

COMPUTATIONAL AUDITORY SCENE RECOGNITION

Vesa Peltonen, Juha Tuomi, Anssi Klapuri

Tampere University of Technology
Signal Processing Laboratory
P.O.Box 553, FIN-33101 Tampere, Finland
{peltonen,tuomi2,klapuri}@cs.tut.fi

Jyri Huopaniemi, Timo Sorsa

Nokia Research Center
Speech and Audio Systems Laboratory
P.O.Box 407, FIN-00045 Nokia Group, Finland
{jyri.huopaniemi,timo.sorsa}@nokia.com

ABSTRACT

In this paper, we address the problem of *computational auditory scene recognition* and describe methods to classify auditory scenes into predefined classes. By auditory scene recognition we mean recognition of an environment using audio information only. The auditory scenes comprised tens of everyday outside and inside environments, such as streets, restaurants, offices, family homes, and cars. Two completely different but almost equally effective classification systems were used: band-energy ratio features with 1-NN classifier and Mel-frequency cepstral coefficients with Gaussian mixture models. The best obtained recognition rate for 17 different scenes out of 26 and for an analysis duration of 30 seconds was 68.4%. For comparison, the recognition accuracy of humans was 70% for 25 different scenes and the average response time was around 20 seconds. The efficiency of different acoustic features and the effect of test sequence length were studied.

1. INTRODUCTION

This paper concerns the problem of *computational auditory scene recognition* (CASR), which is closely related to *computational auditory scene analysis* (CASA) [1, 2]. CASA refers to the computational analysis of an acoustic environment and the recognition of distinct sound events in it. However, in CASR, the focus is in recognizing the context, or environment, instead of analyzing and interpreting discrete sound events.

Practical applications of CASR include wearable devices that sense the environment of their users at a given time. Information about the environment enables the device to provide better service to users' needs, e.g. by adjusting the mode of operation according to the context. For example, a modern hearing-aid may choose an appropriate equalization filter automatically, instead of the user having to switch it manually.

CASR has been studied very little compared to speech recognition, for example. However, there are many research fields related to CASR that have been studied to varying extent. The related research comprises different audio classification problems, such as speech/music discrimination [3], sound source classification [4], noise classification [5], content-based audio classification [6], and classification of general audio [7]. We managed to find only a few classification systems that have been proposed to recognize auditory scenes. Clarkson and his colleagues proposed a system in which an auditory scene is recognized by classifying a temporal sequence of detected and identified sound events [8].

The study described in this paper concerns three main issues. First, we compare the discrimination ability of several different features by applying two standard classification procedures, *k*-nearest neighbors (*k*-NN) classifier and Gaussian mixture models (GMM). Second, the recognition accuracy as a function of the test sequence length is studied using several selected features, and finally, a short experiment of recognizing more general classes is described.

2. DATA COLLECTION

Real-world recordings from a variety of different auditory scenes were made. In Table 1, the different scenes and number of recordings from each scene are listed. The recordings are categorized into six more general classes according to common characteristics of the scenes (outdoors, vehicles, public, offices, home, and reverberant places). The categorization of the scenes was somewhat ambiguous; some of the recordings can be associated with more than one higher-level class.

A total of 226 measurements were recorded using two different configurations. The first configuration consisted of a binaural setup (Brüel&Kjær 4128 head and torso simulator), stereo setup (AKG C460B microphones), and B-format setup (Sound-Field MkV microphone). The acoustic material was recorded into a digital multitrack recorder in 16-bit and 48kHz sampling rate format. 55 recordings were made with this setup.

The remaining 171 measurements were made with a stereo setup (AKG C460B). The recordings were stored using a Sony (TCD-D10) digital audio tape recorder in 16-bit and 48kHz sampling rate format. All tested scenes, except bus and subway train, included recordings made using the both setups.

3. FEATURES

Ten fundamental acoustic features were investigated for classification of auditory scenes. In addition, the *variance* and *delta* features of the basic features were also studied. We provide here a very short description of each feature, more detailed descriptions can be found in [9]. The features are grouped into three categories according to their processing domain.

Time-domain features

- *Zero-crossing rate (ZCR)* is defined as the number of zero voltage crossings within a frame.
- *Short-time average energy* is the energy of a frame.

Frequency-domain features

Let $X_i(n)$ be the n^{th} frequency sample of the discrete Fourier transform of i^{th} time frame.

Table 1. List of the recorded auditory scenes, and number of recordings from each. The test set scenes are in boldface.

Main context	Scene	No. of recordings
Outdoors (53)	Street	16
	Road	12
	Nature	12
	Construction site	11
	Market place	1
	Amusement park	1
Vehicles (54)	Car	27
	Bus	11
	Train	10
	Subway train	6
Public/ Social places (40)	Restaurant, Café	23
	Pub	1
	Supermarket	13
	Lecture pause	1
	Crowd/indoors	2
Offices/ meeting rooms/ quiet places (39)	Office	12
	Lecture, Meeting	16
	Library	11
Home (14)	Living room	2
	Kitchen	4
	Bathroom	6
	Music	2
Reverberant (26)	Church	5
	Railway station	11
	Subway station	7
	Hall	3
Total		226

- *Band-energy ratio* is the ratio of the energy in a certain frequency-band to the total energy. The band-energy ratio of the n^{th} subband is calculated as the sum of power spectra samples $X_i(n)$ belonging to that subband divided by the total energy. In this work, we used 4 and 10 logarithmically divided subbands. The boundaries of the 4 subbands were 0, 3, 6, 12, and 24kHz, and of the 10 subbands 0.023, 0.046, 0.093, 0.187, 0.375, 0.750, 1.5, 3, 6, 12, and 24kHz for sampling rate of 48kHz.
- *Spectral centroid* represents the balancing point of the spectral power distribution. It is calculated as the sum of the frequencies weighted by the amplitudes, divided by the sum of the amplitudes, which is the 1st moment of spectrum with respect to frequency.
- *Bandwidth* is defined as the width of the range of frequencies that the signal occupies. In this work, bandwidth is calculated as

$$BW = \sqrt{\frac{\sum_{n=0}^N (n - SC)^2 \cdot |X_i(n)|^2}{\sum_{n=0}^N |X_i(n)|^2}}, \quad (1)$$

where SC is the spectral centroid and N is the index of highest frequency sample.

- *Spectral roll-off point* measures the frequency below which a certain amount of the power spectrum resides. It is calculated by summing up the power spectrum samples until the desired percentage (threshold) of the total energy is reached. The threshold in our experiments was 0.93.
- *Spectral flux* measures the change in the shape of the power spectrum by calculating the difference between power spectra of successive frames.

Linear prediction and cepstral features. These features are used for estimating the rough shape of the spectrum of a signal.

- *Linear prediction coefficients* (LPC) were extracted using the autocorrelation method.
- *Cepstral coefficients* were derived from the LPC.
- *Mel-frequency cepstral coefficients* (MFCC) were extracted applying the discrete cosine transform to the log-energy outputs of mel-scaling filter-bank.

The analysis window length for all features was 30 ms and the used windowing function was hanning. The overlap between successive frames was 50% of the frame length. Based on preliminary experiments, we noticed that varying the short-time signal processing parameters had only minor effect on the performance.

4. CLASSIFICATION FRAMEWORKS AND FEATURE VECTOR FORMATION

Two different classification frameworks were examined: one based on a k-NN classifier and the other on a GMM. The k-NN classifier performs a class vote among the k nearest neighbors to a point to be classified. In our implementation, the distance between the points was measured using the Mahalanobis distance metric with equal covariance matrix for all classes. The GMMs model the probability density function (pdf) of the data of each class as a mixture of several Gaussian pdfs. A GMM is completely represented with three parameters: mean vectors, covariance matrices, and the mixture weights. The parameters are estimated with the well-known Expectation Maximization (EM) algorithm [10]. The classification is done by estimating the probability of each class given the observation, and the class that gives the highest probability is chosen as the classification result.

In the k-NN classification method, we estimated the mean and standard deviation (std) of the features over one second windows with an intention to model the slow-changing attributes of the auditory scenes such as finite-length acoustic events. These values were used as new feature vectors and each one-second frame was classified using a 1-NN classifier. For clips longer than one second, the final result was chosen by the majority rule. We tried several window lengths for estimating the mean and std (0.05, 0.125, 0.25, 0.5, 1, 2, 4, and 8 seconds). The best result was obtained using a one second window, although there was no great difference to window lengths of 0.25 and 0.5 seconds. Increasing the number of neighbors had only a minor effect on the performance; the mean difference of the recognition rate between 1-NN and 5-NN was -0.39% (variance 2.35%), between 1-NN and 11-NN it was -0.69% (5.30%), and between 1-NN and 25-NN it was 1.75% (4.88%). The mean difference was calculated using band-energy ratio features and a test sequence length of 160 seconds.

The second classification approach was based on the GMM classifier. In this case, we used the feature vectors in each time frame as such without any manipulation. We tested the GMM with varying number of Gaussians, and found that the optimal order for the given amount of data was five. The number of iterations of the EM algorithm was fixed to 40.

5. EXPERIMENTAL SETUP

The training set included all the recorded audio material (26 different scenes, 226 samples), among which 17 scenes were classified. The subset of the scenes included in the test set is highlighted in Table 1. The classified scenes were chosen according to the criterion that each of them had to have at least five samples from

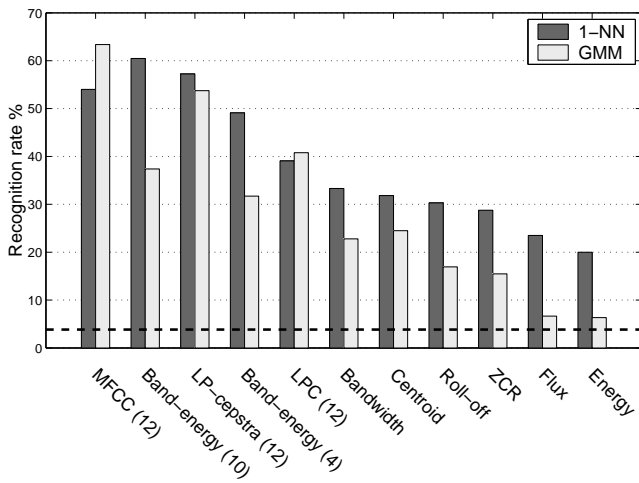


Fig. 1. Recognition rates obtained with different features for test sequence length of 30 seconds using the 1-NN and GMM. The dash line indicates the random guess rate.

different recording sessions. We did not allow multiple class labels for one recording. For example, a recording from a restaurant car of a train was labeled as train, although, it could be labeled as well as restaurant.

The classification performance was evaluated using leave-one-out cross-validation, where a classifier is trained with all instances except the one that is left out for classification. In this way, the training data is maximally utilized, but the system has never heard that particular recording before. The overall recognition rates were calculated as the sample mean of the recognition rates of the individual scenes.

6. PERFORMANCE EVALUATION

6.1. Comparison of different features

The recognition rates obtained for the 17 scenes using individual features with the two classification methods (1-NN and GMM) are shown in Figure 1. The test sequence duration was 30 seconds and the duration of each training instance was 160 seconds for all the cases. The number of trained scenes was 26, which gives an approximate random guess rate of 4% (indicated with dash line in the figure). From Figure 1, we notice that 1-NN (mean+std) classification method performs on the average better than the GMM. We also see that the temporal and spectral features having only one coefficient do not perform very well. With 12 MFCC features, we obtained a recognition accuracy of 63.4% using the GMM classifier, and with 10 band-energy features the recognition accuracy was 61.5% using the 1-NN classifier.

The basic difference between the band-energy ratio and the MFCC features is that each MFCC feature encodes the shape of the overall spectrum, thus being suitable for modeling single sound sources, whereas band-energy ratio features represent separate subbands, and are to represent, to some accuracy, different sound events occurring on different frequency ranges.

We examined only a few combinations of the presented features, and did not fully explore the optimal subset of the features due to the required computation time. A recognition accuracy of 68.4% was obtained for a test sequence length of 30 seconds by us-

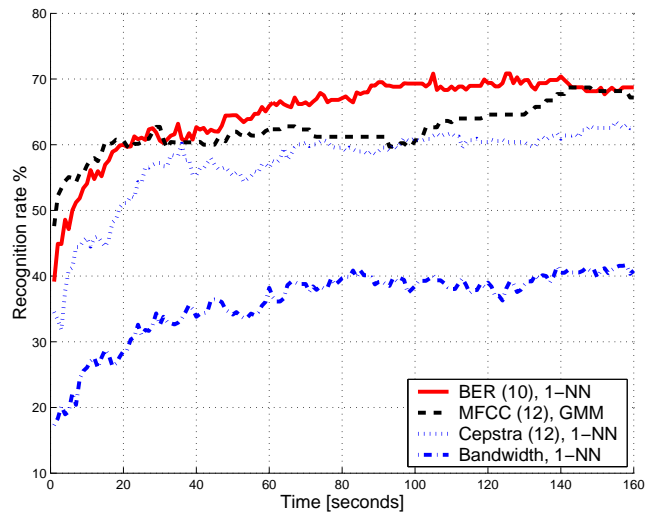


Fig. 2. Recognition rates as a function of test sequence length for the following features: band-energy ratio, LP-cepstra, and bandwidth classified using 1-NN and MFCC classified using GMM.

ing a feature vector consisting of band-energy (10), flux, roll-off, centroid, and ZCR classified using the 1-NN (mean+std) method.

Based on preliminary experiments, the delta features did not perform very well. For example, the accuracy rate obtained for the delta of the band-energy (10) was 48% and the accuracy rate obtained for the band-energy combined with its delta was 59.5%.

6.2. Recognition rate as a function of test sequence length

In Figure 2, the recognition rates of several features are presented as a function of test sequence length. We calculated the recognition rates for three features (band-energy ratio (10), LP-cepstra (12), and bandwidth) using the 1-NN (mean+std) classification method and for the MFCCs using the GMM classifier. The maximum length of test sequence was 160 seconds.

As expected, increasing the length of test sequence improves the overall recognition rate. The trend of all curves is ascending on average. An interesting observation is that the curves resemble the human response time as described in [11].

6.3. Overall recognition accuracy and confusions

The confusion matrix for 17 classified scenes using MFCC of order 12 and GMM with 5 Gaussians is presented in Table 2. The average recognition rate was 63.4% and the analysis duration was 30 seconds. The rectangular boxes enclose the more general contexts as presented in Table 1. The recognition accuracy of individual scenes ranged from 43% (subway station) to 100% (road). Restaurant was the most common target for misclassifications.

6.4. Recognition of metaclasses

We did an experiment of recognizing more general classes using the described 1-NN classification method. The test sequence duration was 30 seconds and the feature vector used consisted of 10 band-energy features. All the recordings were included both in the test and training sets. The evaluation was done using the leave-one-out testing method. The first partition of the scenes consisted of the six main contexts shown in Table 1. The recognition rate

