EURASIP Journal on Applied Signal Processing 2005:18, 2991–3002
© 2005 Michael Büchler et al.

Sound Classification in Hearing Aids Inspired by Auditory Scene Analysis

Michael Büchler
ENT Department, University Hospital Zurich, CH-8091 Zurich, Switzerland
Email: michael.buechler@usz.ch

Silvia Allegro
Phonak AG, CH-8712 Staefa, Switzerland
Email: silvia.allegro@phonak.ch

Stefan Launer
Phonak AG, CH-8712 Staefa, Switzerland
Email: stefan.launer@phonak.ch

Norbert Dillier
ENT Department, University Hospital Zurich, CH-8091 Zurich, Switzerland
Email: norbert.dillier@usz.ch

Received 29 April 2004; Revised 16 November 2004

A sound classification system for the automatic recognition of the acoustic environment in a hearing aid is discussed. The system distinguishes the four sound classes “clean speech,” “speech in noise,” “noise,” and “music.” A number of features that are inspired by auditory scene analysis are extracted from the sound signal. These features describe amplitude modulations, spectral profile, harmonicity, amplitude onsets, and rhythm. They are evaluated together with different pattern classifiers. Simple classifiers, such as rule-based and minimum-distance classifiers, are compared with more complex approaches, such as Bayes classifier, neural network, and hidden Markov model. Sounds from a large database are employed for both training and testing of the system. The achieved recognition rates are very high except for the class “speech in noise.” Problems arise in the classification of compressed pop music, strongly reverberated speech, and tonal or fluctuating noises.

Keywords and phrases: hearing aids, sound classification, auditory scene analysis.

1. INTRODUCTION

It was shown in the past that one single setting of the frequency response or of compression parameters in the hearing aid is not satisfying for the user. Kates [1] presented a summary of a number of studies where it was shown that different hearing aid characteristics are desired under different listening conditions. Therefore, modern hearing aids provide typically several hearing programs to account for different acoustic situations, such as quiet environment, noisy environment, music, and so forth. These hearing programs can be activated either by means of a switch at the hearing aid or with a remote control. The manual switching between different hearing programs is however annoying, as the user has the bothersome task of recognizing the acoustic environment and then switching to the program that best fits this situation. Automatic sensing of the current acoustic situation and automatic switching to the best fitting program would therefore greatly improve the utility of today’s hearing aids.

There exist already simple approaches to automatic sound classification in hearing aids, and even though today their performance is not faultless in every listening situation, a field study with one of these approaches has shown that an automatic program selection system in the hearing aid is appreciated very much by the user [2]. It was shown in this study that the automatic switching mode of the test instrument was deemed useful by a majority of test subjects (75%), even if its performance was not always perfect. These results were a strong motivation for the research described in this paper.

This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

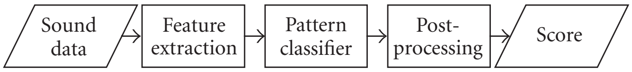


Figure 1: Basic structure of a sound classification system comprising feature extraction, classification, and postprocessing.

There are several commercially available hearing aids which make use of sound classification techniques. Most existing techniques are employed to control noise cleaning means (i.e., noise canceller and/or beamformer). In an approach that is based on an algorithm by Ludvigsen [3], impulse-like sounds are distinguished from continuous sounds by means of amplitude statistics. Ludvigsen states that the amplitude histogram of more or less continuous signals, like background noise and certain kinds of music, shows a narrow and symmetrical distribution, whereas the distribution is broad and asymmetric for speech or knocking noises.

Ostendorf et al. [4] propose a system in which the three sound classes "clean speech," "speech in noise," and "noise without speech" are distinguished by means of modulation frequency analysis. Due to the speech pauses, the modulation depth of speech is large, with a maximum at modulation frequencies between 2 and 8 Hz. By way of contrast, noise shows often weaker but faster modulations and has therefore its maximum at higher modulation frequencies. Ostendorf found that clean speech is very well identified on the basis of the modulation spectra, while noise and speech in noise are confused more often.

A sound classification is also described by Phonak [5]. The algorithm is based on the analysis of the temporal level fluctuations and the form of the spectrum as originally proposed by Kates [1]. Kates used the algorithm for the classification of some everyday background noises, whereas Phonak exploited it to reliably distinguish speech in noise signals from all other sound kinds.

In an approach of Nordqvist [6], the sound is classified into clean speech and different kinds of background noises by means of linear prediction coefficients (LPC) and hidden Markov models. Feldbusch [7] identifies clean speech, speech babble, and traffic noise by means of various time- and frequency-domain features and a neural network.

All of the above mentioned approaches allow a robust separation of clean speech signals from other signals. Music however cannot be distinguished at all, and it is only partly possible to separate noise from speech in noise.

Another application of sound classification which has recently gained importance is the automatic data segmentation and indexing in multimedia databases. For example, Zhang and Kuo [8] describe a system where the audio signal is segmented and classified into twelve essential scenes using four signal features and a rule-based heuristic procedure. Signals containing only one basic audio type (e.g., pure speech or music) were classified robustly, whereas, for hybrid sounds (e.g., speech in noise or singing), misclassification occurred more often. Mixtures of sounds, however, are characteristic of many everyday listening situations. For hearing aid users, especially the situation "speech in noise" is a critical situation, a class that was not included by Zhang and Kuo.

Other typical sound classification systems operate usually on much less universal target signals than the above mentioned applications. Examples of such systems are the recognition of different music styles [9, 10] and the identification of different instruments [11], the differentiation of speech and music signals [12], or the classification of different noise types [13] or alarm signals [14]. Some of these algorithms try to identify classes that contain only one distinct sound, such as a barking dog or a flute tone, and are therefore on a much more detailed layer than is initially desired for an application in hearing aids. The sound classes that are important for hearing aid users typically contain many different sounds, such as the class "music," which consists of various music styles. Nevertheless, concepts from these algorithms may be used for a more specific classification in the future.

The objective of the work described in this paper is the detection of the general classes "speech," "noise," "speech in noise," and "music." With the algorithms mentioned above, a robust recognition of the class "speech in noise" was not possible, and none of the approaches for hearing aids allows to classify music so far. Thus, a combination of existing features with features that are inspired by auditory scene analysis is performed to achieve a robust classification system. These features are evaluated together with different types of pattern classifiers.

2. SOUND CLASSIFICATION INSPIRED BY AUDITORY SCENE ANALYSIS

The basic structure of a sound classification system is illustrated in Figure 1. The classifier separates the desired classes based on the features extracted from the input signal. Postprocessing is employed to correct possible classification errors and to control the transient behavior of the sound classification system. Considering the approaches described in the introduction, it can be stated that in most algorithms, the emphasis lies on the feature extraction stage. Without good features, a sophisticated pattern classifier is of little use. Thus, the main goal is to find appropriate features before evaluating them with different pattern classifier architectures. In order to find such features, it is considered how the human auditory system performs the analysis of an acoustic scene.

2.1. *Auditory scene analysis*

Auditory scene analysis [15] describes mechanisms and processing strategies on which the auditory system relies in the analysis of the acoustic environment. Although this whole process is not yet completely understood, it is known that the auditory system extracts characteristic features from the acoustic signals. The features are analyzed based on grouping rules and possibly also on prior knowledge and hypotheses to form acoustic events. These events are then combined and respectively segregated into multiple sound sources.

The features which are known to play a key role in auditory grouping, the so-called auditory features, are *spectral separation*, *spectral profile*, *harmonicity*, *onsets* and *offsets*, *coherent amplitude* and *frequency variations*, *spatial separation*, and *temporal separation*. For more details on auditory features in particular and auditory scene analysis in general the reader is referred to the literature, for example, Mellinger and Mont-Reynaud [16] or Yost [17].

Note that the auditory system attempts to separate and identify the individual sound sources, whereas sound classification does not necessarily require the separation of the sources. The same is the case for computational models of auditory scene analysis, for example, from Brown and Cooke [18], Ellis [19], or Mellinger and Mont-Reynaud [16]. The aim of these models is to separate sources, rather than to classify them, and they use only little prior knowledge up to now, that is, they simply rely on primitive grouping rules. However, it seems that especially the computation of feature maps in the models can be adapted to gain measures for different signal characteristics, like occurrence of onsets and offsets, autocorrelation for pitch determination, and so forth. In the next section, some of the feature calculation is implemented following these models.

2.2. Features for sound classification

In this approach to sound classification, the aim is to mimic the human auditory system at least partially by making use of auditory features as known from auditory scene analysis. So far, four auditory feature groups are used: amplitude modulations, spectral profile, harmonicity, and amplitude onsets.

Amplitude modulations are characteristic of various natural sound sources, and they differ in strength and frequency for many of these sources. They are described in three different ways here in order to later evaluate those features that perform best for sound classification purposes.

The amplitude histogram of the sounds can be modeled by means of percentiles. The *width* of the amplitude histogram is used to characterize the modulation depth in the signal. This concept is illustrated in Figure 2. A similar kind of amplitude statistics was already used by Ludvigsen [3] for the differentiation of impulse-like sounds from continuous sounds.

The amplitude modulations might also be determined in a similar way as described by Ostendorf et al. [4]. The modulation spectrum of the signal envelope is calculated in three modulation frequency ranges: 0–4 Hz, 4–16 Hz, and 16–64 Hz. A value for the modulation depth of each of the three channels is obtained, the modulation features $m1$, $m2$, and $m3$ (Figure 3).

Furthermore, the approach of Kates [1] is chosen for the description of the amount of level fluctuations. The mean level fluctuation strength MLFS is defined as the logarithmic ratio of the mean to the standard deviation of the magnitude in an observation interval. It was approximated with the following formula:

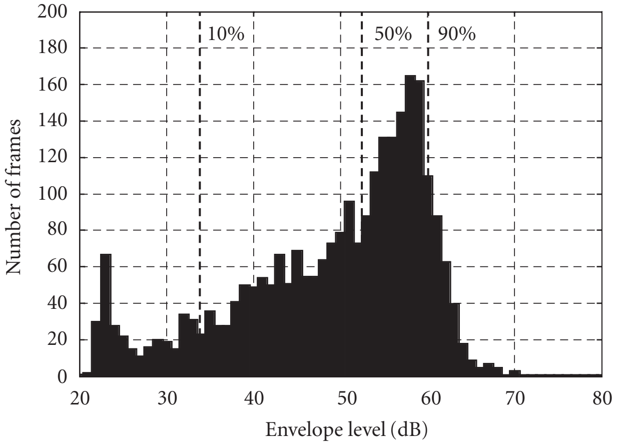


Figure 2: Envelope histogram and percentiles for 30 seconds of clean speech. The amplitude modulations are characterized by means of the amplitude statistics. The amplitude histogram (bar graph) is approximated by several percentiles (dashed lines). The width of the amplitude histogram is defined as the difference of the 90% and 10% percentiles. For clean speech, the width becomes large, whereas, for continuous signals like background noise, the width remains small.

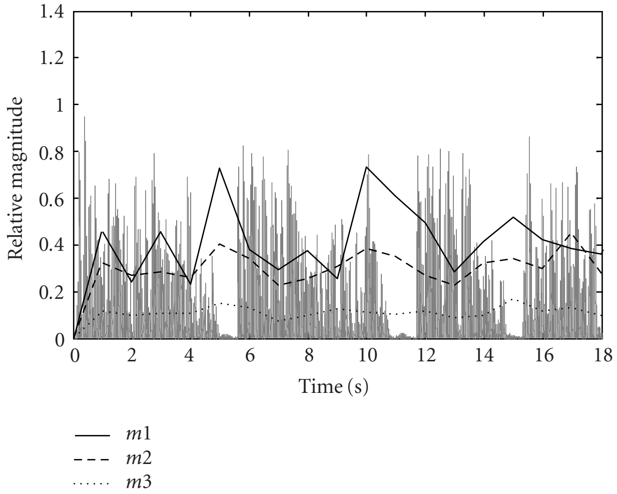


Figure 3: Description of amplitude modulations by extraction of the modulation depth of the signal envelope in three modulation frequency ranges (0–4 Hz, 4–16 Hz, 16–64 Hz), resulting in the features $m1$ (solid line), $m2$ (dashed line), and $m3$ (dotted line).

$$\begin{aligned}\text{MLFS} &= 10 \cdot \log \frac{E(\text{ObsInterval})}{\text{STD}(\text{ObsInterval})} \\ &\approx \log \frac{\text{MLAV}}{(1/3)(\text{ML\_max} - \text{MLAV})}\end{aligned} \qquad (1)$$

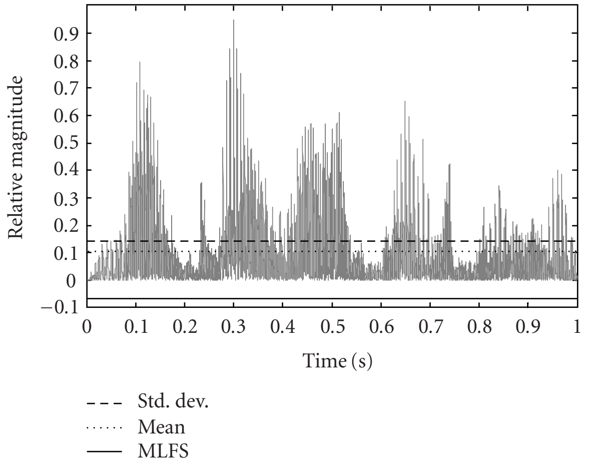


Figure 4: Extraction of amplitude modulations by computation of the mean level fluctuation strength MLFS. Within a time frame of one second, the logarithmic ratio of the mean to the standard deviation is computed. The MLFS gets small or negative values for strongly fluctuating signals like clean speech, and large values for smooth continuous signals like background noise.

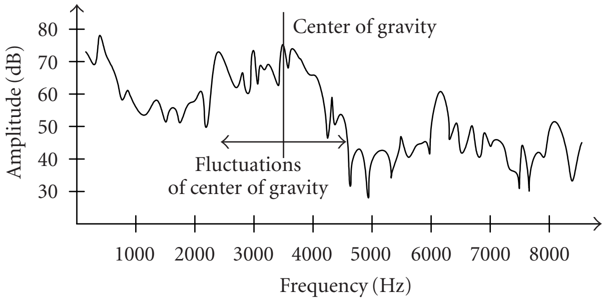


Figure 5: Spectral center of gravity and fluctuations of the spectral center of gravity as employed in the description of the spectral profile.

with

$$\text{MLAV} = \frac{1}{T_{\text{mean}}} \sum_{T_{\text{mean}}} \left(\frac{1}{N} \sum_{n=1}^{N} \log\left(|P(n)| \right) \right),$$
$$\text{ML\_max} = \max_{T_{\text{Mean}}} \left(\frac{1}{N} \sum_{n=1}^{N} \log\left(|P(n)| \right) \right). \tag{2}$$

The mean level average MLAV is calculated out of the sum of the log magnitude P of $N = 20$ Bark bands averaged over a time $T_{\text{mean}} = 1$ second. The standard deviation is approximated by a third of the difference of the maximum and the mean within the observation time T_{mean}, assuming that the amplitude spectrum has a Gaussian distribution, which might not necessarily be the case. The logarithm is calculated because it makes the MLFS more convenient to handle and display. Large values stand for smooth signals, while small or even negative values indicate a signal with large level fluctuations. An example of the determination of the MLFS within a 1-second time frame is given in Figure 4.

The *spectral profile* of a sound can contribute to the classification in that its form may differ for different sound classes, such as for music or noise. Moreover, the shape of the spectrum of most sound sources remains constant as the overall level of the sound is changed. Thus, the auditory system monitors the *relative* differences of the amplitudes of the spectral components as the overall level changes. The spectral profile is modeled here in a very rudimentary way by means of two features, the *spectral center of gravity* CGAV, and the *fluctuations of the spectral center of gravity* CGFS. The CGAV is a static characterization of the spectral profile determined by calculating the first moment of the Bark spectrum and averaging it over an observation interval of $T_{\text{mean}} = 1$ second, with $N = 20$ Barks, k_n FFT bins, and magnitude P in Bark n (using a 128 point FFT at 22 kHz sampling rate):

$$\text{CGAV} = \frac{1}{T_{\text{Mean}}} \sum_{T_{\text{mean}}} \text{CG} \quad \text{with } \text{CG} = \frac{\sum_{n=1}^{N} n \cdot |P(n)| \cdot k_n}{\sum_{n=1}^{N} |P(n)| \cdot k_n},$$
$$k_{1-20} = [1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 5, 6, 7, 9, 12]. \tag{3}$$

The CGFS describes dynamic properties of the spectral profile; it is defined as the logarithmic ratio of the mean to the standard deviation of the center of gravity within the observation interval, equivalent to the calculation of the MLFS:

$$\text{CGFS} = \log \frac{E(\text{CG})}{\text{STD}(\text{CG})} \approx \log \frac{\text{CGAV}}{(1/3)(\text{CG\_max} - \text{CGAV})} \tag{4}$$

with

$$\text{CG\_max} = \max_{T_{\text{Mean}}}(\text{CG}). \tag{5}$$

CGAV and CGFS have already been employed in an earlier sound classifier of Phonak [5]. Figure 5 illustrates the approach.

To describe *harmonicity*, the pitch of the sound is usually employed. Pitch perception is regarded as being an important feature in auditory scene analysis; the existence or the absence of a pitch as well as the temporal behavior of the pitch gives much information about the nature of the signal. The pitch is extracted by calculating a quasi-autocorrelation function as shown in Figure 6, following a simplified algorithm from Karjalainen and Tolonen [20]. If no peak can be detected above a threshold of 20% of the signal energy (0.2 · ACF(0)) and within the range of 50–500 Hz, the pitch frequency is set to 0 Hz. Figure 7 shows the temporal behavior of the extracted pitch of clean speech and classical music. In the present approach, two features are computed that describe the harmonicity: the *tonality* of the sound and the *pitch variance*. The *tonality* is defined as the ratio of harmonic to inharmonic (i.e., pitch frequency = 0 Hz) parts in the sound in an interval of 1 second. The same interval applies for the *pitch variance*. Note that the pitch value itself is not employed as a feature.

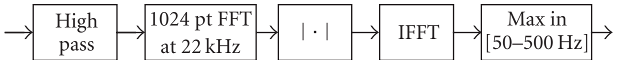
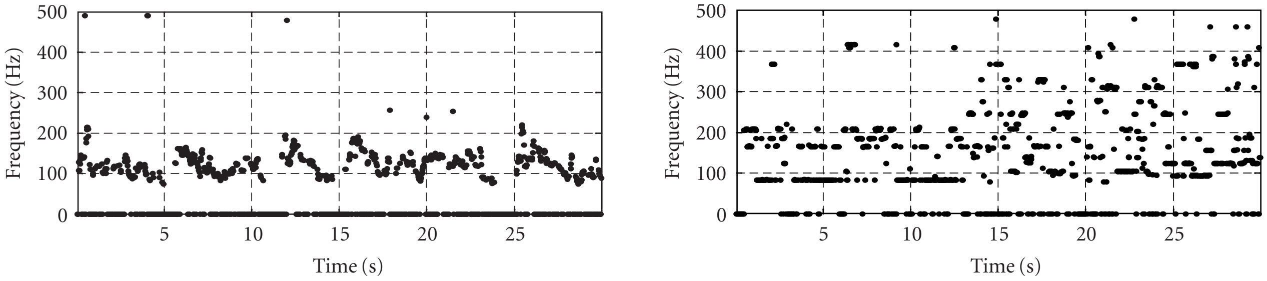


Figure 6: Block diagram of the pitch extractor. A quasi-autocorrelation function is calculated by applying an IFFT to the amplitude spectrum, and the pitch is determined by the maximum within a range of 50–500 Hz.



(a)

(b)

Figure 7: Typical temporal behavior of the pitch frequency for (a) clean speech and (b) classical music. Where a pitch exists, the sound is harmonic; otherwise, where no pitch can be extracted (indicated by 0 Hz), the sound is defined as inharmonic. In speech, the pitch is determined by its prosody, in music by single tones as well as chords. Tonality is the ratio of harmonic to inharmonic periods over time. Pitch variance refers to the harmonic parts of the sound only.

Common *amplitude onsets* of partials are a strong grouping feature in auditory scene analysis. If synchronous onsets of partials occur, they fuse together to one source, whereas asynchronous onsets indicate the presence of more than one source and are therefore used for segregation. Small onset asynchronies between partials are an important contribution to the perceived timbre of the sound source (i.e., musical instrument).

For modeling amplitude onsets, the approach from Brown and Cooke [18] is simplified. The envelope of the signal is calculated in twenty Bark bands with a time constant of 10 milliseconds. This removes the glottal pulses of speech stimuli, but detects fast onsets from plosives. Then, the difference in dB from one frame of 5.8 milliseconds to the next is determined. The outputs of the algorithm are spectrotemporal onset maps as shown in Figure 8 for clean speech and speech in traffic noise. Onsets above 7 dB/frame are displayed as small dots, and above 10 dB/frame as large dots. Four different features are then extracted from the onset map: the mean and the variance of the onset strength in an observation interval of 1 second (*onsetm* and *onsetv*), the number of common onsets across bands (*onsetc*), averaged over the observation interval, and the relation of high to low frequent onsets in the observation interval (*onsethl*).

Finally, the onset maps reveal also some information about the rhythm in the signal. Figure 9 shows the onsets of a pop music sample with a strong beat, which can clearly be seen in the onset pattern. Thus, a feature is extracted from the onset map that describes the strength of the *beat* in the signal. In each Bark band, the onset values are quasi-autocorrelated (similar to the pitch extraction in Figure 6) over a time window of 5.9 seconds (1024 frames of 5.8 milliseconds), and the summary autocorrelation function is then computed (an approach similar to what was chosen by Scheirer [21]). A number of consecutive summary autocorrelation functions are then again summed up to observe a longer time interval of some 30 seconds. This emphasizes beats that remain stable over a longer period of time, assuming that this is more the case for music than for speech. The *beat* feature is then the value of the highest peak of the output within a time interval of 200 to 600 milliseconds (1.7 to 5 Hz, or 100 to 300 bpm).

2.3. Pattern classifiers

After the extraction of feature vectors out of the signal, a decision is made about the class that the signal belongs to. This process is performed in the pattern classifier block. The principle of pattern classification is a mapping from the feature space to a decision space: for every point in the feature space a corresponding class is defined. The borders between the classes are found by performing some sort of training. This is accomplished with a suitable set of sound data. Once the borders are fixed with a set of training sounds, the performance of the classifier is tested with a set of test sounds that is independent of the training set.

There exist a huge number of approaches for pattern classifiers, many of which require quite a lot of computing power and/or memory (for an overview, see, e.g., Schürmann [22], or Kil and Shin [23]). For the application in hearing aids, it is crucial to keep the need for computing time and memory low. Thus, in this paper, the evaluation concentrates on classifiers of low to moderate complexity. Six different classifiers were chosen for this purpose as follows.

A straightforward approach is to define boundaries for every feature itself, that is, some rules are established based

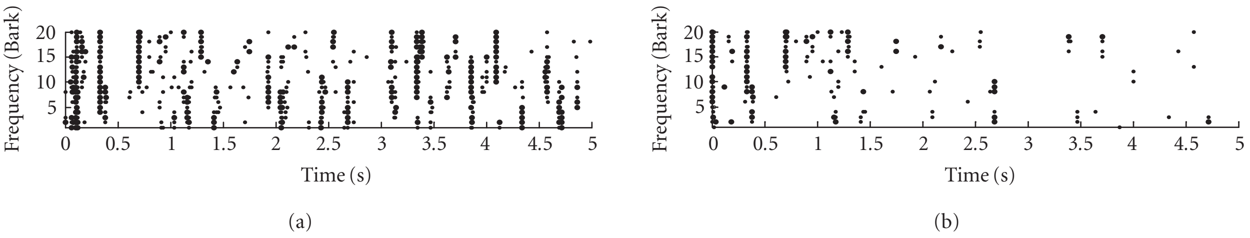


Figure 8: Amplitude onsets in twenty Bark bands for (a) clean speech and (b) speech in traffic noise. Dark areas of the image indicate regions of strong onsets. In speech, many strong onsets occur simultaneously over the bands. If the speech is masked by noise, the onsets are much weaker.

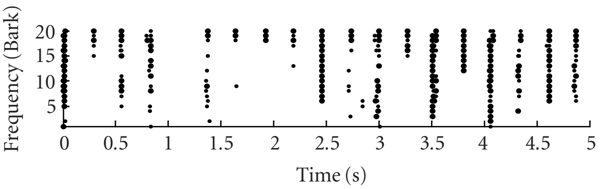


Figure 9: Amplitude onsets in twenty Bark bands for pop music. The strong rhythmic beat in the sample is very well described by the onsets and can be identified by autocorrelating the onsets in each frequency band over a longer period of time.

on the training data and on the a priori knowledge. With such a *rule-based classifier*, the boundaries between the classes are lines orthogonal to the feature axes. For many cases, these straight lines will certainly not be the optimal boundaries. In the present approach, it was tried to choose only few features that allow to divide the feature space in a simple way: the amplitude modulation features $m1$, $m2$, and $m3$ were used to discriminate between speech and other signals, and the two pitch features between music, noise, and speech in noise.

With a *minimum-distance classifier*, the distance of an observation to all classes is measured, and the class with the shortest distance is chosen. For a spherical distribution of the classes, the Euclidean distance is chosen; for a nonspherical distribution, the Mahalanobis distance may give better results, as it takes also the variances of the features into consideration.

The *Bayes classifier* does the classification with the help of histograms of the class-specific probabilities: the class-specific distribution is approximated with multidimensional histograms. Different numbers of histogram intervals were considered to find the optimal number.

The multilayer perceptron, a sort of *neural network*, allows to approximate any discriminant function to arbitrary accuracy. A two-layer perceptron with different numbers of neurons in the hidden layer and sigmoid activation functions was chosen for the evaluation.

Hidden Markov models (HMMs) are a widely used statistical method for speech recognition. One major advantage of HMMs over the previously described classifiers is that they account for the temporal statistics of the occurrence of different states in the features. Typically, LPC are chosen as input of an HMM (see, e.g., Rabiner and Juang [24]). In the present study, however, the features inspired by auditory scene analysis were used. The same basic HMM structure was chosen for all classes, an ergodic HMM with two or more states (ergodic meaning that each state can be reached from any other state).

Finally, a simple two-stage approach was evaluated. The idea of a multistage strategy is to verify the output of a classifier with a priori information of the signal and to correct the classification if necessary. The HMM classifier was used as first stage together with the feature set that had turned out to be optimal (see Section 3). Its output was then verified with a rule-based classifier and a second feature set. This second stage could also be regarded as a special form of Postprocessing.

2.4. Postprocessing

As mentioned earlier, the purpose of postprocessing is to correct possible classification errors and to control the transient behavior of the sound classification system. This is achieved in a very simple manner by observing the classification outcomes over a certain time (e.g., 10 seconds) and taking as a result the class which has occurred most often. By varying the length of the observation interval the transient behavior of the classification result is controlled. However, for the evaluations described in the next section, only the static behavior was tested, and the Postprocessing was left away.

3. EVALUATION OF DIFFERENT CLASSIFICATION SYSTEMS

3.1. Motivation

In the previous section, various features and pattern classifiers have been described, but it was not obvious which combination of features would be optimal with which classifier. For application in hearing aids, “optimal” does not only

Table 1: The four main classes of the sound database and their subclasses.

60 speech	80 speech in noise	80 noise	80 music
40 clean or slightly reverberated	23 in social noise	23 social	19 classic
	7 in the car	7 in the car	19 pop/rock
10 compressed	14 in traffic noise	14 traffic	19 single instruments
10 strongly reverberated	16 in industrial noise	16 industrial	16 singing
	20 in other noise	20 other	7 other

mean to achieve a high recognition rate, but also to keep the need for computing time and memory low. Thus, the goal was to evaluate feature sets and classifiers of different complexity to find a classification system that gives a good score for reasonable computational effort.

3.2. Sound database

The sound database used for the evaluations contains some 300 real-world sounds of 30-second length each, sampled at 22 kHz/16 bit. All of the four desired sound classes ("speech," "speech in noise," "noise," and "music") are represented with various examples (Table 1). The sounds were composed and manually labeled by the authors; they were either recorded in the real world (e.g., in a train station) or in a soundproof room, or taken from other media. The class "speech" is comprised of different speakers speaking different languages, with different vocal efforts, at different speeds, and with different amounts of reverberation and compression. The class "noise" is the most widely varying sound class, comprising social noises, traffic noise, industrial noise, and various other noises such as household and office noises. "Speech in noise" sounds consist of speech signals mixed with noise signals at signal-to-noise ratios (SNRs) between +2 and −9 dB. The class "music," finally, comprises music styles from classics over pop and rock up to single instruments and singing.

3.3. Procedure

For the evaluation, different combinations of the described features and classifiers were considered. If all combinations of the features had been evaluated, it would have resulted in about 2^{14} different feature sets. Thus, an iterative strategy was developed heuristically to find the best feature set, by primarily trying to combine features that describe different attributes of the signal. The strategy followed six steps.

(1) The pitch feature *tonality* was used together with features describing the amplitude modulations (AM), that is, *histogram width*, $m1$, $m2$, $m3$, and MLFS.

(2) The best AM set of step (1) was used without the *tonality*, but together with the second pitch feature, *pitch variance*.

(3) The best set of step (2) was enriched with the spectral features CGAV and CGFS.

(4) The *onset* features were added to the best set of step (3).

(5) The best set of step (4) was reduced in succession by the AM feature(s) and by the spectral feature(s) of steps (1) and (3).

(6) The *beat* feature was added to the best set of steps (4) and (5).

In addition to this iterative approach, the classification was performed with all of the above features. This resulted in about thirty feature sets that had to be processed for each classifier in order to find the optimal combination. Note that this procedure was not performed for the rule-based classifier and for the second stage of the two-stage approach. The structure of these classifiers was explicitly defined by the features that were chosen, and training was performed empirically by observing the distribution of all sounds in the respective feature space and setting the boundaries manually.

In addition, different structures were evaluated for some of the classifiers. For the minimum-distance classifier, both the Euclidean and the Mahalanobis distances were used. In the Bayes classifier, the number of histogram intervals ranged from 5 to 50. The number of hidden neurons in the two-layer perceptron was chosen between 2 and 12. Finally, the number of states in the HMM was changed.

For each sound of 30-second length, classification was calculated once per second, and the class that occurred most frequently was taken as an output for that sound. The dynamic behavior within a sound was not evaluated, but informal tests showed that, for most sounds, classification was either stable (although not necessarily correct) after two or three seconds, or fluctuating between two classes, such as "speech" and "speech in noise."

About 80% of the sounds were used for the training of the classifier, and 20% for testing (for the rule-based classifier, only the test set was used, as it was implicitly trained with a priori knowledge). The sounds for the training and the test sets were chosen at random, and this random choice was repeated 100 times. The actual score was then the mean of the scores of these 100 training and test cycles.

The trained classifier was not only tested with the test data, but also with the training data, in order to check how well the classifier is able to generalize. If the score for the training set is much better than that for the test set, then the classifier is overfitted to the training data; it behaves well for known data, but cannot cope with new data. This can happen when the classifier has many free parameters and only few training data, or when the training data does not represent the whole range of each class homogeneously.

Table 2: Classification results for the six classifiers. For each classifier, the score for the best parameter and the best feature set is given. The simpler approaches achieve around 80% overall test hit rate, which can be improved to some 90% with the more complicated systems.

Classifier type, best parameters	Best feature set	Training set score (%)	Test set score (%)								
		Overall hit rate	Overall hit rate	Speech		Speech in noise		Noise		Music	
				Hit	False	Hit	False	Hit	False	Hit	False
Rule-based	*Tonality, pitch variance, m1, m2, m3*	—	78	79	1.3	67	9.9	88	11.8	78	7.0
Minimum distance, Euclidean	*Tonality, pitch variance, m1, m2, m3, CGFS, onsetv, beat*	84	83	86	3.5	83	11.2	80	7.0	85	1.0
Bayes, 15 intervals	*Tonality, pitch variance, m1, m2, m3, CGFS, onsetm, onsetc*	86	85	90	4.3	83	10.1	84	5.0	82	1.5
Two-layer perceptron, 8 hidden nodes	*Tonality, width, CGAV, CGFS, onsetc, beat*	89	87	86	1.7	86	7.0	89	5.9	87	2.8
Ergodic HMM, 2 states	*Tonality, width, CGAV, CGFS, onsetc, onsetm*	90	88	92	2.2	84	7.0	84	5.3	91	2.2
Two-stage (best HMM and rule-based)	*Tonality, pitch variance, width, CGAV, CGFS, onsetc, onsetm*	—	91	92	1.4	87	5.3	91	3.8	93	2.3

3.4. Results

The performance of each evaluated classification system is expressed by hit and false alarm rates. The hit rate is defined as the percentage of correctly recognized sounds of a particular class; the corresponding false alarm rate is the percentage of sounds which were falsely classified as this class. The results are summarized in Table 2. For each of the classifiers, the best score is shown, that is, the classifier parameter and the feature set for which the overall test hit rate achieves a maximum.

The hit rate of 78% achieved with the simple rule-based approach is not really convincing, but only pitch and amplitude modulation features were needed to get this score. Many sounds of the class "speech in noise" were misclassified. Reverberated and compressed speech was classified either as "music" or "speech in noise," and pop music as "noise." On the other hand, "noise" was never misclassified as "music" or "speech," only as "speech in noise."

The minimum-distance classifier achieved similar results for the Euclidean and the Mahalanobis distances, which indicates that the class distribution in the feature space is fairly ("hyper")spherical. The best hit rate of 83% was obtained using features of each category. The false alarm rates show that many files were misclassified as "speech in noise," especially reverberated speech and cafeteria noises. This is also why the hit rate for "noise" is not so high. The beat feature helped separate rhythmic noises and pop music from the class "speech in noise." There was no danger of overfitting (the difference of the training and test hit rates remained small for all feature sets), because it is not possible to divide the feature space in a complex way with this sort of classifier.

For the Bayes classifier, an optimum was reached with about 15 histogram intervals. More intervals only lead to overfitting. Overfitting could also occur if too many features were used. For the best hit rate of 85%, features of each category except the beat feature were employed. The misclassified sounds were mainly reverberated speech, fluctuating noises, and pop music. They were all regarded as being "speech in noise." Tonal noises, such as a vacuum cleaner, were mostly classified as "music."

The neural network performed best with about 8 hidden nodes and features of each category, achieving a hit rate of 87%. More hidden nodes caused overfitting. Most confusions concerned "noise" and "speech in noise," which included fluctuating noises like cafeteria noise, a passing train, or a weaving machine, and speech in noise with poor SNR. In addition, a few of the reverberated speech sounds were classified as "speech in noise," and a few of the pop music sounds as "noise," especially if they were compressed (i.e., recorded from the radio). Finally, a few tonal noises were regarded as "music."

For the ergodic HMM, it was not possible to use more than two states; otherwise not all parameters had data assigned during training and the training did not converge. One reason for this is certainly that the sounds of a class can be very different, especially regarding their temporal structure. Thus, the best hit rate of 88% was achieved with a two-state HMM and six features of each except the beat category. Using more features, especially the beat feature, only leads to overfitting. Most confusions occurred in the classes "noise" and "speech in noise." Fluctuating noises were often regarded as "speech in noise" and "speech in noise" with poor SNR

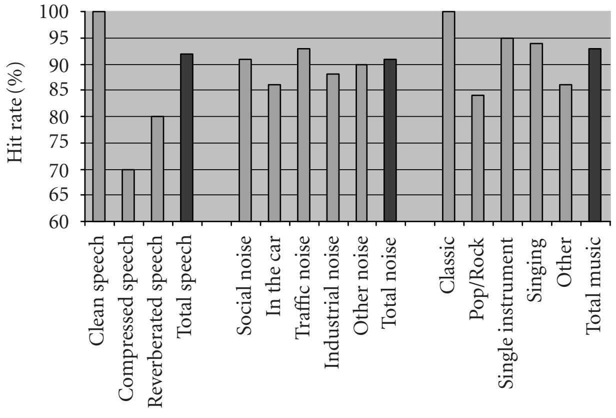


Figure 10: Hit rates of the two-stage classifier for the subclasses of “speech,” “noise,” and “music.” Compressed and reverberated speech as well as pop and rock music are still not robustly classified.

as “noise.” Some compressed speech sounds as well as a few of the compressed pop music sounds were misclassified as “speech in noise.”

In the two-stage approach, we tried to correct some classification errors that the above described HMM system had made. The second rule-based stage considered the classification output of the HMM together with the two pitch and the two spectral features to perform this. The aim was especially to correctly classify compressed or reverberated speech, fluctuating noises and pop music. The highest improvement of 8% was achieved for the class “noise,” because most fluctuating noises that had been misclassified by the HMM could be moved to the correct class in the second stage. Furthermore, at least some of the previously misclassified pop music sounds were classified correctly in the second stage, which is why the false alarm rate of “speech in noise” was slightly improved. However, it was not possible to identify the misclassified compressed or reverberated speech sounds. If a correction had been tried here, many true “speech in noise” sounds would have been moved to the class “speech.” Finally, “speech in noise” with very high SNR was still classified as “speech,” and with very low SNR as “noise.” This, however, is not necessarily wrong. It shows how the boundaries between “speech,” “speech in noise,” and “noise” are somewhat fuzzy.

In Figure 10, the hit rates for the subclasses of “speech,” “noise,” and “music” are also depicted. It can clearly be seen that problems arise in the classification of compressed and strongly reverberated speech as well as pop and rock music.

3.5. Discussion

The single stage approach that performed best was an HMM classifier, followed by the neural network, which performed only slightly worse. The Bayes and minimum-distance classifiers performed a little worse. However, the Bayes classifier could especially be suited if computing time is more limited than memory. If also the memory is restricted, the minimum-distance classifier may be a good choice, because it needs about four times less computing time and memory compared to the HMM or neural network. The rule-based approach might be improved if more features are added, but then it will become difficult to handle. After all, trainable classifiers have been developed so that the training is no longer needed to be done manually. Finally, the HMM score was enhanced by about 3% when a simple rule-based stage was added. This second stage can be regarded as a special form of Postprocessing, or also as a different way of weighting the features (compared to the HMM). This stage especially improved the hit rate for the class “noise” in that many fluctuating noises were then correctly classified.

There is obviously only little temporal information in the features that can be modeled with an HMM. HMMs are commonly more used for the identification of transient sounds, where the HMM states model the onset, the stationary, and the offset parts of a sound (see, e.g., Oberle [14], or Zhang and Kuo [8]). In continuous sounds, as they occur in the sound classes desired in this paper, the states represent different stationary parts that occur in random order, for example, parts with speech and parts with silence, or parts with speech and parts with noise. However, the problem of these sound classes is that the sounds within a class can differ very much (e.g., stationary noises versus impulse-like noise, rock music versus classical music). This means that there might not be a common temporal structure in a class that can be modeled by an HMM.

The features needed for good performance were at least the pitch feature *tonality* and one of the *amplitude modulation* features. When the spectral features CGAV and CGFS as well as one of the *onset* features were added, the score could be further improved. The *beat* feature, however, was of little use.

It is also important to see that the scores are partly a result of the sound database that was used. It was intended to include a great variety of sounds in each class to cover the whole range of the class homogeneously. However, some of these sounds may be quite exotic. A pile driver, for example, is not a noise to which hearing impaired persons are exposed in everyday life. If such sounds are left away, the hit rates will improve. On the other hand, there are everyday sounds that are mostly misclassified, for example, compressed and strongly reverberated speech. How many of these sounds and how many clean speech sounds will be put into the sound database? The hit rate will indeed only be determined by this choice; it will be 100% if only clean speech is taken, and near 0% if only compressed and strongly reverberated speech is used. This example illustrates that classification scores always have to be interpreted with caution.

Another issue related to this is the labeling of the sounds at the time of composing the sound database. Does strongly reverberated speech really belong to the class "speech," or is it already "speech in noise"? Is hard rock music not perceived as being "noise" by some people? And which SNR is the boundary between "speech in noise" and "noise"? If our perception tells us already that the sound does not really sound as it has been labeled, can it be astonishing that its physical properties let the classifier put it into the "wrong" class?

Thus, the choice and labelling of the sounds used influences the classification performance considerably through both training and testing. On the other hand, one and the same signal might be classified differently depending on the context. Speech babble, for example, could either be a "noise" signal (several speakers talking all at once) or a "speech in noise" signal (e.g., a dialog with interfering speakers). Again, the outcome of the classifier in such ambiguous situations depends on the labelling of the sound data.

Ultimately, the perception of a listener also depends on what he wants to hear. For example, in a bar where music plays and people are talking, music may either be the target signal (the listener wants to sit and enjoy) or a background signal (the listener is talking to somebody). This shows the fundamental limitations of any artificial sound classification system. No artificial classifier can read the listener's mind, and therefore there will always exist ambiguities in classification.

4. CONCLUSIONS

For the general classification of the acoustic environment in a hearing aid application, the four main sound classes "speech," "speech in noise," "noise," and "music" will be distinguished. For this purpose, a number of auditory features were extracted from the acoustic signal and classified by means of several pattern classifiers, to assess which combination of features was optimal with which classifier. The employed features describe level fluctuations, the spectral form, harmonicity, onsets, and rhythm.

The results achieved so far are promising. All sound classes except the class "speech in noise" were identified with hit rates over 90%. For "speech in noise" signals, the hit rate was slightly lower (87%). Many sounds of the four classes were very robustly recognized: clean and slightly reverberated speech, speech in noise with moderate SNR, traffic and social noise, and classical music, single instruments, and singing. The misclassified sounds consist of four groups: "speech in noise" with very low or very high SNR, which was classified as "noise" or "speech," respectively, compressed and strongly reverberated speech, a few tonal and fluctuating noises, and compressed pop music, which were all classified as "speech in noise."

During the evaluation of the described sound classification system, insight was gained in how the system could be improved in the future. So far, the performance of the sound classification system has only been tested on sounds from the sound database. One of the next steps should therefore be to evaluate such a system in a field experiment to gain more practical experience. For this purpose, a portable system seems mandatory, as it will enable to carry out the evaluation of the sound classification system in real time. This approach should most probably also provide new ideas about possible optimization strategies.

Taking into account further or ameliorate existing features will be an important aspect for improving and refining the classification. This includes the following.

(a) A better modeling of the spectral profile: so far, the spectral profile has only been modeled in a rudimentary way. It was not possible to describe the tone color of the sound (which contributes to the perceived timbre) in a detailed form. This seems to be a difficult task, because the intraclass variance of the tone color may be very high. Zhang and Kuo [25] analyzed the spectral profile for the classification of some specific environmental sounds, although on a more detailed layer than desired here.

(b) A feature that describes the amount of reverberation in the signal, for example, after approaches presented by Shoda and Ando [26] and Ando et al. [27].

(c) A feature that determines the SNR of the signal, for a more gradual classification of signals containing speech and/or noise.

(d) Spatial features, to analyze where the signal and where the noise come from, or to check both front and back signals on speech content. The latter would, for example, allow to distinguish between "speech signal in speech noise" (speech from the front and from the back), "speech signal in other noise" (speech from the front, noise from the back), and "speech noise only" (speech from the back). Directional microphone and noise reduction in the hearing aid could be set accordingly.

However, we are far from achieving similar performance in a hearing aid as with our auditory system. Today's limitations are on the one hand the ambiguity and context dependence of a large part of the acoustic situations. On the other hand, it is still lacking to understand many of the processes involved in auditory perception. It is striking to realize what complex tasks have to be solved in these processes. However, in contrast to a hearing aid, the human auditory system has a lifetime for training, and it gets also substantial feedback from other senses, such as the visual system. With this, the auditory system can fall back upon invaluable a priori knowledge—the visual system will announce that music will be heard soon even before the orchestra has played a single note.

ACKNOWLEDGMENTS

This work was supported by Phonak AG, Switzerland. Part of it was done as Ph.D. thesis assisted by Professor Dr. Peter Niederer (Institute of Biomedical Engineering, Swiss Federal Institute of Technology Zurich).

REFERENCES

[1] J. M. Kates, "Classification of background noises for hearing-aid applications," *Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 461–470, 1995.

[2] M. Büchler, "How good are automatic program selection features? A look at the usefulness and acceptance of an automatic program selection mode," *Hear. Rev.*, vol. 9, pp. 50–54, 84, 2001.

[3] C. Ludvigsen, *Schaltungsanordnung für eine automatische Regelung von Hörhilfsgeräten* [*An algorithm for an automatic regulation of hearing aids*], Deutsches Patent Nr. DE 43 40 817 A1, 1993.

[4] M. Ostendorf, V. Hohmann, and B. Kollmeier, "Klassifikation von akustischen Signalen basierend auf der Analyse von Modulationsspektren zur Anwendung in digitalen Hörgeräten [Classification of acoustical signals based on the analysis of modulation spectra for the application in digital hearing aids]," in *Fortschritte der Akustik - DAGA '98*, pp. 402–403, Oldenburg, Germany, 1998.

[5] Phonak Hearing Systems, *Claro AutoSelect*, company brochure no. 028-0148-02, 1999.

[6] P. Nordqvist, "Automatic classification of different listening environments in a generalized adaptive hearing aid," in *Proc. International Hearing Aid Research Conference (IHCON '00)*, Lake Tahoe, Calif, USA, August 2000.

[7] F. Feldbusch, "Geräuscherkennung mittels Neuronaler Netze [Noise recognition by means of neural networks]," *Zeitschrift für Audiologie*, vol. 37, no. 1, pp. 30–36, 1998.

[8] T. Zhang and C.-C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 4, pp. 441–457, 2001.

[9] H. Soltau, T. Schultz, M. Westphal, and A. Waibel, "Recognition of music types," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '98)*, vol. 2, pp. 1137–1140, Seattle, Wash, USA, May 1998.

[10] T. Lambrou, P. Kudumakis, R. Speller, M. Sandler, and A. Linney, "Classification of audio signals using statistical features on time and wavelet transform domains," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '98)*, vol. 6, pp. 3621–3624, Seattle, Wash, USA, May 1998.

[11] K. D. Martin and Y. E. Kim, "Musical instrument identification: a pattern-recognition approach," in *Proc. 136th Meeting of the Acoustical Society of America (ASA '98)*, Norfolk, Va, USA, October 1998.

[12] E. D. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '97)*, vol. 2, pp. 1331–1334, Munich, Germany, April 1997.

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

[14] S. Oberle and A. Kaelin, "Recognition of acoustical alarm signals for the profoundly deaf using hidden Markov models," in *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS '95)*, vol. 3, pp. 2285–2288, Seattle, Wash, USA, April–May 1995.

[15] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, Cambridge, Mass, USA, 1990.

[16] D. K. Mellinger and B. M. Mont-Reynaud, "Scene Analysis," in *Auditory Computation*, H. L. Hawkins, T. A. McMullen, A. N. Popper, and R. R. Fay, Eds., pp. 271–331, Springer, New York, NY, USA, 1996.

[17] W. A. Yost, "Auditory image perception and analysis: the basis for hearing," *Hearing Research*, vol. 56, no. 1-2, pp. 8–18, 1991.

[18] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech & Language*, vol. 8, no. 4, pp. 297–336, 1994.

[19] D. P. W. Ellis, *Prediction-driven computational auditory scene analysis*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, Mass, USA, 1996.

[20] M. Karjalainen and T. Tolonen, "Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '99)*, vol. 2, pp. 929–932, Phoenix, Ariz, USA, March 1999.

[21] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.

[22] J. Schürmann, *Pattern Classification: A Unified View of Statistical and Neural Approaches*, John Wiley & Sons, New York, NY, USA, 1996.

[23] D. H. Kil and F. B. Shin, *Pattern Recognition and Prediction with Applications to Signal Characterization*, American Institute of Physics, Woodbury, NY, USA, 1996.

[24] L. Rabiner and B. H. Juang, "Theory and implementation of hidden Markov models," in *Fundamentals of Speech Recognition*, chapter 6, pp. 312–389, Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.

[25] T. Zhang and C.-C. J. Kuo, "Hierarchical classification of audio data for archiving and retrieving," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '99)*, vol. 6, pp. 3001–3004, Phoenix, Ariz, USA, March 1999.

[26] T. Shoda and Y. Ando, "Calculation of speech intelligibility using four orthogonal factors extracted from ACF of source and sound field signals," in *Proc. 16th International Congress on Acoustics and 135th Meeting of the Acoustical Society of America (ICA/ASA '98)*, pp. 2163–2164, Seattle, Wash, USA, June 1998.

[27] Y. Ando, S. Sato, and H. Sakai, "Fundamental subjective attributes of sound fields based on the model of auditory-brain system," in *Computational Acoustics in Architecture*, J. J. Sendra, Ed., chapter 4, pp. 63–99, WIT Press, Southampton, UK, 1999.

Michael Büchler was born in Zurich, Switzerland, in 1967. In 1993, he received a Diploma in electrical engineering from the Swiss Federal Institute of Technology, Zurich. Thereafter, he joined the R&D Department of Feller, a Swiss company for electrical house installations. In 1997, he switched over to the signal processing group of Phonak hearing systems, Switzerland. From 1998 to 2002, he completed a Ph.D. thesis about algorithms for sound classification in hearing aids, also in collaboration with Phonak. Since then he is affiliated with the Laboratory for Experimental Audiology at the University Hospital Zurich, Switzerland, currently doing research in music perception for cochlear implants. His research interests are in the field of signal processing for hearing aids and cochlear implants.

Silvia Allegro holds a Diploma in electrical engineering from the Swiss Federal Institute of Technology Zurich, Switzerland, (1992), an M.S. degree in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, NY (1994), and a Ph.D. degree in technical sciences from the Swiss Federal Institute of Technology Lausanne, Switzerland (1998). During her studies she has been working on research topics in robotics, image processing, and automatic microassembly. In 1998, she joined the Research and Development Department of Phonak hearing systems in Switzerland, where she worked on various signal processing topics for hearing instruments. One of her main areas of activities is the intelligence in modern hearing aids, and in particular the automatic classification of the acoustic environment. Nowadays she manages Phonak's research activities in signal processing.

Stefan Launer is the Director of Research and Technology at Phonak hearing systems, Switzerland. He was born in Würzburg, Germany, in 1966. He studied physics at the University of Würzburg from November 1986 to November 1991. Afterwards, he joined the group "Medical Physics" of Professor Birger Kollmeier in Göttingen and later on in Oldenburg. He finished his Ph.D. thesis on loudness perception in hearing-impaired subjects with the development of a loudness model which accounts for hearing impairment. Since June 1995 he has been in various functions with Phonak's Research and Development, currently responsible for coordinating Phonak's corporate research and technology program. Before that, he was heading the Department "Signal Processing" where he was in charge of research and development related to audiological and clinical procedures of hearing instrument fitting and evaluation as well as the development of future digital signal processing algorithms for hearing instruments.

Norbert Dillier is the Head of the Laboratory of Experimental Audiology at the ENT Department of the University Hospital Zurich, Switzerland. He holds a Diploma in electrical engineering (1974) as well as a Ph.D. degree in technical sciences (1978) from the Swiss Federal Institute of Technology Zurich (ETHZ). He is a Lecturer at the University of Zurich (Habilitation, 1996) and at the ETHZ and member of several international professional societies, associations, and journals' editorial boards. His research focuses on improvements of the function of auditory prostheses such as cochlear implants, auditory brainstem implants, as well as conventional and implantable hearing aids. Major goals are to enhance the speech discrimination performance, especially in noisy environments and to improve the sound quality for music perception with these devices. New methods for programing and speech processor fitting especially for very young children using objective electrophysiologic measurement procedures and the use of bilateral electrical or the combined electrical acoustical stimulation for improved localization and speech recognition in noise are other important areas of his research.