SoundButton: Design of a Low Power Wearable Audio Classification System

Mathias Stäger[1], Paul Lukowicz[1], Niroshan Perera[1],
Thomas von Büren[1], Gerhard Tröster[1], Thad Starner[2]

[1]Wearable Computing Lab, ETH Zürich, Switzerland
[2]Georgia Institute of Technology, Atlanta, USA

Abstract

The paper deals with the design of a sound recognition system focused on an ultra low power hardware implementation in a button like miniature form factor. We present the results of the first design phase focused on selection and experimental evaluation of sound classes and algorithms suitable for low power realization. We also present the VHDL model of the hardware showing that our method can be implemented with minimal resources. Our approach is based on spectrum analysis to distinguish between a subset of sound sources with a clear audio signature. It also uses intensity analysis from microphones placed at different locations to correlate the sounds with user activity.

1. Introduction

After vision, sound is the second most important source of information for human beings. The amount of information contained in a sound signal is best illustrated by the fact that blind people can often get around using audio information alone, in many cases developing a near perfect understanding of the situation. In addition, sound has the advantage of manageable data rates and much smaller processing complexity than image recognition.

Related Work To date, the potential of wearable context recognition based on sound has been studied in some detail by two research groups: Auditory scene analysis focused on detecting distinct auditory events and classifying them has been done by MIT's Media Lab [4, 5]. The Audio Research Group at Tampere University, Finland, works on auditory scene recognition, which focuses on recognizing the context or environment, instead of analyzing discrete sound events [13]. In addition, the classification of sound types has been investigated in the context of hearing aid improvements [2].

Paper Contributions and Organization The sound recognition work presented in this paper is part of our ongoing research on using arrays of simple, ultra low power sensor nodes distributed over the user's body for context recognition. Our final objective is the design of the *SoundButton*: an ultra low power, miniaturized sound recognition device that could be integrated into the user's clothing, watch or other accessory and could operate for weeks on a small battery or even from energy extracted from the environment.

In this paper, we report the results of the first phase of the sound button design. As our focus is on low power, we begin with the selection of sounds and algorithms that provide information not accessible with other sensors while allowing reliable recognition with minimum computational resources. This includes a novel method for detecting sounds caused by the user's hand through intensity analysis of signals from SoundButton devices placed on different locations of the body. We then present the results of a experimental analysis based on realistic scenarios to verify our recognition method and to optimize our algorithms for the best tradeoff between recognition accuracy and computational complexity. Finally, we describe the VHDL design of the corresponding recognition hardware showing that our method can be implemented with minimal resources.

2. Design Considerations

The SoundButton is meant to be part of a wearable context recognition system illustrated in Fig. 1. A number of miniaturized sensor nodes are invisibly integrated into the users outfit, where they can best extract the relevant information (e.g. motion sensors on different limbs, external light sensors on the shoulders, etc.) [8, 11]. The sensors provide context information wirelessly to a central wearable computing node which is also responsible for sensor configuration and control. Ideally, the sensor nodes should be fully autonomous operating for months or years on a miniature battery or even better extracting the energy from the environment. Therefore, the power consumption is a central topic in our work. This has two implications. First the sensor nodes may need to perform certain amount of local processing, to minimize the amount of data that needs to be transmitted (wireless transmission is much more energy consuming than computation). This is particularly impor-

Proceedings of the Seventh IEEE International Symposium on Wearable Computers (ISWC'03)
1530-0811/03 $ 17.00 © 2003 IEEE

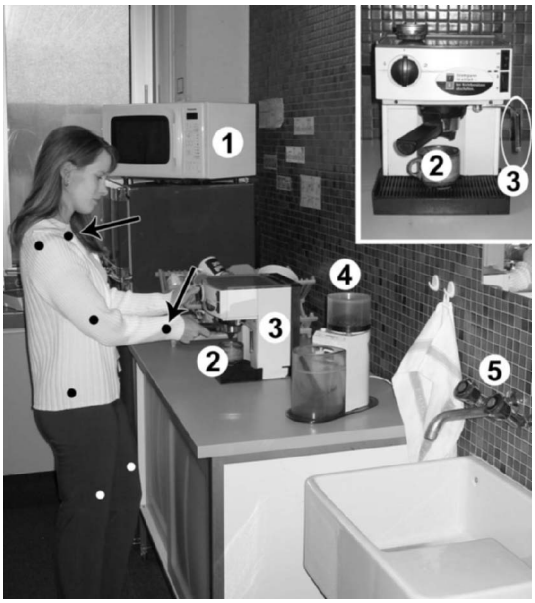

Figure 1. Kitchenette scenario: Dots indicate sensors (SoundButtons with arrows), numbered circles indicate household appliances

tant for sound sensors due to the high sampling rate. Secondly, for the individual sensors the simplicity of the processing algorithms running on the sensor nodes needs to be given priority over the accuracy and generalization capability.

Initial Constraints Regarding the second implication, we constrain ourselves to sounds and algorithms that don't require continuous operation of the sounds sensors and can be used with a low duty cycle (approx. 5%). The relevance and limitations of this approach will be discussed in section 3.3.

The sounds we are concentrating on are *dominant, quasi stationary* in a *known environment*.

By dominant, we mean that the sound in question is the loudest sound source received by the system. Thus the recognition does not have to deal with separating the relevant signal from background noise.

Stationary refers to the temporal evolution and implies that the sound is essentially constant over time. This means that, except for windowing effects, the spectrum of the sound is identical in all time slots regardless of their position and length. Sound classification is thus reduced to pattern matching of the spectrum acquired from an arbitrary sample window. Neither signal segmentation nor time series analysis of the different phases of a sound (such as the phonemes of a spoken word) are required.

We speak of quasi stationary sounds because most relevant sounds have negligible initial and terminal phases and have a fairly long (at least about a second) main phase that can be described as an essentially stationary sound with added noise. The departure from strictly stationary sounds means that instead of having exactly identical spectra, different time windows from the same signal will have similar, but slightly varying spectral signatures.

Finally, by known environment, we mean that other sources of information are used to constrain the number of sounds that we have to discriminate in any given situation.

3. Algorithms

3.1. Sound Classification

The task of sound recognition can be divided into 3 subproblems: (1) feature generation, (2) dimensionality reduction through feature selection and (3) the actual classification. For each phase, we compared different algorithms to attain an optimal tradeoff between recognition rate and computational complexity.

Features Targeting sounds that remain stationary for a time period in the range of at least one second means that we do not have to perform a continuous acquisition. Instead, short frames or time windows t_w can be periodically recorded and analyzed to reduce the duty cycle and power consumption of the system. As with all parameters, the optimal value for t_w was determined empirically (see section 4.1). To make the system more robust to noise and varying distance between microphone and the sound source, the samples were normalized. The first set of features consists of the magnitude of half of the FFT components $|F[0]| \ldots |F[\frac{N}{2} - 1]|$, retrieved from a N-point Fast Fourier Transformation (FFT). The second set of features consists of features which have been used by other groups for specific, complex recognition problems, in particular, speech recognition [9, 10, 13]. In the time domain the following 3 features were calculated: zero crossing rate, fluctuation of amplitude and 90%-10% width of amplitude histogram distribution. The remaining 5 features were applied in the frequency domain: frequency centroid, bandwidth, spectral roll-off point, fluctuation of amplitude-spectra and band energy ratio in 4 logarithmically divided subbands. As a third set 6 cepstral coefficients (CEP) were added to the set of 8 audio features.

Feature Selection Three different feature selection methods were investigated: Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA) [6] and a correlation matrix which allows to discard all correlated features and to keep only non-correlated features. Additionally, we also investigated the possibility to classify the features directly, without reducing the dimensionality first. Since we only need to recognize a limited number of sounds per location, the LDA is applicable in our case. Therefore, in the 'training phase' a transformation matrix can be calculated with the class-dependent transformation form of the LDA [1].

Classification Two different classifiers were evaluated to classify the features that were selected in the previous step: A k-nearest neighbor (k-nn) classifier (with k from 1 to 10) and a nearest class center classifier were used. In the latter case, the mean values of each class are used as a class center. A test point is assigned to the class associated with the minimum Euclidean distance to the class center.

Combining Features, Feature Selection and Classifiers Tab. 1 gives an overview of the 13 different combinations of features and feature selectors we used to determine the optimal recognition method. All 13 combinations were classified using both the nearest class center classifier and the k-nn classifier. Fig. 2 and Fig. 4 give some examples of the resulting recognition rates. The 13 pairs of bars correspond to the 13 different combinations, the brighter bars indicate the results of the nearest class center classifier while the darker bars indicate the maximum of the k-nn classification result (for k from 1 to 10).

Table 1. Features and feature selectors

No.	Feature	Feature selector
1	FFT	LDA
2	FFT	PCA
3	FFT	keep all
4	8 audio features	LDA
5	8 audio features	PCA
6	8 audio features	keep all
7	8 audio features	keep uncorrelated
8	8 audio features	keep uncorrelated + LDA
9	8 audio features + 6 CEP	LDA
10	8 audio features + 6 CEP	PCA
11	8 audio features + 6 CEP	keep all
12	8 audio features + 6 CEP	keep uncorrelated
13	8 audio features + 6 CEP	keep uncorrelated + LDA

3.2. Intensity Analysis with two Microphones

In general, two microphones 1 and 2 placed at different locations on the body will have different distances to the sound source. Thus the signal intensities I_1 and I_2 will be different. Interestingly, since the intensity of a sound signal is inversely proportional to the square of the distance from its source, the ratio of the two intensities I_1/I_2 depends on the absolute distance of the source from the user. Assuming that the sound source is located at a distance d_1 from the first microphone and $d_1 + \delta$ from the second, the ratio of the intensities is proportional to

$$\frac{I_1}{I_2} \propto \frac{(d_1+\delta)^2}{d_1^2} = \frac{d_1^2 + 2d_1\delta + \delta^2}{d_1^2} = 1 + \frac{2\delta}{d_1} + \frac{\delta^2}{d_1^2}$$

For sound sources located far from both microphones (and thus from the user), d_1 will be much larger then δ (since δ can not be larger than the distance between the body locations on which the microphones are placed). As a consequence, the quotient will be close to 1. On the other hand, if the source is very close to the first microphone we have in general $d_1 < \delta$ and with it $I_1/I_2 \gg 1$.

Thus putting the first microphone on the wrist and the second one on the chest, we can use a large quotient as a sign that the sound was generated close to the user's hand. Often this means that the sound was caused by something that the user did with his hand. In terms of computing complexity, the calculation of the intensity is free since its computation (sum of the squares of all samples) is included in the normalization of the audio samples (section 3.1): The relation between the RMS (Root Mean Square), which is used to normalize the N samples of the time window t_w, and the intensity I is given by $RMS = \sqrt{I/N}$.

3.3. Relevance

Restricting the analysis to a few dominant and quasi stationary sounds at a time and relying on few milliseconds samples each second might seem too strict to be useful. However, an analysis of a number of scenarios such as Household Monitoring, Assembly and Maintenance, Office Assistance and Outdoor Guidance has shown that many events occurring in the environment are accompanied by a loud sound that is clearly distinguishable from the background. In addition, the majority of such sounds fall within our definition of quasi stationary (see Tab. 2).

In most cases, other sensors (GPS, inertial navigation, network location) can restrict the users whereabouts to a room or a particular outdoor location. We have found that in most locations there are just a few (between 5 and 10) frequently occurring, relevant sounds. Thus focusing on small groups of sounds is legitimate.

Since we assume the SoundButton to have limited memory it is able to store the transformation matrices and classifier parameters for at most a few sound groups at a time. Fortunately in most everyday situations people tend to spend considerable amount of time at limited set of locations. When at work one would move predominantly between a few offices, the lab, and the cafeteria. Thus it is possible to organize relevant sound groups into sets, with each set corresponding to a certain higher level location and being relevant during a different part of the day. As a consequence at any given time the SoundButton contains only the parameters for the currently relevant sound group set. Whenever there is a change in high level location the corresponding set is downloaded from the central computer. Since in general such a change in high level location will happen only a few times a day, it is not relevant for the overall power consumption.

Interestingly, if we desist from a low power design and start to sample continuously, the intensity analysis with two

microphones helps us to identify even very short sounds that were caused by the user (e.g. opening and closing of a drawer, banging of a door). In [12] we have shown, that with this method and just using FFT, LDA and a simple classifier a high recognition rate of the user's activity can be achieved. Fig. 2 shows that averaging features over several consecutive frames (in this case twenty 53ms frames) helps to improve recognition for 'difficult' classes. The sounds were taken from [7] and contained a street, a restaurant, a lecture scene and a conversation.

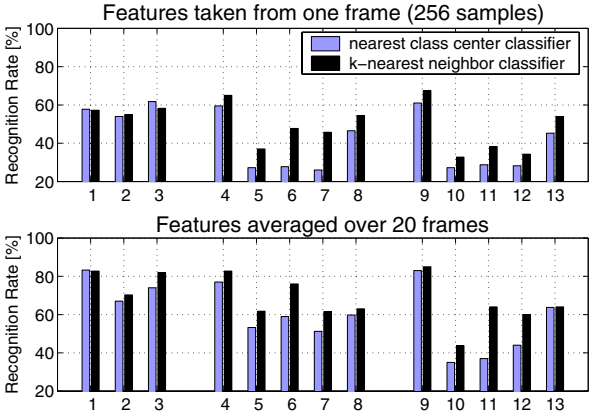

Figure 2. Recognition rates, frame averaging

4. Experiments

To evaluate our approach and to determine the optimal combination of features, feature selectors and classifiers we have performed an experimental analysis on a selected set of sounds. The sounds were chosen to represent typical settings for Household, Maintenance, Office, and Outdoor scenarios as shown in Tab. 2.

For each scenario several 10 to 30 seconds long samples of 5 (4 for the outdoor scenario) relevant sounds were recorded with 16 bit resolution and 48kHz sampling frequency using a Sony microphone (type ECM-T145). Since the user was standing in front of the appliance, the distance from the sound source and the microphone is in the range of 10-50cm for indoor scenarios.

Table 2. Sounds recorded for experiments

Sound group	Sounds
Kitchenette (Household)	microwave, coffee grinder, coffee maker, hot water nozzle, water from water tap
Office	printer, copy machine, telephone, typing on keyboard, ventilator
Workshop (Maintenance)	sawing, drilling, hammering, grinding, filing
Outdoor	inside tram and bus, passing cars, raindrops on umbrella

4.1. Recognition Experiments with one Microphone

Parameter Optimization The sound samples were used to determine how the three parameters crucial for low power consumption; length of the sampling window t_w, sampling rate f_s and resolution can be reduced without incurring an unacceptable penalty on the recognition rate. t_w was varied between 10ms and 110ms, the sampling frequency f_s was changed between 1kHz and 10kHz by resampling the 48 kHz recording. After resampling, samples were converted to 8 bit resolution.

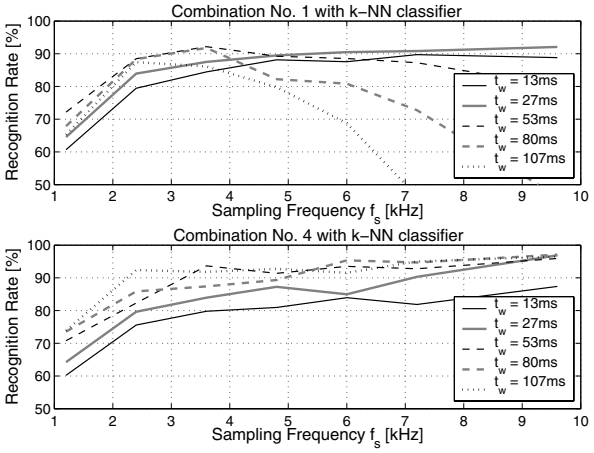

Figure 3. Recognition rate of outdoor sounds as function of f_s and t_w

An investigation of a selection of combinations from Tab. 1 (see Fig. 3 for two examples) revealed that a sampling rate around 5kHz still gives a good recognition rate. For later experiments we used $f_s = 4.8$kHz and 256 samples, which results in $t_w = 53.3$ms. In terms of resolution, it was found that there was hardly any difference in recognition rates between the 16 and the 8 bit signals.

Recognition Rates Fig. 4 shows the recognition rates of all sound groups using the extracted parameters. To further validate our approach all 19 sounds were classified together.

From Fig. 4 it can be concluded that a simple recognition method with just FFT components as features, an LDA matrix transformation for dimensionality reduction and an nearest class center classifier (first bar on the left) gives sufficient high recognition rates. Moreover, since some of the other features are not well adapted to our problem, this method results in some of the highest recognition rates. In terms of complexity, this algorithm is one of the best for a low power implementation.

4.2. Experiments with two Microphones

To evaluate the feasibility of using intensity differences between two microphones, the workshop and kitchenette

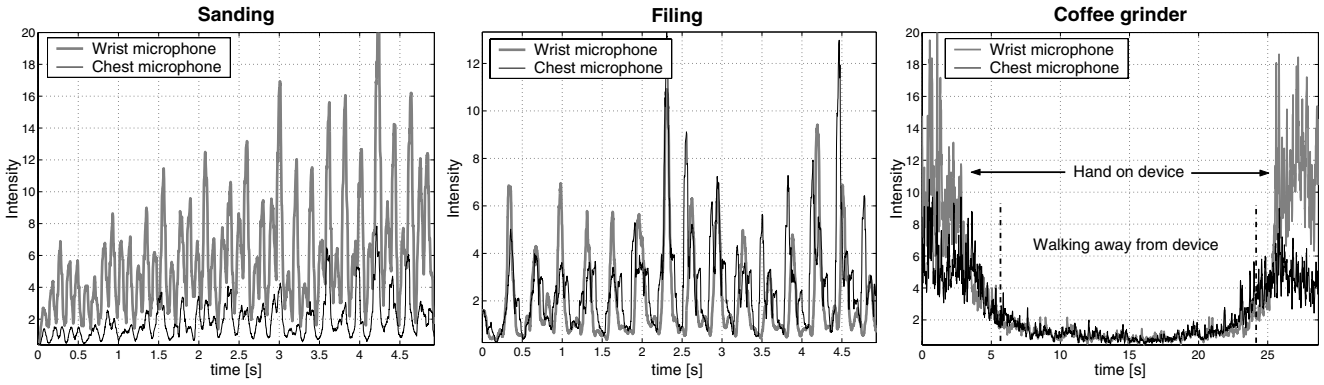

Figure 5. Sound intensity summed over a 51.2 ms sliding window for three different sounds

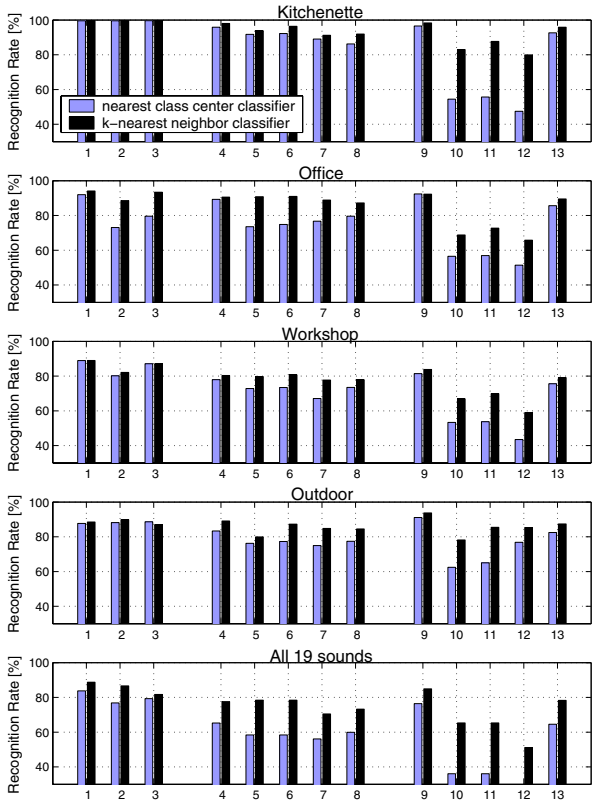

Figure 4. Recognition rates

sounds were recorded using two mono microphones: one worn on the wrist and the other on the chest.

The two scenarios were selected because they are representative of two slightly different situations: (1) the user directly causing a sound through a certain motion of his hand (by using a tool e.g. sawing), and (2) the user being next to the appliance he is operating, possibly having his hand on the switch activating/deactivating it (e.g. switching on the coffee grinder). In all cases, the signal intensity was summed over a sliding 51.2 ms window.

Except for filing, it has been found that in the workshop sounds the sound intensity of the wrist microphone is more than twice the intensity of the chest microphone. As an example, a plot of the sliding window intensity for the sound caused by smoothing a surface with sand paper is shown in the leftmost part of Fig. 5. Since the user's hand is directly on the source of the sound the intensity difference is large.

In the filing example (Fig. 5 middle), the intensity analysis doesn't work well (although a trend can be seen) for two reasons: first, the user's hand shielded the sound to the wrist microphone and second, the user was bent over the workbench bringing the chest microphone within the same distance to the sound as the wrist microphone.

For the coffee grinder in the kitchenette, it has been found that the intensity on the wrist microphone is up to two times larger when the hand was on the switch, becoming equal as soon as the hand was removed and falling on both microphones as the user was moving away (Fig. 5 right).

5. SoundButton Hardware

As a first step towards the implementation of the SoundButton, we have implemented the digital signal processing part in VHDL code. This provides a first estimate of the system complexity and can be used as a basis for power consumption estimation. At a later stage, the VHDL code will be used to implement an application specific integrated circuit (ASIC). The outline of the system is shown in Figure 6. The components outside the box (transceiver and on chip MEMS microphone [14]) will be integrated with the ASIC in a high density electronic package.

All data paths of the ASIC are 16 bit and ALUs (arithmetic-logic units) use floating point arithmetic. The VHDL model was simulated using ModelSim and functional units were synthesized using Synopsys. In the following we give a brief description of the functional units:

The input audio stream is sampled at 5 kHz with a 8 bit **A/D converter**. A successive approximation A/D converter is chosen because of its low power consumption and aver-

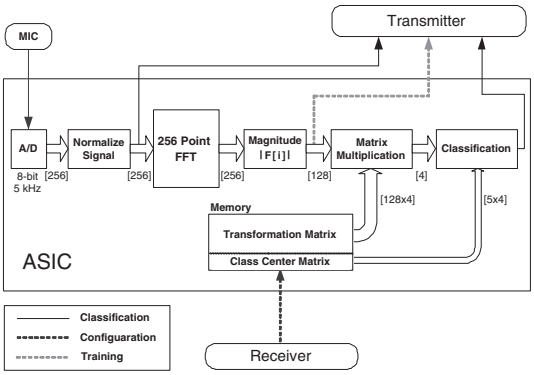

Figure 6. Diagram of the SoundButton including functional units of the ASIC

age hardware complexity. Since we found out that the same recognition rate can be achieved with 8 bit or 16 bit audio signals we preferred a less power consuming 8 bit A/D converter over a 16 bit version, even though the rest of the data paths are 16 bit. The sampled audio signals are stored in memory. For **signal normalization** the RMS of all the samples in the time window is computed and then each sample is divided by the RMS. The normalized samples are written back to memory. The first stage of the RMS calculation also provides the signal intensity (see section 3.2). The **FFT** unit reads the normalized audio samples from memory and performs a 256-point FFT. The FFT core uses a bit parallel radix 2 butterfly [3]. **Magnitudes** are computed only for the first half of the FFT outputs (128 out of 256). Real and imaginary parts of each output are squared, added and then its square root is computed. Then the 128 element magnitude vector is multiplied with the **transformation matrix** (128x4) to generate the feature vector. In the **classification** stage, the feature vector is used to calculate the Euclidean minimum distances to the class centers. The **memory** is a 10kByte SRAM (Static RAM) which holds the transformation and class center matrices.

6. Conclusion

We have shown that simple algorithms optimized for low power implementation can be used to derive important context information from the sound signal. This includes the recognition of different sound classes as well as the use of distributed microphones to correlate sounds with user hand activity. The paper has also demonstrated how power consumption considerations can be included in the entire design process of a recognition system, starting with scenario analysis, through algorithm selection to hardware design.

Initial studies based on our VHDL design and literature values for the power consumption of wireless transmission systems indicate that a device that can perform signal acquisition, preprocessing, feature extraction, classification and wireless result transmission with an expected power consumption of less then 100μW is feasible.

Exact estimation of the power consumption through gate level simulations is the next step in our work to be followed by putting the VHDL design into hardware and perform experimental measurements.

Acknowledgements

The authors would like to thank Bernt Schiele and Nicky Kern, from the Perceptual Computing and Computer Vision group, ETH Zurich, for the input given in discussions.

References

[1] S. Balakrishnama, A. Ganapathiraju, and J. Picone. Linear discriminant analysis for signal processing problems. In *Proceedings of IEEE Southeastcon*, pages 78–81, 1999.

[2] M. C. Büchler. *Algorithms for Sound Classification in Hearing Instruments*. PhD thesis, ETH Zurich, 2002.

[3] E. Cetin, R. C. Morling, and I. Kale. An integrated 256-point complex FFT processor for real-time spectrum analysis and measurement. In *IEEE Proc. of Instrumentation and Measurement Technology Conf.,*, pages 96–101, May 1997.

[4] B. Clarkson and A. Pentland. Extracting context from environmental audio. In *ISWC'98*, pages 154–155, Oct. 1998.

[5] B. Clarkson, N. Sawhney, and A. Pentland. Auditory context awareness in wearable computing. In *Workshop on Perceptual User Interfaces*, Nov. 1998.

[6] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2001.

[7] N. Kern and B. Schiele. Context-aware notification for wearable computers. In *ISWC'03*, Oct. 2003.

[8] N. Kern, B. Schiele, H. Junker, P. Lukowicz, and G. Tröster. Wearable sensing to annotate meeting recordings. In *ISWC'02*, pages 186–193, Oct. 2002.

[9] D. Li, I. Sethi, N. Dimitrova, and T. McGee. Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, 22(5):533–544, 2001.

[10] S. Z. Li. Content-based audio classification and retrieval using the nearest feature line method. *IEEE Transactions on Speech and Audio Processing*, 8(5):619–625, Sept. 2000.

[11] P. Lukowicz, H. Junker, M. Stäger, T. von Büren, and G. Tröster. WearNET: A distributed multi-sensor system for context aware wearables. In *Proceedings of the 4th UbiComp*, pages 361–370, Sept. 2002.

[12] P. Lukowicz, J. Ward, H. Junker, M. Stäger, G. Tröster, A. Atrash, and T. Starner. Recognizing workshop activity using body worn microphones and accelerometers. In *submitted to ICCV2003: 9th Int'l Conf. on Computer Vision*, Oct. 2003.

[13] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa. Computational auditory scene recognition. In *IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, volume 2, pages 1941–1944, May 2002.

[14] P. Rombach, M. Mullenborn, U. Klein, and K. Rasmussen. The first low voltage, low noise differential silicon microphone, technology development and measurement results. In *14th IEEE Int'l Conf. on Micro Electro Mechanical Systems*, pages 42–45, 2001.