# Audio-Based Context Recognition

Antti J. Eronen, Vesa T. Peltonen, Juha T. Tuomi, Anssi P. Klapuri, Seppo Fagerlund, Timo Sorsa, Gaëtan Lorho, and Jyri Huopaniemi, *Member, IEEE*

***Abstract*—The aim of this paper is to investigate the feasibility of an audio-based context recognition system. Here, context recognition refers to the automatic classification of the context or an environment around a device. A system is developed and compared to the accuracy of human listeners in the same task. Particular emphasis is placed on the computational complexity of the methods, since the application is of particular interest in resource-constrained portable devices. Simplistic low-dimensional feature vectors are evaluated against more standard spectral features. Using discriminative training, competitive recognition accuracies are achieved with very low-order hidden Markov models (1–3 Gaussian components). Slight improvement in recognition accuracy is observed when linear data-driven feature transformations are applied to mel-cepstral features. The recognition rate of the system as a function of the test sequence length appears to converge only after about 30 to 60 s. Some degree of accuracy can be achieved even with less than 1-s test sequence lengths. The average reaction time of the human listeners was 14 s, i.e., somewhat smaller, but of the same order as that of the system. The average recognition accuracy of the system was 58% against 69%, obtained in the listening tests in recognizing between 24 everyday contexts. The accuracies in recognizing six high-level classes were 82% for the system and 88% for the subjects.**

***Index Terms*—Audio classification, context awareness, feature extraction, hidden Markov models (HMMs).**

## I. Introduction

CONTEXT recognition is defined as the process of automatically determining the context around a device. Information about the context would enable wearable devices to provide better service to users' needs, e.g., by adjusting the mode of operation accordingly. A mobile phone can automatically go into an appropriate profile while in a meeting, refuse to receive calls, or a portable digital assistant can provide information customized to the location of the user [1].

Many sources of information for sensing the context are available, such as luminance, acceleration, or temperature. Audio provides a rich source of context-related information, and recognition of a context based on sound is possible for humans to some extent. Moreover, there already exist suitable sensors, i.e., microphones, in many portable devices.

In this paper, we consider context recognition using acoustic information only. Within this scope, a context denotes a location with different acoustic characteristics, such as a restaurant, marketplace, or a quiet room. Differences in the acoustic characteristics can be due either to the physical environment or the activity of humans and nature. We describe the collection of evaluation data representing the common everyday sound environment of urban people, allowing us to assess the feasibility of building context aware applications using audio. Using this data, a comprehensive evaluation is made of different features and classifiers. The main focus is on finding methods suitable for implementation on a mobile device. Therefore, we evaluate linear feature transforms and discriminative training to improve the accuracy obtained with very low-order HMMs.

An experiment was conducted to facilitate the direct comparison of the system's performance with that of human subjects. A forced-choice test with identical test samples and reference classes for the subjects and the system was used. We also made a qualitative test to assess the information on which the human subjects base their decision. To our knowledge, this study is the first attempt to present a comprehensive evaluation of a computer and human performance in audio-based context recognition. Some preliminary results on context recognition using audio have been described in [2], [3].

This paper is organized as follows. Section II reviews previous work. Section III presents the feature extraction algorithms used in this study. In Section IV, the classification methods are described. Section V presents an assessment of the computer system. In Section VI, a test on human perception of audio contexts is described. Finally, in Section VII, these results are compared to the performance of the system.

## II. Previous Work

The research on context awareness is still at its early stages and very few applications have been constructed that make use of other context information than global positioning system (GPS) location [4]. One of the earliest prototypes of a context-aware system was the ParcTab developed at the Xerox Palo Alto Research Center [5]. The ParcTab featured, e.g., contextual information and commands, automatic contextual reconfiguration and context-triggered actions.

In many cases, the context-awareness functionality is build upon an array of different sensors sensing the context. In [6], the authors used accelerometers, photodiodes, temperature sensors, touch sensors, and microphones, from which simple low-level features were extracted. Another approach is to transform the

Manuscript received December 31, 2003; revised January 26, 2005. This work was supported by Nokia Research Center and TISE Graduate School. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shoji Makino.

A. J. Eronen, T. Sorsa, G. Lorho, and J. Huopaniemi are with the Nokia Research Center, FIN-33721 Tampere, Finland (e-mail: antti.eronen@nokia.com; timo.sorsa@nokia.com; gaetan.lorho@nokia.com; jyri.huopaniemi@nokia.com).

V. T. Peltonen is with the Nokia Mobile Phones, FIN-33721 Tampere, Finland (e-mail: vesa.peltonen@nokia.com).

J. T. Tuomi and A. P. Klapuri are with the Institute of Signal Processing, Tampere University of Technology, FIN-33101 Tampere, Finland (e-mail: juha.tuomi@tut.fi; anssi.klapuri@tut.fi).

S. Fagerlund is with the Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, FIN-02015, Espoo, Finland (e-mail: seppo.fagerlund@hut.fi).

Digital Object Identifier 10.1109/TSA.2005.854103

raw input into a low-dimensional representation using principal component analysis (PCA) or independent component analysis (ICA) [7], [8].

In general, the process of context recognition is very similar regardless of the sensors or data sources used for the recognition. The feature vectors obtained from sensors are fed to classifiers that try to identify the context the particular feature vectors present. As classifiers, e.g., hidden Markov models (HMMs) [9], or a combination of a self-organizing map and a Markov chain, have been used [6].

Only few studies have attempted to classify contexts using acoustic information. Clarkson has classified seven contexts using spectral energies from the output of a filter bank and a HMM classifier [9]. In [10], Sawhney describes preliminary experiments with different features and classifiers in classifying between voice, traffic, subway, people, and others. The most successful system utilized frequency-band energies as features and a nearest-neighbor classifier.

El-Maleh *et al.* classified five environmental noise classes (a car, street, babble, factory, and bus) using line spectral features and a Gaussian classifier [11]. Couvreur *et al.* used HMMs to recognize five types of environmental noise events: car, truck, moped, aircraft, and train, using linear prediction cepstral coefficients as features and discrete HMMs [12]. The authors also described an informal listening test, which showed that, on the average, humans were inferior in classifying these categories compared to the system.

The features we are using are similar to those used in different audio information retrieval tasks [13]. Scheirer and Slaney described a speech/music discrimination system, which used a combination of several features [14]. More recent studies include that of Lu *et al.* [15] and Li *et al.* [16] who also included environmental noise as one of the categories. Zhang and Kuo [17] classified between harmonic environmental sound, nonharmonic environmental sound, environmental sound with music, pure music, song, speech with music, and pure speech.

Casey has used a front-end where log-spectral energies are transformed into a low-dimensional representation with singular-value decomposition and ICA [18]. The classifier uses single-Gaussian continuous-density HMMs with full covariance matrices trained with Bayesian maximum *a posteriori* (MAP) estimation. Casey's system was evaluated on a database consisting e.g., of musical instrument sounds, sound effects, and animal sounds.

To our knowledge, context recognition using audio has not been studied to this extent before. The results existing in the literature have used only a limited number of categories, often focusing into a certain noise type such as vehicle sounds. In this paper, we present results using comprehensive data measured from several everyday contexts. The most promising features presented in the literature are compared on this data. We propose a linear transformation of the concatenated cepstral and delta cepstral coefficients using PCA or ICA and show that this slightly improves the classification accuracy. Moreover, we demonstrate that compact diagonal-covariance Gaussian HMMs and discriminative training are an effective classifier for this task. To our knowledge, discriminatively trained HMMs have not been used for audio-based context recognition before.

## III. Acoustic Measurements and Feature Extraction

### A. Recording Procedure

To obtain a realistic estimate of the feasibility of building context-aware applications using audio input, we paid special attention to gathering a data set that would be representing of the everyday sound environment encountered by urban people. The recording procedure has been described in [19] and is summarized here. A total of 225 real-world recordings from a variety of different contexts were made using two different recording configurations. The first configuration has been developed by Zacharov and Koivuniemi [20]. It consists of a head-and-torso simulator with multiple microphones and is capable of storing multiple audio formats simultaneously. For the purpose of this study, we only utilized the binaural recordings (two channels) and stereo recordings (two channels). The microphones mounted in the ears of the dummy head enable a realistic binaural reproduction of an auditory scene. The stereo setup consisted of two omnidirectional microphones (AKG C460B), separated by a distance of one meter. This construction was attached to the dummy head. The acoustic material was recorded into a digital multitrack recorder in 16-bit and 48-kHz sampling rate format. A total of 55 recordings were made with this setup. The remaining measurements were made with an easily portable stereo setup using AKG C460B microphones.

The recording of spatial sound material was done for subjective evaluations. In computer simulations, we only used the left channel from the stereo setup. Table I shows the division of recordings into different categories.

### B. Feature Extraction

A wide set of feature extractors was implemented for this study in order to evaluate the accuracy obtained with each, and to select a suitable feature set for the system.

All features are measured in short analysis frames. A typical analysis frame length in this study was 30 ms with 15-ms overlap. The hanning window function was used. The following features were evaluated in this study.

*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 and is computed as the sum of squared amplitudes within a frame.

*Mel-frequency cepstral coefficients (MFCC)* are a perceptually motivated representation of the coarse shape of the spectrum [21]. We used 11 or 12 MFCC coefficients calculated from the outputs of a 40-channel filterbank.

*Mel-frequency delta cepstral coefficients ($\Delta$MFCC)* are used to describe the dynamic properties of the cepstrum. We used a three-point linear fit to approximate the first time derivative of each cepstral coefficient.

*Band-energy* refers to the energies of subbands normalized with the total energy of the signal. We experimented with four and ten logarithmically-distributed subbands.

*Spectral centroid* represents the balancing point of the spectral power distribution.

*Bandwidth* is defined as the estimated bandwidth of the input signal [16].

TABLE I
STATISTICS OF THE AUDIO MEASUREMENTS

| High level category | Context | Number of Recordings |
|---|---|---|
| Outdoors | Street | 16 |
| | Road | 13 |
| | Nature | 12 |
| | Construction | 11 |
| | Marketplace | 1 |
| | Fun Park | 1 |
| Vehicles | Car | 27 |
| | Bus | 11 |
| | Train | 10 |
| | Subway Train | 6 |
| Public / Social Places | Restaurant | 13 |
| | Cafeteria | 10 |
| | Pub | 1 |
| | Shop | 13 |
| | Lecture Pause | 1 |
| Offices / Meetings / Quiet Places | Office | 12 |
| | Lecture | 12 |
| | Meeting | 4 |
| | Library | 11 |
| Home | Living Room | 2 |
| | Kitchen | 4 |
| | Bathroom | 6 |
| | Music | 2 |
| Reverberant Places | Church | 4 |
| | Railway Station | 11 |
| | Subway Station | 7 |
| | Hall | 4 |
| **Total** | | 225 |

*Spectral roll-off* [16] measures the frequency below which a certain amount of spectral energy resides. It measures the "skewness" of the spectral shape.

*Spectral flux (SF)* is defined as the difference between the magnitude spectra of successive frames [14].

*Linear prediction coefficients (LPCs)* were extracted using the autocorrelation method [22, p. 103]. The number of LPC coefficients extracted was 12.

*Linear prediction cepstral coefficients* are obtained using a direct recursion from the LPC coefficients [22, p. 115]. The number of cepstral coefficients was 12 after discarding the zeroth coefficient.

All the features were mean and variance normalized using global estimates measured over the training data.

### C. Feature Transforms

The main idea of linear data-driven feature-transformations is to project the original feature space into a space with a lower dimensionality and more feasible statistical properties, such as uncorrelatedness. In this work, three different techniques were used. The PCA finds a decorrelating transform [25, p. 115], ICA results in a base with statistical independence [25, p. 570], and the linear discriminant analysis (LDA) tries to maximize class separability [25, p. 120].

PCA projects the original data into a lower dimensional space such that the reconstruction error is as small as possible, measured as the mean-square error between the data vectors in the original space and in the projection space. Projection onto a lower dimensional space reduces the amount of parameters to be estimated in the classifier training stage, and uncorrelated features are efficiently modeled with diagonal-covariance Gaussians.

The goal of ICA is to find directions of minimum mutual information, i.e., to extract a set of statistically independent vectors from the training data. Here, the FastICA algorithm was used for finding the ICA basis transformation [23].

Himberg *et al.* have used PCA and ICA to project multidimensional sensor data from different contexts into a lower dimensional representation, but reported only qualitative results [4]. In speech recognition, the use of an ICA transformation has been reported to improve the recognition accuracy [24]. In the MPEG-7 generalized audio descriptors, ICA is proposed as an optional transformation for the spectrum basis obtained with singular value decomposition, and Casey's results have shown the success of this method on a wide variety of sounds [18]. Our approach is different from all these studies, since we perform ICA on concatenated MFCC and $\Delta$MFCC features. Including the delta coefficients is a way to include information on temporal dependencies of features, which is ignored if the transform is applied on static coefficients only. In [18] and [24], delta coefficients were not considered.

The third feature transform technique tested here, LDA, differs from PCA and ICA by utilizing class information. The goal is to find basis vectors that maximize the ratio of between-class variance to within-class variance.

It should be noted that the extra computational load caused by applying any of these transformations occurs mainly in the off-line training phase. The test phase consists of computing the features in the usual way plus an additional multiplication once per analysis frame with the transform matrix derived off-line using the training data.

## IV. CLASSIFICATION METHODS

### A. $K$-Nearest Neighbors

The most straightforward classification method is nearest neighbor classification. The $K$-nearest-neighbors ($K$-NN) classifier performs a class vote among the $k$ nearest training-data feature vectors to a point to be classified [25, p. 182]. In our implementation, the feature vectors were first decorrelated using PCA and the Euclidean distance metric was used in the transformed space. Averaging over 1-s-long segments was used to reduce the amount of calculations and required storage space.

### B. HMM

*1) Description of the Model:* A HMM [22, pp. 321–386] is an effective parametric representation for a time-series of observations, such as feature vectors measured from natural sounds. In this work, HMMs are used for classification by training a HMM for each class, and by selecting the class with the largest *a posteriori* probability.

*2) Model Initialization:* We used the maximum-likelihood based Baum–Welch algorithm to train the "baseline" HMMs for each class separately. The number of states (NS) and the number of component densities per state (NC) was varied. The models were initialized with a single Gaussian at each state, and the component with the largest weight was then split until the

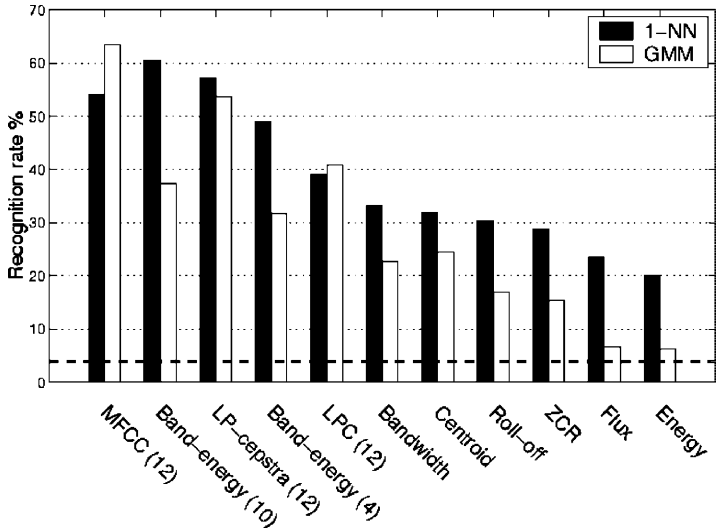


Fig. 1. Recognition accuracy obtained with different features using the GMM and 1-NN classifiers and 30 s of each test signal.

desired value of NC was obtained. Each component split was followed by 15 Baum–Welch iterations, or until the likelihood converged.

*3) Discriminative Training of HMMs:* In applications where computational resources are limited, we are forced to use models with as few Gaussians as possible, since their evaluation poses the computationally most demanding step in the recognition phase. In these cases the HMM is not able to fully represent the feature statistics and other approaches than maximum likelihood parameter estimation may lead into better recognition results. Discriminative training methods such as the maximum mutual information (MMI) aim at maximizing the ability to distinguish between the observation sequences generated by the model of the correct class and those generated by models of other classes [22, p. 363].

We used a discriminative training algorithm recently proposed by Ben–Yishai and Burshtein [26]. The algorithm is based on an approximation of the MMI. It starts from a "baseline" model set trained with the Baum–Welch algorithm, followed by an iterative discriminative training phase. At each discriminative training iteration, new statistics for the model parameters are accumulated not only from the observations of the correct class, but also from a set of confusing classes. The set of confusing classes is obtained by MAP classification performed on the training set. An interested reader should refer to [26] for more details of the algorithm.

## V. Evaluation

### A. Experimental Setup

Two training and testing setups were formed from the samples. Setup 1 consisted of 155 recordings of 24 contexts that were used for training and 70 recordings of 16 contexts were tested. Random division of recordings into the training and tests sets was done 100 times. The contexts selected into the test set had to have at least five recordings from different locations at different times. Setup 2 was used in the listening test and in the direct comparison, and had two nonoverlapping sets of 45 samples from 18 different contexts in the test set.

A higher level of abstraction may be sufficient for some applications. Hence, the recordings were also categorized into six high-level classes that are more general according to some common characteristics. These classes are: 1) outdoors, 2) vehicles, 3) public/social, 4) offices/meetings/quiet, 5) home, and 6) reverberant places. It should be noted that the allocation of individual contexts into high-level classes is ambiguous; many contexts can be associated with more than one high-level class.

### B. Results

*1) Comparison of Features:* In the first experiment, we compared the accuracy obtained with different features. In this experiment, classification performance was evaluated using leave-one-out cross-validation on all the recorded data. The classifiers were trained with all recordings except the one that was left out for classification. In this way, the training data is maximally utilized but the system has never heard the test recording before. The overall recognition rate was calculated as the sample mean of the recognition rates of the individual contexts.

The recognition rates obtained at the context level using individual features with two different classifiers, the 1-NN and a one-state HMM (a GMM), are shown in Fig. 1. The test sequence duration was 30 s taken from the beginning of each test recording and the duration of each training recording was 160 s. The random guess rate for 24 classes is shown with the dashed line in Fig. 1. The 1-NN classifier performs on the average better than the GMM. This is indicative of complicated distributions of many features, which are not well modeled with a GMM with five diagonal-covariance Gaussians. The MFCC coefficients are well modeled with a GMM. With 12 MFCC features, we obtained a recognition accuracy of 63% using the GMM classifier, and with ten band-energy features the recognition accuracy was 61% using the 1-NN classifier.

*2) Discriminative Training:* The second experiment studied the HMM and the MFCC features in more detail. The MFCC coefficients were augmented with the delta coefficients. We trained models with different NSs and NCs, and varied the

TABLE II
RECOGNITION ACCURACY USING ONE-STATE HMMS WITH VARYING NUMBER OF COMPONENT DENSITIES

| # Components | Baum-Welch | Discriminative |
|---|---|---|
| $NC = 1$ | 57 ± 4 | 60 ± 4 |
| $NC = 2$ | 62 ± 4 | 63 ± 4 |
| $NC = 3$ | 64 ± 4 | 65 ± 4 |
| $NC = 4$ | 65 ± 4 | 66 ± 4 |
| $NC = 5$ | 65 ± 4 | 66 ± 4 |

TABLE III
RECOGNITION ACCURACY (%) AND STANDARD DEVIATION USING HMMS WITH VARYING TOPOLOGIES AND NUMBER OF STATES

| | Fully-Connected | | Left-Right with Skips | |
|---|---|---|---|---|
| # States | Baum-Welch | Discriminative | Baum-Welch | Discriminative |
| $NS = 2$ | 60 ± 4 | 62 ± 4 | - [a] | - [a] |
| $NS = 3$ | 61 ± 5 | 64 ± 5 | 62 ± 4 | 64 ± 5 |
| $NS = 4$ | 63 ± 5 | 65 ± 5 | 63 ± 5 | 65 ± 5 |

[a]This topology is identical to the fully-connected with two states.

TABLE IV
RECOGNITION ACCURACY (%) AND STANDARD DEVIATION WHEN CONFUSIONS WITHIN THE SIX HIGHER LEVEL CLASSES ALLOWED

| # States | Baum-Welch | Discriminative |
|---|---|---|
| $NS = 2$ | 75 ± 3 | 77 ± 3 |
| $NS = 3$ | 77 ± 3 | 79 ± 3 |
| $NS = 4$ | 77 ± 4 | 79 ± 4 |

model topology. The second aim was to compare the baseline maximum-likelihood training using the Baum–Welch algorithm and discriminative training. The division into training and test data was done according to Setup 1. The amount of training data used from each recording was 160 s. In order to obtain reliable accuracy estimates and to utilize the test data efficiently, the recognition was performed in adjacent 30-s windows with 25% overlap, and the final recognition result has been averaged over the different train/test divisions, recognition windows, recordings, and classes.

Tables II–IV show the results from this experiment. The baseline models were obtained after 15 Baum–Welch iterations. Three iterations of discriminative training were then applied on the models obtained from Baum–Welch re-estimation. Using an HMM with two or three states, or a one-state HMM with two or three component densities gives acceptable accuracies especially when discriminative training is used, taking into account the low computational demand of having to evaluate just a few diagonal covariance Gaussians.

*3) Linear Feature Transforms:* In the next experiment, we evaluated the use of the three linear feature transforms: PCA, ICA, and LDA. Table IV shows the recognition accuracies when the different transforms were applied on a feature vector consisting of concatenated MFCCs and their derivatives. On the average, applying the ICA or PCA transforms gives a slight improvement in recognition accuracy (Table V). In these experiments, we used a two-state HMM with one component density per state.

TABLE V
RECOGNITION ACCURACY USING LINEAR FEATURE TRANSFORMS

| | No Transform | PCA | ICA | LDA |
|---|---|---|---|---|
| Context | 61 ± 3 | 62 ± 3 | 62 ± 4 | 60 ± 4 |
| Higher level | 75 ± 3 | 76 ± 2 | 77 ± 3 | 76 ± 3 |

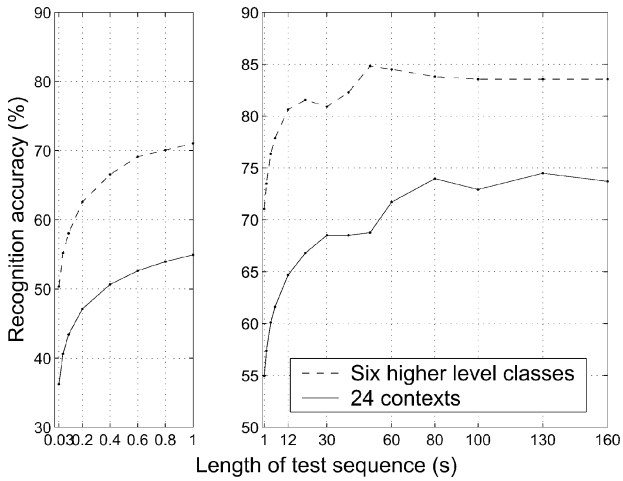


Fig. 2. Recognition accuracy as a function of test sequence length for the individual contexts and the six high-level classes. The left panel shows details of a test sequence length less than 1 s; the shortest length 0.03 corresponds to a single frame.

In [24], the authors reported improvements in speech recognition over the baseline using MFCC coefficients without a transform when these same transforms were applied either to the log-energy outputs of the MFCC filter bank, or the static MFCC coefficients. We made experiments also with these methods but improvement over the baseline was observed only when the concatenated MFCCs and deltas were transformed.

*4) Effect of Test Sequence Length:* In Fig. 2, the recognition rates obtained using the ICA transformed MFCC features and two-state HMMs are presented when the length of the test sequence was varied. The results for the six high-level classes have been derived from the results at the context level when confusions within the higher level categories are allowed.

As expected, increasing the length of test sequence improves the overall recognition rate. However, it takes rather long for the result to converge (around 60 s). With less than 20 s of test data, the recognition accuracy drops fast. Thus, this amount can be regarded as the lower limit for reliable recognition. The left panel shows the details with very short recognition sequence lengths ranging from just a single frame (30 ms) to 1 s. Even with these very short analysis segments some degree of accuracy can be obtained.

| Presented \ Responded | Street | Road | Nature | Constr. site | Market place | Amusement park | Car | Bus | Train | Metro train | Restaurant | Cafe | Pub | Supermarket | Lecture pause | Office | Lecture | Meeting | Library | Living room | Kitchen | Bathroom | Music | Church | Railway station | Metro station | Hall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Street | **56** | 35 | 3.7 | 1.9 | 1.9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Road | 24 | **70** | 0 | 0 | 0 | 0 | 1.9 | 3.7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Nature | 0 | 1.9 | **96** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Constr. site | 3.7 | 1.9 | 0 | **91** | 0 | 1.9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.9 |
| Car | 3.7 | 0 | 0 | 0 | 0 | 0 | **74** | 7.4 | 7.4 | 5.6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Bus | 1.9 | 1.9 | 0 | 1.9 | 1.9 | 1.9 | 3.7 | **67** | 5.6 | 11 | 0 | 0 | 3.7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Train | 0 | 0 | 0 | 0 | 0 | 0 | 9.3 | 11 | **65** | 1.9 | 1.9 | 0 | 0 | 0 | 0 | 3.7 | 0 | 0 | 3.7 | 0 | 0 | 0 | 0 | 0 | 0 | 1.9 | 1.9 |
| Metro train | 0 | 0 | 0 | 0 | 0 | 1.9 | 1.9 | 0 | 20 | **72** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3.7 | 0 |
| Restaurant | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **70** | 24 | 0 | 1.9 | 0 | 0 | 0 | 0 | 0 | 0 | 3.7 | 0 | 0 | 0 | 0 | 0 | 0 |
| Cafe | 1.9 | 0 | 0 | 0 | 1.9 | 3.7 | 0 | 0 | 0 | 0 | 17 | **48** | 5.6 | 3.7 | 3.7 | 1.9 | 0 | 0 | 0 | 3.7 | 3.7 | 0 | 0 | 0 | 1.9 | 1.9 | 1.9 |
| Supermarket | 0 | 0 | 0 | 0 | 0 | 1.9 | 0 | 0 | 0 | 0 | 1.9 | 9.3 | 1.9 | **59** | 0 | 0 | 0 | 0 | 0 | 3.7 | 0 | 0 | 3.7 | 0 | 3.7 | 1.9 | 13 |
| Office | 0 | 0 | 0 | 0 | 0 | 1.9 | 1.9 | 0 | 1.9 | 0 | 1.9 | 3.7 | 7.4 | 0 | 0 | **59** | 0 | 5.6 | **11** | 5.6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Lecture | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **91** | 9.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Meeting | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.9 | 3.7 | 1.9 | 0 | 9.3 | 17 | 1.9 | **56** | 0 | 1.9 | 7.4 | 0 | 0 | 0 | 0 | 0 | 0 |
| Library | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5.6 | 0 | 17 | 7.4 | 17 | 1.9 | 0 | **35** | 7.4 | 3.7 | 0 | 0 | 1.9 | 1.9 | 0 | 1.9 |
| Bathroom | 0 | 0 | 0 | 3.7 | 0 | 0 | 0 | 1.9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.9 | 0 | 0 | 0 | 0 | 17 | **76** | 0 | 0 | 0 | 0 | 0 |
| Railway station | 0 | 0 | 0 | 1.9 | 3.7 | 3.7 | 0 | 0 | 3.7 | 1.9 | 0 | 0 | 0 | 1.9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **52** | 1.9 | 30 |
| Metro station | 1.9 | 0 | 0 | 1.9 | 0 | 0 | 0 | 0 | 11 | 7.4 | 0 | 0 | 0 | 0 | 1.9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **69** | 7.4 |

Fig. 3. Confusion matrix of the listening test experiment using stereo samples. The boxes indicate the high-level classes, which are (from left to right, top to bottom) outdoors, vehicles, public/social, offices/meetings/quiet, home, and reverberant.

## VI. Human Perception of Audio Contexts

### A. Setup of the Experiment

We also carried out an experiment on human recognition of audio contexts in order to obtain a performance baseline for the assessment of the system. This experiment was organized in three listening tests.

*1) Stimuli, Reproduction System, and Listening Conditions:* The stimuli for the listening tests were the recordings from the Setup 2 as described in Section V-A. All stimuli employed in this experiment were 1-min-long samples and were defined using two levels of categorization: context and high-level context.

All tests were performed in an ITU-R BS.1116-1 compliant listening room [27]. Audio samples were reproduced at a natural sound level over a stereophonic setup using Genelec 1031A loudspeakers placed at $\pm 30^\circ$ in front of the listener. The test design and administration were performed using the Presentation software [28]. This system allows very accurate monitoring of the reaction time between sample replay and subject responses.

*2) Description of the Three Listening Tests:* The focus of the main test was in studying the accuracy and reaction time of humans in audio context recognition. The second test compared the human ability in recognition with three different sound configurations, namely, the monophonic, stereophonic, and binaural reproduction techniques, in an assumed order of increasing degree of spaciousness. A subset of 18 samples from nine different contexts was selected for each configuration in this part of the experiment. For the binaural samples, crosstalk cancellation filters were designed based on the MIT KEMAR HRTF measurements [29] in order to obtain appropriate reproduction of the signal over loudspeakers (i.e., a binaural to transaural conversion).

The aim of the third test was to obtain a qualitative description of the recognition of auditory scenes. Subjects were asked to listen to nine samples and rate the information they used in the recognition process. After each stimulus, listeners filled in a form in which they were asked to evaluate and rate on a six-point discrete scale, how important different cues were in recognition (0 accounted for a cue not used and 5 for a cue considered very important).

In the three tests, subjects were instructed to try to recognize the context as fast as possible. A list of possible contexts was given to the test subjects. The list included also contexts not presented during the test. Recognition time was measured from the starting time of the stimulus presentation to the first keyboard press, after which the subject could select the context recognized by an additional keyboard press.

Eighteen subjects participated in the test, which was designed for two groups, each including the same number of stimuli and identical contexts. This permitted the use of more samples from the database, still keeping the total duration of the test within 1 h. The listening test started with a training session including nine samples not included in the actual test to familiarize the subjects with the user interface and the test setup.

### B. Results of the Listening Test

Two measures were analyzed from this listening test, the recognition rate and the reaction time for each stimulus. Statistical methods employed were different due to the different nature of the two measures. First, recognition rate was analyzed as a set of right or wrong answers using a nonparametric statistical procedure, i.e., the Friedman and Kruskal–Wallis tests. For the reaction time, the statistical analysis was performed with a classical parametric statistical procedure (ANOVA), after discarding data considered as outliers.

*1) Stereo Test:* Rate was calculated for both context and high-level context recognition. As a result, the average recognition rate was 69% for contexts and 88% for the high-level contexts. Fig. 3. presents the confusion matrix for this experiment averaged over all listeners (differences between the two groups are not significant). Context and high-level context with

| Presented \ Responded | Street | Road | Nature | Constr. site | Market place | Amusement park | Car | Bus | Train | Metro train | Restaurant | Cafe | Pub | Supermarket | Lecture pause | Office | Lecture | Meeting | Library | Living room | Kitchen | Bathroom | Music | Church | Railway station | Metro station | Hall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Street | **0** | 0 | 0 | 75 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Road | 0 | **100** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Nature | 0 | 0 | **83** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Constr. site | 0 | 0 | 0 | **25** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 50 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 25 | 0 | 0 | 0 | 0 | 0 |
| Car | 0 | 0 | 0 | 0 | 0 | 0 | **100** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Bus | 0 | 0 | 0 | 0 | 0 | 0 | 17 | **67** | 0 | 0 | 0 | 0 | 0 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Train | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **50** | 0 | 0 | 25 | 0 | 0 | 0 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Metro train | 0 | 0 | 0 | 0 | 0 | 17 | 0 | 0 | 0 | **50** | 17 | 0 | 13 | 3.7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Restaurant | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **25** | 50 | 0 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Cafe | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 | 46 | **16** | 4.7 | 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Supermarket | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 6 | 0 | **66** | 0 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Office | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **75** | 0 | 0 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Lecture | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **100** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Meeting | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 | 33 | 50 | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Library | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 | 0 | 0 | 0 | 15 | **68** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Bathroom | 0.3 | 0 | 0.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 16 | 0 | 2 | **81** | 0 | 0 | 0 | 0 | 0 |
| Railway station | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **75** | 25 | 0 |
| Metro station | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 25 | 0 | 0 | **74** | 0 |

Fig. 4. Confusion matrix when the system was tested on the samples from the listening test. Compare this to Fig. 3. The boxes indicate the high-level classes, which are (from left to right, top to bottom): Outdoors, vehicles, public/social, offices/meetings/quiet, home, and reverberant.

TABLE VI
RECOGNITION ACCURACY (%) FOR THE DIFFERENT PRESENTATION TECHNIQUES

| | Mono | Stereo | Binaural | Average |
|---|---|---|---|---|
| Context | 63 | 70 | 62 | 66 |
| Higher-level | 86 | 89 | 90 | 88 |

TABLE VII
CUES USED FOR AUDIO CONTEXT RECOGNITION

| | Human activity | Spatial information | Prominent event | Continuous voice | Vehicle noise | Nature sounds |
|---|---|---|---|---|---|---|
| Cues used | 67% | 67% | 64% | 47% | 32% | 8% |
| Importance of the cue | 2.55 | 1.88 | 2.50 | 1.89 | 1.77 | 2.26 |

the highest recognition rate were respectively *nature* (96%) and *outdoors* (97%), whereas those with the lowest rate were *library* (35%) and *office/other quiet places* (76%). Reaction time was also compared for the 18 contexts. Overall, the average reaction time was 13 s, ranging from 5 s (*nature*) to 21 s (*library*).

*2) Mono/Stereo/Binaural Test:* In the analysis of the second test, recognition rates were compared for monophonic, stereophonic, and binaural presentations. The average rate for context recognition with the three presentation techniques is shown in Table VI. The recognition rate averaged over the three techniques was 66% for context and it increased to 88% for high-level contexts. Differences in recognition accuracy can be observed between the different presentation techniques, especially with the stereo configuration in the case of context recognition, but this is not statistically significant overall. An average recognition time of 14 s was found for all stimuli. Comparing now the three presentation techniques, a significant difference was found with lower average recognition time for the stereo and binaural presentation (13 s) than the mono one (15 s).

*3) Qualitative Test:* In the last test, data on the qualitative assessment of recognition cues was collected and analyzed. The two measures computed from the questionnaire were a percentage of specific cues used in recognition (i.e., cue not used for a 0 rating and cue used otherwise) and its importance for the recognition process (i.e., an average of rates over subjects), as shown in Table VII. As a result, it was found that human activity and spatial information cues are most often used (67% of cases), with a lower importance for spatial information, however (1.88 rating against 2.55 for human activity). Prominent events were also mentioned as an important cue for recognition with a rate of 2.50.

### C. Conclusion of the Subjective Test

This listening test showed that humans are able to recognize contexts in 69% of cases. The recognition rate increases to 88%, when considering high-level categorization of contexts only. Recognition time was 13 s on average. It should be noted, however, that reaction time for high-level context detection alone would probably be significantly faster. Indeed, some of the subjects reported that they could exclude most of the contexts fast, but the final decision between specific contexts from the same high-level context class took more time. Differences between the different reproduction techniques were also found, but these were not statistically significant. The presentation technique was only found to be significant for the reaction time.

### D. Performance Comparison Between the System and Human Listeners

A direct comparison between the system and the human ability was made using exactly the same test samples and reference classes as in the listening test. Figs. 3 and 4 show the

averaged confusion matrices for the subjects and the system on this test setup, respectively. The boxes indicate the six high-level categories. The amount of test data given to the system was 30 s, since the human subjects did not usually listen through the whole 60 s. The averaged recognition accuracies of the computer system are 58% and 82% against the accuracies 69% and 88% obtained in the listening test for contexts and high-level classes, respectively.

## VII. Conclusion

Building context aware applications using audio is feasible, especially when high-level contexts are concerned. In comparison with the human ability, the proposed system performs rather well (58% versus 69% for contexts and 82% versus 88% for high-level classes for the system and humans, respectively). Both the system and humans tend to make similar confusions mainly within the high-level categories.

Computationally efficient recognition methods were evaluated. Quite reliable recognition can be achieved using only a four-dimensional feature vector that represents subband energies, and even simplistic one-dimensional features achieve recognition accuracy significantly beyond chance rate. Discriminative training leads to slightly but consistently better recognition accuracies particularly for low-order HMM models. Slight increase in recognition accuracy can also be obtained by using PCA or ICA transformation of the mel-cepstral features.

The recognition rate as a function of the test sequence length appears to converge only after about 30 to 60 s. Some degree of accuracy can be achieved even in analysis frames below 1 s. The average reaction time of human listeners was 14 s, i.e., somewhat smaller but of the same order as that of the system.

## References


[1] J. Mäntyjärvi, P. Huuskonen, and J. Himberg, "Collaborative context determination to support mobile terminal applications," *IEEE Wireless Commun.*, no. 5, pp. 39–45, Oct. 2002.

[2] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, May 2002, pp. 1941–1944.

[3] A. Eronen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context awareness – Acoustic modeling and perceptual evaluation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 5, Apr. 2002, pp. 529–532.

[4] G. Chen and D. Kotz, "A survey of context-aware mobile computing research," Dept. Comp. Sci., Dartmouth College, Hanover, NH, Tech. Rep. TR2000-381, 2000.

[5] B. N. Schilit, N. Adams, R. Gold, M. Tso, and R. Want, "The PARCTAB mobile computing system," in *Proc. IEEE 4th Workshop on Workstation Operating Systems*, Oct. 1993, pp. 34–39.

[6] K. Van Laerhoven, K. Aidoo, and S. Lowette, "Real-time analysis of data from many sensors with neural networks," in *Proc. 5th Int. Symp. Wearable Computers*, 2001, pp. 115–123.

[7] J. Himberg, J. Mäntyjärvi, and P. Korpipää, "Using PCA and ICA for exploratory data analysis in situation awareness," in *Proc. IEEE Int. Conf. Multisensor Fusion and Integration for Intelligent Systems*, Sep. 2001, pp. 127–131.

[8] F. M. Salam and G. Erten, "Sensor fusion by principal and independent component decomposition using neural networks," in *Proc. IEEE Int. Conf. Multisensor Fusion and Integration for Intelligent Systems*, Aug. 1999, pp. 127–131.

[9] B. Clarkson and A. Pentland, "Unsupervised clustering of ambulatory audio and video," Perceptual Computing Group, MIT Media Lab, Cambridge, MA, Tech. Rep. 471.

[10] N. Sawhney, "Situational awareness from environmental sounds," Project Rep., Speech Interface Group, MIT Media Lab, Cambridge, MA, 1997.

[11] K. El-Maleh, A. Samouelian, and P. Kabal, "Frame level noise classification in mobile environments," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, Mar. 1999, pp. 237–240.

[12] C. Couvreur, V. Fontaine, P. Gaunard, and C. G. Mubikangiey, "Automatic classification of environmental noise events by hidden Markov models," *Appl. Acoust.*, vol. 54, no. 3, pp. 187–206, 1998.

[13] J. Foote, "An overview of audio information retrieval," *Multimedia Syst.*, vol. 7, pp. 2–10, 1999.

[14] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, Apr. 1997, pp. 1331–1334.

[15] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 504–516, Sep. 2002.

[16] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern Recognit. Lett.*, no. 22, pp. 533–544, 2001.

[17] T. Zhang and C.-C. J. Kuo, *Content-Based Audio Classification and Retrieval for Audiovisual Data Parsing*. Norwell, MA: Kluwer, 2000.

[18] M. Casey, "Generalized sound classification and similarity in MPEG-7," *Org. Sound*, vol. 6, no. 2, 2002.

[19] V. Peltonen, "Computational auditory scene recognition," M.S. thesis, Dept. Inf. Technol., Tampere Univ. Technol., Tampere, Finland, 2001.

[20] N. Zacharov and K. Koivuniemi, "Unraveling the perception of spatial sound reproduction: Techniques and experimental design," presented at the Audio Eng. Soc. 19th Int. Conf. Surround Sound, Techniques, Technology and Perception, Jun. 2001.

[21] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[22] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall, 1993.

[23] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, May 1999.

[24] I. Potamitis, N. Fakotakis, and G. Kokkinakis, "Independent component analysis applied to feature extraction for robust automatic speech recognition," *Electron. Lett.*, vol. 36, no. 23, Nov. 2000.

[25] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[26] A. Ben-Yishai and D. Burshtein, "A discriminative training algorithm for hidden Markov models," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 3, pp. 204–217, May 2004.

[27] "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," Int. Telecommun. Union Radiocommun. Assembly, ITU-R, Recommendation BS.1116-1, 1997.

[28] "Presentation" [Software]. Neurobehavioral Systems. [Online]. Available: http://www.neurobehavioralsystems.com/software/presentation/

[29] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR dummy-head microphone," Perceptual Computing Group, MIT Media Lab, Cambridge, MA, Tech. Rep. 280, 1994.



**Antti J. Eronen** was born in Ilomantsi, Finland, in 1977. He received the M.S. degree in information technology from the Tampere University of Technology (TUT), Tampere, Finland, in 2001. He is currently pursuing the Ph.D. degree.

From 1998 to 2003, he was with the Institute of Signal Processing, TUT. In 2003, he joined Nokia Research Center, Tampere. His research interests include content recognition, analysis, and synthesis of audio and music.

**Vesa T. Peltonen** was born in Virolahti, Finland, in 1974. He received the M.Sc. degree in information technology from the Tampere University of Technology (TUT), Tampere, Finland, in August 2001.

From 2000 to 2002, he was a Researcher with the Digital Media Institute, TUT. Since 2002, he has been with Nokia Mobile Phones, Tampere.

**Juha T. Tuomi** was born in Seinäjoki, Finland, in 1979. He is currently pursuing the M.S. degree in audio signal processing at the Tampere University of Technology (TUT), Tampere, Finland, with a thesis on auditory context tracking.

He joined the Institute of Signal Processing, TUT, in 2001, and has since been working on auditory context awareness. His principal focus is robust auditory context transition detection. His other research interests include mobile context awareness and the human perception of auditory contexts.

**Anssi P. Klapuri** was born in Kälviä, Finland, in 1973. He received the M.Sc. degree in information technology and the Ph.D. degree from the Tampere University of Technology (TUT), Tampere, Finland, in June 1998 and April 2004, respectively.

He has been with the Institute of Signal Processing, TUT, since 1996. His research interests include audio signal processing and, particularly, automatic transcription of music.

**Seppo Fagerlund** was born in Pori, Finland, in 1978. He is currently pursuing the M.S. degree at the Helsinki University of Technology (HUT), Espoo, Finland, with a thesis on automatic recognition of bird species by their sounds.

In 2002, he was a Research Assistant at Nokia Research Center, Tampere, Finland. In 2004, he became a Research Assistant at the Laboratory of Acoustics and Audio Signal Processing, HUT. His research interests include signal processing of bioacoustic signals and pattern recognition algorithms.

**Timo Sorsa** was born in Helsinki, Finland, in 1973. He received the M.Sc. degree in electrical engineering from the Helsinki University of Technology, Espoo, Finland, in 2000.

In 1999, he joined the Nokia Research Center, Helsinki, where he is currently with the Multimedia Technologies Laboratory. His current research interests include perceptual audio quality, audio content analysis, and audio signal processing.

**Gaëtan Lorho** was born in Vannes, France, in 1972. He received the M.S. degree in fundamental physics from the University of Paris VII, Paris, France, in 1996, and the M.S. degree in sound and vibration studies from the Institute of Sound and Vibration Research, University of Southampton, Southampton, U.K., in 1998.

Since 1999, he has been a Research Engineer at the Nokia Research Center, Helsinki, Finland. His main research interests are in the subjective evaluation of audio quality, spatial sound reproduction, and audio user interfaces.

**Jyri Huopaniemi** (M'99) was born in Helsinki, Finland, in 1968. He received the M.Sc., Lic.Tech., and D.Sc. (Tech.) degrees in electrical and communications engineering from the Helsinki University of Technology (HUT), Espoo, Finland, in 1995, 1997, and 1999, respectively. His doctoral thesis was on the topic of virtual acoustics and 3-D audio.

He was a Research Scientist at the Laboratory of Acoustics and Audio Signal Processing, HUT, from 1993 to 1998. In 1998, he was a Visiting Scholar at the Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, Stanford, CA. Since 1998, he has been with Nokia Research Center, Helsinki, where his current position is Head of Mobile Applications Research. His professional interests include multimedia, software platforms, virtual audio-visual environments, digital audio signal processing, room and musical acoustics, and audio content analysis and processing. He is author or coauthor of over 55 technical papers published in international journals and conferences, and he has been actively involved in MPEG and Java standardization work.

Dr. Huopaniemi is a member of the AES and the Acoustical Society of Finland.