be used to more easily interpret outcomes.

In this vein, this paper contributes a re-examination of the dataset captured by Ma et al, comparing their HMM-based approach against an SVM-approach proposed by Lu et al. In addition, we consider the effect of feature choice and audio segment length on performance, evaluating various combinations of features and lengths as recommended by Eronen. Our results demonstrate improved accuracy on audio context recognition tasks over previously reported approaches using HMMs, through the use of SVMs and averaged Mel scaled log amplitudes of the spectrum (hereafter refer to as the averaged Mel spectrum). To make our results comparable with those of Ma et al [5], we reproduce their scene classification experiments as described in their paper and employ their datasets[1] in all of our experiments. However, we discovered several problems with this dataset which lead to unexpected performance results, which we discuss in Section 4.

The rest of this paper is organized as follows: In section 2 the dataset and feature extraction procedure are described. Section 3 goes through our experiments and results. Section 4 is dedicated to further examination, connecting results to our procedure and dataset.

2. Data and methods

2.1. Dataset

To be able to directly compare our results to the state of the art, we use the dataset from Ma et al. [5]. The dataset consists of recordings from 12 different auditory contexts recorded using a mobile device (8 kHz, 8-bit, mono). The dataset is summarized in Table 1 [5]. Because Ma et al. used one 5 minute recording of each environment for training and one for testing, we follow the same setup. We call the set of recordings used for training 'dataset1' and the set of recordings used for testing 'dataset2'. The database contains an additional set of 5 min recordings, but unfortunately it lacks the recording from one of the environments (building site). However, we use this set, called 'dataset3', with the remaining 11 classes. Thus the dataset as a whole contains 175min of audio.

Table 1. Recordings from 12 different environments (from [5])

Number	Routine	Environment
1	Walk to bus stop	Street (traffic)
2	Take bus to office	Bus
3	Pass a building site	Building site
4	Work in office	Office
5	Listen to a presentation	Presentation
6	Urban driving	Car (city)
7	Shopping in mall	Shopping mall
8	Walk in city	Street (people)
9	Shopping in supermarket	Supermarket
10	Laundrette	Laundrette
11	Driving (long distance)	Car (highway)
12	Local or express train	Train

As mentioned in section 1, Ma et al. reported 96% accuracy when HMM-based classifier was trained using dataset1 and tested on dataset2. Nevertheless, they also reported that when using dataset1 for training, and testing using half of dataset3, the accuracy fell to 75%. This showed that dataset1 is considerably more similar to dataset2 than to dataset3.

2.2. Feature extraction

For each segment, the 8kHz source audio signal is framed without pre-emphasis into 3 second non-overlapping segments; each segment is further windowed (using a Hamming window) into 512-sample frames with a 384-sample overlap between them. From each frame, a 40-element Mel spectrum is computed and used to derive 12-element MFCCs for that frame. In addition to this baseline feature set, several additional features were computed: the overall log energy of the segment, the zero crossing rate, spectral centroid, and spectral flux [8]. Each of these features is averaged across all the frames in each segment to yield the features for each segment. The extraction procedure for the averaged MFCCs is described e.g. in [10]. For this set of experiments, we used Roger Jang's audio toolbox to extract features [9]. Additionally, we chose not to use the standard deviation of the averages because in early experiments we noticed it did not

[1] http://fizz.cmp.uea.ac.uk/Research/noise_db

improve results significantly, and because we wanted to keep the size of the feature vector similar to that of [5]; the feature sets are summarized in Table 2.

2.3. Classifiers

WEKA 3.5.5 [11] was used for training and testing the SVM-based classifiers. All of WEKA's default settings for SVMs were used except for kernel parameters, which were hand-tuned. To perform multiclass classification using their binary SVM-classifier, we employed a "one-against-one" (i.e., pairwise) voting scheme, because it seemed to perform the best for this problem [12].

3. Results

In this section we describe the performance of our classifier on Ma et al.'s 12-scene auditory context classification task. We evaluate 8 different choices for features, and analyze the impact of sliding segmentation, segment duration and temporal smoothing on recognition accuracy.

3.1. Feature set

The recognition accuracies achieved with different feature sets are compared in this section. In preliminary tests, we tuned the kernel degree for polynomial kernel, the gamma-parameter for the RBF kernel, and the regularization parameters for each. This was done by training on dataset1 and testing with dataset2. The difference in accuracy between the best-performing polynomial kernel (of degrees in the range 1 to 15) and the best RBF kernel was less than 1 percent. Therefore we chose to stick with the polynomial kernel for all tests described here. In all tests, a segment duration of 3 seconds was used. The regularization parameter was increased from 1 up to 100000 with decade-steps for all kernel degrees and gammas. In the following, results are reported for the best-performing parameters.

To compute the accuracy we trained the classifier using the particular choice of features on dataset1. Then, this trained classifier was run on all of the test examples for each class in datasets 2 and 3 in turn. Since the examples (in both test set and training set) of a class come from a single continuous audio recording, they cannot be considered entirely statistically independent, however this is how Ma et al. evaluated their classifiers and thus we chose to repeat the same procedure in our evaluation.

Table 2 summarizes the results. The best overall accuracy across test datasets, 92.8%, was achieved using feature set 3, consisting of the 40-element Mel spectrum. Interestingly, this is slightly better than that of feature set 4, which and adds the MFCCs derived from the Mel spectrum. Hence, adding features can have detrimental effects on performance.

When using dataset3 for testing, the classifier was trained using twelve classes, but tested presenting examples from eleven classes, because dataset3 lacks the building site recording (see section 2.1). The best accuracy, 87.1%, for dataset3 was achieved using feature set 3. Ma et al. reported 75% accuracy for their HMM-based classifier [5] for this setup. Some of the 12% difference may amount to the test setup just described, although we assume Ma et al. have used a similar test setup for dataset3.

3.2. Sliding window segmentation

Limited audio data for training in CASR can cause key, short-lived acoustic events which may be useful for identifying a scene but happen rarely in the signal to be underrepresented. For example, the sound of a door closing might be key to identifying an office scene, or a bus's brakes to identifying a bus scene. One proposed approach to combat this scarcity suggested by [6], is to try to re-use some of these acoustic events across multiple training examples by overlapping segments in the training set.

We therefore conducted experiments studying the effect of changing the segmentation to a sliding-window approach on classifier performance. A fixed inter-segment hop length of 0.1s (800 samples) was chosen and held constant across experiments; thus the number of examples from every class increased from 100 to 2970. Otherwise the feature extraction was performed as described in section 2.2. In both tests, polynomial kernel SVMs were trained using the best performing kernel degree and regularization parameters in the same manner as our previous experiment.

When testing this classifier on the same data (no overlapping, 100 examples per class) as in section 3.1, the accuracy rises to 88.3%. Thus, the more effective use of training data due to the segment overlapping improves accuracy only 1.2% at 3s (over the 87.1% shown in Table 2)

Table 2. Summary of classification performance using different combinations of features. The highest performing feature combinations for each dataset are highlighted in bold

Feature set	Features	#of features	Dataset2	Dataset3	Dataset2 -30dB	Average
Feature set 1	12-element MFCCs	12	87.0%	79.2%	87.0%	84.4%
Feature set 2	12-element MFCCs and log energy term	13	93.1%	83.0%	27.6%	67.9%
Feature set 3	40-element Mel spectrum	40	*95.7%*	***87.1%***	***95.7%***	**92.8%**
Feature set 4	12-element MFCCs and 40-element Mel spectrum	52	95.2%	84.3%	95.2%	91.6%
Feature set 5	12-element MFCCs, log energy term, and 40-element Mel spectrum	53	**96.5%**	85.8%	47.1%	76.5%
Feature set 6	40-element Mel spectrum and zero crossing rate	41	95.5%	87.0%	95.5%	92.7%
Feature set 7	40-element Mel spectrum and spectral centroid	41	96.1%	86.1%	96.1%	92.8%
Feature set 8	40-element Mel spectrum and spectral flux	41	95.5%	83.4%	95.5%	91.5%

3.3 Segment duration

Next, we examined the effect of varying segment duration on classifier performance. We maintained sliding window segmentation described in the last section, with a hop length of 0.1s.

First, the green curve (with triangles) in Figure 1 shows the results using dataset1 for training and dataset2 for testing, varying segment duration from 0.1s to 10s. The recognition rate increases from 0.1s until 3s and then plateaus. A possible cause is our feature extraction and classification scheme; averaging may cause the difficulty of discriminating between two classes to vary with segment duration. Second, the classifier trained on dataset1 was evaluated on dataset3. The results are shown in Figure 1 as the blue curve (with diamonds). We examine these results in section 4.2.

3.4. Temporal smoothing

In an effort to increase classifier robustness, we evaluated an approach which combined multiple predictions for individual 0.1s segments (computed as described earlier) into a single prediction through majority voting. The purple (crosses) and the red (squares) curves in Figure 1 depict the performance achieved with this approach, when predictions are smoothed using windows containing 3-100 segment predictions, corresponding to overall audio durations of 0.3s-10s. The test examples are presented to the classifier one class at a time, so the correct label changes 11 times (10 times with dataset3). Figure 1 shows that temporal smoothing improves results for only the short segments. This is likely caused by the "filtering effect" over the noisy predictions of individual classifications of the 0.1s segments.

3.5 Audio volume

In practical daily use, audio captured from device microphones might be attenuated randomly for various physical reasons. For example, the difference between a mobile phone being placed on a surface in open air versus in a person's pocket or purse could easily cause a 30dB or greater attenuation. Since performance degradation from such activities could impact the use of CASR in practice, we examined robustness of each our classifiers to signal attenuation. We did this simply by applying a 30dB attenuation to the signal captured in dataset2. Note the major differences in attenuated and unattenuated performance for some feature sets and no difference in others. As (perhaps) could be expected, the feature sets that contain a log energy term (2 and 5) do not perform well. This limits the applicability of such features in practice. For example, Ma et al. [5] include a log energy feature in their feature set, but did not study robustness with respect to variable signal power.

4. Analysis

In this section, we revisit our results, comparing them to those of Ma et al. [5], and investigate sources of performance degradations that we observed.

4.1. Feature sets

In Table 2 the accuracies of the feature sets from 3 to 8 are comparable to the reported 96% of Ma et al. for dataset2. They used 12 MFCCs and a log energy term with their first and second-order deltas, resulting in a 39-element feature vector, and classified the 3s segment using 9-state left-to-right HMMs. Our feature set 3 is of comparable length.

Next, we revisit the results of section 3.1 for feature set 3. The accuracies of individual auditory contexts

are shown in Table 3 (polynomial kernel degree 2, one-against-one scheme, 95.7% accuracy). Comparing these accuracies to the ones presented by Ma et al. in [5], a significant difference is that in [5] *bus* is recognized at 81% accuracy, whereas here it is 100%. As opposed to that, here *supermarket* is recognized with accuracy 83.8%, but Ma et al. report 100% for it. Finally, the accuracy for street (traffic) is 100% with our scheme, but 93% with the HMM-based scheme of Ma et al. For both schemes, *launderette* and *shopping mall* are among the most difficult to recognize. Either the different feature sets or the different classification schemes may cause these dissimilarities.

Table 3. Accuracies of auditory contexts for 3s segment duration and feature set 3, using dataset1 for training and dataset2 for testing

Auditory context	Acc. (%)	Auditory context	Acc. (%)
Street (traffic)	100.0	Shopping mall	77.8
Bus	100.0	Street (people)	97.0
Building site	100.0	Supermarket	83.8
Office	100.0	Launderette	92.0
Presentation	99.0	Car (highway)	99.0
Car (city)	100.0	Train	100.0

4.2 Segment duration

Kernel degrees 1 and 2 provided the best results for most segment durations. With a few exceptions, accuracy fell for all segment durations as a function of kernel degree, for degrees larger than 2, indicating over-fitting of higher degree kernels.

To analyze the unexpected drop-off in overall performance with segment lengths longer than 4s for dataset3 depicted in Figure 1, we studied it a bit further. Figure 2 decomposes the aggregate performance to accuracies for classifying each of the individual classes. It can be seen that the peaks of the *supermarket* (squares, light blue) and *bus* (smooth line, red) largely cause the peak between 3s and 4s in Figure 1. Examination of the confusion matrices revealed that the peak of *supermarket* is due to its confusions with *street (people)*, *launderette*, and *car* having a minimum at 4s. Similarly, the drop-off in performance with *bus* after 4s was due to a significant increase in confusion with *presentation*. As mentioned above, some of these variances in accuracies may be related to our averaging-based feature extraction scheme. However, we believe that a number of these issues were caused by characteristics of the original recordings, as we discuss in the next section. Further study with different datasets is needed to rule out any of these possibilities.

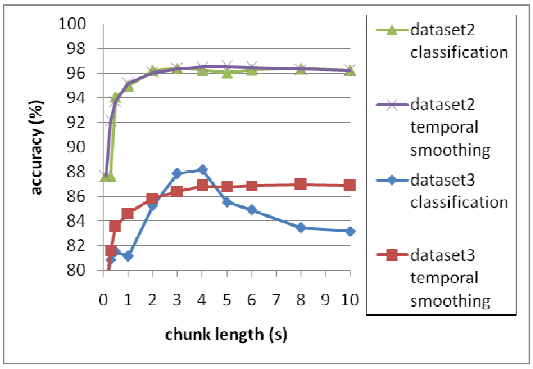

Figure 1. Accuracy as a function of segment duration, using overlapping segments from dataset1 in training. Temporal smoothing is applied over the SVM predictions from 0.1s segments in a sliding window corresponding to the x-axis value

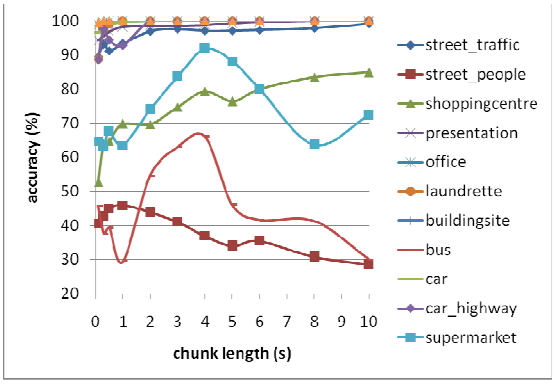

Figure 2. Accuracies of classes as a function of segment duration; overlapping segments, dataset3 as test set

With regard to choice of segment duration for a system implementation, the arbitrary choice of using 3s segments made by Ma et al. [5] seems to suit these dataset well, at least using our classification schemes. Using our schemes, accuracy seems to increase with segment duration until approximately 4s. However, shorter segment durations starting from 0.5s provide reasonable accuracy-latency trade-offs. We plan to confirm these results with another dataset in our future work.

4.3 Dataset difficulties

To get an idea of the cause for the lower accuracy of our classifier on dataset3 compared to dataset2, we examined the spectrograms of the recordings from the datasets (Figure 3). Figure 3 shows frequencies up to 4kHz for the first 60s segment from each of the

selected auditory contexts (as opposed to only the first 3 seconds reported in [5]).

Note the large difference between the launderette recording from dataset1 and dataset2 in the high frequencies. Similarly, the frequency content of bus recordings from dataset1 and dataset2 are similar, whereas in dataset3 the energy is concentrated at lower frequencies. These differences confirm the intuition that not all launderettes or busses sound the same – and to ensure generalization to new environments of each class, datasets should contain more examples from each environment type.

An additional difficulty in the source recordings surrounded gain issues in several of the samples. In all three datasets, we noticed that there was considerable "clipping" in some of the recordings (e.g., the bus) – giving them a very harsh and noisy texture that made them barely recognizable to the experimenters. On the one hand, several of the recordings had virtually no human-audible signal (e.g., the office), which similarly made them hard to differentiate from any other near-silent environment. We believe that these cases could have been mitigated by more careful gain control during the recording process, and may have artificially skewed results against several classes in the dataset.

5. Summary

In this paper, we presented results on using averaged spectral features with SVM-based classifier for auditory context recognition. The results show that despite the natural temporal continuity of mixtures of audio signals forming auditory contexts, SVM classifiers perform well compared to HMM-based classifiers in auditory context classification. In particular, using a feature vector of comparable size, the accuracy of our SVM-based classifier is about 13% higher than the reported accuracy of a HMM-based classifier for the same classification task (train:dataset1 , test:dataset3). For the other setup (train:dataset1, test:dataset2) our system achieved equal accuracy, but without using attenuation-sensitive log energy. In general, we think that energy-based features should not be used as such to recognize auditory contexts.

Considering kernel degree 1 and regularization parameter set to 1 as a baseline, tuning the degree and the regularization parameter had no significant affect on recognition accuracy when using feature set 3.

We studied also how the duration of analyzed segment affect recognition accuracy. While in our tests highest accuracy is obtained from segments longer than 3s, durations starting from 0.5s provide reasonable accuracy-latency trade-offs. Considering the quality of the used datasets, we plan to confirm these results using another set of recordings.

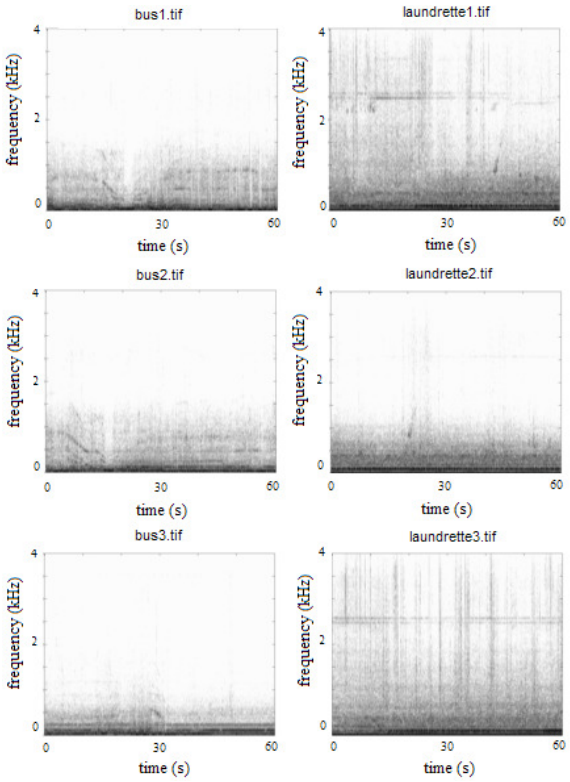

Figure 3. Spectrograms of recordings from the environments *laundrette*, *bus*, and *shopping centre* from dataset1, dataset2, and dataset3. Each spectrogram spans 60s of audio from the beginning of the recording. Frequencies up to 4kHz are shown

6. References

[1] Dey,A., (2001) "Understanding and Using Context", Personal and Ubiquitous Computing, Vol. 5, 1, pp. 4-7.

[2] Martin,K., (1999) "Sound Source Recognition: A Theory and Computational Model".

[3] Peltonen,V., Tuomi,J., Klapuri,A., Huopaniemi,J. & Sorsa,T., (2002) "Computational auditory scene recognition", in proc: *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, Vol. 2, pp. 1941-1944.

[4] Sawhney,N., (1997) "Situational Awareness from Environmental Sounds".

[5] Ma,L., Milner,B. & Smith,D., (2006) "Acoustic environment classification", *ACM Trans. Speech Lang. Process.*, Vol. 3, 2, pp. 1-22.

[6] Eronen,A.J., Peltonen,V., Tuomi,J., Klapuri,A., Fagerlund,S., Sorsa,T., Lorho,G. & Huopaniemi,J., (2006)

"Audio-based context recognition", *IEEE Trans. Audio, Speech, and Language Processing,* Vol. 14, pp. 321-329.

[7] Lu,L., Zhang,H. & Li,S.Z., (2003) "Content-based audio classification and segmentation by using support vector machines", *Multimedia Systems*, Vol. 8, 6, pp. 482-492.

[8] Scheirer,E. & Slaney,M., (1997) "Construction and evaluation of a robust multifeature speech/music discriminator", in proc: *Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 1331-1334.

[9] Jang,R., "Audio Processing Toolbox", URL: http://www.cs.nthu.edu.tw/~jang.

[10] Lee,C., Chou,C., Han,C. & Huang,R., (2006) "Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis", *Pattern Recognition Letters*, Vol. 27, 2, pp. 93-101.

[11] Witten,I.H., & Frank,E., (2005) "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann.

[12] Rifkin,R. & Klautau,A., (2004) "In Defense of One-Vs-All Classification", J.Mach.Learn.Res., Vol. 5, pp. 101-141.