Diss. ETH no 14498

Algorithms for Sound Classification in Hearing Instruments

A dissertation submitted to the

SWISS FEDERAL INSTITUTE OF TECHNOLOGY ZURICH

for the degree of
Doctor of Technical Sciences

presented by
Michael Christoph Büchler
dipl. El. Ing. ETH
born April 26, 1967
citizen of Brugg/AG

accepted on the recommendation of
Prof. Dr. Peter Niederer, examiner
PD Dr. Norbert Dillier, co-examiner
Dr. Stefan Launer, co-examiner

2002

Meinen Eltern, in Dankbarkeit

Abstract

Hearing instrument users often prefer different instrument-settings in different acoustic environments. Thus, modern hearing instruments allow the user to select between several hearing programs for different situations, to change the frequency response and compression parameters, or to activate a directional microphone, noise reduction or feedback suppression. However, the user has the bothersome task of recognizing the acoustic environment and then switching to the program that best fits this situation, using a switch on the hearing instrument or a remote control. 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 instruments.

The above assumption was confirmed by practical experiences. In a study with hearing impaired subjects, the usefulness and acceptance of an automatic program selection mode in the hearing instrument was investigated from the point of view of the user. It was shown that the automatic switching mode of the test instrument was deemed useful by a majority of test subjects, even if its performance was not perfect. These results were a strong motivation for the research described in the present work.

In this thesis, an automatic sound classification system for application in hearing instruments is developed. Our goal was on the one hand to develop a robust classification algorithm for at least the four classes 'speech', 'speech in noise', 'noise', and 'music', on the other hand to collect fundamental knowledge as a basis for a more detailed classification. The refined classes may for example be different noise types and music styles, or clean and reverberated speech.

So far, existing sound classification algorithms designed for hearing instruments have particularly been able to separate speech signals from other signals. Musical sounds however could not be recognized, and it was only partly possible to separate noise from speech in noise. Other algorithms, designed primarily for multimedia applications, allowed the recognition of sounds on a more specific level, such as certain distinct alarm signals or certain kinds of music. Thus, for the recognition of the four more general sound classes mentioned above, a new approach had to be developed. Generally, a sound classification system consists of the extraction of appropriate features from the signal, followed by a pattern classifier and an optional post processing step. This architecture was also chosen for the system described in this thesis.

In order to study how the human auditory system classifies sound, the mechanisms of Auditory Scene Analysis were investigated. The extraction of auditory features was shown to be an important step in the process of sound segmentation performed by the human auditory

system. Thus, one of the main goals in this thesis was to find appropriate features. A number of adequate auditory features have been modeled, including amplitude modulations, harmonicity, spectral profile, amplitude onsets, and rhythm.

These auditory features were evaluated together with different pattern classifiers. Considering the application in hearing instruments, where computing time and memory are limited, simple classifiers (rule-based and minimum-distance classifiers) have been compared with more complex ones (Bayes classifier, neural network, hidden Markov model, and a multistage approach). A hit rate of about 80 % was achieved with the simpler classifiers, which could be increased up to some 90 % when a more complex classifier was used. However, both the computing time and memory requirements are about four times larger with the more complex than with the simpler approaches.

The best classification system contains two stages: The first stage consists of a first feature set and a hidden Markov model. In the second stage, comprising a second feature set and a rule-based approach, the output of the hidden Markov model is verified and the classification is corrected if necessary. This approach worked well for most sounds within the four classes, resulting in a hit rate of 91 %. There are a number of sounds in each of the four classes that were recognized very robustly: Clean and slightly reverberated speech, speech in noise with moderate SNR, traffic and social noise, and classical music, single instruments and singing. However, some sounds were problematic and were mostly misclassified: Speech in noise with very low or very high SNR was classified as 'noise' or 'speech', respectively; compressed and strongly reverberated speech, a few tonal and fluctuating noises, and compressed pop music were all classified as 'speech in noise'.

Thus, for a more detailed sound classification in hearing instruments, further research is required. However, some fundamental limitations are already evident: the hearing instrument will not always be able to recognize whether a sound will be regarded as a desirable signal or as noise by the user.

Zusammenfassung

Benutzer von Hörgeräten bevorzugen abhängig von der akustischen Umgebung oft unterschiedliche Geräteeinstellungen. Darum ermöglichen es moderne Hörgeräte auch, mehrere Programme für verschiedene akustische Situationen zu wählen, um den Frequenzgang und Kompressionsparameter zu ändern, oder Richtmikrofon, Störgeräusch-Reduktion oder Feedback-Unterdrückung zu aktivieren. Allerdings müssen die Hörgeräteträger bei herkömmlichen Mehr-Programm-Hörgeräten selbständig die akustische Hörsituation beurteilen und entscheiden, welches Programm für diese Situation optimal ist, um dann per Schalter am Hörgerät oder über eine Fernbedienung in das entsprechende Programm umzuschalten. Eine automatische Erkennung der aktuellen akustischen Situation und automatisches Umschalten in das geeignetste Programm würden deshalb den Komfort von solchen modernen Hörgeräten verbessern.

Diese Annahme wurde durch praktische Erfahrungen bestätigt. In einer Studie mit Schwerhörigen wurden die Nützlichkeit und Akzeptanz einer automatischen Programmwahl in Hörgeräten aus der Sicht der Benutzer untersucht. Es konnte gezeigt werden, dass die Automatik des Testinstruments von einer Mehrheit der Versuchspersonen als nützlich beurteilt wurde, auch wenn die Programmwahl nicht immer korrekt war. Diese Resultate waren eine grosse Motivation für die in dieser Dissertation beschriebene Forschung.

In der vorliegenden Arbeit wird ein System zur Klangklassifizierung für die Anwendung in Hörgeräten entwickelt. Das Ziel war einerseits, einen robusten Klassifizierungs-Algorithmus für mindestens die vier Geräuschklassen 'Sprache', 'Sprache im Störgeräusch', 'Störgeräusch' und 'Musik' zu entwickeln, andererseits grundlegende Erfahrungen als Basis für eine detailliertere Klassifizierung zu sammeln. Eine mögliche feinere Unterteilung der genannten Klassen sind zum Beispiel verschiedene Typen von Störgeräuschen, verschiedene Musikstile oder reine und verhallte Sprache.

Bisherige Ansätze zur Klang-Klassifizierung in Hörgeräten waren insbesondere dazu fähig, Sprache von anderen Signalen zu trennen. Musik jedoch konnte nicht erkannt werden, und es war nur beschränkt möglich, Störgeräusch von Sprache im Störgeräusch zu trennen. Weitere Geräuschklassifizierungs-Algorithmen wurden vor allem für Multimedia-Applikationen entwickelt; sie erlauben es, spezifischere Klänge zu erkennen, wie zum Beispiel einzelne Alarmsignale oder gewisse Musikstile. Für die automatische Erkennung der vier oben erwähnten Geräuschklassen musste daher ein neuer Ansatz entwickelt werden. Im allgemeinen besteht ein System zur Geräuschklassifizierung aus der Extraktion von geeigneten Merkmalen aus dem Signal, gefolgt von einem Klassifizierer und einem optionalen Nachverarbeitungsblock. Diese Architektur wurde auch für das in der vorliegenden Arbeit entwickelte System verwendet.

Um darüber Kenntnis zu erlangen, wie das menschliche auditorische System Klangklassifizierung vornimmt, wurden die Mechanismen der auditorischen Szenenanalyse studiert. Es zeigte sich, dass die Extraktion von auditorischen Merkmalen eine wichtige Stufe im Prozess der Geräuscherkennung im auditorischen System ist. Daher war eines der Hauptziele der vorliegenden Arbeit die Suche nach geeigneten Merkmalen. Eine Reihe von zweckmässigen auditorischen Merkmalen wurde modelliert. Dazu gehören Amplitudenmodulationen, Harmonizität, spektrales Profil, Amplituden-Onsets und Rhythmus.

Diese auditorischen Merkmale wurden zusammen mit verschiedenen Klassifizierern evaluiert. Im Hinblick auf eine Anwendung in Hörgeräten, wo Rechenzeit und Speicherplatz beschränkt sind, wurden einfachere Klassifizierer (regelbasierter und Minimum-Distance Klassifizierer) mit komplexeren verglichen (Bayes Klassifizierer, neuronales Netz, hidden Markov Modell, und ein mehrstufiger Ansatz). Mit den einfacheren Ansätzen konnte eine Trefferquote von etwa 80 % erreicht werden, die auf bis zu 90 % erhöht werden konnte, wenn ein komplexerer Klassifizierer gewählt wurde. Sowohl der Bedarf an Rechenzeit als auch an Speicherplatz ist jedoch für die komplexeren Ansätze etwa vier mal höher als für die einfachen.

Das beste System umfasst zwei Stufen zur Klassifizierung: Die erste Stufe besteht aus einem ersten Merkmalsset und einem hidden Markov Modell. In der zweiten Stufe, bestehend aus einem zweiten Merkmalsset und einem regelbasierten Klassifizierer, wird das Resultat des hidden Markov Modells verifiziert und die Klassifizierung wenn nötig korrigiert. Dieses System erzielte eine Trefferquote von 91 %. Einige Klänge in jeder der vier Klassen wurden damit sehr robust erkannt. Dies sind reine und leicht verhallte Sprache, Sprache im Störgeräusch mit moderatem SNR, Verkehrs- und Partygeräusch, klassische Musik, einzelne Instrumente und Gesang. Einige Klänge waren jedoch problematisch und wurden meist falsch klassifiziert: Sprache im Störgeräusch mit ziemlich kleinem oder grossem SNR wurden als 'Störgeräusch' beziehungsweise 'Sprache' klassifiziert, komprimierte und stark verhallte Sprache, einige tonale und fluktuierende Störgeräusche und komprimierte Popmusik wurden alle als 'Sprache im Störgeräusch' betrachtet.

Für eine detailliertere Geräuschklassifizierung in Hörgeräteanwendungen ist somit weiterer Forschungsbedarf vorhanden. Dabei sind jedoch einige grundsätzliche Grenzen vorgezeichnet; das Hörgerät wird nicht immer erkennen können, ob der Benutzer ein Geräusch als Nutzsignal oder als Störsignal empfindet.

Contents

1 Introduction **1**

1.1 Background and Motivation ... 1
1.2 Objectives and Approach ... 2
1.2.1 Sound Classes to Be Recognized ... 2
1.2.2 Objectives ... 4
1.2.3 Approach ... 4
1.3 Contributions ... 5
1.4 Thesis Outline ... 5

2 State of the Art in Sound Classification **7**

2.1 Introduction ... 7
2.2 Existing Algorithms for Hearing Instruments ... 8
2.2.1 Amplitude Statistics ... 8
2.2.2 Temporal Fluctuations and Spectral Form ... 13
2.2.3 Modulation Frequency Analysis ... 16
2.2.4 Evaluation and Comparison of the Three Algorithms ... 19
2.2.5 Noise Classification with Neural Networks by Feldbusch ... 26
2.2.6 Noise Classification with HMMs by Nordqvist ... 26
2.3 Other Sound Classification Algorithms ... 27
2.3.1 Environmental Noises and Alarm Signals ... 27
2.3.2 Music and Multimedia ... 27
2.4 Discussion and Conclusions ... 29
2.4.1 Overview ... 29
2.4.2 Feature Extraction ... 30
2.4.3 Classifiers ... 31
2.4.4 Post Processing ... 31
2.4.5 Further Work ... 31

3 A Field Study with Hearing Impaired Subjects **33**

3.1 Introduction ... 33
3.2 Method ... 33
3.3 Results ... 34
3.4 Discussion and Conclusions ... 36

4 Principles of Auditory Scene Analysis **39**

4.1 Introduction ... 39

4.2 Definitions ... 40
4.2.1 Feature – Event – Source ... 40
4.2.2 Fusion and Segregation ... 40
4.3 Overview of Scene Analysis in the Auditory System ... 40
4.3.1 Peripheral Ear Filtering and Transduction ... 41
4.3.2 Feature Filtering ... 43
4.3.3 Event Formation ... 44
4.3.4 Source Formation ... 44
4.4 Features for Event Formation ... 44
4.4.1 Spectral Separation ... 44
4.4.2 Spectral Profile ... 44
4.4.3 Harmonicity and Pitch ... 46
4.4.4 Spatial Separation ... 49
4.4.5 Temporal Onsets and Offsets ... 50
4.4.6 Amplitude Modulation ... 52
4.4.7 Frequency Modulation ... 53
4.5 Event and Source Formation ... 54
4.5.1 Grouping Principles ... 54
4.5.2 Event Formation ... 55
4.5.3 Source Formation ... 55
4.6 Application for Sound Classification ... 57
4.6.1 Auditory Features ... 57
4.6.2 Grouping ... 58
4.6.3 Existing Models of Auditory Scene Analysis ... 59

5 Classification Systems I: Features for Sound Classification 61
5.1 Introduction ... 61
5.2 Features Motivated by Auditory Scene Analysis ... 62
5.2.1 Spectral Separation ... 62
5.2.2 Spectral Profile ... 62
5.2.3 Harmonicity / Pitch ... 64
5.2.4 Spatial Separation ... 70
5.2.5 Temporal Onsets and Offsets ... 70
5.2.6 Rhythm or Beat Extraction ... 77
5.2.7 Amplitude Modulation ... 80
5.2.8 Frequency Modulation ... 81
5.3 Summary and Conclusions ... 81

6 Classification Systems II: Pattern Classifiers for Sound Classification 85
6.1 Introduction ... 85
6.2 Preprocessing of the Feature Vectors ... 86
6.3 Bayes Decision Theory ... 87
6.4 A Selection of Classifiers for Sound Classification ... 87
6.4.1 Heuristic Rule-Based Classifier ... 87
6.4.2 Minimum Distance Classifier ... 88
6.4.3 Bayes Classifier ... 89
6.4.4 Multilayer Perceptron ... 90
6.4.5 Hidden Markov Models ... 91
6.5 Discussion and Conclusions ... 94

7 Evaluation of Different Classification Systems 95
7.1 Introduction ... 95
7.2 Procedure ... 95
7.2.1 Overview ... 95
7.2.2 Sound Data ... 96
7.2.3 Feature Sets ... 97
7.2.4 Pattern Classifier ... 97
7.2.5 Post Processing ... 98
7.2.6 Scores ... 98
7.3 Results ... 99
7.3.1 Rule-Based Classifier ... 99
7.3.2 Minimum Distance Classifier ... 101
7.3.3 Bayes Classifier ... 103
7.3.4 Multilayer Perceptron ... 105
7.3.5 Hidden Markov Model ... 107
7.3.6 A Simple Multistage Strategy ... 109
7.4 Discussion ... 111
7.5 Conclusions ... 114

8 Summary and Conclusions 117
8.1 Summary of Achievements and Conclusions ... 117
8.2 Future Work ... 119
8.3 Outlook ... 121

Appendix A: Sound Database 123

Bibliography 129

Dankeswort 135

Curriculum Vitae 137

1 Introduction

1.1 Background and Motivation

About 12 % of the population of the industrialized countries experience a significant hearing loss and need to be supplied with a hearing instrument. The aim of a hearing instrument is essentially to amplify the sound acoustically to a level at which the user can hear it again and understand the information it contains. However, simple amplification of the sound may not be sufficient, since sensorineural hearing loss, that is, hearing loss of cochlear or retrocochlear origin, does not only cause the sound to be perceived lower; it is also being distorted. There are mainly two reasons for the distortion (see for example Moore, 1997): First, the hearing threshold is higher in the pathological than in the physiological case, but the uncomfortable level stays the same. This means that the dynamic range has become much smaller: the loudness perception at low and medium level is altered, whereas it appears to be normal at high levels. This phenomenon is called *recruitment*. Second, the frequency resolution in the cochlea is degraded in the pathological case; the spectrum will be "smeared". These distortions can make conversations a difficult matter, especially if they are held in a noisy environment. A hearing instrument cannot make up for all of this damage, especially not for the frequency smearing; it can only try to filter out the noisy parts of the sound that do not contain information, to optimize both speech intelligibility and sound quality.

Conventional hearing instruments are fit to an individual's hearing loss based on audiometric data such as the audiogram and possibly also loudness scaling. The required amplification for each frequency region is evaluated using a formula, as for example the Desired Sensation Level prescription method (DSL [i/o], Cornelisse et al., 1995) or the revised National Acoustics Laboratories procedure (NAL-RP, Byrne & Dillon, 1986). Thus, a single setting is supposed to cover all listening situations. The resulting fitting is most likely a compromise; it is not the optimal setting for *all* listening environments and signal characteristics, as in one situation speech intelligibility may be regarded to be most important, but sound quality and comfort in another.

Modern hearing instruments allow the user to switch manually between different programs, that is between different frequency responses or other processing options, such as directional microphone, noise reduction, feedback canceler, compression methods or limiters. In the situation 'speech in noise', for example, the system tries to optimize speech intelligibility, and a hearing program is chosen where a directional microphone is activated. In 'music', on the other hand, an omnidirectional sound field is preferred to achieve best sound quality. However, the user has the difficult task of recognizing the acoustic environment and then

switching to the program that fits best to this situation. Even for normal hearing people, it is not always clear which program should be selected for best performance, and considering that elderly people are the main group of hearing instrument users, it cannot be expected that they always handle this task correctly.

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 instruments. A simple approach to perform sound classification in a hearing instrument has been introduced recently (Phonak, 1999). This hearing instrument can distinguish between the acoustic situation 'speech in noise' and other situations.

The goal of this thesis is to perform a more detailed classification of the acoustic environment. In the next section, the acoustic environments to be recognized will be defined, and the approach that is chosen to perform such classification will be described.

1.2 Objectives and Approach

1.2.1 Sound Classes to Be Recognized

To find out which acoustic environments are critical in everyday life and thus most important to be recognized, it seems a good idea to ask hearing impaired persons. This has been done for example in a study from Fedtke et al. (1991). Subjects with a moderate hearing loss were asked to judge how important it is to hear well in 52 different situations in the area of home life, work, culture, leisure time and traffic. The situations judged most important can be roughly divided into four classes:

- Speech (dialogue, lectures, theater, cinema, phone calls, television)
- Speech in noise (cocktail party situation, announcement at train station or airport, speech in a car)
- Alarm signals (ringing phone, door bell)
- Nature (chirping birds).

The first three of these classes seem quite obvious, as they contain essential information for the people. The class 'nature', however, shows their desire for a certain listening comfort. In this context, it is a bit astonishing that 'music' is not named as an important sound class as well. This is probably due to the sounds that were presented in the study; no musical signals had to be judged apart from 'singing in a theatre', which was regarded as being quite important. Thus, it is assumed that music sounds also belong to the more important situations.

Haubold et al. (1993) developed a new hearing instrument fitting procedure based on natural acoustical patterns. They propose to use eight different classes for the fitting, which included 'speech', 'speech in noise', 'noise', 'warning signals', 'nature', and 'music'.

The classes 'speech' and 'speech in noise' are again situations that are apparently important for communication. It is of course also important that warning signals such as car horns, phone or door bells etc. can be heard. These sounds can be very short. The concept of an automatic program switch in the hearing instrument, however, will probably be that it reacts to events that remain stationary over a longer period of time, in the order of ten seconds or longer. The class 'alarm' is therefore a special situation which will be omitted in a first approach, assuming also that the hearing instrument will anyway amplify such sounds by default.

For listening comfort, 'noise', 'music' and 'nature' sounds are obviously important. Of course, the behavior of the hearing instrument will be different for 'noise' than for 'music' and 'nature': The noise shall be attenuated, whereas the other signals shall be amplified without any distortion. The class 'nature', however, suffers from the same problems as the class 'alarm': The sounds can be very short, which will be difficult to handle. Therefore, it is also left away in a first approach, and the focus lies on situations with continuous music, such as listening to a concert etc.

Finally, it seems reasonable to add a class 'silence', to identify quiet situations adequately.

This leads us to the following five main acoustic situations that are to be classified by a hearing instrument in a first approach:

- Silence
- Speech
- Speech in Noise
- Noise
- Music

In a second approach, these classes might be refined in subclasses: A special kind of speech is distorted speech in a reverberating room. For noise, cocktail party situations and traffic noise are important subclasses, but also in-the-car noise and industrial noises. For music, one might want to separate classical music from pop and rock music, or from a single instrument being played, or from one's own voice, when singing.

How shall the signal to noise ratio (SNR) for the class 'speech in noise' be chosen? On the one hand the noise shall have a significant influence on the intelligibility, on the other hand it shall still be possible to understand parts of the speech signal. The speech reception threshold (SRT) for 50 % intelligibility is a reasonable value for the average SNR. However, the SRT is not the same for people with normal and impaired hearing; the difference is around 5 dB for a moderate hearing loss (Killion, 1997). Furthermore, the SRT depends on the kind of noise. Thus, different SNRs must be chosen according to the different background noises, in the range of +2 to –9 dB. For further details, see appendix A describing the soundset.

Considering these reflections, the "must" and "wish" classes can be formulated as shown in the table on the next page.

Sound class	Priority	
	Must	Wish
Silence	×	
Speech without noise	×	
• Reverberated speech		×
Speech in noise	×	
• Speech in social noise		×
• Speech in a car		×
• Speech in traffic noise		×
Noise	×	
• Social noise		×
• In the car noise		×
• Traffic noise		×
• Industrial noise		×
Music	×	
• Classical music		×
• Pop and rock music		×
• Single instrument		×
• Singing		×
Nature		×
Alarm signals		×

1.2.2 Objectives

The primary goal of this thesis is to develop a robust algorithm for the acoustic environment. In particular, the "must" classes defined in section 1.2.1 shall be distinguished. A further goal is to build up fundamental knowledge in sound classification as a basis for a more detailed classification of sounds. The recognition of the "wish" sound classes as defined in 1.2.1 is not within the scope of this thesis.

1.2.3 Approach

The development of the sound classification system will be started with a review of existing algorithms, most of which are designed for other – for example multimedia – applications rather than for hearing instruments.

Another approach to the topic will be to investigate how the auditory system performs sound classification. It is known from the theory of Auditory Scene Analysis (Bregman, 1990), that the human hearing system possesses an amazing ability to adapt to various acoustic situations, which shows that the actual situation is somehow recognized. If it can be found out which mechanisms of Auditory Scene Analysis are relevant for the recognition of the acoustic environment, and if these mechanisms can be modeled, then an effective classification system may be established. Indeed, a number of models exist already for sound separation rather than for classification, and some of this work may be adapted accordingly.

Of course, if a sound classification system is designed for the application in hearing instruments, there are some limitations. First of all, the space in a hearing instrument is very

restricted, and so is memory. Second, the algorithm has to run in real-time, that is, with a maximum delay of some milliseconds between input and output; hence also computing time is limited. Finally, it has to be considered that hearing instruments are worn in various everyday situations, so the system will not be in a protected environment. Most of the existing sound classification approaches or Auditory Scene Analysis models do not yet account for these limitations; many of them are designed for laboratory experiments in controlled acoustic situations, and they require computing time that is many times longer than real-time, even on very fast computers.

1.3 Contributions

The primary contribution of the research described in this thesis is the conceptual development, implementation and evaluation of a sound classification algorithm that is inspired by Auditory Scene Analysis. In particular, this includes:

- Evaluation of existing classification algorithms for hearing instrument applications.
- Composition of a large sound database containing hundreds of different everyday sounds for evaluation purposes.
- Following principles of Auditory Scene Analysis for sound classification; in particular determination and implementation of features that are based on Auditory Scene Analysis.
- Evaluation and comparison of different pattern classifier types, such as rule-based and statistical classifiers, with respect to the application of sound classification in hearing instruments.
- Evaluation of the combination of various feature sets and different pattern classifiers.
- Design, implementation and evaluation of a sound classification system using selected features from Auditory Scene Analysis and a two stage classifier consisting of a hidden Markov model and a rule-based classifier, achieving a high recognition rate for the four main sound classes 'speech', 'noise', 'speech in noise', and 'music'.

1.4 Thesis Outline

In chapter 2, five existing classification algorithms for hearing instruments are reviewed and compared. This is followed by a review of sound classification systems that are designed for other applications.

In chapter 3, a study will be presented that investigated the usefulness and acceptance of an automatic program switch in a hearing instrument as judged by the users. The conclusions of this study will contribute to the design of the classification system, and are a strong motivation for the research done in this thesis.

Chapter 4 gives an introduction to Auditory Scene Analysis. It will be discussed which features and mechanisms of auditory perception can be applied for technical sound classification.

Chapter 5 deals with features that are motivated by Auditory Scene Analysis. This includes in particular features describing harmonicity, onsets, and rhythm in the sound signal.

Chapter 6 is dedicated to an overview of pattern classifiers. A number of classifiers will be presented that may be suited for the application in hearing instruments, both from a point of view of performance and complexity.

In chapter 7, the different features and pattern classifiers are evaluated as to find the optimal combination of features as well as the best classifier. The best sound classification system will be presented and discussed.

Chapter 8, finally, summarizes the contributions of this thesis and recommends future directions.

2 State of the Art in Sound Classification

2.1 Introduction

In this chapter, an overview is presented of the state of the art in sound classification. In the literature, many sound classification algorithms are described, but only few are designed for hearing instrument applications. Most of them are determined for other applications, such as multimedia, and are only able to classify subsets of the classes that are desired for hearing instruments, for example different music types, background noises or alarm signals.

The general structure of a sound classification system can be described with a block diagram, as it is shown in Figure 2.1. From the sound data, a number of characteristic features are extracted, which are then classified with some sort of pattern classifier. An optional post processing step may correct possible classification outliers and control the transient behavior of the algorithm. The output of the algorithms are the recognized sound classes.

Figure 2.1: General block diagram of a sound classification system.

In the following, five currently known methods for sound classification in hearing instruments will be presented. Three of them are already exploited in commercial hearing instruments, the analysis of the amplitude statistics by Ludvigsen (1993), the classification based on temporal fluctuations and spectral form by Kates (1995) and further developed by Phonak (1999), and the analysis of the modulation spectrum (Ostendorf et al., 1997). The two other algorithms are also designed for hearing instruments, but not exploited so far (Feldbusch, 1998, and Nordqvist, 2000).

The feature extraction blocks of the three already exploited approaches will be evaluated and compared. It will be shown that they are related in that most of the features described in these algorithms represent the amplitude modulations in the signal, and that this enables the discrimination of speech signals from other sounds very well. A more detailed classification of the acoustic environment is however hardly possible with these approaches.

Finally, a review of other sound classification algorithms is given. This includes procedures for the classification of environmental noises, alarm signals, musical signals, music types, as well as multimedia sounds. One of the multimedia classifiers, from Zhang and Kuo (1998,

1999, and 2001) seems to be a very interesting approach, in that it performs very well in classifying about the same classes as are intended in this thesis.

2.2 Existing Algorithms for Hearing Instruments

2.2.1 Amplitude Statistics

Ludvigsen (1993) proposes to automatically control the amplification and/or the frequency response of a hearing instrument by investigating the continuity of the input signal; that is, by discriminating impulse-like and continuous signals. He does this by investigating the amplitude statistics of the signal.

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. The examples in Figure 2.2 to Figure 2.4 show the amplitude histogram of speech, party noise and speech in party noise. The histograms were built over thirty seconds of the envelope of each signal.

Due to the pauses in the speech signal, its level varies very much over time, resulting in a broad and asymmetrical amplitude histogram. The level of the party noise is much more constant, that is, the amplitude histogram has a narrow and symmetrical form. The speech in party noise signal is a bit broader, but still symmetric; the two modes are not typical for speech in noise sounds.

In addition to the histograms, some percentiles are also drawn in the figures. The 30 % percentile, for example, shows the level below which the envelope is 30 % of the time. The asymmetrical distribution in the speech signal results in a much larger distance between the 10 % and the 50 % percentile than between the 50 % and 90 % percentile, or, in other words, the 50 % percentile is far away from the arithmetical mean of the 10 % and 90 % percentile. For the noise and the speech in noise signals, the 50 % percentile is more or less in the middle of the 10 % and the 90 % percentile, representing the symmetrical distribution.

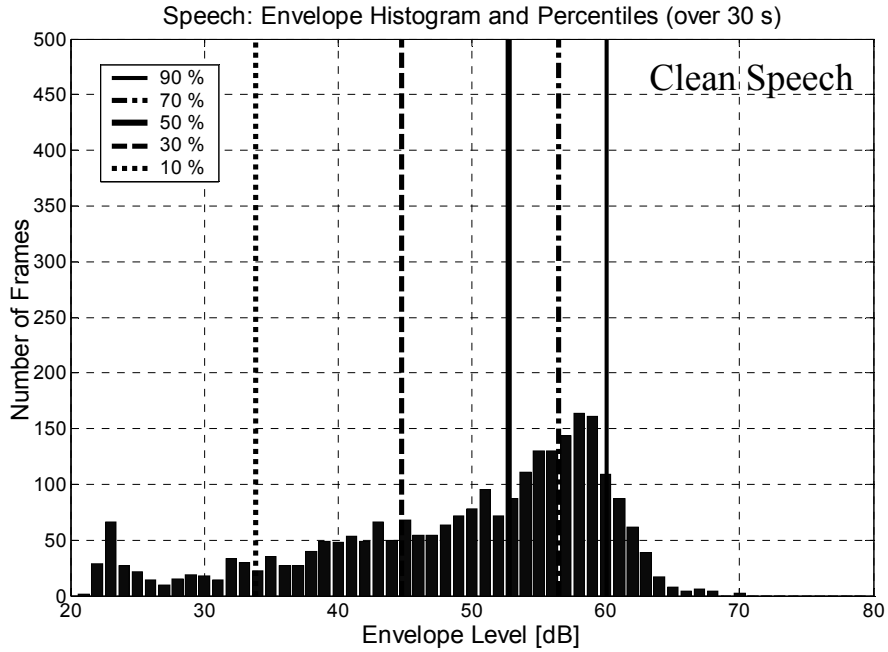


Figure 2.2: Amplitude envelope histogram of clean speech acquired over thirty seconds. Due to the pauses in the speech signal, the histogram is very broad and asymmetric. In addition, some percentiles are plotted. The distance between the 10 % and 50 % percentiles (or 30 % and 50 %) is much larger than the one between the 50 % and the 90 % percentile (or 50 % and 70 %).

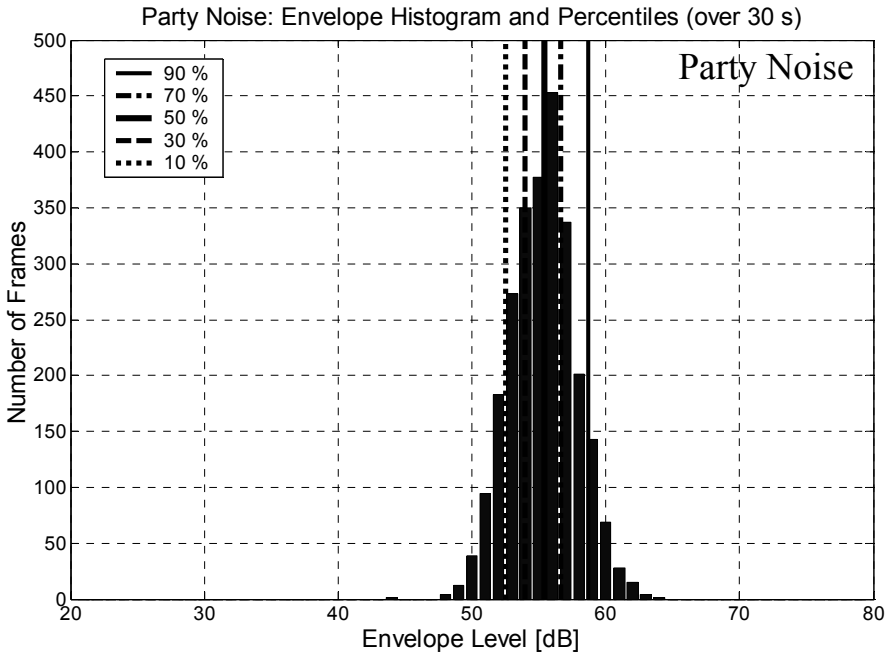


Figure 2.3: Amplitude envelope histogram of party noise acquired over thirty seconds. Due to the quasi-stationary character of the signal, the histogram is quite narrow and symmetric. In addition, some percentiles are plotted. The distance between the 10 % and 50 % percentiles (or 30 % and 50 %) has about the same size as the one between the 50 % and the 90 % percentile (or 50 % and 70 %).

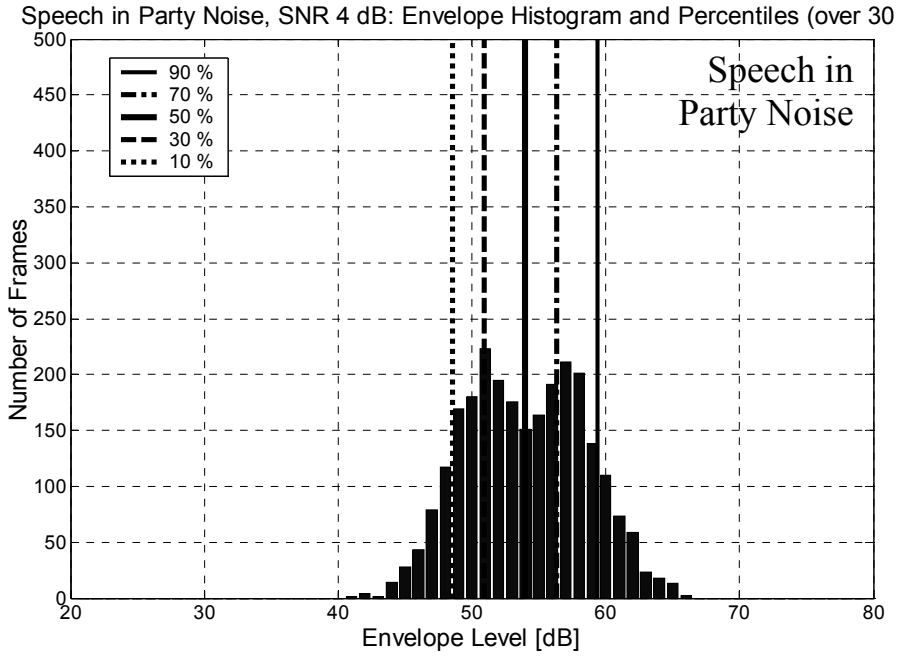


Figure 2.4: Amplitude envelope histogram of speech in party noise acquired over thirty seconds. It is a bit broader than the one of noise, but still symmetric; the two modes are not typical for speech in noise. In addition, some percentiles are plotted. The distance between the 10 % and 50 % percentiles (or 30 % and 50 %) has still about the same size as the one between the 50 % and the 90 % percentile (or 50 % and 70 %).

As Ludvigsen (1993) states, it is sufficient to calculate some of the percentiles to decide whether the signal is more of a continuous or of an impulse-like character, instead of calculating the whole histogram. He suggests to take the 10 % , 50 % and 90 % percentiles and some relations between these percentiles. Possible relations using these percentiles as well as the 30 % and 70 % percentile are shown in Figure 2.5.

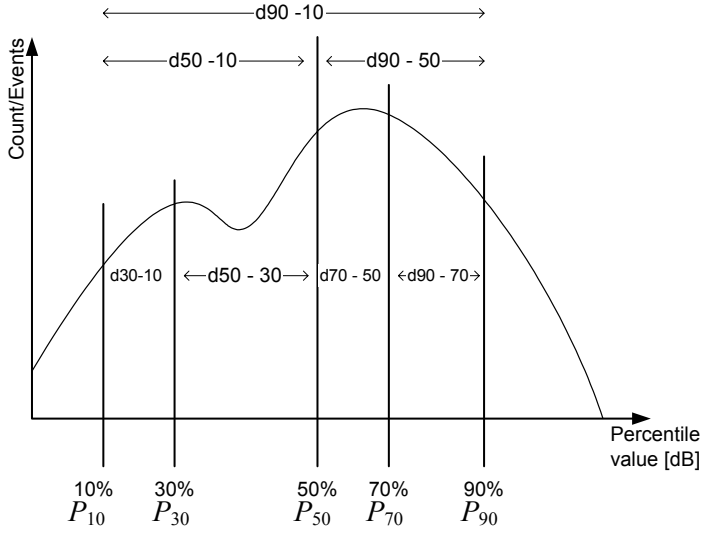


Figure 2.5: Some inter-histogram relations. The width of the histogram is well described by the distance of the 90 % and the 10 % percentiles, $d_{90\text{-}10}$; the symmetry by the difference $(d_{50\text{-}10})\text{-}(d_{90\text{-}50})$.

The distances between the percentiles may thus be the basis for more complex features. They are normalized to the 50 % percentile; the distance between the 10 % and the 90 % percentiles for example is calculated as follows:

$$d_{90-10} = \frac{P_{90} - P_{10}}{P_{50}} \tag{2.1}$$

A number of features are presented in the following that appear to be valuable for the description of the form of the histogram:

- *Width*:
  The width of the histogram is well described by the distance between the 90 % and the 10 % percentile:

  $$width = d_{90\text{-}10} \tag{2.2}$$

  Alternatively, $d_{70\text{-}30}$ could be taken.

- *Symmetry*:
  The symmetry can be investigated by looking at the difference $(d_{90\text{-}50})$-$(d_{50\text{-}10})$, or at $(d_{70\text{-}50})$-$(d_{50\text{-}30})$ respectively:

  $$symmetry = (d_{90\text{-}50})\text{-}(d_{50\text{-}10}) \tag{2.3}$$

  The symmetry is near zero for symmetrical distributions, positive for left sided distributions, and negative for right sided distributions. Impulse-like signals are asymmetric right-sided due to the signal pauses.

- *Skewness*:
  The skewness of the histogram can be regarded as the difference between the 50 % percentile and the median:

  $$skewness = P_{50} - \tilde{x} \tag{2.4}$$

  with the median being estimated by the mean between the 10 % and the 90 % percentile

  $$\tilde{x} = \frac{(P_{90} + P_{10})}{2} \tag{2.5}$$

  For asymmetrical distributions the difference between P_{50} and the approximated median should be large, for symmetrical distributions approximately zero.
  Note that the feature *skewness* is equivalent to the *symmetry* of the distribution without normalization: From equation (2.3), the unnormalized symmetry is $P_{90} + P_{10} - 2 \cdot P_{50}$, which equals $-2 \cdot (P_{50} - Median)$.

- *Kurtosis*:
  The kurtosis corresponds to the approximation

  $$kurtosis = \frac{P_{70} - P_{30}}{2(P_{90} - P_{10})} \tag{2.6}$$

which sets the middle 50 % interval in relation to the range of the distribution, indicating whether the distribution has a narrow or a broad peak.

- *Lower half*:
  The distributions in the lower half of the histogram are expressed by the difference ($d_{50\text{-}30}$)-($d_{30\text{-}10}$):

$$lower\ half = (d_{50\text{-}30})\text{-}(d_{30\text{-}10}) \tag{2.7}$$

  The lower half of the distribution allows to characterize right-sided distributions by encoding the relations between the lower and upper half (that is, below and above P_{30}) of the lower half of the total distribution (that is, below P_{50}). For impulse-like signals, this feature will have a large value, for continuous signals it will be approximately zero.

The benefit of these features for classification will be investigated in section 2.2.4.

In an implementation, the percentile generator consists mainly of an envelope detector, a comparator and an integrator for each percentile, as Figure 2.6 shows. The comparator compares the actual envelope value with the integrated values of the previous frames and increments or decrements the integrator accordingly.

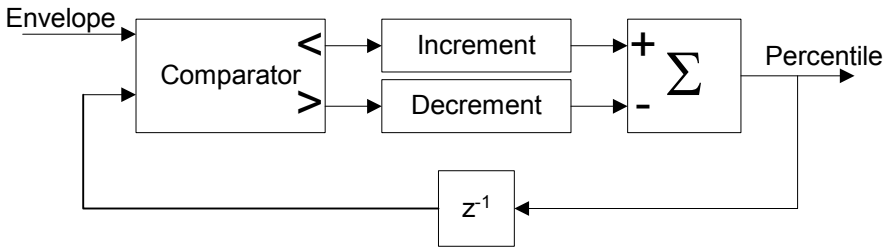


Figure 2.6: Detailed structure of a percentile generator according to Ludvigsen. The output is compared with the current envelope value and is incremented or decremented accordingly.

The percentile generator can be applied broadband or in multiple frequency bands. Ludvigsen recommends three bands, without specifying the boundaries of these bands.

When looking at the histograms in Figure 2.2 to Figure 2.4, it can be assumed that speech can be separated from quasi-stationary noise very well, whereas it will be more difficult to isolate the class speech in noise. If the signal has a good SNR, its histogram form will be more like the one of speech, if the SNR is poor, the form will more resemble the one of noise. If the noise is not continuous but fluctuating in level, the histogram may become broader and asymmetric, which might make it difficult to distinguish it from a speech histogram. It is not described by Ludvigsen what the histograms of musical signals look like. For the application in hearing instruments, music is an important class, as has been discussed earlier.

The behavior of the algorithm for musical signals, various noises and speech in noise with different SNRs will be evaluated and compared to the algorithms from Kates (1995) and Ostendorf et al. (1997) in section 2.2.4.

2.2.2 Temporal Fluctuations and Spectral Form

2.2.2.1 Original Algorithm

Kates (1995) tried to classify some everyday background noises, like multitalker babble, dinner with conversation and clattering dishes, dishwashing, printer, traffic, gaussian noise and others. He included a signal of a single male talker, despite the fact that this is not a background noise.

He started his investigations by analyzing the envelope modulation spectra in auditory critical bands, as proposed by Kollmeier and Koch (1994). The results show that the interactions of the envelope modulation and the signal spectrum do not contain significant amounts of information, that is, there are no important peaks or valleys at specific combinations of envelope modulation frequency and critical band for most of the test signals. Thus, it does not seem to be necessary to analyze the envelope in different frequency bands, and a simplified procedure is proposed, extracting four signal features.

The first feature is the amount of level fluctuations. It is described as the logarithmic ratio of the mean magnitude to the standard deviation:

$$MLFS = \log \frac{E(Amplitude)}{STD(Amplitude)} \tag{2.8}$$

The mean level fluctuation strength $MLFS$ will have a large value for smooth signals as for example a continuous sinusoid, while a small or even negative value indicates a signal with large fluctuations as would result from a set of widely spaced impulses.

The three remaining features describe the form of the spectrum. The mean frequency denoted by the center of gravity CG is determined by calculating the first moment of the spectrum on a logarithmic frequency scale, expressed as an FFT bin index k, and with the total number of bins K:

$$CG = \frac{\sum_{k=1}^{K} f(k)}{\sum_{k=1}^{K} \frac{1}{k} f(k)} \tag{2.9}$$

The portions in the spectrum below and above the center of gravity CG are each fit with a line segment using linear regression on a logarithmic frequency scale. The slopes of these lines expressed in dB/oct give two additional features describing the spectrum.

Kates performed a cluster analysis on the four features that had been extracted from the sound data. The classification accuracy was better than 90 % for seven or fewer clusters. The single talker and the dinner conversation signal always formed separate clusters each; these two clusters being very near together however. This already shows that speech is well separated from noise.

An open question is how the features are suitable for other classes, like speech in different noises, or music.

2.2.2.2 Modified Algorithm

The algorithm as proposed by Kates was altered by Phonak (1999), with the aim of simplifying it for the implementation in a hearing instrument.

The level fluctuation strength *MLFS* is approximated with the following formula:

$$MLFS = 10 \cdot \log \frac{E(ObsInterval)}{STD(ObsInterval)} \approx \log \frac{MLAV}{\frac{1}{3}(ML\_\max - MLAV)} \tag{2.10}$$

with

$$MLAV = \frac{1}{T_{mean}} \sum_{T_{mean}} \left(\frac{1}{N} \sum_{n=1}^{N} \log(|P(n)|) \right) \tag{2.11}$$

and

$$ML\_\max = \max_{T_{Mean}} \left(\frac{1}{N} \sum_{n=1}^{N} \log(|P(n)|) \right) \tag{2.12}$$

The mean level average *MLAV* is calculated out of the sum of the log magnitude *P* of *N* Bark bands averaged over a time T_{mean}, which is typically one 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. Note that the *MLFS* is level-dependent[1].

The center of gravity of the spectrum *CG* is calculated as in (2.9), except that the formula is applied to the Bark spectrum instead of the linear spectrum. From a psychoacoustical point of view, it does not seem to be correct to introduce an additional nonlinear weighting $1/k$, as the Bark spectrum is already nonlinearly built. A weighting of the *N* Bark bands according to k_n, the number of FFT bins in Bark n, seems more justified:

$$CG = \frac{\sum_{n=1}^{N} n \cdot f(n) \cdot k_n}{\sum_{n=1}^{N} f(n) \cdot k_n} \tag{2.13}$$

For the investigations in section 2.2.4, however, (2.9) was used together with the Bark spectrum. The results will have slightly different numerical values, but their expressiveness will remain the same.

The center of gravity *CG* is also averaged over some time T_{Mean}, which is again typically one second. This results in the average center of gravity, *CGAV*:

$$CGAV = \frac{1}{T_{Mean}} \sum_{T_{mean}} CG \tag{2.14}$$

Instead of fitting two lines below and above the *CG*, its temporal fluctuations are estimated in a similar way as the level fluctuations in (2.10):

[1] In further implementations, the level-dependence was compensated.

$$CGFS = \log \frac{E(CG)}{STD(CG)} \approx \log \frac{CGAV}{\frac{1}{3}(CG\_\max - CGAV)} \tag{2.15}$$

with

$$CG\_\max = \max_{T_{Mean}}(CG) \tag{2.16}$$

The center of gravity fluctuation strength *CGFS* indicates how strongly the center of gravity *CG* changes over time. Continuous signals like a sinusoid will have small spectral fluctuations, whereas the spectral fluctuations of other signals like speech will be quite large. Note that large fluctuations will result in small *CGFS* values and vice versa.

An additional feature proposed by Phonak (1999) is the overall level of the signal, as shown in (2.17). To simplify the computation, the amplitude is taken instead of the power, and the values are not normalized to the broadness of the *N* Bark bands. This means that the high frequencies will be overestimated, which can clearly be seen with signals containing high frequency energy.

$$TotPowdB = 20 \cdot \log \left\{ \frac{1}{T_{tot}} \sum_{T_{tot}} \sum_{n=1}^{N} \left| P(n)(t) \right| \right\} \tag{2.17}$$

The total power is averaged over typically one second. It gives some general information about the energy in the signal. In quiet, when the energy is very low, there is no need for further investigations of the other features.

A first evaluation of the performance of the combination of these features was done with a cluster analysis. 42 different signals of the classes speech, speech in noise, noise and music were processed. The use of the feature *TotPowdB* did not make sense, however, because the signals were calibrated to a fixed level before processing. So, the cluster analysis was performed using the features *MLFS*, *CGAV* and *CGFS*.

The results showed that three different clusters can be well determined:

1. clean speech,
2. high frequency signals, which include typewriter noise, white noise, rain noise as well as speech in these noises,
3. the rest of the investigated signals, that is strongly reverberated speech, babble noise, traffic noise and in-the-car noise as well as speech in these noises, classical music, pop music and some instruments.

Clean speech can be very well identified via the level fluctuations *MLFS*, and also via the spectral fluctuations *CGFS*. The more the speech signal is distorted with reverberation or noise, the less it can be identified as speech. The second cluster containing the high frequency signals is only identified via the spectral center of gravity *CGAV*. It is not possible to separate music signals from noise or speech in noise.

Further evaluations of the algorithm will follow in section 2.2.4 for various signals of the classes speech, speech in noise, noise and music.

2.2.3 Modulation Frequency Analysis

Ostendorf et al. (1997) investigated modulation spectra of different signals and confirmed that they show systematic differences. The goal was to distinguish speech, speech and noise, and noise signals to allow the automatic control of the compressor settings of a hearing instrument.

The modulations of a signal, described by the signal envelope, are characterized by the modulation frequencies and the corresponding modulation depths. The modulation frequency denotes the velocity of the modulations, and the modulation depth denotes the strength of the modulation. It has been shown that different signal classes exhibit different characteristics in their modulation frequency spectrum. The envelope of speech for example is determined by the phonemes, the syllables, the words, and the sentences. Normally we articulate about 12 phonemes, 5 syllables, and 2.5 words per second. To formulate sentences, several seconds are required. Thus, speech has modulation frequencies of approximately 12 Hz (phonemes), 5 Hz (syllables), 2.5 Hz (words), and < 1 Hz (sentences). Due to the speech pauses, the modulation depth of speech is large (Holube, 1998). The maximum in the modulation spectrum of clean speech is in the area of 2 to 8 Hz. Note that this corresponds very well to psychoacoustical findings: the maximum sensitivity to fluctuation strength[2] occurs at 4 Hz and indicates the excellent correlation between the speech and the auditory system (Zwicker and Fastl, 1990). By way of contrast, noise shows often weaker but faster modulations and has therefore its maximum at higher modulation frequencies. Hence, modulation frequencies and the corresponding modulation depths represent a powerful feature for the perception and discrimination of sounds.

The modulation spectrum was first calculated as shown in Figure 2.7. The envelope of the signal is scaled to its root mean square and Fourier transformed. The third spectrum of the absolute values of the FFT bins is then calculated.

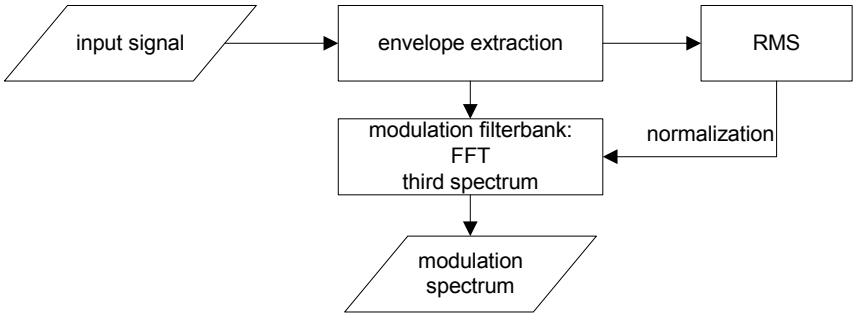


Figure 2.7: Block diagram for computing the modulation spectrum. The envelope of the signal is scaled to its RMS and Fourier transformed. The absolute values of the FFT bins are combined to the third spectrum.

The investigation of the modulation spectrum was limited to frequencies below 64 Hz. This is reasonable since speech is an important acoustic situation to be recognized, and modulation frequencies above approximately 70 Hz may already contain the pitch of normal male voices.

[2] Modulated sounds elicit two different kinds of hearing sensations: at low modulation frequencies up to about 20 Hz, fluctuation strength is produced. At higher modulation frequencies, the sensation of roughness occurs (Zwicker and Fastl, 1990).

Note, however, that the pitch of male voices can even go as low as 50 Hz (Deller et al., 1993), which will then appear as modulations and falsify the results.

In a first step, Ostendorf averaged the modulation frequencies over the whole signal length, which was 25 seconds. 43 signals of the classes speech, noise, and speech in noise were used.

The speech signals showed the expected clear peak around 4 Hz, and only few modulations at higher frequencies. This peak was absent for most of the noise signals; these signals may have peaks at higher modulation frequencies, if they have a rhythmic structure, as for example a clattering machine. The speech in noise signals tended to show higher values at low modulation frequencies, but no clear peaks around 4 Hz.

Ostendorf states that it seems therefore to be possible to distinguish between these three sound classes on the basis of the modulation spectra. In a next step, the modulation spectrum was subdivided into three channels with the frequency ranges 0 .. 4 Hz, 4 .. 16 Hz, and 16 .. 64 Hz, and the values in each channel were summed up. The resulting three modulation depths were called m_1, m_2 and m_3. The size of these three parameters allows to make statements about the affiliation of a sound to a particular class of signals, such as clean speech, speech and noise, or noise only.

An example of the values in the three modulation frequency bands (before summing them up to m_1, m_2 and m_3) is shown for clean speech, party noise and speech in party noise in Figure 2.8 to Figure 2.10. The peak around 4 Hz in the speech signal does not occur in the noise signal. In the speech in noise signal, it depends on the SNR; with 4 dB SNR, a small peak is visible, indicating the presence of speech.

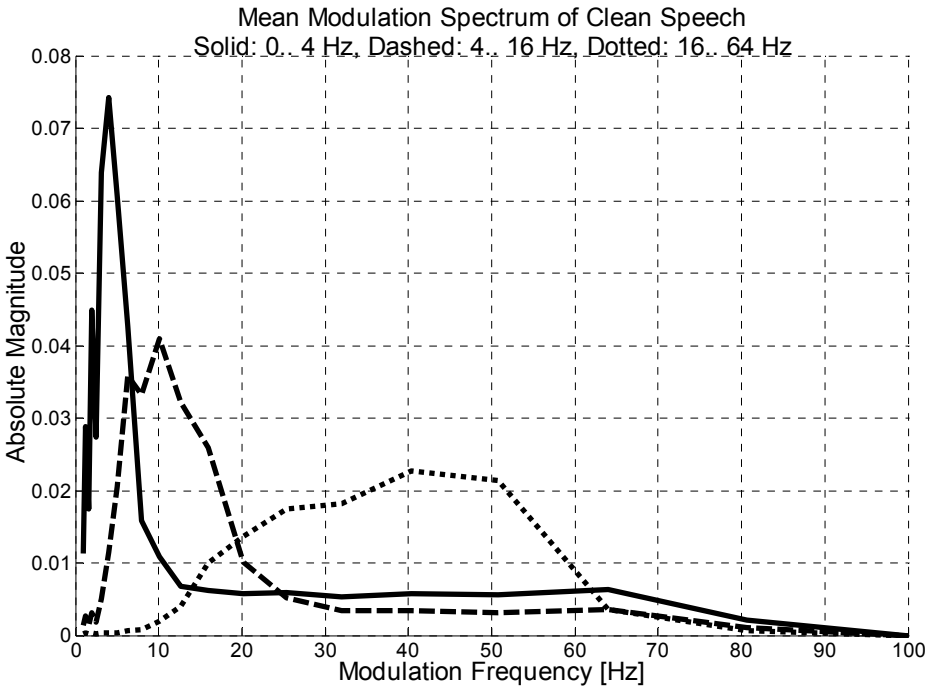


Figure 2.8: Modulation spectrum of clean speech. The modulation spectrum is subdivided into three modulation frequency bands: 0.. 4 Hz (solid), 4.. 16 Hz (dashed), and 16.. 64 Hz (dotted). Thy typical modulations for speech result in a clear peak around 4 Hz.

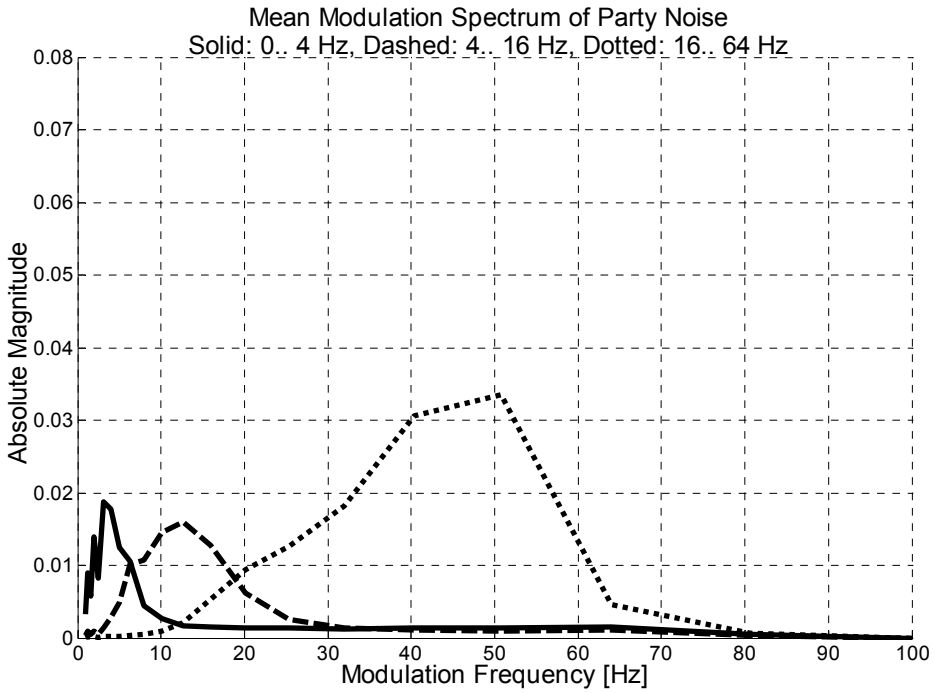


Figure 2.9: Modulation spectrum of party noise. The modulation spectrum is subdivided into three modulation frequency bands: 0.. 4 Hz (solid), 4.. 16 Hz (dashed), and 16.. 64 Hz (dotted). The modulations are quite small; there might be larger modulations in the higher frequencies for other noises, like a clattering machine.

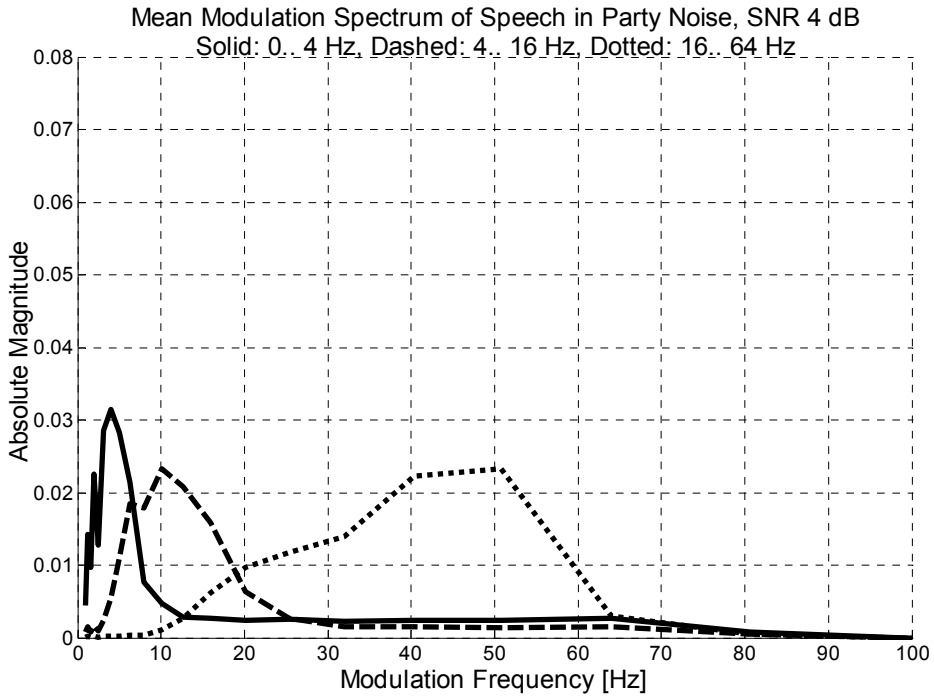


Figure 2.10: Modulation spectrum of speech in party noise, SNR 4 dB. The modulation spectrum is subdivided into three modulation frequency bands: 0.. 4 Hz (solid), 4.. 16 Hz (dashed), and 16.. 64 Hz (dotted). The small peak round 4 Hz indicates the presence of speech in the signal; it depends on the SNR.

As Ostendorf states, the detection of speech and of noise can be done reliably. The situation speech in noise however lies somewhere between these signals and can therefore not be identified easily.

In further investigations, Ostendorf made use of a Bayes classifier to study the classification using the three modulation depths m_1, m_2 and m_3 (Ostendorf et al., 1998). The scores show that clean speech is identified very well (hit rate over 90 %), whereas noise and speech in noise are confused more often (hit rate about 85 % each). Dividing the class speech in noise into two subclasses, speech in fluctuating noise and speech in stationary noise, gave only poor results.

It is not described how the three modulation depths m_1, m_2 and m_3 behave for musical signals. This, as well as the behavior of the algorithm for various other signals will be evaluated in the following section.

2.2.4 Evaluation and Comparison of the Three Algorithms

The description of the three algorithms that have been presented so far clearly shows that the emphasis is on the feature extraction stage. The choice of a pattern classifier is not in the focus at the moment – it will only follow when good features have been evaluated (see chapter 6 and 7). Thus, the evaluation in this section is restricted to the features provided by the three algorithms. A summary of these features is given in the table below.

Feature	**Describes**
Ludvigsen (1993)	
Width	Width of the amplitude envelope histogram
Symmetry	Symmetry of the amplitude envelope histogram
Skewness	Skewness of the amplitude envelope histogram
Kurtosis	Width of the peak of the amplitude envelope histogram
Lower half	Distribution of the lower part of the amplitude envelope histogram
Kates (1995) and Phonak (1999)	
TotPowdB	Mean level of the signal
MLFS	Mean level fluctuation strength
CGAV	Average of the spectral center of gravity
CGFS	Mean fluctuation strength of the spectral center of gravity
Ostendorf et al. (1997)	
m1	Modulation depth for the modulation frequencies from 0 to 4 Hz
m2	Modulation depth for the modulation frequencies from 4 to 16 Hz
m3	Modulation depth for the modulation frequencies from 16 to 64 Hz

The benefit of these features shall be investigated and compared in this section regarding the following aspects:

1. Classification capability of the features for the four sound classes speech, noise, speech in noise, music, using a comprehensive soundset.
2. Physical comparison of the features.

2.2.4.1 Sound Classification Using a Comprehensive Soundset

For the classes defined in chapter 1, a comprehensive soundset has been put together, on the one hand to evaluate the classification capability of a single feature, on the other hand to be able to train a pattern classifier, such as a neural network. The soundset for the four classes contains 287 different sounds. Each sound has a length of 30 seconds and belongs to one class; that is, no class changes occur within a sound. The table below summarizes the soundset:

Class	# of Sounds	Includes
Speech	60	Clean speech, raised voice, fast speech, dialogues, compressed speech from the radio, reverberated speech
Speech in Noise	74	Speech in social (party) noise, in in-the-car noise, in traffic noise, in industrial noise, in other noise, SNR 2.. –9 dB
Noise	80	Social (party) noise, in-the-car noise, traffic noise, industrial noise, other noise
Music	73	Classical music, pop and rock music, single instruments, singing
Total	**287**	

A more detailed description of the soundset is found in appendix A.

Each feature was computed for all 287 sounds. The feature values within one sound were averaged over the sound length, so that one feature value is obtained per sound. This means of course that any temporal information within the feature values for a sound is lost. However, the 287 values obtained for one feature allow a good first estimate of the classification ability of this feature, if they are plotted in a bar diagram. In Figure 2.11, the mean feature values per sound for the feature *width* are plotted as an example.

In this example, the feature range can roughly be divided into two areas, even if the boundaries are somewhat fuzzy. In the first area, clean speech and a few music signals can be found. The second area is occupied by speech in noise with poor SNR, classical and pop/rock music (no. 215-248), and most of the noises. The transitional zone between the two areas contains compressed speech (no. 41-50) and partly also reverberated speech (no. 51-60), speech in noise with high SNR, single instruments (no. 249-267), singing (no. 268-280) and a few non-continuous noises.

Thus, musical signals cover especially the second area and the transitional zone and overlap therefore with the classes 'speech in noise' or 'noise'. For distorted speech, the situation is similar: The more reverberation is added to the speech, the nearer it is to the class 'speech in noise'. This is equivalent for compressed speech. On the other hand, the occurrence of the class 'speech in noise' depends strongly on the SNR, its extremes are the 'speech' area and the 'noise' area. The transitional zone is therefore not a well-defined zone, and clear statements can only be made if the feature values are in area one or two.

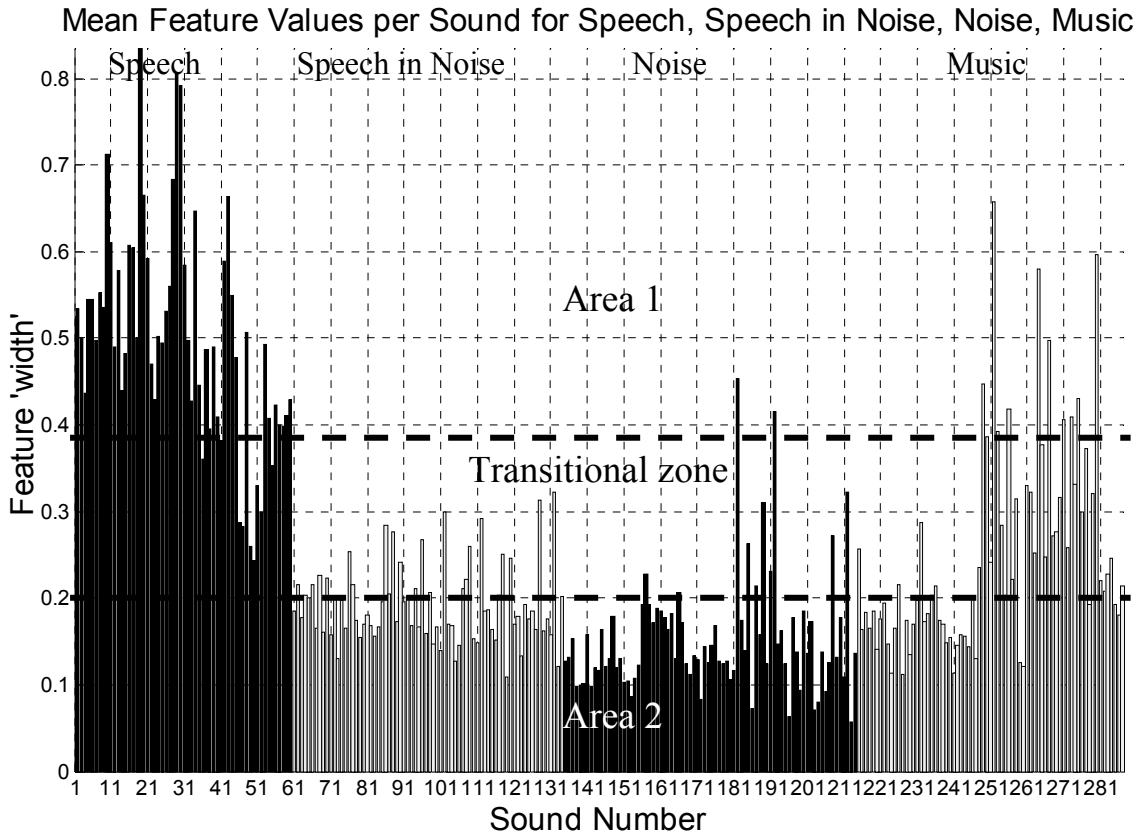


Figure 2.11: Feature "width" for 287 different sounds of the classes speech, speech in noise, noise, music, averaged over the signal length of thirty seconds. The feature range can roughly be divided into two well-defined zones and a transitional zone that contains a bit of each class.

The mean feature values per sound of the rest of the features have been interpreted in the same way, and the corresponding areas are listed in the table below. For some of the features, the transitional zone is included in the second area, indicating that the sounds of area one stand out against the rest, and no clear differentiations can be made within the rest.

Feature	Area	Sounds in this area
Width	1	Clean and slightly reverberated speech Few musical signals (some instruments and partly singing)
	Trans. zone	Some compressed or reverberated speech sounds Speech in noise, high SNR Single instrument Singing Few non-continuous noises
	2	Noise Speech in noise, low SNR Classical and pop/rock music

Feature	Area	Sounds in this area
Symmetry	1	Clean and slightly reverberated speech Singing
	Trans. zone	Some compressed or reverberated speech sounds Few non-continuous noises Speech in noise, high SNR Single instrument
	2	Speech in noise, low SNR Noise Classical and pop/rock music
Skewness	1	Clean and slightly reverberated speech
	Trans. zone	Some compressed or reverberated speech sounds Single instrument Singing Few non-continuous noises
	2	Speech in noise Noise Classical and pop/rock music
Kurtosis	1	Speech in noise *(only slightly higher than the rest)*
	2	Rest
Lower half	1	Clean and slightly reverberated speech
	2	Rest
TotPowdB	-	*General indicator low level – high level*
MLFS	1	Clean and slightly reverberated speech
	Trans. zone	Some compressed or reverberated speech sounds Speech in noise, high SNR
	2	Rest
CGAV	1	Low frequency signals, e.g. in-the-car noise, bass instruments Speech in in-the-car noise
	Trans. zone	Rest
	2	High frequency signals, e.g. frying, printer, rain Speech in these noises
CGFS	1	Speech Speech in noise
	2	Noise Music *(the two areas are quite overlapping)*
M1	1	Clean and slightly reverberated speech
	Trans. zone	Some compressed or reverberated speech sounds Speech in noise, high-medium SNR Some instruments Singing Few non-continuous noises
	2	Speech in noise, low SNR Noise Classical and pop/rock music
M2	-	*Similar to M1, but less distinct, except in-car noise*
M3	-	*No clear picture, except noises with fundamental frequencies < 64 Hz:* in-car noise, speech in the car, typewriter noise, jackhammer noise

The following conclusions can be drawn from these results:

- Clean speech can be well isolated by many of the above features.
- Speech in noise lies for all features somewhere in the transitional zone between speech and noise; signals with high SNR are more on the speech side, signals with low SNR more on the noise side. This dispersion makes it difficult to identify the class speech in noise.
- Musical signals cannot be isolated with any of the above features. It is however partly possible to separate music with many instruments from music with a single instrument or from singing. The former lies in the area of noise, the latter in the area of speech in noise or speech.
- Continuous noise signals cluster together for many of the above features. The feature values of non-continuous noises tend to move towards speech in noise or speech, due to the pauses or level fluctuations in the signal.
- Most of the features behave very similar; they seem to describe the same signal characteristics. This will be confirmed in section 2.2.4.2, where a physical comparison of the three algorithms is made.
- Two of the features describe really different signal characteristics. These are the features *TotPowdB*, which describes the overall level, and *CGAV*, which enables to identify low- and high-frequency signals.

2.2.4.2 Physical Interpretation of Features and Qualitative Comparison of the Three Algorithms

It has been shown that all of the three algorithms allow a reliable discrimination of clean speech from other sounds. Yet, none of the algorithms allows robust further sound classification such as for example the discrimination of music from noise, etc. It was therefore suspected that the three algorithms – even though they seem to be very different at first sight – might all describe the same signal characteristics.

In order to understand which auditory mechanisms are represented by each of the algorithms and to carry out a physical comparison of the algorithms, a physical interpretation of the features computed by each algorithm is required. In the following, an interpretation of all features and a qualitative comparison of the algorithms based on this interpretation is attempted.

The features of the Ostendorf algorithm (Ostendorf et al., 1997) are easiest to interpret. The Ostendorf algorithm effectively attempts to partly model the temporal signal processing of the auditory system by extracting the strength of modulation in particular modulation frequency bands (Dau, 1996). Modulations are inherent in the signal envelope; as described earlier Ostendorf computes her features from the magnitude and frequency content of the signal envelope.

Interestingly, also the Ludvigsen algorithm (Ludvigsen, 1993) investigates the envelope of the signal. The amplitude distribution described by means of the extracted percentiles refers to the signal envelope, and thus, in particular the width of the distribution (2.2) is just another means for describing modulation strength.

With the mean level fluctuation strength *MLFS*, the Kates algorithm (Kates, 1995) also operates on a kind of envelope (see equation (2.11), the mean level is the average spectral

amplitude over all frequencies). Remember also that the original idea of Kates for the discrimination of different noises was to use a variant of the Kollmeier and Koch modulation spectrum (Kollmeier and Koch, 1994), which is also the basis of the Ostendorf algorithm. With the proposed simplification, however, the resulting *MLFS* feature (2.10) becomes strongly similar to Ludvigsen's *Width* feature, since both *Width* and *MLFS* describe the amplitude distribution of the envelope. Thus, Kates' *MLFS* feature can also be considered as a kind of modulation strength.

The reason why all of the three algorithms allow the identification of clean speech so well is now clear: all of the algorithms provide a means for *measuring amplitude modulations* of the investigated signal. Due to the pauses and to the frequencies of phonemes, syllables, words, and sentences (see also section 2.2.3), speech signals are characterized by very strong amplitude modulations, in particular at low modulation frequencies. Amplitude modulations are apparently a key feature for the discrimination of speech from other sounds. While the Ostendorf algorithm directly detects such *amplitude modulations*, the Kates and Ludvigsen algorithms take the detour of computing the *amplitude distribution* of the signal envelope. The pauses of highly fluctuating signals such as speech lead to broad and asymmetrical amplitude distributions since they tie the distribution to very low values which are not present in continuous signals. Note that these pauses are the main reason why speech signals incorporate such strong modulations.

Hence, we can qualitatively say that *Ostendorf's modulation depths, Ludvigsen's width, and Kates' mean level fluctuation strength are equivalent in terms of signal classification*. This is also explained graphically in Figure 2.12.

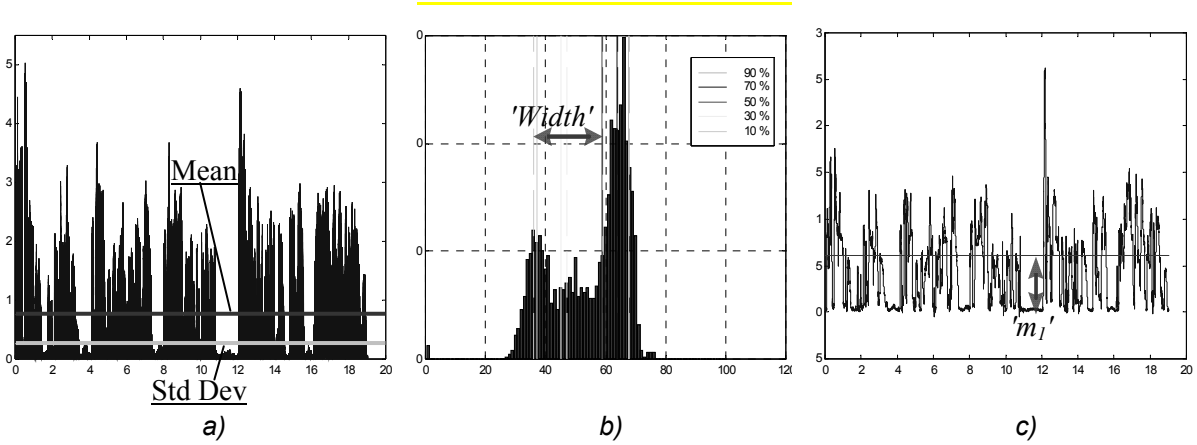


Figure 2.12: Equivalence of Kates MLFS, Ludvigsen width, and Ostendorf modulation depths shown on a clean speech signal. a) Kates MLFS feature determined according to (2.10): dark line: mean of rectified amplitude, light line: standard deviation estimated by (2.10). b) Arrow: Ludvigsen width feature determined according to (2.2). c) Arrow: Ostendorf modulation depth m_1.

The table below lists a qualitative physical comparison of the three algorithms comprising all the knowledge gathered so far.

	Ludvigsen (1993)	**Kates (1995), Phonak (1999)**	**Ostendorf et al. (1997)**
goal of approach	discrimination between impulse-like and continuous signals	classification of different background noises	discrimination between clean speech, speech mixed with noise, and noise
underlying **idea**	investigation of the amplitude distribution	simplified investigation of the modulation spectrum, leading to a description of the amplitude distribution and the frequency spectrum	investigation of the modulation spectrum
described **auditory mechanisms** and corresponding **features**	temporal signal processing (P10 ... P90, *Width*, *Symmetry*, *Skewness*, *Kurtosis*, *Lower half*)	- 'loudness' (*TotPowdB*) - temporal signal processing (*MLFS*) - spectral profile (*CGAV*, *CGFS*)	temporal signal processing (m_1, m_2, m_3)
modulation **strength** and **speed**	modulation depth across all modulation frequencies ('broadband modulations')	modulation depth across all modulation frequencies ('broadband modulations')	- m_1: modulation depth across low modulation frequencies (0 .. 4 Hz) - m_2: ... medium mod. frequ. (4 .. 16 Hz) - m_3: ... high mod. frequ. (16 .. 64 Hz)

To summarize, the following conclusions can be drawn from the qualitative comparison of the three algorithms:

- All of the three algorithms model temporal auditory signal processing (temporal envelope) by describing amplitude modulations in one way or another.
- Only the (modified) Kates algorithm describes further signal characteristics, namely the energy content of the signal (*TotPowdB*), and information about the spectrum (features *CGAV* and *CGFS*), but this in only a very rudimentary way.
- The *modulation depth* (Ostendorf), the *width* of the amplitude distribution of the signal envelope (Ludvigsen), and the *ratio of mean to standard deviation* of the spectral amplitudes (Kates) all describe amplitude modulations (that is, temporal auditory signal processing) and are hence equivalent in terms of signal classification.
- The Ostendorf algorithm describes the amplitude modulations with most detail by computing information about both the *modulation depth* and the corresponding *modulation frequencies*.
- The Kates and the Ludvigsen algorithm describe modulations by means of the *amplitude distribution* of the envelope. No information about the frequencies of the amplitude modulations is available.
- The Ludvigsen algorithm allows to correctly model the amplitude distribution in some detail (width, symmetry, etc. depending on the features formed from the extracted

percentiles), while the Kates algorithm only provides an estimation of the width of the distribution.

Amplitude modulations and spectral profile are just two of several features which serve to analyze an auditory scene, as will be shown in chapter 4, where the principles of Auditory Scene Analysis will be explained. It can be assumed that automatic sound classification will be improved if more of these features will be used.

2.2.5 Noise Classification with Neural Networks by Feldbusch

An algorithm for automatic switching between different hearing programs is presented by Feldbusch (1998). The main classes to be identified are speech, babble noise and traffic noise. Further classes shall be speech in babble noise, speech in traffic noise, music, nature, and possibly alarm signals.

A large set of features is calculated both in the time domain and in the frequency domain. In the time domain, the zero-crossing rate, the maximum of change of the zero-crossing rate and some derivatives are computed, in the frequency domain, either Fourier or wavelet coefficients are taken as features, together with some features that are extracted from each of the coefficients, such as the mean and maximum of each coefficient in a certain time window. Feldbusch states that for practical applications, the features should be independent of the signal level.

This results in over 150 different features that are fed into a neural network. Several network topologies have been tried; a so called Time Delayed Neural Network (TDNN) was used to analyze the temporal structure in the features. The classification with the TDNN was quite poor, however. Feldbusch assumes that some of the features express already the temporal dynamics of the signal, and that the temporal changes of the other features do not contribute much to the classification.

The best results were obtained with a neural network with one hidden layer. The classification of the main classes was quite good, for music and nature however bad. Feldbusch states that these classes may contain very different signals, which makes it difficult to cluster the signals into a class. The large number of features might make it difficult to train the network properly; a preselection of the best features is currently in progress (Feldbusch, 2001). Feldbusch recommends to apply a post processing stage at the output of the network to make the system more inert and robust.

2.2.6 Noise Classification with HMMs by Nordqvist

Nordqvist (2000) designed another automatic classifier for different listening situations, which shall enable the hearing instrument to switch between different filters, look-up tables or other settings, such as directional microphone, noise reduction and feedback suppression. The classes are speech, babble noise, traffic noise, subway noise and outdoor noise.

The algorithm is based on hidden Markov models (HMM). As features, LPC coefficients are used, which are vector quantised before being fed into the HMM. A post processing block controls the switching from class to class (that is, between the HMMs), in order to allow different switching delays for different class transitions.

The long-term classification error is close to zero. At first sight, it seems however quite easy to recognize two of the classes: Subway noise contains very low frequencies, and outdoor noise probably just means that the average signal level is low. Anyway, the use of a HMM as

classifier appears to be a good way of identifying the temporal structure that lies in the features. It would be interesting to know how well other noises or musical signals can be recognized with this approach.

2.3 Other Sound Classification Algorithms

2.3.1 Environmental Noises and Alarm Signals

Goldhor (1993) calculated cepstral coefficients for 23 familiar environmental sounds, such as door bell, ringing phone, car engine, vacuum cleaner, running water, closing door, etc. He then performed a cluster analysis and found that only the cepstral coefficients representing low frequency spectral and temporal variations were required in order to obtain accurate classification. Obviously, cepstral coefficients are useful to separate transient noises with quite different spectral and temporal variations. This might not be the case for more stationary sounds, like speech in noise and music.

Gaunard et al. (1998) and Couvreur et al. (1998) present a method for classification of five types of noise events: Car, truck, moped, aircraft, and train. These noises all have a transient nature, as the recording was made with the vehicles passing by. The best performance was achieved with LPC coefficients and a five-state left-right HMM. For this type of noise events, a HMM seems indeed the best solution, as it is able to model the temporal structure, that is the different phases in the transient noises.

Oberle and Kaelin (1995) and Oberle (1999) tried to identify four different alarm signals: Car horns, streetcar bells, streetcar rings, and phone rings. Cepstral LPC coefficients and the energy were taken as features. An ergodic HMM with four states outperformed a minimum-distance classifier and a neural network[3]. Again, the sounds have a transient nature and are quite short, with a silent phase at the beginning and the end of the alarm signal.

It would be interesting to see how a HMM is suited to model more stationary noises and sounds of other classes. There, the temporal structure has quite a different character; it is more the fluctuations within the sound itself that contributes to the structure than the sound appearing and fading out again.

2.3.2 Music and Multimedia

Martin (1998) and Martin et al. (1998) present some reflections about the analysis of musical content through models of audition. The practical goal is to develop a robust automatic recognition system for musical multimedia applications, such as automatic labeling of radio music pieces. They investigated the use of different features for three prototype instruments: Violin, trumpet and flute. Features that appear to be important include: Pitch, frequency modulation (jitter), spectral envelope, spectral centroid, intensity, amplitude envelope, amplitude modulation, onset asynchrony, and inharmonicity. As will be discussed in chapter 4 about Auditory Scene Analysis, the interaction of these features is crucial for scene analysis. Martin and Kim (1998) further elaborated the ideas and constructed an algorithm to identify 14 orchestral instruments of the families string, woodwind and brass. They extracted the features named above from each instrument tone, plus other features that are derived from the initial features, for example the variance of the pitch or of the spectral centroid. Initially, the

[3] The same sound set was investigated with neural networks by Leber (1992).

classification was done for the instrument families and the particular instruments. The woodwind family, however, was quite heterogeneous, and some of the instruments were more similar to the brass family than to the rest of the woodwind family. For this reason, the instrument taxonomy was revised, dividing the woodwinds into more homogenous groups: String, flute and piccolo, brass and reeds. This resulted in a better classification. This shows that it is obviously essential to incorporate a priori knowledge about the signals into the structure of the classification algorithm.

Soltau et al. (1998) describe a music type recognition system that can be used to index and search in multimedia databases. The music types that are required are rock, pop, techno and classic. A so called ETM-NN (Explicit Time Modeling with Neural Network) was compared with a hidden Markov model, using cepstral coefficients as features. With an ETM-NN, the temporal structure in the signal is taken into account by extracting a sequence of abstract acoustic events and analyzing how often an event, or event pairs or triplets, occur in the sequence. A neural network that is trained with this information outperforms a HMM clearly. The use of the acoustic event sequence is probably a better way to explore the rhythms of different music classes than explicitly searching for the rhythm in the signal, for example by autocorrelating the signal. Changes in the rhythm or weak beats in the signal might make it difficult to see the desired information in the autocorrelated signal.

A sound source identification system for music is presented by Kashino and Murase (1999). Piano, flute and violin notes that are played simultaneously are identified by comparing the input signal to a bank of waveforms that is subdivided according to the fundamental frequency. In a next step, the music streams are recovered by linking the consecutive notes. A possible application of such a system is an automatic transcription system. Although the recognition of notes is a layer lower than the classification intended in this thesis, the use of a template matching procedure could be useful for example to classify different spectral profiles.

Lambrou et al. (1998) classify the three musical styles rock, piano, and jazz, using statistical features in time and wavelet transform domain. The features include first order statistics, that is, mean, variance, skewness and kurtosis, and second order statistics, that is, angular second moment, correlation, and entropy. Additionally, the zero crossing rate was calculated. Together with a minimum distance classifier, a high hit rate was achieved. However, only four samples of each class were used for the training, which is certainly not enough for a general description of each class. The intra-class variance will probably be much higher when more training samples are used, which might reduce the hit rate.

Scheirer and Slaney (1997) report on the construction of a real-time system capable of distinguishing speech signals from music signals. A possible application is in a system that performs automatic speech recognition on soundtrack data. For such a system, it is important to be able to distinguish which segments of the soundtrack contain speech. Thirteen different features were used together with three different statistical classifiers, but the univariate discrimination performance of each single feature was also investigated. The best three features were: 1) the modulation energy at 4 Hz, where speech has a characteristic peak, 2) the variance of the spectral flux, which is the difference of the spectrum magnitude from frame to frame; music has a much higher rate of change, and goes through more drastic frame-to-frame changes than speech does, 3) the pulse metric, a feature which uses long-time bandpassed autocorrelations to determine the amount of "rhythmicness" in a 5-second window. Strong beats without tempo changes are detected quite well. The feature values for

music signals without beat, however, will be in the same range as for speech, and will therefore not be useful. If the speech signal is not clean but distorted by reverberation or noise, the features values might be more similar to those of music; this remains to be investigated.

Last but not least, Zhang and Kuo (1998, 2001) propose a system for online segmentation and classification of audiovisual data based on audio content analysis. The audio signal from movies or TV programs is segmented and classified into twelve scenes using only four features. These features are energy, zero crossing rate, fundamental frequency, and spectral peak tracks. The last feature, being the most expensive in computation, is only calculated when desired, that is for the purpose of characterizing sounds of song and speech. A rule-based heuristic procedure is used to segment and classify the signal. In a first step, it is distinguished between silence and signals with and without music components. The segments with music components are further classified into pure music, song, speech with music background, environmental sound with music background, and harmonic environmental sound. The segments without music components are divided into pure speech and non-harmonic environmental sound. The rule-based procedure, together with the four features, works astonishingly well for most of the classes. It seems that this heuristic approach has a better performance than a pattern classifier that is automatically trained, due to the fact, that not only the feature values, but also their change patterns over time and the relationship among the features are taken into account. Furthermore, the rule-based classifier is fast and easy to compute.

The system was extended by a second stage that further classifies the sounds into finer classes, such as rain, bark, explosion, foot step, laugh, river, thunder, and windstorm (Zhang and Kuo, 1999). This fine-level classifier is performed by feeding the 65 bins of the Fourier-transformed signal into a HMM. A state of the HMM represents then a certain spectral profile (Zhang and Kuo call this timbre, a bit incorrectly), and the transitions between the states stand for the change of spectral profile (called rhythm, if the changes occur at regular periods of time). Unfortunately, only few samples were used for training and testing of the HMM, which might not result in a good generalization of the classes, that is in a robust classifier. Nevertheless, it is a good idea to use a HMM for modeling the different spectral profiles in a sound, even if this might not work so well in more general sound classes, such as simply music.

2.4 Discussion and Conclusions

2.4.1 Overview

From the point of view of hearing instruments application, which demands the main sound classes 'speech', 'noise', 'speech in noise', and 'music', the following conclusions of the state of the art in sound classification can be drawn:

- There are currently three algorithms that are exploited in hearing instruments (Ludvigsen, 1993; Phonak, 1999; Ostendorf et al., 1997/1998). These algorithms allow a robust classification of clean speech signals from other signals. Music however can not be distinguished, and it is only partly possible to separate noise from speech in noise.

- The two additional algorithms that are designed for hearing instruments (Nordqvist, 2000; Feldbusch, 1998) are neither able to classify music accurately, and are only partly designed to detect speech in noise.
- There is one algorithm that is able to classify the essential classes (Zhang and Kuo, 1998, 1999, 2001). It was designed for online segmentation and audio classification of movies or TV programs. Unfortunately, the publications from Zhang and Kuo were only available at the end of this thesis, which means that the investigations described there are not yet taken into account. They will, however, contribute to the concepts for future work in this topic.
- Many other algorithms exist that are able to classify subsets of the desired classes, such as noise types (Goldhor, 1993; Gaunard et al., 1998; Couvreur et al., 1998), alarm signal types (Oberle and Kaelin, 1995; Oberle 1999), music types (Martin, 1998; Martin et al., 1998; Martin and Kim, 1998; Soltau, 1998; Kashino and Murase, 1999; Lambrou et al., 1998), or speech and noise (Scheirer and Slaney, 1997). 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 more detailed layer than is initially desired in this thesis, that is, for classes with many different sounds, such as different noises. Nevertheless, concepts from these algorithms could be used for finer classification; first however, the detection of the main classes (speech, noise, speech in noise, music) has to be performed robustly.
- If the general block diagram of a sound classification system (Figure 2.1) is considered again, it can be stated that in most algorithms, the emphasis lies in the feature extraction stage. Without good features, a sophisticated pattern classifier is of little use. Thus, the main goal in this thesis will be to find appropriate features before considering different pattern classifier architectures.

A discussion about relevant aspects is now made separately for the three blocks "feature extraction", "classifier" and "post processing".

2.4.2 Feature Extraction

Many features have been described in the various sound classification algorithms, and several of them are related to each other in that they describe similar signal characteristics. The table below gives an overview of the feature categories proposed so far:

Feature	**Comment**
Energy / Power / Amplitude	General information if signal is silent or loud
Amplitude modulations / Level fluctuations / Zero crossing rate	Information about breaks in the signal, e.g. speech pauses
Amplitude onset (a)synchrony	Low-frequency components may have a different rise time than high-frequency components, e.g. for musical instruments
Frequency bins / Spectral profile	For single sounds; probably not suited for classes with many different sounds (large intraclass variance)
Spectral centroid	General information if signal is low- or high-frequent
Frequency modulation / Fluctuations of spectral centroid	Detection of vibrato / jitter

Feature	Comment
Fundamental frequency / Pitch / (In)harmonicity	Detection of musical sounds (can be music or noise)
Cepstral coefficients / LPC coefficients	For short, transient sounds, together with HMM or the like
Rhythm / Beat / Pulse	Detection of rhythmic sounds (can be music or noise)
Mean / Variance / etc.	Statistical features can be derived from most of the above features; e.g. mean modulations, variance of pitch, etc.

The next step in feature selection will be the choice of features from above that seem to be valuable for the classification of the desired classes. The choice of good features depends of course on the desired classes. Thus we can not just take all the features that have been described in the classification algorithms so far, because many of these algorithms aim to identify distinct sounds or subclasses of the sound classes desired in this thesis. Furthermore, good features should possess large interclass mean distance and small intraclass variance (see Kil and Shin, 1996). The latter may be quite difficult if the desired classes are of a general nature and contain various signals with different structures, like in the class 'noise', for example. The features should generally also be independent of the signal level (unless they are intended to describe the level itself, of course).

A promising approach is to consider how the auditory system performs sound classification. This is why in chapter 4, an overview of Auditory Scene Analysis will be presented, and motivated by this, further features will be chosen in chapter 5.

2.4.3 Classifiers

Different pattern classifiers are exploited in the sound classification algorithms described above. This includes minimum distance and Bayes classifiers, neural networks, hidden Markov models, and heuristic rule-based approaches. However, the pros and contras of these classifiers are not yet obvious. Thus, an overview of pattern classifiers will be presented in chapter 6, comparing different types as well as their complexity.

The classifier block may, however, also consist of different stages, for example of a neural network for coarse classification and a heuristic approach for the finer classes (or the other way round, as Zhang and Kuo, 1999, have done it). Again, Auditory Scene Analysis may give some hints how to perform a multistage approach.

2.4.4 Post Processing

The post processing block is intended to control the transient behavior of the sound classification system. It is an open question how fast or slowly the classification shall react to signal changes. This depends on the preferences of the hearing instrument users. Thus, in the next chapter, a study is presented that tries, among other things, to give an answer to this question (see the section below).

2.4.5 Further Work

In the following, the knowledge gained in the overview of the state of the art in sound classification shall be extended by the following investigations:

- The principles of Auditory Scene Analysis will be reviewed, in order to find features and mechanisms that are relevant for sound classification.
- Based on this and on the features that have been described earlier, a selection of features will be made according to the a priori knowledge of the signals and classes to be identified.
- After that, a number of different pattern classifiers will be compared and evaluated together with the extracted features.

Before this is carried out, an evaluation study is presented in the next chapter. In this study, the usefulness and acceptance of an automatic program selection mode in hearing instruments from the point of view of the user were investigated.

3 A Field Study with Hearing Impaired Subjects

3.1 Introduction

It has been described in chapter 1 that hearing impaired persons prefer different settings for different listening situations. Furthermore, they named a number of situations in which they judged it most important to hear well. The latter was the motivation for the sound classes to be recognized (at least the classes 'speech', 'speech in noise', 'noise', 'music'). It did, however, not show how an *automatic* switching system that is based on the sound classification would be perceived by the hearing impaired subjects. This is why a study was carried out about the usefulness, acceptance and problems of such an automatic program switching system.

The modified Kates algorithm that has been described in the previous chapter (Phonak, 1999) was selected as the basis for an automatic program switch in a commercially available hearing instrument. The hearing instrument automatically switches between two programs depending on the acoustic environment, with the aim of achieving the highest possible speech intelligibility in all listening situations. For quiet environments, that is also clean speech, or when there is only background noise, the first program is activated by the automatic program selection, and for speech in background noise, the second program. Thus, the task of the classification system is to identify speech within noise from all other acoustic situations. For speech in noise, the use of adaptive directional microphone technology, noise suppression, and appropriate frequency response and compression parameters aim to improve speech intelligibility in adverse environments.

With the help of a questionnaire and subjective reports made by hearing instruments users, the study aimed to establish whether the automatic program switching mode of the test instrument changed between programs in the desired way and if it was judged helpful by the hearing-impaired subjects. The study was carried out at three different sites; see Boretzki et al. (2001), Büchler (2001), Gabriel (2001).

3.2 Method

63 subjects with moderate hearing losses were binaurally fitted with in the ear or behind the ear instruments. After several weeks of acclimatization and experience, a questionnaire had to be filled out. The table below summarizes the questions that had to be answered:

Questionnaire	
Process of switching	(1) Are you aware of program changes?
	(2) How do you rate the frequency of switching?
	(3) Does the instrument change programs when you expect it to?
Program choice	(4) Is the program choice suitable to the situation?
	(5) When does the program choice seem to be unsuitable?
Usefulness	(6) How often was the automatic program choice used?
	(7) How useful is the automatic program choice?

For each question, answers with five to seven scaled categories were offered (for example, not useful, of little use, a bit useful, quite useful, very useful). For question (5), ten specific situations were presented for rating[4]. In addition, the subjects were allowed to provide their own comments.

3.3 Results

Question (1) was important to check whether the rest of the questions could be answered at all. Two third of the users noted program switching regularly and were therefore able to answer the other six questions.

Question (2), regarding the switching frequency, was perceived quite differently by the individual subjects (Figure 3.1). For the majority, however, it was considered to be within acceptable limits. About three-quarters (74 %) of all participants found the frequency "exactly right", "a bit too often" or "a bit slow". For the remaining quarter (26 %) of the respondents, the frequency of switching was "too seldom" or "too often". This indicates that the possibility of adjusting the frequency of switching would be useful sometimes to satisfy the individual preferences of the user.

Regarding question (3), about 70 % of the subjects found that the instrument switched programs mostly as expected. For the rest the switching seemed more arbitrary. This may be related to the delay in the switching mechanism (the same classification has to be met for at least ten seconds in order to avoid too high a switching rate between programs), as well as to individual user differences in the expectation of program choice for certain acoustic environments (see Figure 3.2).

About 75 % found that the program choice matched the situation quite well (question (4)).

[4] Question (5) was only asked in this form at one of the three sites and was therefore only answered by 22 subjects.

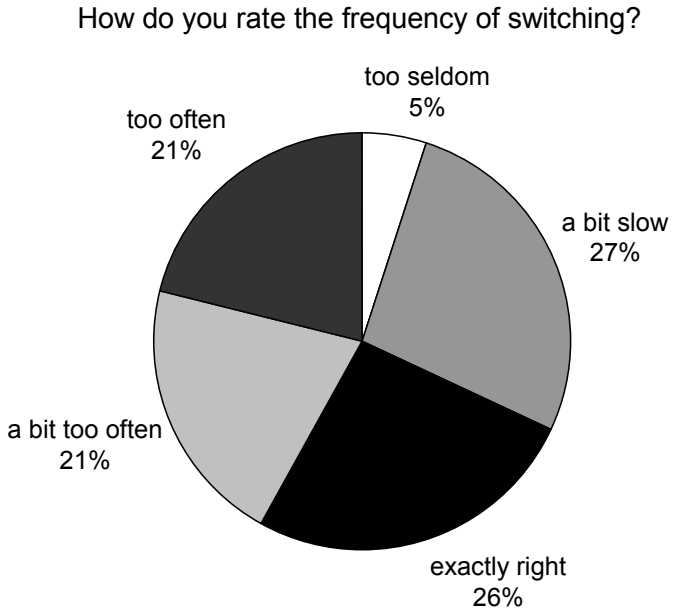


Figure 3.1: The switching frequency was judged quite differently, which means that an individual setting is desired.

The answers to question (5) are shown in Figure 3.2. On the whole, there were three situations in which the automated system did not always comply with the expectations of the subjects:

- Speech in noise: With speech in the presence of noise, the instrument will switch on the adaptive directional microphone and noise suppression. Some subjects preferred that only the speech of the person they were looking at be amplified fully. Others found this unsatisfactory at times because they missed what other speakers in a group were saying. In this situation, however, it is unlikely that even a "perfect" algorithm could predict the wishes of the individual user.
- Traffic noise: In background noise without speech the directional microphone and noise suppression are usually not activated. The reason for this is that, in a traffic situation for example, important noises such as approaching vehicles usually should be perceived. The disadvantage is that also many "unwanted" sounds are heard, and some users find this annoying. Additionally, as with speech-in-noise situations, individual preferences can differ.
- Music and singing: The signal identification system used is less robust in these situations. Music is sometimes classified as 'speech in noise'. The activation of the directional microphone and background noise suppression is not suitable when it is music the user wishes to enjoy. A more reliable distinction between speech in noise and music would be desirable.

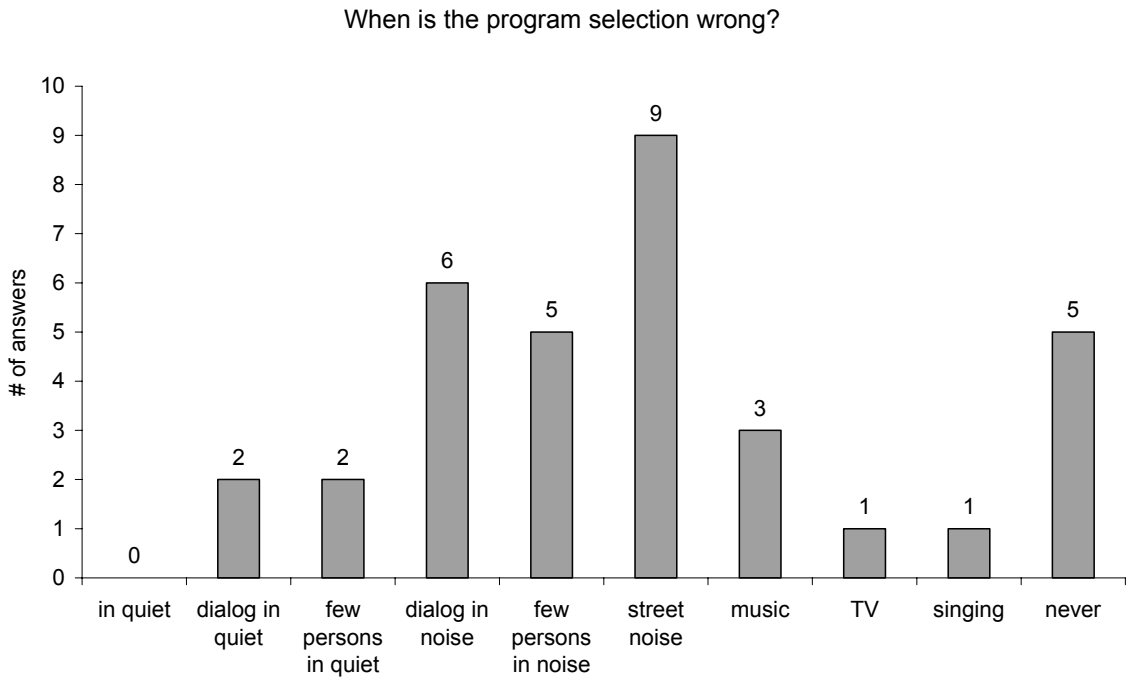


Figure 3.2: The program selection is perceived as being wrong in mainly three situations: Conversations with few persons in noise, traffic noise, and musical signals.

The automatic program choice was used in about 75 % of the time that the hearing instrument was worn (question (6)).

Regarding the usefulness of the automatic program choice (question (7)), three quarters (75 %) of the test subjects found the automatic system "quite useful" or "very useful" and only 20 % reported they would rather not use it. Therefore, the automatic program selection was found to be a valuable and desirable function. This result is consistent with the finding that the automatic program selection was usually suitable for a particular environment (question (4)).

3.4 Discussion and Conclusions

This study investigated the usefulness and acceptance of an automatic program selection mode in hearing instruments from the point of view of the hearing instrument user. It was shown that the automatic switching mode of the test instrument was deemed useful by a majority (75 %) of test subjects. The program choice at any given time was found to be appropriate for the majority of situations.

With regard to the switching frequency of the automatic program selection, there were clear individual differences. For some subjects the system switched too often, for others not often enough. A fine-tuning parameter for the acoustical criteria of the program selection feature would be useful in order to better satisfy the individual preferences of hearing instrument users.

'Speech in noise' was a situation in which the choice of program by the automatic program selection mode often did not concur with the wishes of the test subject. Optimization of the automatic selection mode may be possible through further analysis of the acoustical

parameters of these environments. The description 'speech in noise' can refer, however, to a multitude of different acoustic situations, and the transition between acoustic environments is variable by nature. It is therefore difficult to delineate clear boundaries.

Even a "perfect" classification of a given acoustic environment still does not solve the problem of individual user preferences at any given time. The suitability of a given program depends on whether the user is actively listening and needing high speech intelligibility from in front, or whether, for example, the user is also wanting good sound clarity from the sides and from behind. For example, one subject preferred the instrument not to switch in traffic noise while conversing with someone walking next to him. In this case, the reduction of sound from the sides is not suitable. The option to use the manual program override is a logical addition to the automatic program selection for these types of situations.

These results show that an automatic sound classification system is appreciated very much by the user, even if its performance is not perfect. This is a strong motivation for the research described in this thesis. Furthermore, the need for a refinement of the classification into at least the classes 'speech', 'speech in noise', 'noise' and 'music' is clearly shown.

4 Principles of Auditory Scene Analysis

4.1 Introduction

A person is talking to another in a noisy situation, such as in a restaurant; the listener is able to understand by somehow focusing on the sound that originates from the talker. An orchestra plays a concert; the audience can hear out many of the different instruments that are being played simultaneously. How does the listener segregate auditory objects from each other? People, when asked how they do it, are inclined to say that they solve the problem by simply paying attention to one of the sounds at the time. In saying this, they imply that the parts of the same sound are somehow a coherent bundle that can be selected by the process of attention. However, the only thing received by the ear is a pattern formed by pressure changes over time. If we look at a graph of a complex signal, there is nothing obvious in it that tells us how many sources there are or how to take them apart. The auditory system is faced with the problem of separating the sources from this complex, composite signal that reaches the ears.

Auditory Scene Analysis tries to answer the questions that arise from the above description: Which mechanisms and processing strategies are used by the auditory system to fuse together events in the spectrotemporal pattern and to segregate them into multiple sources? Which is the relevant information in the signal that allows to perform this task? Where in the auditory system do the processing stages take place? Which are the more peripheral, which the more central processes?

In this chapter, the mechanisms that seem to be responsible for Auditory Scene Analysis are investigated in a hierarchical way. First, some definitions are given, followed by an overview of the auditory system and a localization of some of the mechanisms. Then, features are described that seem to play a key role in auditory grouping, such as spectral separation, spectral profile, harmonicity (pitch), onset and offset (timbre), coherent amplitude and frequency variation, spatial and temporal separation. After that, formation and segregation principles are discussed, which serve as basis for the grouping of events and the separation of sounds. This chapter follows in a great extent the reports of Mellinger & Mont-Reynaud (1996), Yost (1991) and Yost & Sheft (1993) who give an excellent introduction into the topic (for an extensive insight in Auditory Scene Analysis, see Bregman, 1990). A number of references have been added, especially to experiments from Darwin & Carlyon (1995) that provide psychoacoustical evidence to the various assumptions that are made. Experiments are a very important means in Auditory Scene Analysis, because there are only two ways, how assumptions can be proven: With psychoacoustical and neurophysiological measurements.

Reflections on how Auditory Scene Analysis can be used for sound classifications conclude the chapter.

4.2 Definitions

Before giving an overview of the stages involved in Auditory Scene Analysis, some important terms that will be used throughout this chapter have to be introduced. Mellinger & Mont-Reynaud (1996) and Yost (1991) give the following definitions:

4.2.1 Feature – Event – Source

The terms feature, event, and source are used to refer to different levels of organization of sounds in the auditory system. A *feature* is a part of the sound signal occurring at a specific time and frequency, for example the onset of sound energy at a specific frequency, or a change in frequency in the harmonic of a pitched sound. Feature filtering in the auditory system is strongly data driven, meaning that the context of higher level objects plays very little role in it. While a feature is instantaneous, an event extends over a range of time and perhaps also over multiple frequencies. We call an *event* an auditory phenomenon of relatively short duration that exhibits constancy or continuity through time. It has an onset and an offset and represents the lowest time-extensive perceptual entity, for example a single note in music or a syllable in speech.

Features and events are auditory phenomena, points or regions in time-frequency space. In contrast, an auditory *source*, or *stream*, is a perceptual object, more permanent than an event, to which an explanation is attached, for example an instrument in an orchestra or a voice.

4.2.2 Fusion and Segregation

Fusion or *formation* processes form each of the auditory images from the neurally coded complex sound field. Event formation groups lower level features into an event, while source formation assigns one or more events to a source. *Segregation* or *separation* processes refer to the separation of one auditory image from other auditory images. While fusion emphasizes the grouping effect within the objects at one level, segregation focuses on the partitioning into distinct groups. The two processes occur in tandem.

Sequential formation is a grouping process that associates entities over *time*; *simultaneous* fusion is the complementary grouping process, a term applied to entities that happen concurrently, across *frequency*. These distinctions can be applied at the levels of features, events, and sources.

4.3 Overview of Scene Analysis in the Auditory System

An overall view of the stages involved in Auditory Scene Analysis shall be given now, and it will be tried to assign some of these mechanisms to structures in the auditory system (based on Warren, 1999). In Figure 4.1, the anatomy of the entire ear is depicted, and Figure 4.2 shows the hierarchical structure of the stages that will be discussed in the following. Generally, it can be stated that the farther away the processes are from the periphery, the less is known of their exact function.

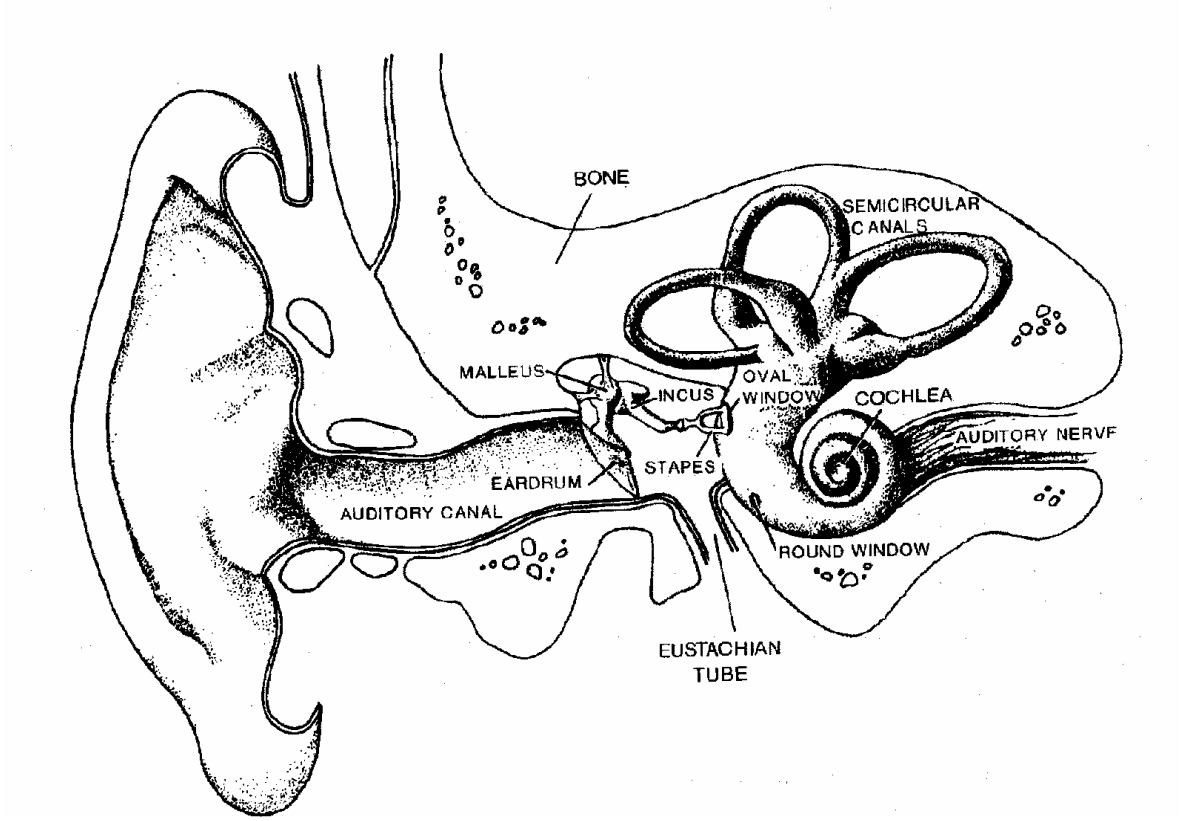


Figure 4.1: Anatomy of the entire ear with outer, middle and inner ear (from Warren, 1999).

4.3.1 Peripheral Ear Filtering and Transduction

The first stage of the auditory system consists of the filtering done by the *outer ear* (and head). The pinna supports localization, as it has a direction-specific effect on the intensity of especially the high frequencies. The *ear canal* is more than just a passive conduit: It forms a resonant tube which amplifies frequencies in the range of 2 to 5.5 kHz, with a maximum at about 4 kHz, where the amplification is about 11 dB. The influence of these structures on the signal is of course not perceived as such, but results in improved spatial localization and perception.

The *middle ear* performs an impedance transformation to adapt the transduction of sound in air from the outer ear to that in liquid in the inner ear. It has roughly the characteristics of a lowpass-filter. Further nonlinear behavior is caused by two muscles, tensor tympani and stapedius, that are able to decrease the transduction in a reflexive manner if the sound level is high, with greatest reduction in the low frequencies. It is also assumed that this mechanism supports directional listening in noisy environment, when the muscles react independently in the two ears, focusing somewhat on the desired source.

The final, and most complex peripheral stage consists of the *cochlea*, where the transduction of the sound to neural impulses takes place. The sound waves travelling along the basilar membrane trigger responses from hair cells at specific frequencies, with high frequencies providing resonance at the basal and low frequencies at the apical end of the basilar membrane. Thus, this is where the spectral separation occurs. There are also mechanisms for gain control and intensity encoding, as the 120 dB range perceivable by the auditory system must be compressed to the much smaller dynamic range of firing rates of neurons. The outer

hair cells play a key role in this active feedback process. However, recent studies claim that not the basilar membrane, but the hair cells themselves act as tuned resonators. The role of the basilar membrane would then be that it absorbs excessive energy at high stimulus levels and thereby contributes to the extremely great dynamic range (Warren, 1999). This stage also confronts the filtering trade-off in which finer time resolution leads to coarser frequency resolution and vice versa.

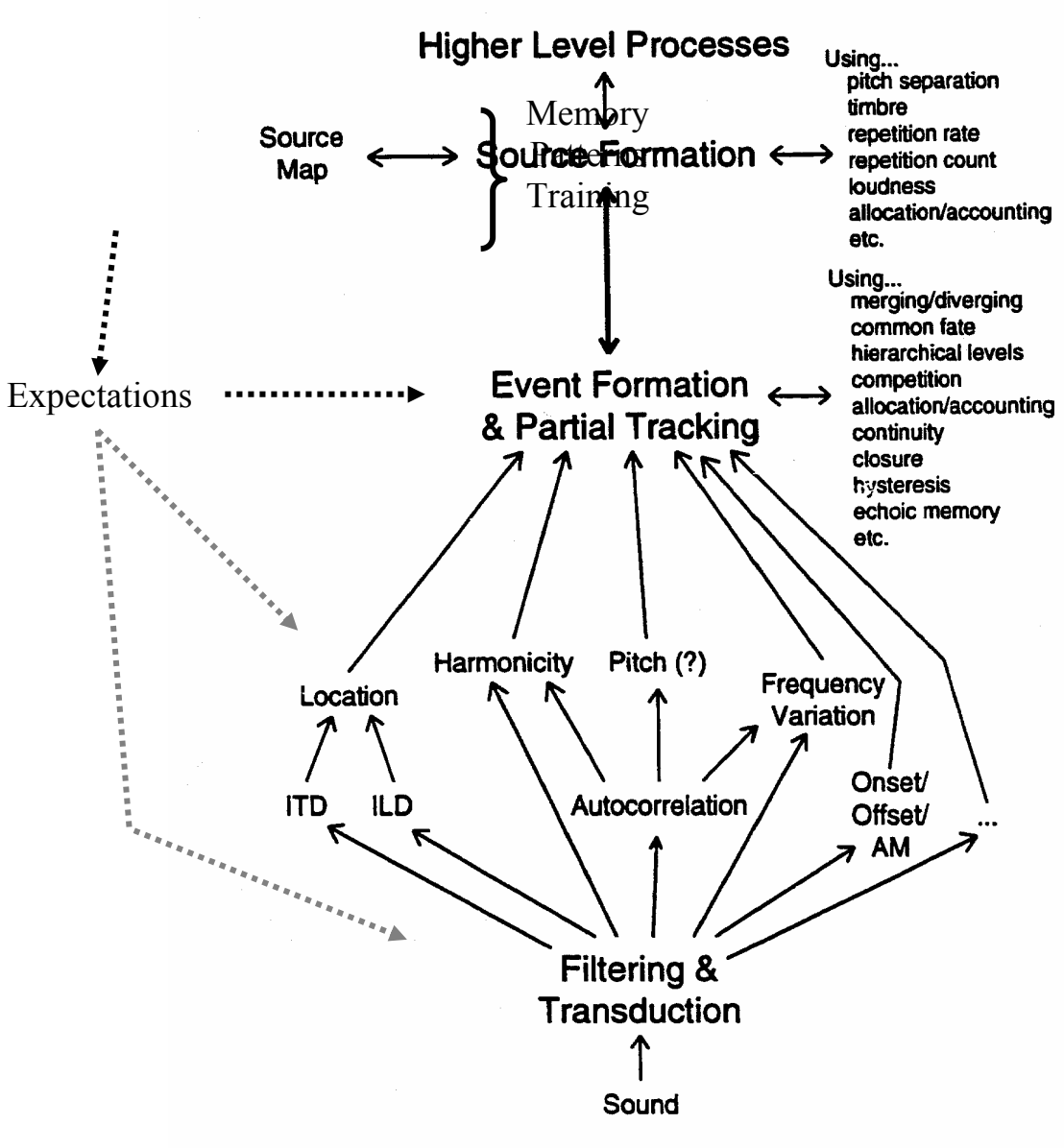


Figure 4.2: Block diagram of the stages in Auditory Scene Analysis (from Mellinger, 1992). Event and source formation are not only data driven, but depend also from previous knowledge and expectations. This may partly also apply for the transduction and filtering stage.

How much of these complex, nonlinear processes will have to be implemented in a *model* performing Auditory Scene Analysis? Modeling the whole peripheral chain in a very precise way will need much computational power. If the time to frequency transformation is performed via a fast Fourier transformation – thereby focusing on fine resolution in the frequency domain – some temporal information is lost in the spectrotempogram which would be present in a cochleagram (a map of neural firing rate as a function of time and place along the cochlea). Temporal cues however seem to be very important for sequential formation that associates entities over time. Periodicity information in the signal, for example, is one of the few cues that are undisturbed by most nonlinear transformations (Slaney & Lyon, 1993). However, there is no evidence that auditory front-ends like cochlea models will lead to performance improvements in speech recognizers. So, a discussion about the needs for cochlea models is required for each application; for sound classification, it will follow in chapter 5.

4.3.2 Feature Filtering

The next stage in Figure 4.2 is a series of filtering processes, each of which extracts or filters a different type of feature that may be present in the data. These filters output a number of *feature maps* in the brain, representations in time and frequency and perhaps other dimensions of these features. Many maps in the auditory system are organized tonotopically, that is with one dimension representing a place along the cochlea. Various maps have been found in the auditory cortex of animals, for example one with axes of frequency (tonotopy) and frequency-change rate, or amplitude modulation maps that encode short-time amplitude fluctuations.

The primary map from which all others are computed is the time versus log-frequency map of firing of cochlear nerve fibers. All other maps are – directly or indirectly – computed from this one. However, if each map is kept in a different place, then there arises the problem of bringing together the data that belong together. The data, for example for a specific frequency, would be scattered over several areas if several maps have frequency as a dimension, and such information must be recombined for computation. It is not fully revealed yet how this task is handled. Fortunately, a computer model does not suffer from this problem, thanks to random access memories, although this might not represent exactly the physiology.

Some of the features found at this level include (Yost, 1991):

- Spectral profile
- Harmonicity
- Interaural time and level difference
- Amplitude onsets and offsets
- Amplitude modulation
- Frequency modulation

Section 4.4 will give a review of the clues that suggest that these features are relevant for Auditory Scene Analysis, especially for simultaneous formation. References to various psychoacoustical tests that prove the importance of the features will be made.

4.3.3 Event Formation

After the features have been filtered, they are analyzed to form events out of the energy represented by neural firing. The formation process obeys a number of rules, like continuity principles. Information about event formation need not be completely data driven, it can also include knowledge about patterns of learned sounds. The human memory is indeed a very important element for the recognition of sounds, such as the timbre of an instrument or of a voice. This is indicated in Figure 4.2 with the dashed arrows that perform a feedback in the system. If some incoming patterns look similar to stored patterns, further (earlier trained) adaptation of the formation processes may occur. This may even apply to lower level processes like the feature extraction or filtering stage.

4.3.4 Source Formation

Source formation assigns events to separate sources. While features are points in time-frequency space, and events typically range from a few tens of milliseconds to a few seconds, sources last arbitrarily long. The source formation process is where the auditory evidence – the events – meets with an explanation. Many of the explanations are based on prior knowledge, although purely auditory explanations of very low specificity are available by default: something is heard as a sound but we do not know what it is. Bregman (1993) states that there are different processes occurring in the human listener that serve to decompose auditory mixtures.

4.4 Features for Event Formation

In this section, each of the features listed above is now described in more detail.

4.4.1 Spectral Separation

The auditory system's remarkable ability to determine the *spectral content* of sound provides one form of information that might be used for sound source processing. When two sound sources contain different spectral components, the auditory system may resolve those components. However, if these components are too close together in frequency, they can not be resolved. This is especially the case for high-frequency components because the cochlear filters are broader in the higher than in the lower frequencies.

Thus spectral separability alone does not appear as an appealing variable to directly control sound source determination, but, as stated above, the spectral map generated by the cochlea is the general basis for all further computation.

4.4.2 Spectral Profile

Most sound sources produce a particular amplitude spectrum that remains relatively constant in terms of its spectral profile as the overall level of the sound is changed. The spectral shape of musical or vocal sounds is in large part determined by the formants, which are relatively fixed in frequency, even though the pitch may be changing, as in singing. Figure 4.3 shows schematically how the spectral profile of a sound remains constant as the overall level is changed. Thus, a key to process such stimuli is for the auditory system to monitor the relative differences of the amplitudes of the spectral components as the overall level changes. Note, however, that speech is an exception (Handel, 1995): If the speaker is murmuring, the higher-

frequency harmonics fall off in intensities at greater rate than normal speech and are replaced by the aspiration noise.

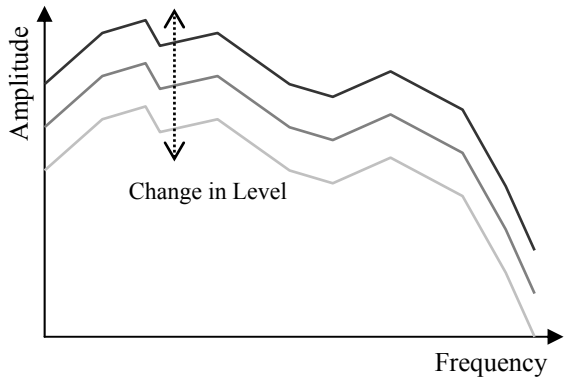


Figure 4.3: Schematic diagram of the spectrum of a complex sound coming from one source. As the sound gets louder, the spectral profile of the source does not change, that is, the relative amplitudes of the frequency components remain the same.

In Figure 4.4, the power spectra of a 1.5 s window of three sounds are displayed: A speech signal, a musical signal (flute) and party noise. The three profiles clearly differ in the overall shape, which can contribute to the recognition of the different sources, although variances of the shapes can be quite high within one sound class. These profiles could provide us with a means for grouping by comparing them with learned patterns. And indeed is the auditory system performing a profile analysis in order to make comparisons across the spectrum of the signal. Hartmann (1998) states that this analysis is responsible for the perceived "tone color", which contributes to that part of the perceived timbre that is attributable to the steady state part of a tone, that is, the tone without transients associated with onset or offset or ongoing aperiodic fluctuations. Timbre perception will be discussed in more detail in section 4.4.5 about onset and offset, because the temporal changes of the profile, that is the onsets and the offsets of the partials are more important cues for timbre. To convince yourself of this fact, play a piano sound backwards, it will more sound like an organ than like a piano. So, although the "color" of a tone is determined only by the shape of the spectrum, the overall perceived timbre of complex sounds is highly dependent on temporal cues rather than just the instantaneous across-frequency profile.

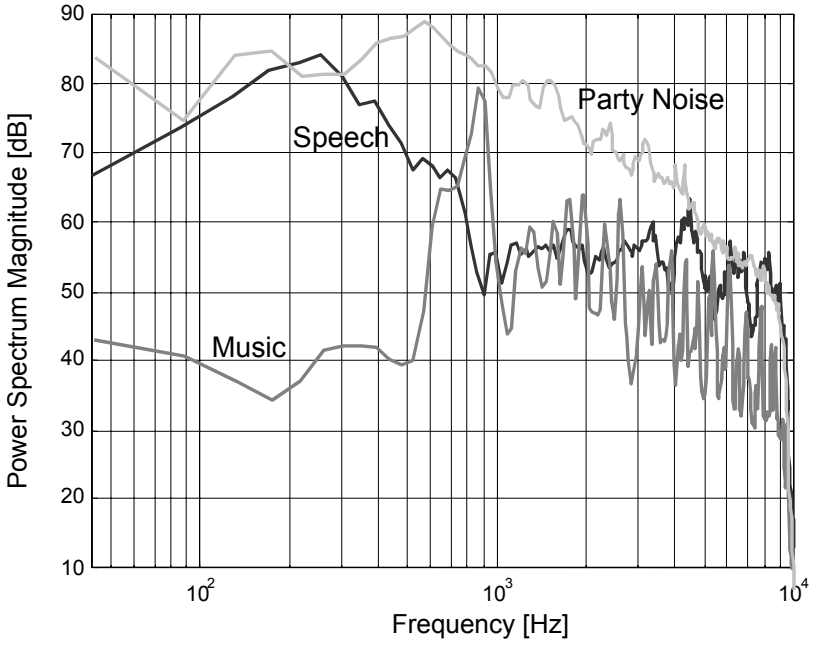


Figure 4.4: Power spectra of a 1.5 s window of a speech signal, a musical signal (flute), and party noise. The spectra have shapes that clearly differ from each other and are somewhat typical for each class, although the variances within a class can be quite large.

4.4.2.1 Psychoacoustical Evidence

Experiments have shown that listeners are quite sensitive at detecting changes in the spectral pattern even when the overall level undergoes large variations. In an experiment described by Yost (1994), a signal consisting of a number of different frequency components of equal level is played, but one frequency component stands out in one of two observation intervals, while the overall level is changed randomly over a 40 dB range. Listeners have to determine which of the two stimuli is the one that contains the more intensive component. This experiment works well as long as the frequency components are not as close together that they fall into the same critical band. Obviously, the *relative* difference in the level is used to detect the peak. An experiment by Green (1993) showed that it is easier to detect a change in the spectral shape of a complex sound than an increase in intensity of just one single sinusoid. It was also shown that the interstimulus interval had a big influence on the single sinusoid experiment, but almost no influence on an experiment with a complex tone consisting of 21 frequency components. Thus the memory of changes in the spectral shape does not appear to deteriorate much with time.

Further experiments by Green (1993) proved that phase changes in the signal are irrelevant for spectral profile perception, which indicates that the auditory system compares changes in the magnitude of the spectrum, not changes in the temporal waveform.

4.4.3 Harmonicity and Pitch

Many sound sources in nature have a harmonic or nearly *harmonic partial structure* resulting from their origin in a vibrating medium; an example is shown in Figure 4.5. Because different vibrational sources usually vibrate at different frequencies, their partials form separate harmonic series.

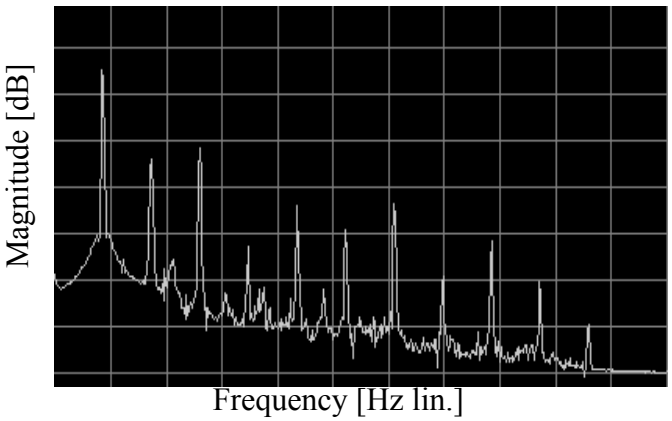


Figure 4.5: Harmonic partial structure of a recorder flute sound.

The definition for harmonicity given by Mellinger & Mont-Reynaud (1996) is: Harmonicity refers to the degree to which a partial falls into a harmonic series with other partials. Bregman (1990) claims that we have two separate mechanisms for hearing harmonicity, one that gives rise to a perception of consonance or dissonance, and a second that produces fusion of partials into a sound source. The single auditory object that is produced by a harmonic series is called *pitch*. Pitch perception is an excellent example of simultaneous fusion, because the partials are grouped across frequency.

Lyon & Shamma (1996) state that pitch is that perceptual property of a sound that can be used to play a melody (while timbre distinguishes musical sounds of the same pitch, see section 4.4.5.1). As pitch is a subjective attribute, it can not be measured directly. The definition given by ANSI (1994) is: "Pitch is that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high. Pitch depends mainly on the frequency content of the sound stimulus, but it also depends on the sound pressure and the waveform of the stimulus."

The pitch of a pure tone is primarily related to its frequency, whereas a tone complex may either evoke a single pitch or a cluster of pitches. Even sounds that are not formed of well-defined discrete partials can provoke pitch sensations, referred to as non-tonal pitch.

The pitch of a pure tone depends also on the intensity. On the average, the pitch of tones below 1 kHz decreases with increasing intensity, remains almost constant between 1 and 2 kHz and increases with decreasing intensity above about 2 kHz. This behavior is popularly known as Stevens rule (Stevens, 1935).

For a harmonic series, the pitch often corresponds to the fundamental frequency. Note, however, that the same pitch is perceived even if the partial at the fundamental frequency is missing in the harmonic series. This is called the *missing fundamental* phenomenon, and it can be quite practical in everyday life: When listening to music on a cheap transistor radio with small loudspeakers, frequencies below some 150 Hz are not played. Because the higher harmonics are present, lower pitches are perceived nevertheless. The same can be observed with speech on the telephone, where the frequencies below some 400 Hz are not transmitted.

As pitch is not identical to frequency, and not even linear in frequency, a scale has been defined which describes the relation of pitch to the frequency of a pure tone. In the mel scale (the unit is derive from *mel*ody), a 1000 Hz tone is arbitrarily assigned a value of 1000 mels. The frequency of a tone that sounds twice as high receives the value 2000 mels. The mel scale is linearly related to the Bark scale (100 mels = 1 Bark). Note that the mel scale has never become very popular; by default, the linear frequency scale is used.

Yost (1994) states that although pitch is a powerful fusing mechanism, it does not seem to aid the auditory system very much in segregating different harmonic series into different sources. Other means of separation are required to fulfil this task satisfactorily. With cues such as different onset and spatial separation, two separate pitches can sometimes be identified. The pitch perception process appears to behave more synthetically (group all harmonics together) than analytically (analyze the sound into its two harmonic parts).

Moore (1993) claims that in principle, there are two ways in which the pitch of a sound may be coded: by the distribution of activity across different auditory neurons and by the temporal patterns of firing within and across neurons. Based on this knowledge, two general forms of pitch perception models have been stated:

- **Spectral Models:**

The spectral models attempt to find the best fitting harmonic series to the spectral components of the sound as they are represented in the auditory periphery. Figure 4.6 shows the idea of this model for a harmonic series of tones yielding a pitch of 850 Hz. This model has a good anatomical justification, because of the tonotopic organization throughout the entire auditory pathway, and it is well established, because it has been confirmed in a number of independent ways. Lyon & Shamma (1996) and Moore (1993) claim that spectral approaches allow to

determine the pitch of complex tones even if the fundamental is missing (by somehow calculating the least common factor of the harmonics present).

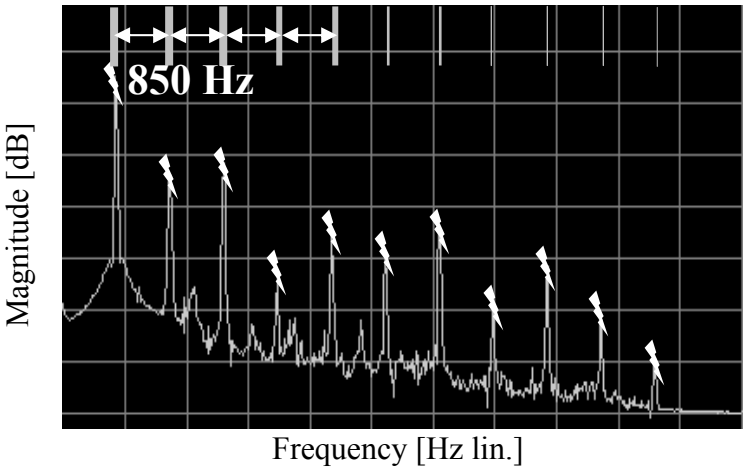


Figure 4.6: Spectral approach to determine the pitch of a sound: The distance between the harmonic partials equals the pitch. If the fundamental is missing, the distance between the higher harmonics nevertheless indicates its frequency.

- **Temporal Models:**

The temporal models assume that the auditory system searches for robust periodicities in the temporal neural pattern at the auditory periphery, and the reciprocal of that period is the pitch of the sound. Figure 4.7 shows the harmonics of Figure 4.6 as a temporal waveform.

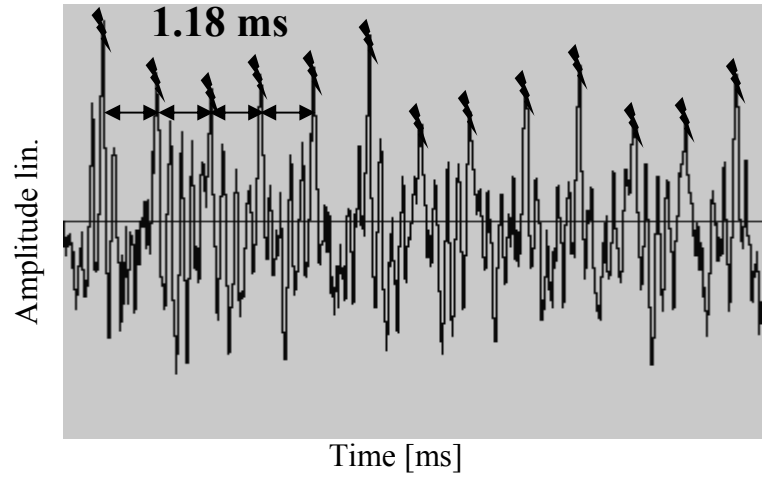


Figure 4.7: Temporal approach to determine the pitch of a sound: Strong periodicities in the signal are somehow determined; for example by peak-detection, which however does not work in the case of missing fundamental.

Hartmann (1998) states that the temporal models have difficulties with pitch-shift effects observed experimentally, but they are also suited for complex tone perception, where the fundamental may be missing. To detect a missing fundamental in the time signal, however, requires to autocorrelate it: There is a peak in the autocorrelogram at $1/f_0$, even when the fundamental is missing. Moore (1993) further finds that pitches can be perceived when the harmonics are too close in frequency to be resolvable and when the stimuli have no well-defined spectral structure (for example interrupted noise), and pitch perception and discrimination can also be affected by the relative phases of the components in a complex tone. A serious limitation is that above 5 kHz, the neurons do not maintain synchrony with the stimulus, and temporal pitch determination becomes impossible. However, there are experiments that show that pitch perception also deteriorates at high frequencies.

Recent models try to combine the spectral and temporal aspects (Moore, 1993). Spectro-temporal models can account for most experimental data on the pitch perception of complex tones. It could be that the pitch of pure tones is determined by temporal mechanisms at low frequencies (< 5 kHz) and place mechanisms at higher frequencies (> 5 kHz), where phase-locking of firing neurons disappears.

Let us conclude the discussion about pitch with a statement from Lyon & Shamma (1996) – to stress once more that the Auditory Scene Analysis problem is multidimensional and complex: "Pitch and timbre are fundamental attributes of any sound signal. But only together with other attributes, such as loudness, localization in space, onset and offset characteristics, does a sound achieve its unitary identity and can be perceived as emanating from a particular source."

4.4.3.1 Psychoacoustical and Neurophysiological Evidence

The strength of harmonicity as a grouping cue is so well established that it is used for psychoacoustical experiments as a basis against which competing forces for segregation are measured. Darwin & Carlyon (1995), for example, describe an experiment where the harmonics of a complex tone are split up in two parts which are applied to the two ears. Under most conditions, they still fuse together to one single auditory object. So, harmonicity seems to be a more compelling factor for object recognition than common spatial location!

A single harmonic can be heard out of a complex only if its level is increased, if it differs in onset, or if it is mistuned relative to the others. From 2 % mistuning on, the harmonic can be heard as a separate event (tone), but the pitch of the complex is also shifted. However, if the mistuning is greater than 8 %, the pitch shift disappears. This shows that fusion, or more specifically pitch perception, is not an all-or-none process. Interestingly a lower mistuning is required for higher harmonics, because a sort of beat is perceived. Harmonics in the higher frequencies can not be resolved, because there, they will fall into the same critical band.

However, so far, no "harmonicity neurons" have been found in the auditory system, that is, neurons that integrate information from widely spaced frequencies and respond only if the frequencies present are in a harmonic series. As stated above, it is possible that associations between harmonics are discovered in the time domain, based on phase-locking (up to 5 kHz). Pitch has to be extracted very early in the brain, as many studies have confirmed that synchrony to repetitive features of a stimulus becomes progressively worse toward the cortex. It is conceivable that a spatial map of pitch can be derived from the spectral profile representation (Lyon & Shamma, 1996).

4.4.4 Spatial Separation

The auditory system performs a remarkable task in deciding from which spatial location a sound originates. Interaural time and level differences and aspects of the head-related transfer function are crucial for spatial location. In 1953, Cherry defined the term "cocktail party effect" to refer to the auditory system's ability to determine the sources of sounds when they are located at different points in space (Cherry, 1953).

Spatial location is a significant cue for sound separation, although not an indispensable one. To convince yourself of this fact, simply cover one ear whilst listening to a concert, or listen to a mono-recording of an orchestra. You will have little difficulty in determining a large number of the instruments being played.

4.4.4.1 Psychoacoustical Evidence

So-called masking level difference (MLD) experiments show that the threshold for detecting a pure tone within a noise is much lower when interaural differences occur in either the signal or the masker. For the case where the phase of the tone differs for the two ears, the MLD is up to 15 dB (Yost & Sheft, 1993).

Darwin & Carlyon (1995) state that pitch perception mechanisms seem not to use binaural information. When parts of the spectrum of a sound are played to opposite ears, they are nevertheless fused in many cases. On the other hand, if in a narrow part of the spectrum of a noise only the phase or the level is shifted, this part is perceived as a pitch lateralized away from the midline (where the noise is). This phenomena was found 1958 by Cramer and Huggins and is referred to as Huggins pitch (Cramer & Huggins, 1958); it proves the existence of binaural processing, since there are no monaural cues that could be used in this experiment.

4.4.5 Temporal Onsets and Offsets

One of the features acting most strongly for grouping related partials is common amplitude onset. Onset asynchronies can aid the perceptual decomposition of a complex stimulus into harmonic subsets. An illustration of the power of onset asynchrony to pull apart a single source to make several is presented in Figure 4.8. If a short tone is played with all harmonics starting synchronously, a single pitch is perceived. If a delay of one second is introduced between the onset of successive partials, each harmonic stands out briefly as a separate tone before merging with the existing sound complex.

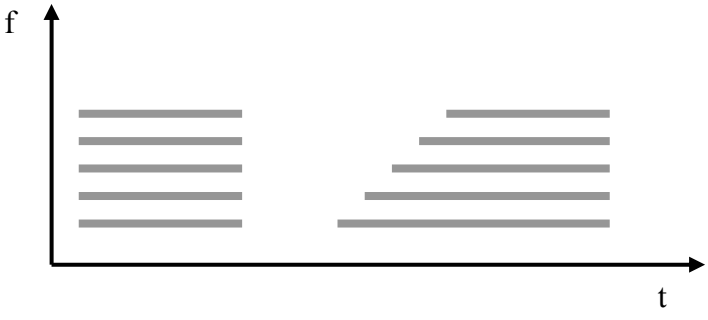


Figure 4.8: The effect of onset asynchronies on fusion. Each horizontal represents a sinusoidal tone. At left, all sinusoids fuse together to a single auditory object. At right, successive tones begin at intervals of 1 s and stand out briefly before merging with the rest of the complex (after Mellinger & Mont-Reynaud, 1996).

In most musical instruments, the start of a note is marked by the rapid rise of all strong partials within a period of about 40 ms. An important characteristic of these onset asynchronies is that it forms an significant part of the timbre, for which reason timbre is discussed in this section.

Common offset is much less important than common onset for grouping the parts of a spectrum produced by one source.

4.4.5.1 Timbre

The term timbre is used to refer to the perceptual qualities of auditory objects, that is, "what it sounds like". It is the timbre which allows us to tell the difference between musical instruments. Handel (1995) compares the perception of timbre to the perception of faces. The definition given by ANSI (1960) is: "The quality of sound by which a listener can tell that two

sounds of the same loudness and pitch are dissimilar." This definition rather tells us what timbre is not than what it is. We have described above that "tone color" is determined by the steady part of the sound, that is, the across-spectral profile or the strength and number of (mostly harmonic) partials present. Temporal cues, however contribute much more to timbre perception: The temporal envelope of a tone, that is, the time course of the tone's amplitude, has a significant influence on the perceived timbre of the tone (Houtsma, 1989).

Especially the character of the beginning and ending transient sounds of a tone play a great role in timbre perception (see the piano example in section 4.4.2). The attack and decay of the sound produced by many musical instruments provide most of the information that allows us to differentiate among different instruments. A major aspect of music synthesizers is to accurately simulate the attack and decay of a note played by the instruments being synthesized (Yost, 1994).

According to the ANSI definition, also frequency and amplitude modulation can contribute to the timbre, for example the vibrato of an instrument or of a voice.

In speech, the timbre of a voiced sound is imposed by the resonances of the oral and nasal cavities, as controlled by the tongue, lips, and velum (Lyon & Shamma, 1996). Two vowels, uttered by the same speaker with the same loudness are easily discriminated. According to the ANSI definition, the two vowels could be said to differ in timbre (Yost & Sheft, 1993). In other words, the unvarying formant frequencies of the voice contribute to the timbre.

Handel (1995) states that it is not only impossible to claim that a particular spectral profile is the cause of identification or timbre of an instrument (because that profile changes across notes), but that it can neither only be the transients that are used for timbre perception, because also the pattern of the transients will vary across sounds. His message is that "the cues to identification and timbre are not invariant and vary across notes, durations, intensities, and tempos. There are no pure cues. Every event exists in context, and the acoustical properties depend on the context. The attack and decay transient, the duration, the spectral shape, the inharmonic noise, and the amplitude and frequency modulation all contribute." Thus, an auditory object has a distinct timbre, but this timbre cannot be related simply to one acoustic feature or to a combination of them. Timbre is partly a function of the acoustical properties and partly a function of the perceptual processes. Listeners may use different features for timbre determination, depending on the context.

4.4.5.2 Rhythm

The perceived rhythm of sounds such as speech and music depends on the peaks in the temporal loudness pattern which can be identified by the amplitude onsets (another way to determine the rhythm could be to investigate the amplitude modulations, which is somewhat the same).

It is often assumed that sound bursts of equal temporal spacing elicit the sensation of a subjectively uniform rhythm. However, this simple rule holds only for very short sound bursts with steep temporal envelopes. Sound bursts with a more gradual rise in the temporal envelope often require systematic deviations from physically uniform spacing, in order to produce a subjectively uniform rhythm. An example is given in Figure 4.9, where the longer burst has to start earlier to evoke a subjectively uniform rhythm.

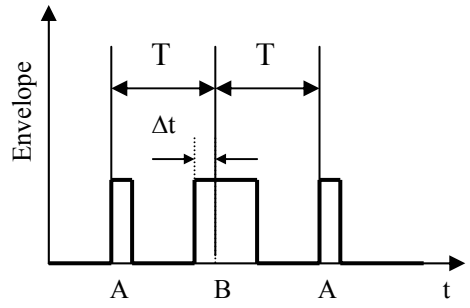


Figure 4.9: Human perception of the temporal characteristics of sound bursts. If the bursts do not have the same length, the longer burst B has to start earlier to give a rhythmic impression. Δt can be up to 20 ms for this example. After Zwicker and Fastl (1990).

Terhardt (1998) found that for most music pieces, the rhythm is between 125 and 500 ms, with a maximum in the distribution around 330 ms, whereas the rhythm of speech has a maximum around 250 ms. Thus, the rhythmic events occur in similar ranges for music and speech.

4.4.5.3 Psychoacoustical Evidence

Onset

Interestingly, the ability to accurately locate a sound source depends to a large extent on the first wave reaching the ears, in that interaural information arriving after the first wavefront appears to be suppressed (Yost, 1994).

As already mentioned in the example of Figure 4.8, delaying the onset of a partial will perceptually highlight it.

Mellinger & Mont-Reynaud (1996) report measurings of onset of notes among different instruments in an ensemble and found that they could differ up to 50 ms. This enables the musicians in an orchestra to hear out their instruments as well as the audience to separate the instruments.

Timbre

White (1991) reports an experiment to show the importance of the transients. The attack of one instrument is spliced onto the steady-state sound of another instrument. In almost all cases, the sound is identified as the instrument that supplied the attack. For instance, an oboe attack spliced onto a violin steady-state sounds much more like an oboe than a violin.

Handel (1995) describes experiments with single notes of musical instruments where either the transients or the spectral information has been made equal for all sounds. The results show that the cues that determine timbre quality are independent, that the attack transient or the steady state alone can provide enough information for identification, and that the cues depend on the context. These results are also valid for sequences of notes.

4.4.6 Amplitude Modulation

Amplitude modulation (AM) is characteristic for many natural sound sources. An important example is speech: The temporal envelope of speech is dominated by periodicities of less than 20 Hz, with a maximum around 4 Hz. The modulations at this frequency are the reason why a rhythm is perceived in speech, as described in section 4.4.5.2.

Common AM is the parallel variation of a number of partials. Most physical processes that change the intensity of a partial will change the intensity of all others at the same time, which is a strong indicator that they come from the same source. The auditory system is able to fuse spectral components that are modulated with the same temporal pattern (simultaneous fusion) or segregate sounds with different patterns, if the modulation frequencies are below 50–100 Hz. This indicates that the auditory system can make comparisons across wide spectral regions and across time.

4.4.6.1 Psychoacoustical Evidence

Von Békésy (1960) found that sine waves at 750 and 800 Hz, when presented to opposite ears, could be made to form a single auditory image by the imposition of coherent AM of 5–10 Hz.

The strongest line of evidence for the power of common AM comes from comodulation masking release (CMR) experiments: Noise bands that vary in amplitude coherently, with the same modulating function, are not as effective at masking other within-band signals as those that are modulated incoherently. On the other hand, if the amplitude of a single component of a tonal complex is continuously modulated, it tends to stand out from the complex. It is suggested that CMR occurs because the auditory system listens to the target signal in the more silent periods of beats between the noise carrier frequencies. This is supported by the fact that CMR fails for frequency-modulated, instead of amplitude-modulated sounds.

Modulation detection interference (MDI) experiments show that detection of modulation depth is difficult when both a probe and a masker signal are comodulated. Maybe comodulation fuses them into one auditory image.

4.4.7 Frequency Modulation

Frequency modulation (FM) refers here to the change in frequency of a partial. If a subset of the spectral components is coherently modulated, fusion will work much better. However, Yost & Sheft (1993) state that this seems to work only for harmonic partials and suggest that the process might not be specific to FM detection, but to covarying cues of harmonicity or envelope coherence.

Mellinger & Mont-Reynaud (1996) describe common FM as one of the weaker grouping cues, which is nevertheless important musically in that other cues are often missing, and it is the only one available to hear out multiple voices. This statement seems to be true only if spatial, onset and AM cues are not present in the signal.

4.4.7.1 Psychoacoustical Evidence

A synthesized vowel sound resembles much more a single voice when identical slight vibrato is added to the partials. Mellinger & Mont-Reynaud (1996) therefore suggest that the auditory system uses vibrato as a grouping feature. On the other hand, vibrato occurs naturally in speech, so a vowel with vibrato is well-known to the auditory system as a single source and will of course also sound more natural.

However, further studies have shown that 4% vibrato in notes against a background of a masking tone improve detectability by about 17 dB compared to the case without vibrato. This amount of vibrato, typical for musical instruments, makes notes stand out even when they are much less intense than a masking tone.

Darwin & Carlyon (1995) state that the presence of FM at least increases the perceived prominence, identificability and naturalness of sounds. For segregation, however, many studies point against an independent role for FM phase differences (see for example Baumann, 1995). There are a number of reasons why this seems to be sensible: First, FM phase differences would be computationally expensive to detect, second, they can be detected via inharmonicity already, and third, the formants of a speech signal often change frequency in opposite directions, that is, there would be a mechanism needed to determine whether a peak was due to a harmonic or to a formant.

4.5 Event and Source Formation

In the following, the event and source formation processes are again discussed in more detail. First, some basic grouping principles are described, which can then be applied on the level of event as well as source formation.

4.5.1 Grouping Principles

4.5.1.1 Primitive Grouping versus Schema Based Grouping

Bregman (1990) distinguishes two types of mechanisms that can be used to decide which components belong to a particular sound source: *Primitive* grouping mechanisms partition the auditory input on the basis of simple stimulus properties, such as common onset time or a harmonic relationship among components, and so do not necessarily depend on specific experience. *Schema governed* mechanisms are presumed to involve the activation of stored knowledge of familiar patterns in the acoustic environment and of search for confirming stimulation in the auditory input; they are generally learned and so depend on the listener's specific experience.

4.5.1.2 Gestalt Principles

At the beginning of the twentieth century, German psychologists began to explore visual perception on the search for natural organization principles. The search resulted in the so-called Gestalt principles, which are first described for visual perception here, but will be applied to acoustical event and source formation in the following. Two general principles concern exclusive allocation and accounting (Bregman, 1990): Any element of the incoming pattern should be assigned to only one source, it cannot do double duty[5]. The complementary principle of accounting requires that all elements be assigned to one source or another. If the element cannot be assigned to any existing source, then it becomes a source itself.

Figure 4.10 shows a visual representation of the Gestalt principles, which are obviously quite simple and intuitive to understand.

[5] The mistuned harmonic example in section 4.4.3 shows, however, that exceptions are possible, where a single element can contribute to more than one perceived object.

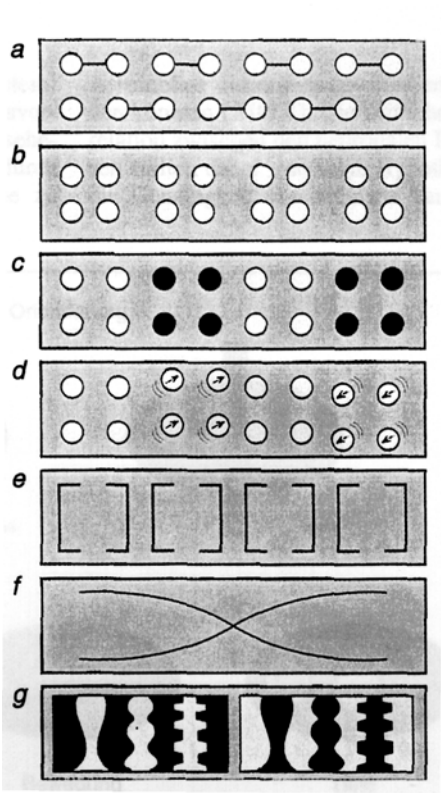


Figure 4.10: Gestalt principles (from Engel & Singer, 1997)

(a) ***Continuity****: Regions of the image that are connected will normally be seen as part of the same figure.*

(b) ***Proximity****: Elements that lie together closely tend to form a single group.*

(c) ***Similarity****: The same is the case for elements that look similar.*

(d) ***Common Fate****: Elements that change coherently in time and space will stand out as one figure.*

(e) ***Closure****: In general, elements will be grouped together that form a closed contour. Thus, four squares will be seen in this representation.*

(f) ***Good Continuation****: Smooth continuation is perceptually preferred over abrupt and rapid continuation. In the example, two crossing lines will be seen instead of two tips touching each other.*

(g) ***Symmetry****: Areas that are enclosed by symmetrical lines will stand out as figures.*

These principles can be applied at different hierarchical levels. In what sense they apply for auditory event and source formation is described in the following two sections (after Mellinger & Mont-Reynaud, 1996, and Baumann, 1995).

4.5.2 Event Formation

Event formation is primarily a spectral organization, placing sound energy into correct groups at each instant. One of the stronger grouping principles is *common fate*: Common onset and offset as well as common amplitude modulation and – less strongly – common frequency modulation are all evidence that several partials belong to the same event.

The principle of *good continuation* can be looked at temporally and spectrally: Partials tend not to change in frequency rapidly over time, and the partials of harmonic sounds are of good continuation with regard to their spectral placement.

Closure helps to preserve continuity of partials across interruptions. At higher level, it shows some astonishing effects, as will be discussed in section 4.5.3.

Note that the absence of cues can be as important as their presence; for instance, the absence of harmonicity for a given partial tends to make it stand out as a separate source.

4.5.3 Source Formation

Source formation refers to grouping events coherently over time. Important principles are *similarity*, for example notes with the same timbre that are assigned to come from the same stream (instrument), and *proximity*, for example successive notes near each other in frequency (pitch) that tend to be placed into a common source by the auditory system and conversely.

Another important principle for source formation is *closure*, which helps to preserve the continuity of events in streams. A speech signal, when interrupted rhythmically will sound more pleasant if the gaps are filled with a noise signal, and even the intelligibility may be improved. With the noise, the speech signal will be perceived as a continuing (but temporarily masked) stream, which is of course more familiar to the auditory system than silent gaps. Figure 4.11 shows a visual representation of this phenomenon. An exact simulation of auditory perception will have to consider this phenomenon; simple models will hardly fuse the separate events to one stream, whether noise is present or not.

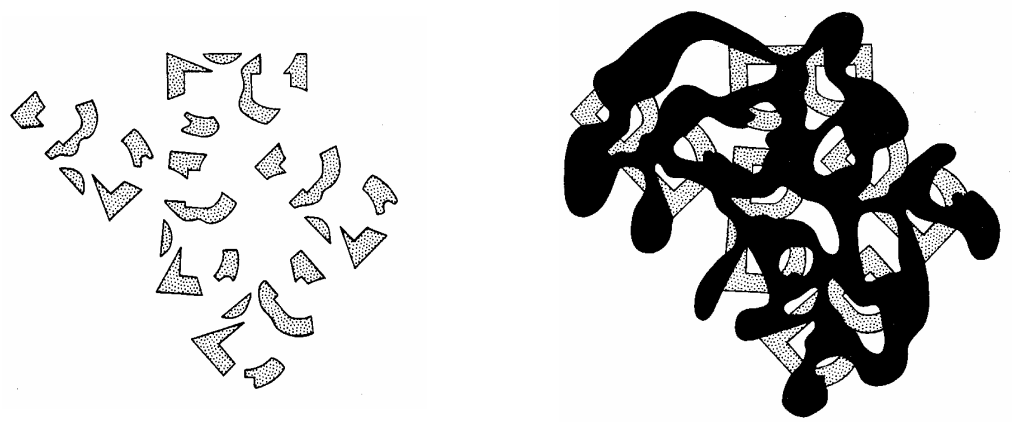


Figure 4.11: If the fragments of a figure are presented as shown on the left, it is hardly possible to recognize any pattern. If the gaps between parts that belong together are filled out with ink (right), the closure principle can apply very easily and the letters „B“ are perceived as being complete, but partly covered (from Bregman, 1990).

Good continuation refers to the fact that most sounds tend not to change in character rapidly: A piano is not likely to suddenly sound like a violin. Interestingly however, the auditory analogy for the visual example of good continuation (Figure 4.10f) is mostly perceived differently: Figure 4.12 shows how an ascending and a descending stream are perceived as streams in the lower and upper frequency range. It seems that the principle *proximity* is instead applied here.

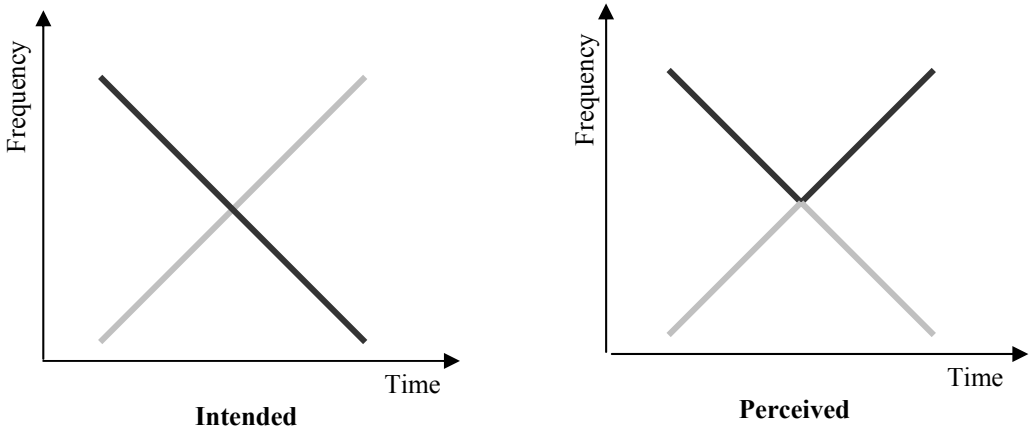


Figure 4.12: Intended and perceived representation of two streams. The two crossing streams on the left are perceived as one stream in the higher and one in the lower frequency range (right): The good continuation principle does not apply here (after Baumann, 1995).

Thus, different formation principles may compete each other for event and source formation. Figure 4.13 shows how a visual pattern can be perceived ambiguously. The ambiguity lies here in the assignment of the foreground and background.

Figure 4.13: Ambiguous perception depending on the foreground-background assignment (from Shepard, 1991).

4.6 Application for Sound Classification

It shall now be discussed which elements from Auditory Scene Analysis may be used for sound classification, and what existing computational Auditory Scene Analysis models look like.

4.6.1 Auditory Features

The table below lists again the auditory features and their possible use for sound classification.

Auditory Feature	**Use for sound classification**
Spectral separation	Like for scene analysis, spectral separability can hardly be used directly for sound classification, but the spectral map serves as an excellent basis for the calculation of further features. However, good temporal resolution is as important as good spectral resolution; a trade-off between the two has to be found.
Spectral profile	How can the difference in the spectral profile of different sounds, that we can see so easily in a spectral plot (see for example Figure 4.4), be extracted computationally? Some ideas to characterize the profile by various statistical measures of the relative amplitudes of the partials have been presented in the literature. However, it seems difficult to derive sound class specific information by the spectral profile, as the variations of the profiles within a sound class can be very high. Nevertheless, "standard" profiles might be found for some specific sound classes.
Harmonicity, Pitch	The harmonicity of a signal can be determined by measuring its pitch, if there is one present at all (actually, it is the fundamental frequency that is determined, as pitch is a subjective measure). The pitch might also be traced

Auditory Feature	Use for sound classification
	over time to calculate some statistical values, like the variance of the pitch etc. Most noises have no harmonic partials and can be identified by that. Some noises, however, can contain harmonics. It is expected that the temporal behavior of the pitch is different for harmonic noise, music and speech.
Spatial separation	At present, the use of binaural cues is not in the focus, as there is currently only a mono signal available for processing in the hearing instrument. However, in the future, it might be very useful to analyze the stereo signal. This could especially make sense for speech signals, to distinguish between one, two, and several speakers, because it is difficult if not impossible to distinguish between one and two speakers using other features. Furthermore, if there is speech from behind as well as from the front of the listener, the one from behind could be rated as disturbing noise.
Amplitude Onset/Offset, Timbre, Rhythm	The timbre, being a subjective and quite complex feature, can hardly be modeled. However, features that contribute to the timbre might be useful, as amplitude onsets and spectral profile. It seems that the amount of onsets (or offsets) within a time window and frequency band as well as the distribution across frequency could provide information for classification purposes. It remains to be seen how much these values differ for different sound classes. If the intervals between the onsets are explored, a feature describing the rhythm may be found, telling whether a beat is present in the signal or not. This might be useful for the recognition of music with strong beats, such as pop music.
Amplitude modulation	It is known that speech signals are much more modulated in amplitude than noise signals, which are mostly steadier over time. Thus, AM provides a powerful cue for the separation of speech and other signals.
Frequency modulation	The amount of frequency changes in several frequency bands could give a measure for the "turbulence" of the signal, and the spectral distribution of the "turbulences" could provide additional information. As for onset, it remains to be seen how much these measures differ for different sound classes.

Thus, promising features are the spectral profile, harmonicity or pitch, amplitude onsets, and – as already seen in chapter 2 – amplitude modulations. Chapter 5 deals with the implementation of these features.

4.6.2 Grouping

The aim of event and source formation is to fuse together parts of the signal that belong to the same auditory object, and to segregate the objects into different sources. The labeling (that is, the classification) of a sound is probably performed after that. For the purpose of computational sound classification, it seems a bit complicated to carry out the whole grouping process – it is believed that it is sufficient to identify the sources in a sound without actually segregating them.

Although little is known about schema-based grouping in the auditory system, a hypothesis-driven approach could be beneficial: A number of stored, general patterns could be compared with the input, and based on this, a first guess could be made about the sound class. Depending on the guess, some specific features would then be investigated to confirm or reject the hypothesis, or to refine the classification.

This process would be more top-down oriented, as opposed to primitive grouping, which would be bottom-up, taking the feature values as input to a classifier straight away. However, because schema-based grouping is probably computationally quite expensive, the first approach shall be to perform primitive grouping only.

4.6.3 Existing Models of Auditory Scene Analysis

A number of computational auditory scene analysis (CASA) models have been presented in the literature, for example from Cooke (1993), Brown and Cooke (1994), Baumann (1995), Mellinger and Mont-Reynaud (1996). The aim of these CASA models is to separate sources, rather than to classify them, and they do not use any learned schemas up to now, that is, they perform only primitive grouping. As an example, Figure 4.14 shows the block diagram of the CASA model from Brown and Cooke (1994).

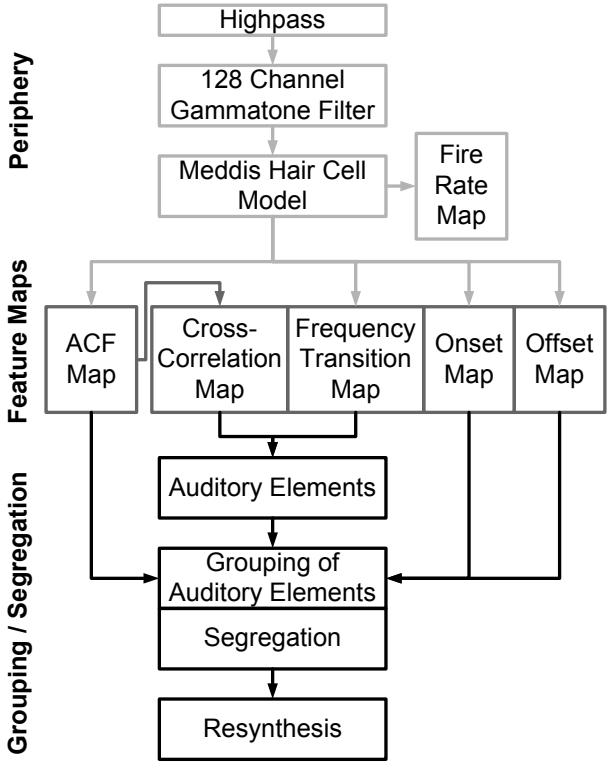


Figure 4.14: CASA model from Brown and Cooke (1994). The periphery is modeled by a large filterbank and a complex hair cell model. After the extraction of some feature maps, the grouping and segregation of the auditory elements is performed in order to resynthesize each source, that is, to separate the sources.

Most models have a complex peripheral stage with outer and middle ear transfer functions, multichannel (for example 128 channel) gammatone filterbank and hair cell model. This is followed by the extraction of feature maps, which are used to first group auditory elements and then to segregate and resynthesize the sources.

An improvement of these models is presented by Ellis (1996). In his *prediction-driven* approach, the analysis is a process of reconciliation between the observed features and the

predictions of an internal model of the sound-producing entities in the environment. In this way, it is for example possible to track sources over time even if there are intervals where they are masked by interfering noise.

The complex peripheral stage in these models makes them quite unsuitable for real-time processing. It is assumed that for sound classification, it is not necessary to model the periphery in this precise form; a simple FFT filterbank and maybe a compression stage might be sufficient.

As stated above, it seems that the last part – the grouping part – of the models is not what is needed for classification, as we do not really want to separate sounds, at least not at this stage of our research. It might however be, that a separation of sources preceding the classification could be beneficial in future work, in order to classify each source separately. Still, at the time of writing this thesis, such an algorithm is far too complex to be implemented in hearing instruments.

This means that especially the middle part of the models – the computation of feature maps – can be adapted to gain measures for different signal characteristics, like occurrence of onsets and offsets, autocorrelation for pitch determination and so on. In the next chapter, some of the feature calculation will be implemented following these models.

5 Classification Systems I: Features for Sound Classification

5.1 Introduction

In the last chapter, an overview of the mechanisms of Auditory Scene Analysis has been given. Obviously, an important stage within this analysis is the extraction of auditory features from the spectrotemporal pattern that is reaching the auditory cortex from the inner ear. These features shall be used in this chapter as a model for the structure of the feature extracting block in our sound classification system, as it is again shown in Figure 5.1.

Figure 5.1: The feature extraction block is the first stage of the sound classification system. Our approach is to use features that are inspired by Auditory Scene Analysis.

Some basic criteria shall however also be considered that are relevant in the feature extraction block: This stage attempts to find and extract the kind of information of the signal that differs between the classes, but stays the same within a class. In other words, good features should possess large interclass mean distance and small intraclass variance. Furthermore, it is crucial to select those features that describe a property of the signal that is not described by the other features. In other words, the features should be as uncorrelated as possible, and the less are needed the better from an implementational point of view. Finally, the features should be insensitive to extraneous variables, such as the signal level (unless they describe the level itself).

It is also important to understand that the feature extraction block acts as a filter for the following classifier, which means that the classifier can not perform satisfyingly if not all essential information is passed on to it – a complex classifier is of no use if the features supply only rubbish.

In this chapter, a number of features are presented that have been investigated and/or implemented. These features motivated by Auditory Scene Analysis have partly been used by other people, mainly for source tracking and source separation, rather than for sound classification.

The extraction of the features is generally done in two steps: First, a feature map is calculated that shows a pattern of a certain property of the signal over time and frequency (for example an onset pattern), second, the actual features are extracted from this map (for example the mean of the onsets over a certain time).

The chapter is concluded by a first guess about the choice of a feature set that is optimal for sound classification with respect to the criteria named above. However, which set is actually the best can only be evaluated together with a pattern classifier. This will be done in chapter 7.

The implementations described here are all performed with the help of Matlab and its signal processing toolbox.

5.2 Features Motivated by Auditory Scene Analysis

5.2.1 Spectral Separation

Spectral separation builds the basis for most of the following features, in the same way as the spectral map generated by the cochlea is the general basis for further auditory processing. It can be performed with some kind of filterbank or Fourier transformation. In the Auditory Scene Analysis model of Brown and Cooke (1994), a 128 channel gammatone filterbank was used together with a hair cell model in order to model the cochlea quite accurately. In a hearing instrument, such a representation is presumably too complex and not needed for sound classification. Thus, the spectral separation is carried out according to the Bark scale; a frequency scale that is motivated by the shape of the auditory filters in the cochlea, with narrow filters in the lower and broader filters in the higher frequencies (see for example Zwicker and Fastl, 1990) .

If a simple FFT is used, a trade-off between good spectral resolution (especially in the low frequencies) and good temporal resolution (especially in the high frequencies) has to be made. A 128 point FFT at 22.05 kHz sampling rate, for example, will give a time window of 5.8 ms, or a frequency resolution of 172 Hz, respectively. This is only slightly worse than is required for a Bark scale in the low frequencies (about 100 Hz). In the higher frequencies, the FFT bins are combined to the width of the Bark bands. The temporal resolution of 5.8 ms is just high enough to account for fast transients in speech or music (around 10 ms). However, for pitch detection, a longer time window will be needed, in order to detect pitch frequencies down to some 50 Hz (see section 5.2.3).

5.2.2 Spectral Profile

It has been shown in chapter 4.4.2 that different sounds mostly have different spectral profiles. Thus, features describing the spectral profile could contribute to sound classification. Two spectral features described earlier are shortly reviewed here, followed by a more extensive approach.

5.2.2.1 Spectral Center of Gravity and Its Fluctuations

A simple way to analyze the spectrum was already described in chapter 2.2.2: The center of gravity of the spectrum *CGAV* (equation 2.14) gives a general idea whether the signal is low or high frequently, and the temporal behavior of the center of gravity is described by the

fluctuation strength *CGFS* (equation 2.15). The spectral fluctuations of speech and partly of music are much higher than of many noises.

5.2.2.2 Spectral Ripple

If a more precise description of the spectral form is desired, an analysis of the spectrum can be performed by the cepstrum (for a detailed description of the cepstrum, see for example Gold and Morgan, 2000). The lower cepstral coefficients (the lower quefrencies) describe the rough form of the spectrum, the higher coefficients (the higher quefrencies) the finer structures. That is, the lower cepstral coefficients are suited to describe different spectral profiles.

It is reasonable to reduce the large number of cepstral coefficients to a smaller number of features before feeding them into a pattern classifier. One idea is to analyze the modulations of the spectrum in different modulation frequency ranges. Figure 5.2 shows the block diagram of an algorithm that it is proposed by Korl (1999).

The cepstrum is calculated by taking the logarithm and the IFFT of the Bark spectrum. The logarithm may be omitted, which leads to a "pseudo" cepstrum. The cepstrum is liftered in three channels, that is, the coefficients are combined in three ranges, and an FFT is applied in each channel to analyze the frequencies in these channels, which describe the coarser and finer "changes" of the spectrum across frequency. Because the zero cepstral coefficient is not considered, the absolute level of the spectrum should not matter. The RMS of the three signals, normalized to the intensity of the spectrum, gives then the three "modulations depths".

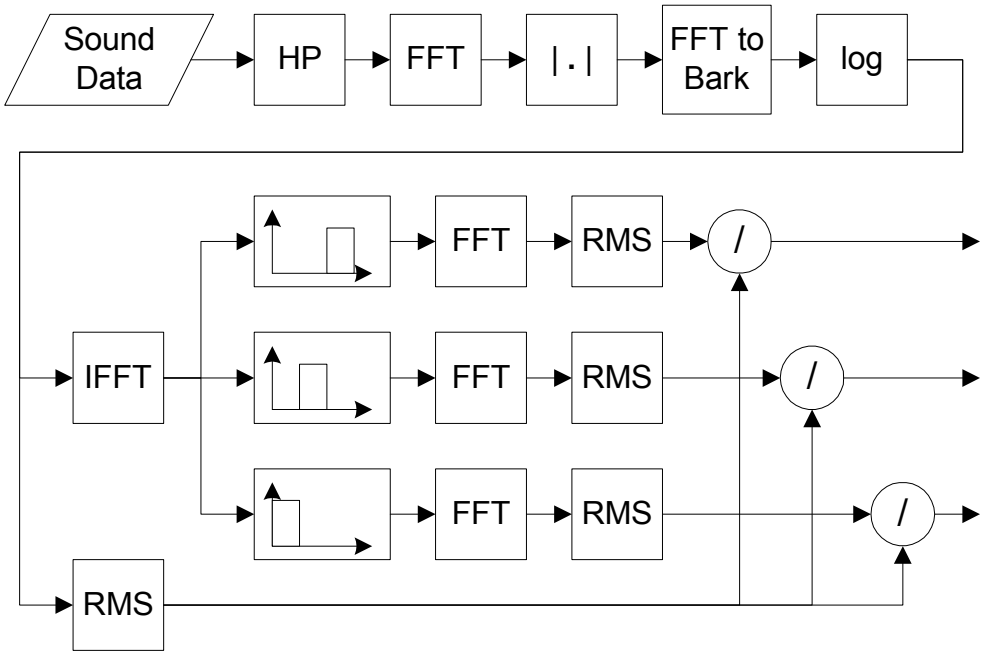


Figure 5.2: Block diagram for the calculation of the spectral ripple, after Korl (1999). The cepstral coefficients are combined into three channels, and the modulations in each channel are calculated.

A factor analysis was performed by Korl (1999) to find out the best organization of cepstral coefficients into the three channels. The best combination is c1-c4, c5-c16 and c17-c31.

Korl (1999) also evaluated this algorithm and found that it is level dependent, despite the normalization. This makes it quite inapplicable for classification.

So far, the algorithm has not been further investigated, partly because of this difficulty, partly because the variance of the spectral shapes in a sound class may be quite high; there are other more promising features, which were explored first, for example the harmonicity, as will be described in the next section.

5.2.3 Harmonicity / Pitch

Harmonicity and pitch perception are regarded as being very important features in Auditory Scene Analysis. The existence or the absence of a pitch as well as the temporal behavior of the pitch give us much information about the nature of the signal.

Some features describing the harmonicity of the signal shall tell us if the signal contains tonal components (that is, a pitch) or only non-tonal noise. It is expected that this helps to distinguish music from noise, and also speech from noise.

It has been stated earlier that pitch is a subjective measure. The existing models can only account for a few of the complex auditory mechanisms that contribute to pitch perception, such as the "missing fundamental" problem (see chapter 4.4.3). So, if the expression "pitch" is used in the following, it may for some cases just mean fundamental frequency.

There are currently two important models of pitch perception: The model from Meddis and Hewitt (1991) and the AMPEX-Model from Martens and Van Immerseel (1992). Both models consist of a large peripheral stage with 128 channel filterbank and hair cell model, autocorrelation of the neural impulses, and the summation of the correlation functions. Thus, they are computationally quite expensive and can hardly be used directly. The model of Meddis and Hewitt has been simplified by Karjalainen and Tolonen (1999), as Figure 5.3 shows.

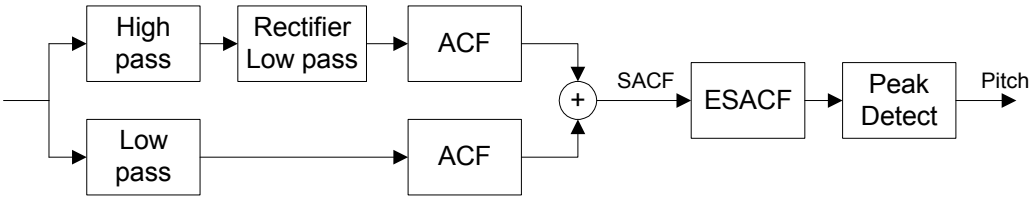


Figure 5.3: Block diagram of the model from Karjalainen and Tolonen (1999). Corresponding to the processing in the auditory system, the computation of the ACF in the low frequencies is accomplished directly, in the high frequencies after building the envelope. The summary ACF is then further enhanced and a peak detector determines the pitch.

The signal is first split up into two channels below and above 1 kHz. Because the auditory system is able to resolve the harmonics in the lower frequencies, the lowpass channel is directly autocorrelated. Higher harmonics cannot be resolved any more by the auditory system and give just an overall impression. This is why in the highpass channel, the envelope is built by rectifying and lowpass filtering, before it is autocorrelated.

The two signals are then summed up to the summary ACF (SACF). The SACF is further processed in order to get rid of redundant peaks, resulting in the enhanced summary ACF (ESACF). Finally, a peak detection block extracts the pitch, if there is one present at all.

Korl (1999) compared this model to a simplified version, as shown in Figure 5.4. The calculation of the ACF in two channels and the enhancing block ESACF are useful if a separation of sources is desired; for simple pitch determination, Korl left them away. Thus, the signal was autocorrelated in one channel by calculating the IFFT of the power spectrum (the power spectrum and the ACF build a Fourier transform pair). The results were comparable when the amplitude spectrum instead of the power spectrum was taken, saving the squaring operation. Note that the quasi-ACF obtained by this is the same as when a quasi-cepstrum is calculated without the logarithm. So, ACF and cepstrum describe both the same, which is the form of the spectrum (see also section 5.2.2.2).

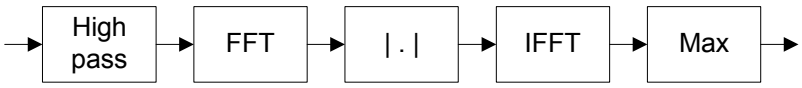


Figure 5.4: Block diagram of the simplified algorithm after Korl (1999). A quasi-ACF is calculated by applying an IFFT to the amplitude spectrum, and the pitch is determined by the maximum within a certain range.

The pitch is then extracted by a simple peak detector within a reasonable pitch range (for example between 50 and 500 Hz).

The highpass at the beginning shall remove the DC part and the frequencies below the resolution of the ACF (determined by the window length).

Figure 5.5 shows some typical samples of the extracted pitch for clean speech, speech in traffic noise, traffic noise, and classical music. If no pitch is detected, it is set to zero. For clean speech, the prosody can be observed very well. If noise is added, this is partly masked. The example with the traffic noise shows the extreme case, where only non-tonal components are present. The pitch in music is determined by single tones or chords and can jump up and down a lot for classical music.

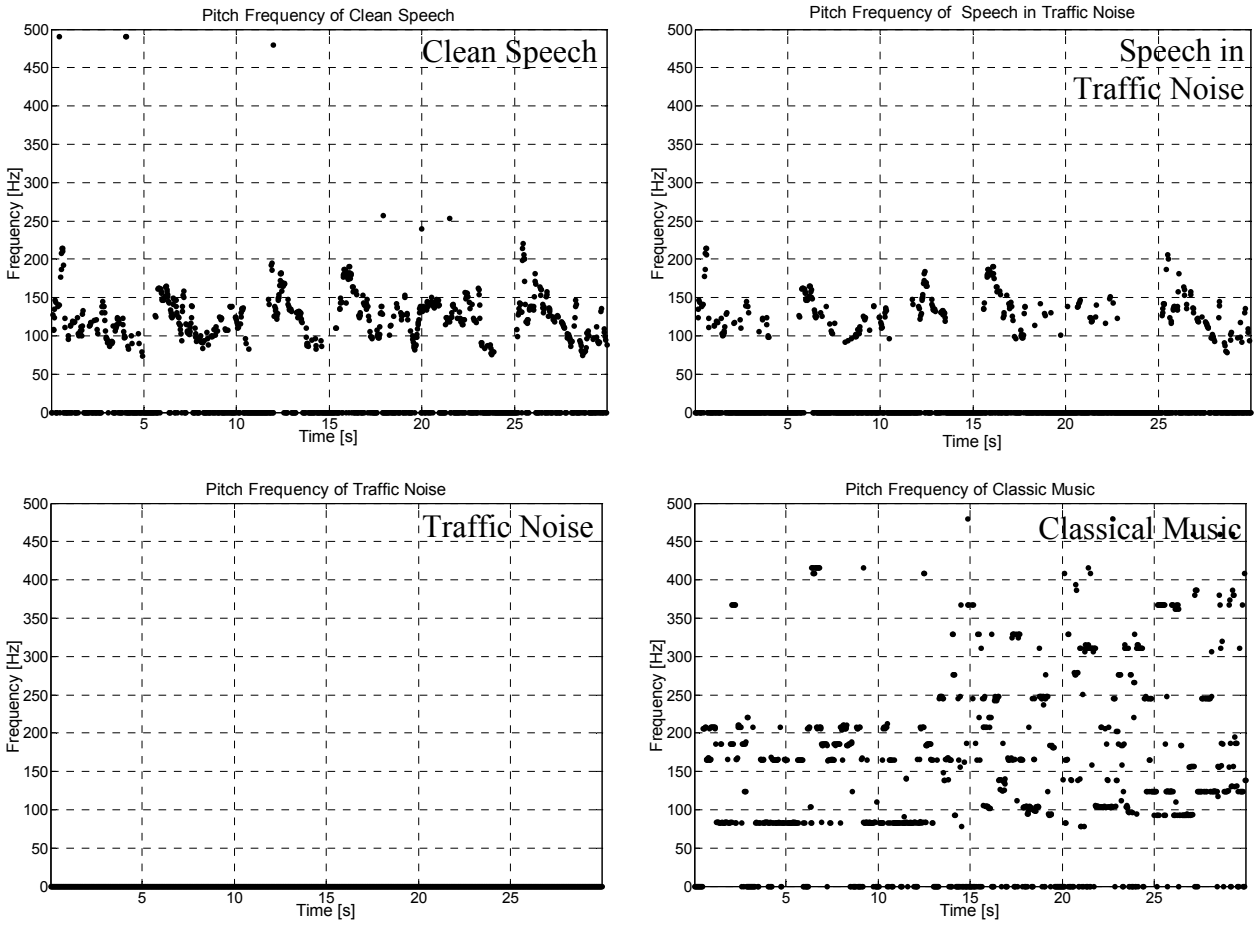


Figure 5.5: Typical temporal behavior of the pitch for clean speech, speech in traffic noise, traffic noise, and classical music. If no pitch is detected, it is set to zero. The prosody of the speech is partly masked when noise is added. Noise alone has significantly less tonal components (here, the extreme case is shown). Single tones as well as chords determine the pitch in music.

5.2.3.1 Pitch Features

The pitch itself is only one possible feature for classification; it was however not explicitly used, because the absolute value of the frequency has not been considered very useful for classification. More important is probably a feature that shows whether a pitch is present or not, that is, whether the signal contains tonal components or not. Another important matter is the temporal variation of the pitch; it can for example be described by its variance, or by the difference between two consecutive pitch values (approximation of the derivative). This leads to the three pitch features that are shown in Figure 5.6 and in the table below.

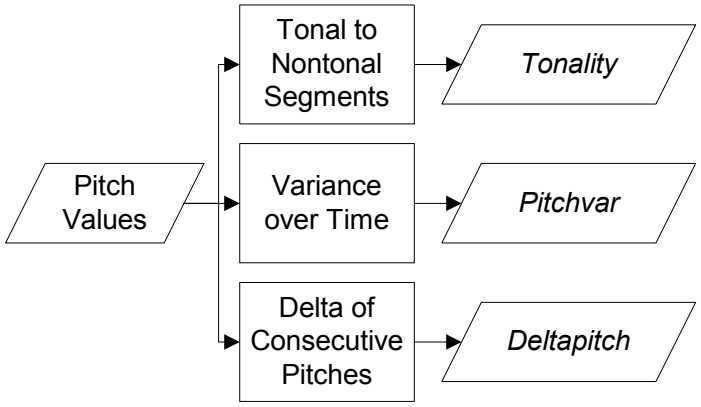


Figure 5.6: Block diagram for the extraction of the three pitch features Tonality, Pitchvar, and Deltapitch.

Pitch Feature	Description
Tonality	Relation between tonal and non-tonal segments in a certain time window
Pitchvar	Variance of pitch in a certain time window
Deltapitch	Absolute value of difference of two consecutive pitch values, averaged over a certain time window

5.2.3.2 Mean Feature Values per Sound

The bar graphs in Figure 5.7 to Figure 5.9 show the mean pitch feature values for the 287 sounds of the soundset described in appendix A. It would need too much space to plot all feature values for all sounds. Thus, the feature values are averaged for each sound over the thirty seconds sound length, resulting in one feature value for each sound. Note that with such a representation, the temporal behavior of the features within a sound can not be observed.

The *Tonality*, depicted in Figure 5.7, is not only the simplest feature, but also the most expressive. It is rather high for most speech files and many music files, a bit lower for speech in noise, and quite low for noise files. However, there are some exceptions:

- In some strongly reverberated speech files (no. 58-60), especially non-tonal hissing sounds are reverberated, which results in a low tonality.
- A few noises contain tonal components, which results in a mediocre to high tonality: Chainsaw (no. 179), electric shaver (no. 204).
- A couple of pop music samples (for example no. 241, 245, 282) obviously resemble more to speech in noise than to music, and their tonality is also in this range. If some drums are dominating the scene, they contribute strongly to a low tonality.

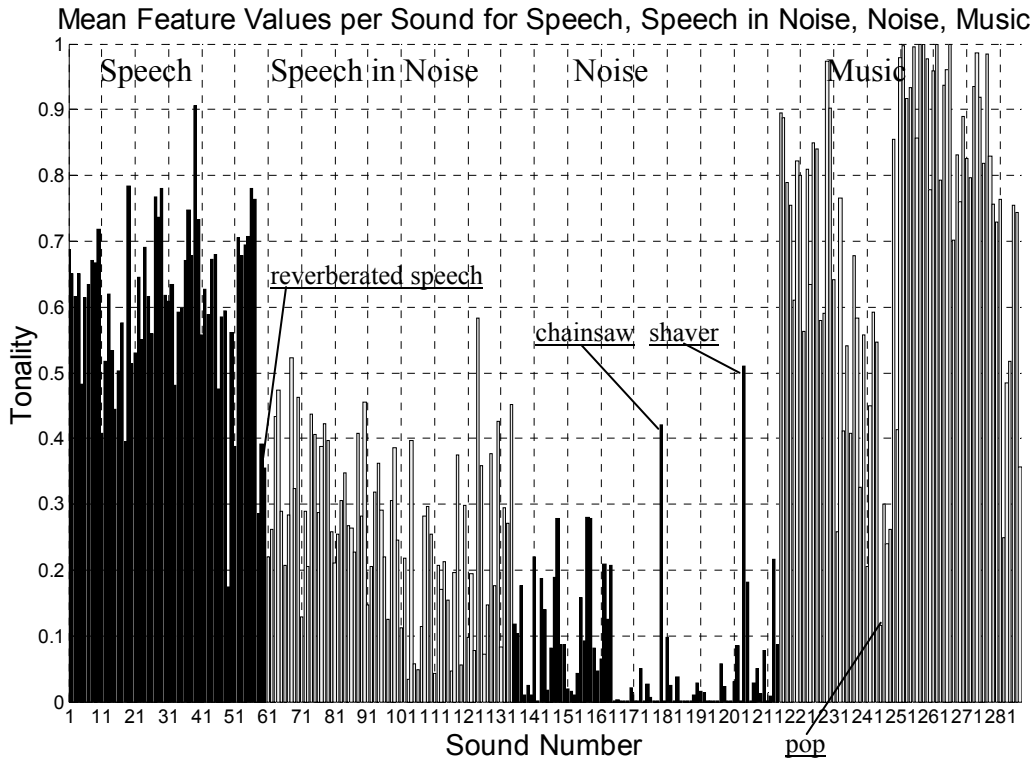


Figure 5.7: Mean feature value per sound for the feature Tonality. For most speech and music sounds, it is quite high, whereas for speech in noise and especially noise sounds, it is rather low. Exceptions are particularly noises with tonal components and pop music with dominating drums or other non-tonal components.

The *Pitchvar*, shown in Figure 5.8, is higher for music than for the rest – the pitch of music changes according to the melody. The variance of this feature within a class however is quite high, which makes it less useful than the *Tonality* feature. These differences can be explained:

- The dialogues in the class speech (no. 31-35) may have a higher pitch variance than the sounds with only one speaker, because there are two pitches that alternate.
- The lower frequencies of the radio speaker signal (no. 48) are attenuated due to the transfer function of the loudspeaker in the radio. Thus, the first harmonic is often higher in level than the fundamental frequency, and the pitch detector jumps between the two peaks.
- In the class speech in noise, two pitches may be present, if the noise contains tonal components; for example speech in a drilling noise (no. 101) or speech in a car (no. 80). Because the pitch detector jumps between the two pitches, the pitch variance increases (this can of course also happen for two tonal noises with different pitches). In real life situations, this might occur very often, making the *Pitchvar* feature rather useless.
- In music, singing slowly (no. 271, 276, 277) can result in a constant pitch, that is, in a low pitch variance.

Maybe the pitch detector should account for the fact that more than one pitch source can be present and detect the number of pitch sources. This however can be a difficult task and is not performed so far; for the moment, the *Pitchvar* feature is used as it is.

The *Deltapitch* (Figure 5.9) correlates very much with the *Pitchvar*, because it describes a similar signal characteristic. It is most probably sufficient to take one of the two features for classification.

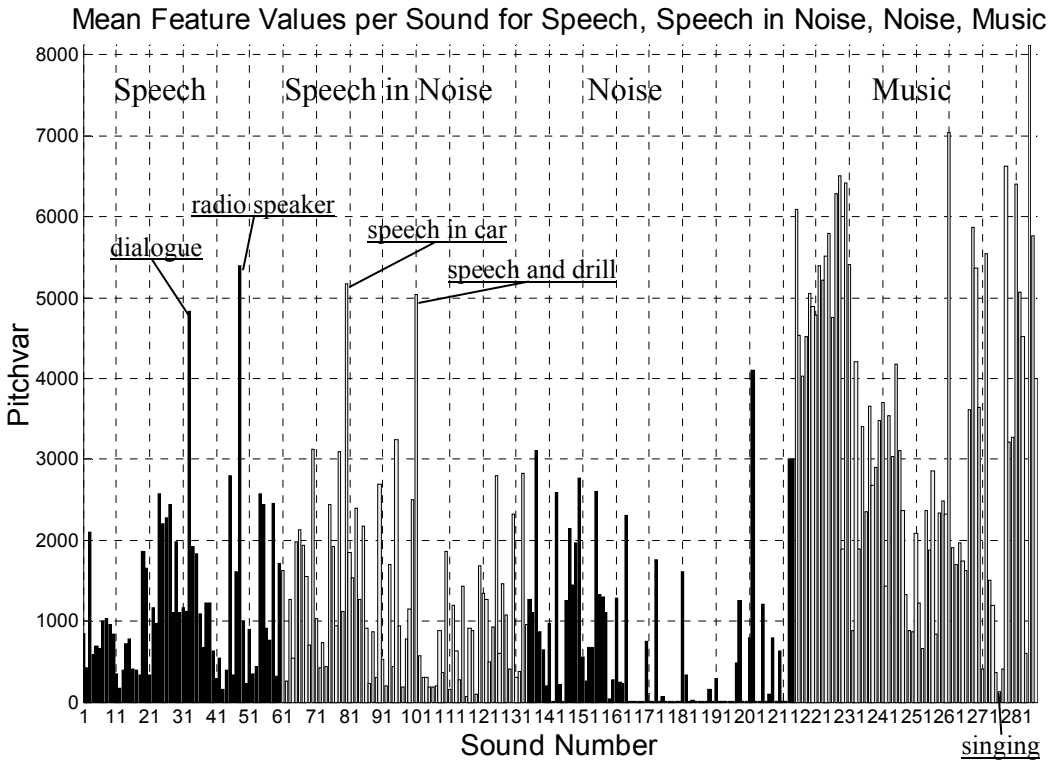


Figure 5.8: Mean feature value per sound for the feature Pitchvar. For music, the values are a bit higher than for the rest, but the variance within the classes is rather high. If more than one pitch source is present in the signal (dialogue, speech in tonal noise), it will result in a high pitch variance.

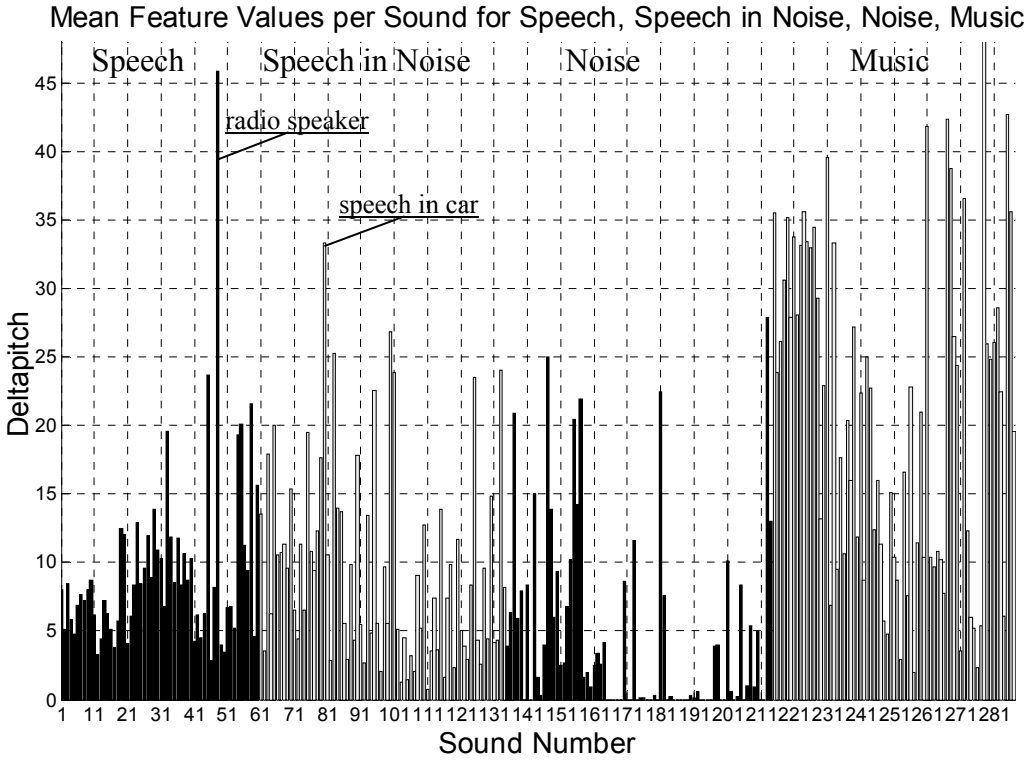


Figure 5.9: Mean feature value per sound of the feature Deltapitch. This feature correlates very much with the Pitchvar feature; it is probably sufficient to use only one of the two features.

5.2.4 Spatial Separation

Spatial location is a significant cue for sound separation and may also be of great importance for the analysis of the acoustic environment. However, the signal to be explored is currently a mono signal, as has been described earlier. Thus, no spatial features are investigated for classification in this thesis.

In the future, the hearing instruments might classify the signal coming from the front and the one from the back separately, or the number of sources might be explored with the help of interaural differences in time and level (or differences between two or more microphones), in order to find out how many speakers are present, or to locate and suppress noise, etc. The latter is already available in modern hearing instruments: The noise is located and the characteristics of the directional microphone are adapted in order to have the highest attenuation in the direction from which the noise originates. See also chapter 8.2 about future work.

5.2.5 Temporal Onsets and Offsets

The overview of Auditory Scene Analysis showed that common onsets of partials are a strong grouping feature (see chapter 4.4.5). If common (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 (that is, musical

instrument). Offsets are regarded as much less important in Auditory Scene Analysis; thus, only onsets are considered here.

In a number of source separation algorithms, spectrotemporal onset maps have been computed (for example Brown & Cooke, 1994, Mellinger & Mont-Reynaud, 1996). They have, however, not been used yet for classification purposes; the aim was to group acoustic components which have the same onset times. The idea of the models is to respond with a short burst at an amplitude onset of a stimulus. Brown and Cooke (1994) integrated the output of the hair cell model over 20 ms to get the envelope. This time constant was chosen so that glottal pulses of speech stimuli are removed, but fast onsets from plosives (20 to 30 ms) will be detected. After the stimulation, an excitation burst is emitted, followed by strong inhibition, to prevent activity throughout the remaining stimulation.

This physiologically motivated model is simplified as shown in Figure 5.10. After the calculation of the power spectrum, the envelope is built in one band or in twenty Bark bands. Each band (one band in the broadband case) is averaged using a first-order lowpass with a time constant of 10 ms. Note that this is a simple approximation for the envelope, which is however preferred to the mathematically correct solution via the Hilbert transform (Hartmann, 1998). The latter can produce envelopes that "may not be slowly varying with respect to the fine structure" (Viemeister & Plack, 1993). Then, the difference (in dB) of the envelope from frame to frame is calculated. If it is above a certain threshold, the difference is output, otherwise the output remains zero. This way, only large onsets will produce an output peak.

The output of the algorithm are spectrotemporal onset maps in one or twenty bands, respectively. They build the basis for further computations, but also for a visual analysis of the data, which is done next.

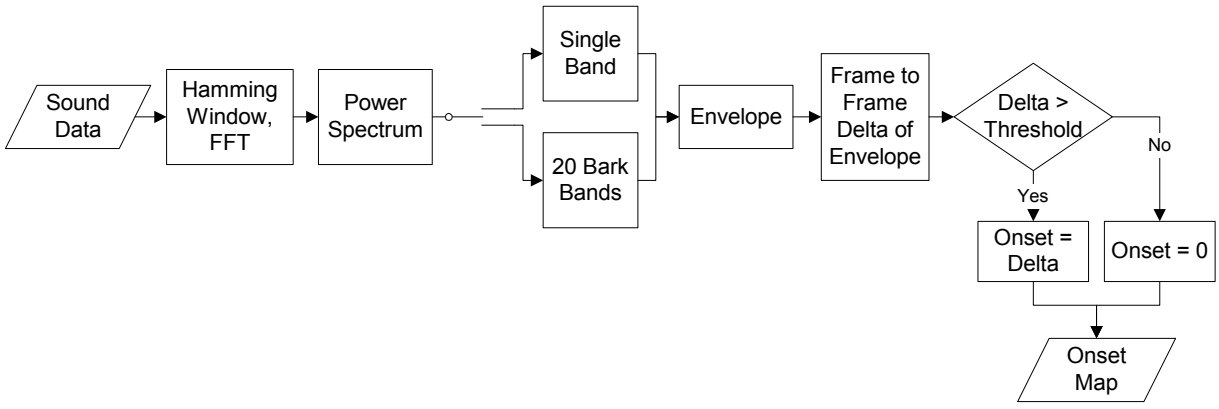


Figure 5.10: Onset extraction of sound data. If the difference of the envelope from frame to frame is above a certain threshold, it is considered as onset. The calculation can be done in one single band or in twenty Bark channels.

5.2.5.1 Visual Analysis of Onset Maps

Figure 5.11 shows singleband onset maps for a nine seconds time window of several different signals. In speech signals, strong (steep and high) onsets alternate with weaker ones, whereas noise and music signals only have weak onsets. Many of the noise and music samples have very similar onset patterns, which makes it hard to distinguish between them. Different music styles tend to give different onset patterns: the pop music sample shows more fluctuations in the onsets as the classical music sample. Note, however, that this characteristic does not occur

for all pop or classical samples, as there may be very smooth pop music, or classical music with many *crescendi*. For many pop music samples, the rhythmic beat can be seen in the onset pattern.

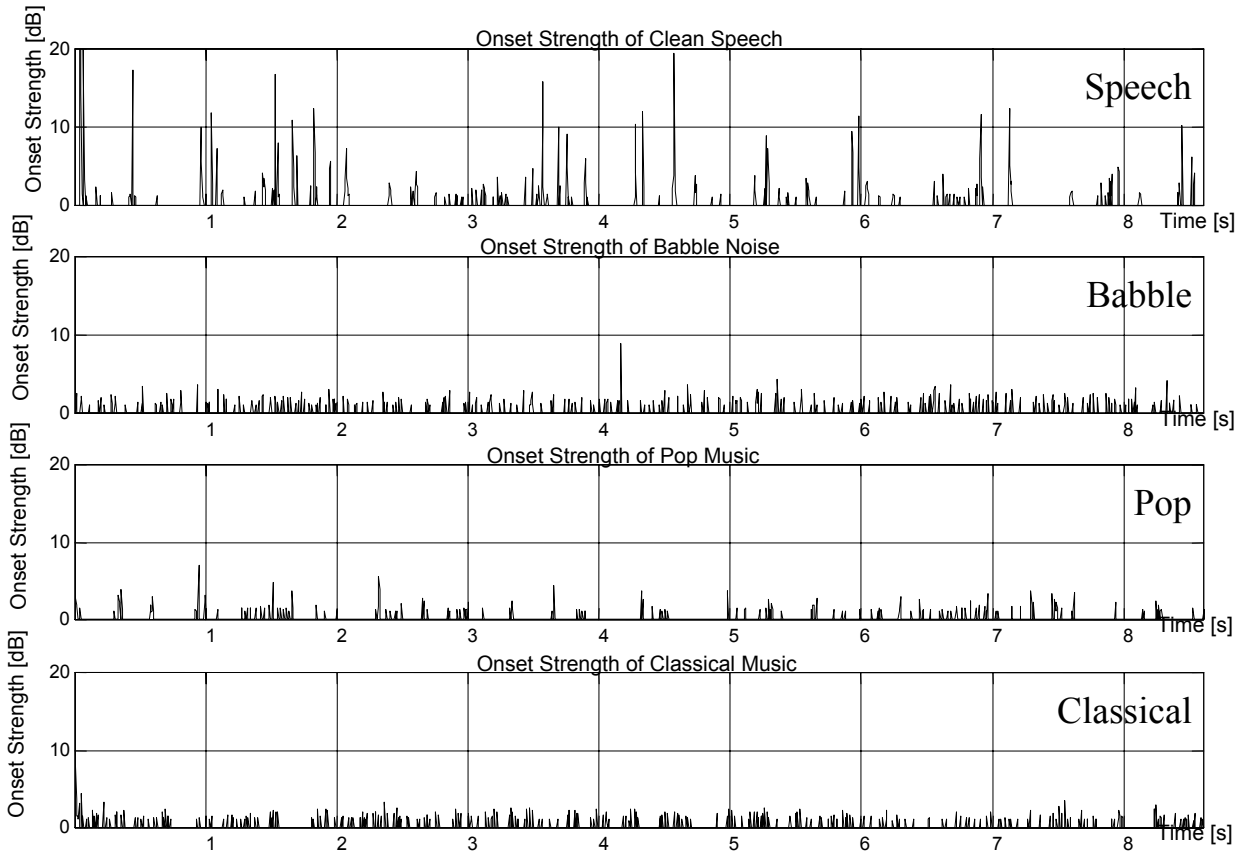


Figure 5.11: Broadband onset strength for speech, speech babble, pop and classical music. Speech signals show much stronger (steeper and higher) onsets than other sounds. Noise samples differ only partly from music samples.

Thus, singleband processing of the onsets is obviously not sufficient for separation of the four classes. Therefore, it shall be investigated now how much information can be found across frequency bands. If the onset is calculated in 20 Bark channels, spectrotemporal onset patterns like in Figure 5.12 are obtained. Onsets that are above 7 dB/frame are displayed as small dots and above 10 dB/frame as large dots. The strong onsets in speech signals can also be seen here, and, there are many onsets simultaneously occurring over many bands, which is especially due to the plosives. For speech in noise, the speech onsets are partly masked by the noise floor. In quasi-stationary noise, the onsets across bands are mostly quite weak and uncorrelated. This is also the case for some music samples. In the figure, however, a pop music signal is shown that contains a strong rhythmic beat. The rhythm can clearly be determined via the onset map, especially in the upper frequency bands, due to the high frequency drums. Note however that in other pop music samples, the drums are more in the low frequency range, or masked by stronger onsets, which means that it might not always be easy to determine the rhythm from the onset map. Nevertheless, the question arises whether a

measure for rhythm can be calculated from the onset pattern, to distinguish between music and noise or different music types. This will be discussed separately in section 5.2.6.

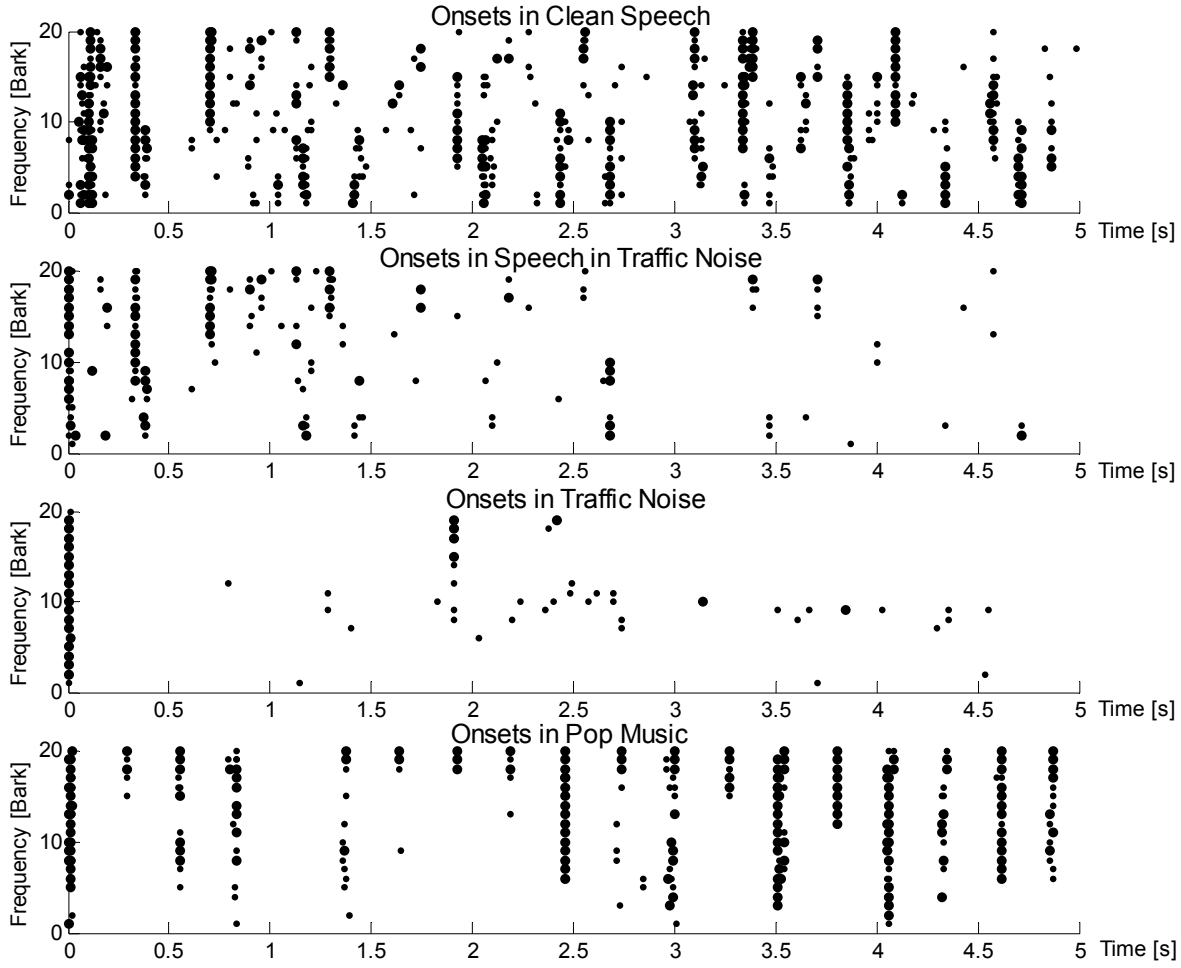


Figure 5.12: Amplitude onsets in twenty Bark bands for clean speech, speech in traffic noise, traffic noise and pop music. Dark areas of the image indicate regions of strong onsets. In speech, many strong onsets occur simultaneously over the bands. If the speech is masked by noise, the onsets are weaker; in quasi-stationary noise, they almost disappear. This is also the case for many music samples. In the pop music sample, the strong rhythmic beat is very well described by the onsets.

Finally, informal investigations showed that in-the-car noise can be identified by its onset pattern. Many common – though not strong – onsets occur in the upper bands, which makes the pattern visually immediately stand out of other patterns. Thus, in the future, a feature might be implemented that identifies in-the-car noise based on the onset pattern.

5.2.5.2 Onset Features

A couple of simple features can be extracted from the spectrotemporal onset map, as is shown in Figure 5.13.

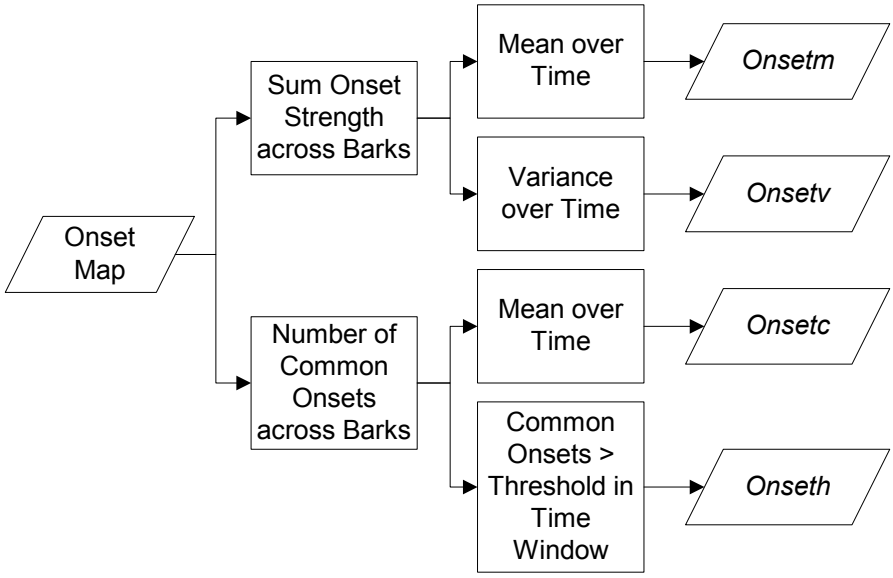


Figure 5.13: Block diagram of the extraction of the onset features from the onset map.

First, the onset strengths are summed up across bands. Then, the mean and the variance of these sums are taken over a certain time window, *Onsetm* and *Onsetv*. This shall give information about the overall onset strength and its fluctuation over time.

A further feature describes how many onsets occur simultaneously across the bands at one time, the so called common onsets. If the onset in a band is above a certain threshold, the common onset counter is increased by one. The common onset feature *Onsetc* is then the mean of the counter values over a certain time window.

Finally, it is counted, how often the common onsets are above a certain number during a certain time, resulting in the feature *Onseth*. Together with the *Onsetc*, this gives some information about the relation of few common onsets to many common onsets across time.

This leads to the following onset features:

Onset Feature	Description
Onsetm	Mean onset strength over all bands in a certain time window
Onsetv	Variance of onset strength in a certain time window
Onsetc	Common onsets across bands, averaged over a certain time window
Onseth	Number of times that a high number of common onsets occurs in a certain time window

5.2.5.3 Mean Feature Values per Sound

Again, the mean features per sound are calculated for the whole soundset and plotted as a bar graph in Figure 5.14 to Figure 5.17.

The mean broadband onset strength *Onsetm* (Figure 5.14) is higher for clean speech than for most other files. Exceptions are noises with strong onsets like teleprinter and typewriter (no.

183 and 212), or speech in these noises (no. 103 and 132), and also music with fast drums (no. 284) or specific instruments like xylophone (no. 267). In-the-car noise (no. 158-164) stands out of other noises only a bit. The onsets in reverberated speech (no. 51-60) are less strong and near the speech in noise region.

The variance of the onsets *Onsetv* (Figure 5.15) gives quite a similar picture as the mean. The contrast is even increased when there are strong onsets alternating with no onsets in a sound, such as in the typewriter noise (no. 212) or the xylophone (no. 267). In-the-car noise (no. 158-164) is not well detectable with this feature.

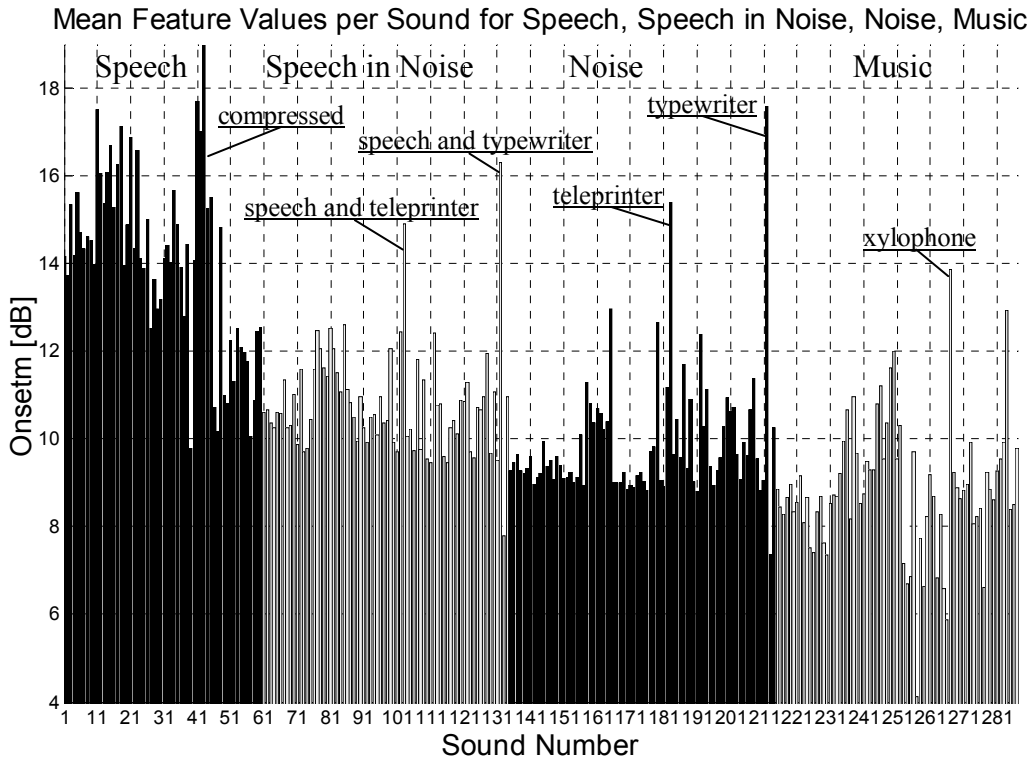


Figure 5.14: Mean feature values per sound for the mean onset strength over all bands (Onsetm). Speech stands out of the rest; reverberated speech is in the area of speech in noise.

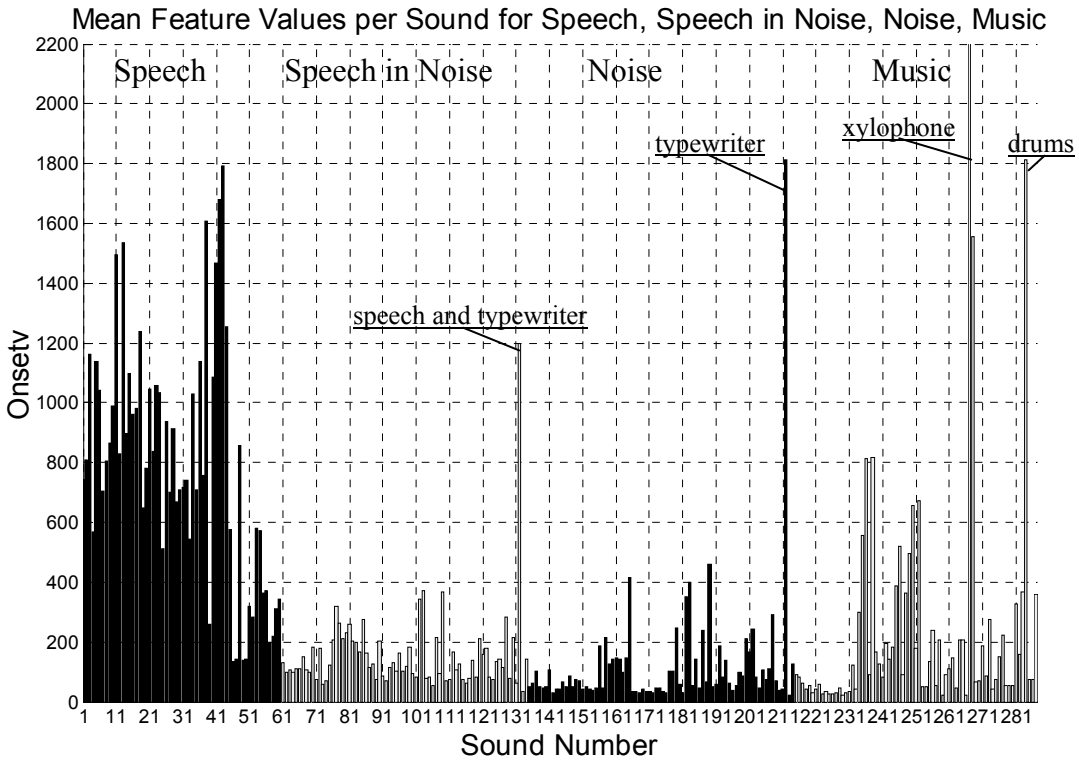


Figure 5.15: The mean feature values per sound for the variance of the onset strength (Onsetv) is similar to the one for the mean. For some sounds, the contrast is even higher.

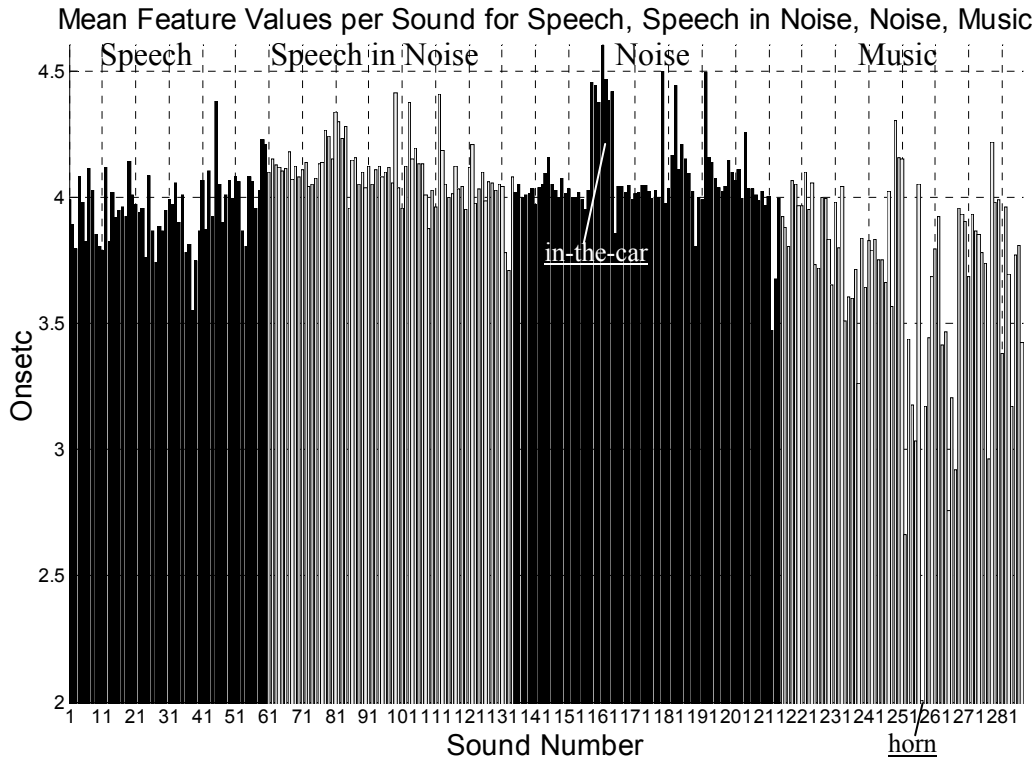


Figure 5.16: Mean feature values per sound for common onsets across bands (Onsetc). In-the-car noise has many common onsets and stands out a bit.

The mean common onsets *Onsetc* (Figure 5.16) do not show much difference across the classes. Exceptions are the in-the-car noises (no. 158-164), which have common onsets in many frames. With some musical instruments, the notes may be played quite long, resulting in very few common onsets over time (for example with a horn, no. 257). Furthermore, this feature will probably reveal more information when its temporal behavior within a sound is investigated, instead of the mean value over time.

The number of times that a high number of onsets occur simultaneously in the bands, *Onseth*, is a bit higher for speech than for the rest (Figure 5.17). The intraclass variance of this feature is however quite large, which makes it not very useful. Only the in-the-car noises (no. 158-164) are all really high.

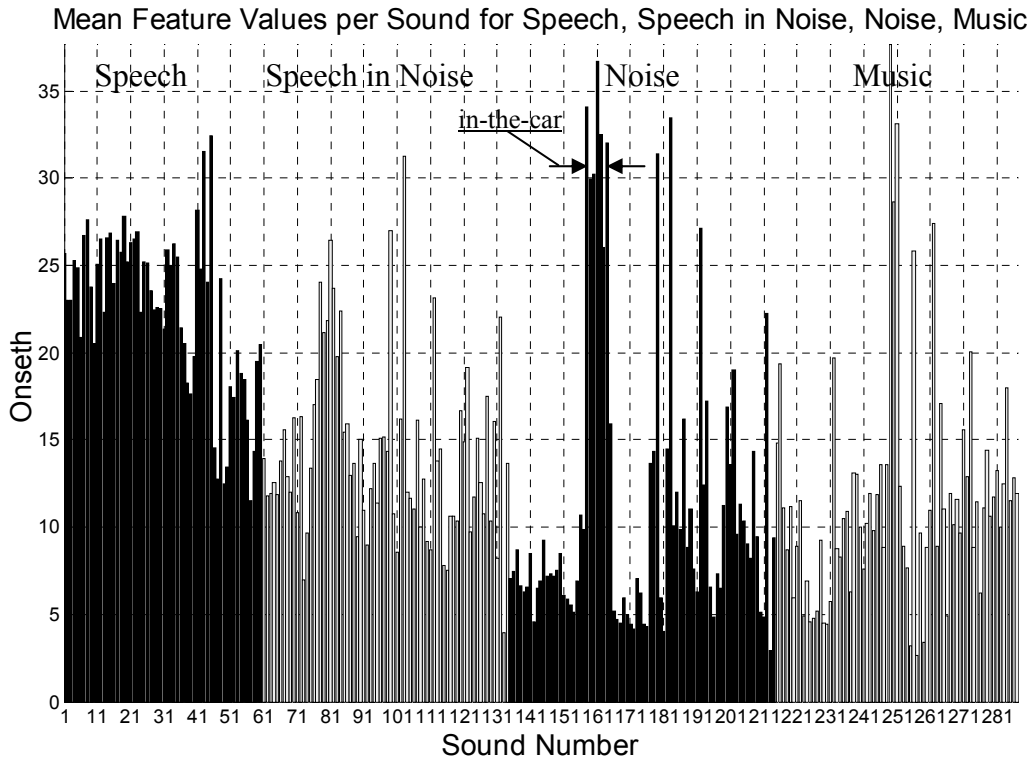


Figure 5.17: Mean feature values per sound for number of times that a high number of onsets occur simultaneously in the bands (Onseth). Speech and in-the-car noise stand again out of the rest, as well as some musical instruments.

5.2.6 Rhythm or Beat Extraction

As we have seen in chapter 4.4.5, rhythm is not a physical, but a subjective measure, that is associated with qualities of grouping and hierarchy. If a measure for equally spaced temporal events shall be described, the words "beat" or "pulse" are used in the literature (see for example Handel, 1989). As shown in Figure 5.12, the beat in a musical signal is well reflected in the spectrotemporal onset pattern. Especially pop music often has a strong beat; thus a feature describing the beat, that is the strength of the rhythmic events, could be helpful for classification, either to distinguish between different music styles or between music and speech in noise. This might solve the problem that most properties of certain pop music

samples can be quite similar to those of speech in noise, if one lead singer is present together with a number of background instruments that may sound quite noisy.

Scheirer (1997, 1998) compared an algorithm for beat and tempo tracking with the pitch tracking algorithm of Meddis and Hewitt (1991), see section 5.2.3. He found that the pitch tracker can also be used for beat analysis, when different time constants (or frequency ranges, respectively) are used, for the range of 60 to 180 beats per minute. Figure 5.18 shows the block diagram of an algorithm that is based on this principle. The onsets are computed in twenty Bark bands as shown in Figure 5.10. Then, the signal is highpassed to remove the DC part and frequencies below the resolution of the following ACF (determined by the window length). A quasi-ACF is computed by applying an FFT, the absolute value and an IFFT to each Bark band. Then, the summary ACF (SACF) is calculated out of the normalized ACFs in the twenty Bark bands.

To reduce the complexity, it is also possible to first sum up the onset channels and compute only one broadband ACF. First investigations showed that the results are similar for many sounds, even if not for all.

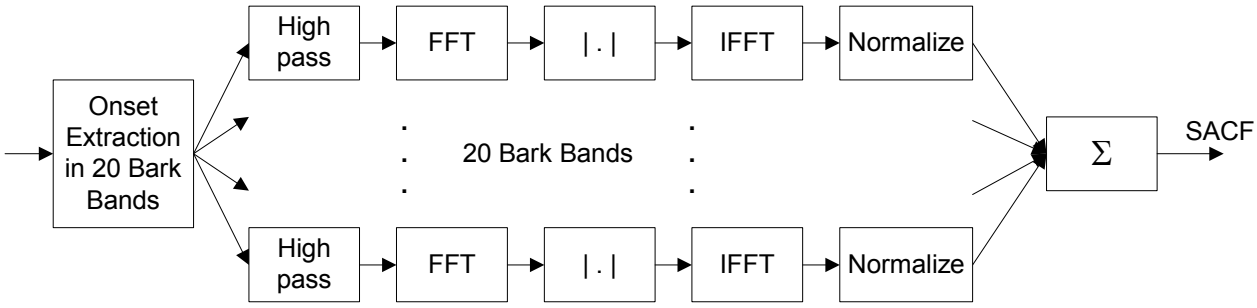


Figure 5.18: The beat is analyzed by computing the quasi autocorrelation function of the onsets in Bark bands. The procedure in a band is equivalent to the pitch extraction algorithm of section 5.2.3 using other time constants, or frequency ranges, respectively. The twenty normalized ACFs are then summed up to a summary ACF (SACF).

The length of the ACF window was chosen so that it contains at least ten beat pulses, which is a couple of seconds. In Figure 5.19, the output of the algorithm is plotted for a six seconds window of pop music and of speech in babble noise. The speech in noise sample has no significant beat, but the peaks at 270 ms and especially at 540 ms in the pop music sample indicate the presence of strong a beat at 111 beats per minute or a weaker beat at 222 beats per minute.

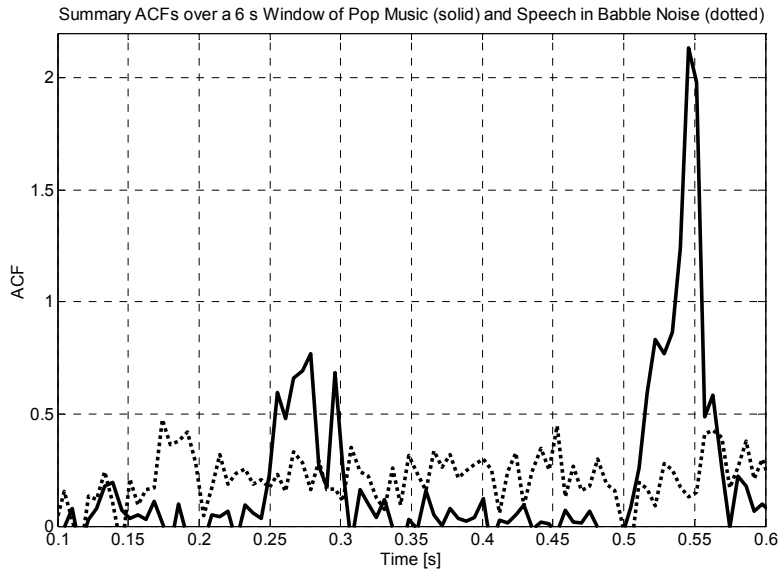


Figure 5.19: Output of the beat analyzer for a 6 s window of pop music (solid) and speech in babble noise (dotted). The high peak at 540 ms in the pop music sample indicates a strong beat at 111 beats per minute; the lower peak around 270 ms shows that there is a weaker beat at double rate (222 beats per minute).

5.2.6.1 Beat Feature

For classification, it does not seem to be very important to know the exact beat frequency, but rather whether a beat is present at all in a certain range. The most simple way to extract a beat feature out of the summary ACF is to determine the peak in the desired time range, which is set to 100.. 600 ms. It has, however, turned out to be useful to observe a couple of consecutive ACFs, to emphasize the beat that remains constant over a longer time period, for example 30 s. The idea is that the beat of music is more persistent than the "beat" of speech signals. This is performed as shown in Figure 5.20. A number of consecutive ACFs are summed up, and then the highest peak in the summary ACF is determined, resulting in the feature *Beat*. Note that this means that a new feature value is obtained only every 30 s or so.

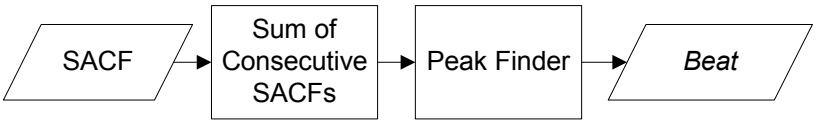


Figure 5.20: For the feature Beat, a number of consecutive ACFs are summed up before determining the maximum in a certain time range. This emphasizes beats that remain stable over a longer period of time, assuming that this is more the case for music than for speech.

5.2.6.2 Mean Feature Values per Sound

The mean feature values per sound for the beat feature are shown in Figure 5.21.

Many signals of the classes speech, speech in noise and noise have low beat feature values, compared to pop music. There are however a number of exceptions:

- One of the speech signals (no. 13) is a rhyme with a rhythmic structure. This results in strong beat features.
- A radio news speaker signal (for example no. 41) may be very continuous, which results also in a certain rhythm.

- Any rhythmic noise (no. 165, tractor; no. 185, lawnmower; no. 189, piledriver; no. 194, weaving machine; no. 204, electric shaver; no. 212, typewriter) or speech in these noises will result in high beat feature values.

The music signals can also be divided into signals with beat (mostly pop music) and without beat.

Concluding it can be stated that the algorithm cannot distinguish between musical or machine beat, as long as the beat frequency is in the same range. Signals with beat and without beat occur in a all four classes. Thus, the beat feature will probably not be suited to distinguish the four main classes, but it might help for a more detailed classification into subclasses.

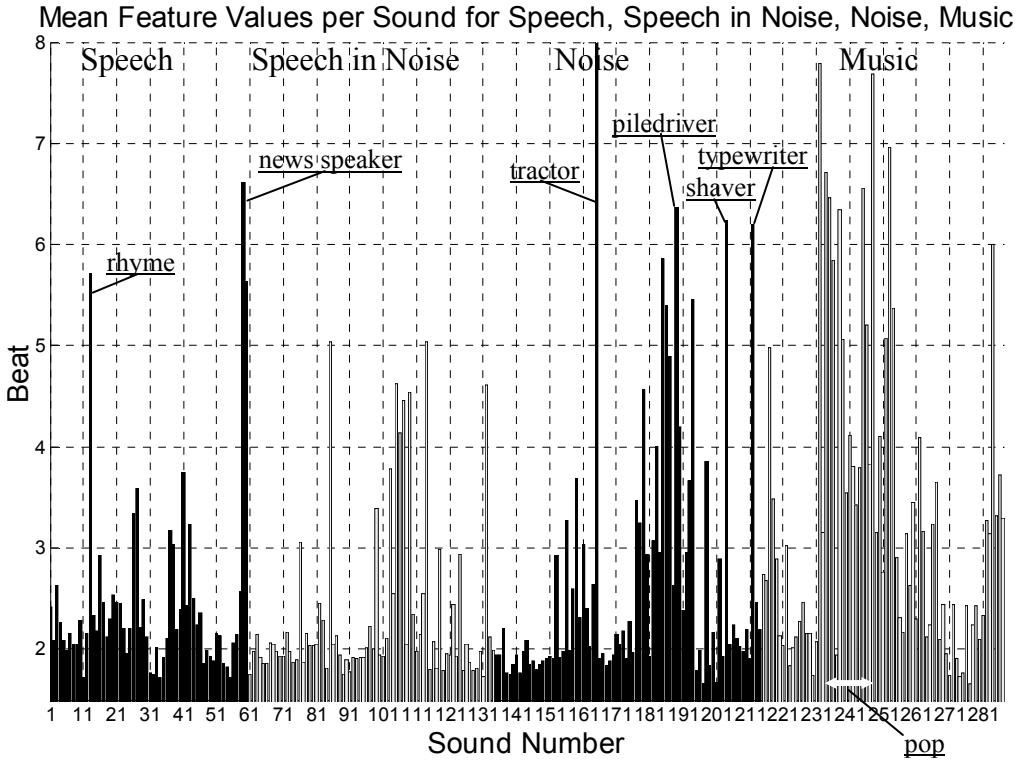


Figure 5.21: Mean feature values per sound for the feature Beat. Signals with a rhythmic content (for example a rhyme spoken, tractor or typewriter noise, but especially pop music) can be distinguished from non-rhythmic signals. It is however not possible to discriminate musical and machine beat.

5.2.7 Amplitude Modulation

The overview of Auditory Scene Analysis showed the importance of amplitude modulations for fusion as well as for segregation. It is also known that the amplitude modulations differ in level and frequency for many sounds.

A number of features describing the amplitude modulation of a signal have already been presented in chapter 2.2 about the state of the art in sound classification for hearing instruments. It was shown that especially speech signals can be separated from other signals, due to the amplitude modulations around 4 Hz that occur in speech.

The histogram features after Ludvigsen (1993), the mean level fluctuation strength after Kates (1995) and the modulation depths after Ostendorf et al. (1997) will be used for classification together with the other auditory features presented in this chapter.

5.2.8 Frequency Modulation

Although frequency modulation appears to be one of the weaker cues in Auditory Scene Analysis, it has been modeled for a number of source separation algorithms. Brown & Cooke (1993, 1994) and Mellinger & Mont-Reynaud (1996) each describe a way to detect frequency modulation in the signal. First, a hair cell model (for example from Meddis and Hewitt, 1991) is applied to the signal, which produces a spectrotemporal pattern at the output, the so called cochleagram. Then, a two-dimensional cross-correlator that operates on rectangular areas of the cochleagram is used. The kernel of this cross-correlation operator is chosen to filter frequency modulation at a certain rate, for example at two octaves/s. A number of different kernel functions are applied, for different transition rates (upward, static, downward). The output is a frequency transition map as depicted in Figure 5.22. The arrows indicate the direction of the transition; in the background, the cochleagram is shown.

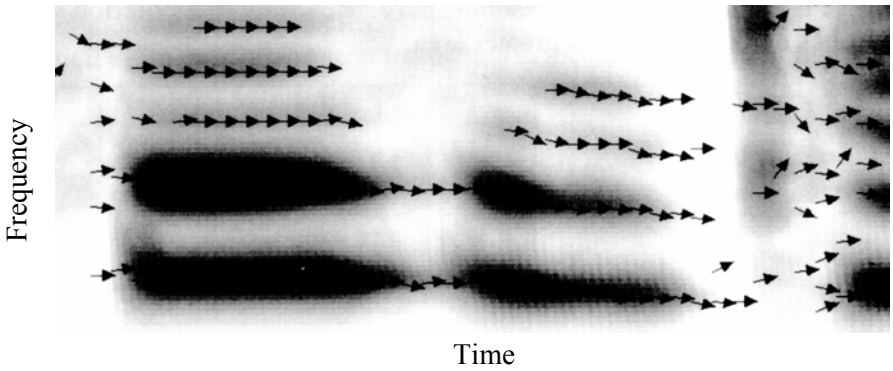


Figure 5.22: Cochleagram and frequency transition map of a speech signal. The arrows indicate the direction of the transition of the partials (from Brown and Cooke, 1993).

The frequency transition map is used for event formation, that is, for tracking the partials over time. For sound classification, it is conceivable to compare the transitions over the frequency bands and over a certain time window, in order to get a feature that gives information about the amount of "turbulence" in the signal.

Up to now, no such feature has been implemented however. The computational burden seems quite high for such an algorithm: On the one hand, the frequency bands must not be too broad, otherwise the transitions are not detected; on the other hand, a number of kernel functions, for different transition rates, must be used for each cross-correlation operation.

5.3 Summary and Conclusions

A number of features motivated by Auditory Scene Analysis have been presented that could be useful for sound classification. The actual benefit of a feature may only be evaluated in combination with a pattern classifier, and, depending on the classifier, different feature sets may be optimal. Furthermore, the representation of the features in this chapter only showed the mean feature values per sound, and the temporal behavior of the features within a sound could not be observed. Temporal information within the features could however also be valuable for classification, if a pattern classifier is used that accounts for this (for example a hidden Markov model, see next chapter). Thus, if a promising feature set is presented here, it is only a first guess that has to be further be verified in chapter 7.

The table below summarizes the features that have been presented in this chapter. The letters in the last column have the following meaning:

R: The feature is implemented and ready to be tested with a pattern classifier.
F: Further developments or improvements must be made before the feature can be used.

Auditory Scene Analysis Feature	**Feature Name**	**Description**	**Status**
Spectral profile	*CGAV*	Spectral center of gravity	**R**
	CGFS	Temporal fluctuations of *CGAV*	**R**
	SpecRip	Ripple in the spectrum	F
Harmonicity / Pitch	*Tonality*	Relation of tonal / non-tonal components	**R**
	Pitchvar	Variance of pitch	**R**
	Deltapitch	Difference between consecutive pitch values	**R**
Spatial separation	(for example)		
	DirSpe	Direction of speech	F
	DirNoi	Direction of noise	F
Amplitude Onsets	*Onsetm*	Mean of onsets	**R**
	Onsetv	Variance of onsets	**R**
	Onsetc	Common onsets across frequency bands	**R**
	Onseth	Occurrence of a high number of common onsets	**R**
Rhythm / Beat	*Beat*	Beat strength	**R**
Amplitude Modulation	*Width*	Amplitude histogram width	**R**
	Symmetry	Amplitude histogram symmetry	**R**
	Skewness	Amplitude histogram skewness	**R**
	Kurtosis	Amplitude histogram kurtosis	**R**
	Lower Half	Shape of lower half of the amplitude histogram	**R**
	MLFS	Level fluctuations	**R**
	M1	Amplitude Modulations 0..4 Hz	**R**
	M2	Amplitude Modulations 4..16 Hz	**R**
	M3	Amplitude Modulations 16..64 Hz	**R**
Frequency Modulation	*Turbulence*	Turbulence of frequency modulations	F

Based on the information gained from the bar graphs showing the mean feature values per sound, a first choice of features can be made that will be employed for the following tasks:

- Clean speech shall be identified by one or more of the amplitude modulation features (for example *Width*, *MLFS*, or *M1*).
- By investigating the harmonicity features (*Tonality*, and maybe *Pitchvar*, *Deltapitch*), music (and speech) can be distinguished from noise.
- Details about the frequency content are revealed by the features *CGAV* and *CGFS*. This serves especially to identify low and high frequency signals, which are mostly noises.
- Rhythmic signals can be separated from non-rhythmic signals using the feature *Beat*.

The examinations done in chapter 7 will show if these assumptions are confirmed. It will be seen on the one hand which features are redundant, and on the other hand, what kind of information is missing for accurate classification into the desired classes. Based on this, it will be proposed what kind of features should be implemented in addition to the existing ones in the future.

First however, the pattern classifiers that will be used for these investigations are discussed in the next chapter.

6 Classification Systems II: Pattern Classifiers for Sound Classification

6.1 Introduction

In the previous chapter, a number of features have been presented that are motivated by Auditory Scene Analysis. A first choice of a promising feature set has been made there. However, a more detailed evaluation of the best features can only be made in combination with a pattern classifier, which is again shown in Figure 6.1.

Figure 6.1: After the extraction of features, a decision has to be taken about the class membership using a pattern classifier.

After the extraction of feature vectors out of the signal, a decision is taken about the class that the signal belongs to. This process is performed in the classifier block. The principle of pattern classification is a mapping from the feature space to a decision space, as shown in Figure 6.2. For every point in the feature space – two dimensions in the example – a corresponding class is defined by mapping the feature space V to the decision space Ω. The borders between the classes (the dashed lines in the feature space) are found by performing some sort of training. In our case, 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 is a huge number of approaches for pattern classifiers, many of which require quite a lot of computing power and/or memory. For the application in hearing instruments, the general rule is "keep it simple", in order to keep the need for computing time and memory low.

An important issue is how to account for the temporal information that may lie in the features. Generally, there are two possibilities to do this: The first is to find additional features that describe the temporal statistics of the other features. In this case, the classifier can be static, that is, the temporal information is already analyzed in the feature extracting block. The second possibility is to use a classifier that explicitly models the temporal statistics in the features. Examples are certain kinds of neural networks, where the inputs are current as well as previous feature values, or where the network output is fed back to the input together with

the following feature vector (see for example Soltau et al., 1998). Another possibility is to use hidden Markov models, which try to model the different states that a feature can attain (see below). It has to be checked for each application which is the better approach; if the first way is chosen, it is not straight off obvious if all relevant temporal information is described with the features.

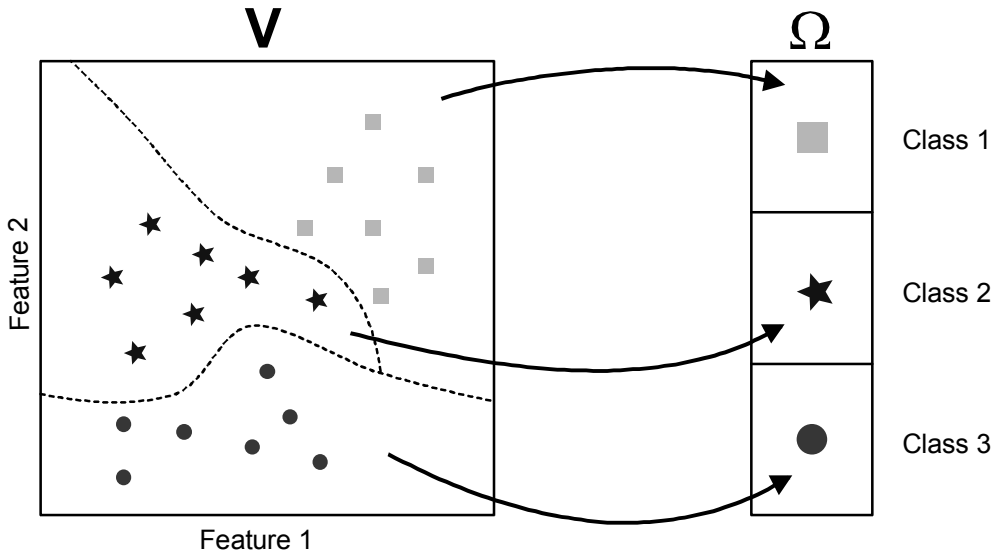


Figure 6.2: Pattern classification viewed as establishing a mapping from feature space V to decision space Ω, here for a two-dimensional feature space.

In the following, a number of pattern classifiers are briefly reviewed that will be evaluated for sound classification in chapter 7. These approaches are all *supervised*, that is, the training data is labeled to indicate its class membership. If unlabeled pattern vectors are used for training, procedures are said to be *unsupervised*: There will be seen what can be done when all that is at hand is a collection of samples without any class membership. A typical example for an unsupervised procedure is a cluster analysis.

It is first discussed how the features have to be normalized to achieve good results, followed by the description of the Bayes rule, which forms the basis of all classifiers. Then, the classifiers will be presented, which includes a rule-based classifier, minimum distance classifiers, the Bayes classifier, the multilayer perceptron (an example of a neural network), and finally hidden Markov models. The hidden Markov model is chosen to check if the temporal information in the features improves classification, as the other approaches do not account for this information.

More detailed information on this topic can be found for example in Duda and Hart (1973), Rabiner and Juang (1993), Bishop (1995), Schürmann (1996), or Kil and Shin (1996).

6.2 Preprocessing of the Feature Vectors

In order to remove biases associated with differently scaled feature values and to preserve good numerical behavior, the features should be normalized so that the mean and standard deviation of each feature are equal to zero and one, respectively. This is performed with the so called z-normalization:

$$x_i' = \frac{x_i - \bar{x}}{s_x} \tag{5.1}$$

where x_i is the original and x'_i the normalized feature value, $\bar{x}$ the mean of the feature and s_x its standard deviation. Only two values for each feature have to be stored for the normalization; they are determined from the training data.

6.3 Bayes Decision Theory

Generally, each pattern classifier is based on the Bayes rule:

$$P(\omega_j|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_j)P(\omega_j)}{p(\mathbf{x})} \qquad \omega \in \Omega, \mathbf{x} \in \mathrm{V} \tag{5.2}$$

The a posteriori probability $P(\omega_j \mid \mathbf{x})$ is the probability that the observation vector **x** belongs to the class ω_j. The classification is done according to this value. $p(\mathbf{x} \mid \omega_j)$ is the distribution of the observations of the class ω_j, $P(\omega_j)$ the a priori probability of a class independently of the observations, that is, the probability that the class occurs generally, and $p(\mathbf{x})$ is the probability that the observation **x** is made independently of the class.

In other words: If we know the a priori probability of each class and the likelihood of class ω_j with respect to **x**, we can compute the a posteriori probability to achieve the minimum probability of error. The most probable class is then determined by the so called maximum a posteriori (MAP) decision rule:

$$\omega = \arg\max_j P(\omega_j \mid \mathbf{x}) = \arg\max_j \left(p(\mathbf{x} \mid \omega_j)P(\omega_j)\right) \tag{5.3}$$

The denominator of equation (5.2) can be left away for the MAP rule, as it remains the same for all classes.

The classifiers can be sorted according to the probabilities they model. In one case it is tried to approximate the a posteriori probability $P(\omega_j \mid \mathbf{x})$ directly. The class ω_j is calculated on the basis of the observations **x** using a function that approximates this probability; an example is the perceptron classifier. Another possibility is to model the class-specific distribution $p(\mathbf{x} \mid \omega_j)$ and then apply equation (5.3). Here, it is distinguished between parametric and non-parametric approaches. In the former, the distribution is described by parameters such as its mean and variance (for example minimum distance classifier or hidden Markov model), in the latter, it is described on a histogram basis (for example Bayes classifier).

6.4 A Selection of Classifiers for Sound Classification

6.4.1 Heuristic Rule-Based Classifier

A straightforward approach is to define boundaries for every feature itself, that is, some rules are settled based on the training data and on the a priori knowledge. This can be regarded as a form of a non-parametric approach. In the example in Figure 6.3, signals with a high amplitude modulation feature *M1* are said to be speech. Additionally, the *Tonality* feature is checked; the harmonic signals are music, the others noise.

In other words, the discriminant functions consist of lines orthogonal to the corresponding feature axis, as shown in the example. For many cases, these straight lines will certainly not be the optimal boundaries. On the other hand, specific exceptions in an otherwise normal distribution can be well handled, as for example the harmonic noises in the class noise. Another feature, such as the variance of the pitch, can be checked to deal with this exception, that is, to distinguish harmonic noise from music.

For large numbers of features and classes however, the rule-based approach can become quite complex and uneasy to handle, because it is difficult to consider more than three dimensions at a time.

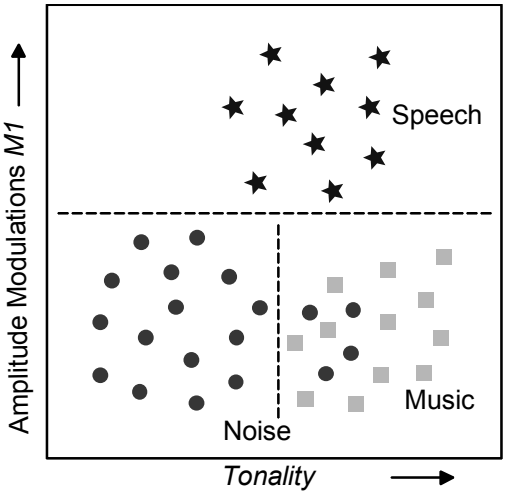


Figure 6.3: Rule-based classification of speech, music and noise with two features. The boundaries (dashed lines) are lines orthogonal to the feature axes. For the identification of the harmonic noises, a third feature is required.

6.4.2 Minimum Distance Classifier

The idea of all minimum distance classifiers is to measure the distance of an observation to some form of representation of each class and to choose the class with the shortest distance. The simplest way is to measure the Euclidean distance

$$d_j(\mathbf{x}) = \left|\mathbf{x} - \mu_j\right|^2 \tag{5.4}$$

where $\mathbf{x}$ is the observation vector and μ_j is the mean vector of class ω_j. Figure 6.4 shows this for two dimensions. In this example, the distance to class 3 is the shortest.

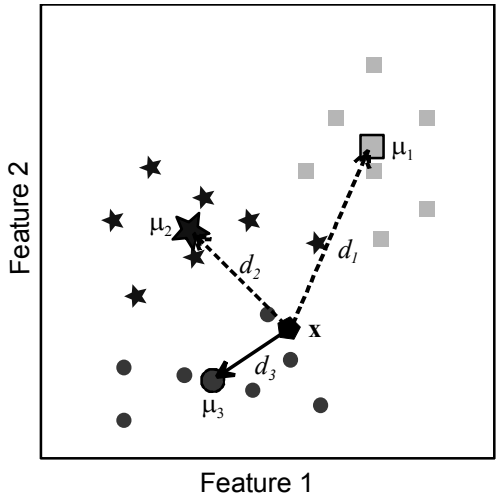


Figure 6.4: Euclidean minimum distance classification in two dimensions. The distance d_3 from the observation $\mathbf{x}$ to the mean μ_3 of the class 3 is the shortest. However, because the distribution of class 3 is not spherical, the variance of the features should be considered using the Mahalanobis distance.

Actually, equation 5.4 is based on equation 5.2 with the premise that the class probability $P(\omega_j)$ is the same for all classes. By taking the Euclidean distance, it has been assumed that the features are statistically independent and have the same variance, that is, the covariance matrix has a diagonal form. The optimum performance of this classifier is reached when the distribution of each class about its mean is in the form of a spherical "hypercloud" in the measurement space, and when the distance between means is large compared to the spread of each class with respect to its mean (that is, the variance is low).

If the distribution of the classes is not spherical (for example class 3 in Figure 6.4), separating them in an Euclidean way is not reasonable. We should look for a way to consider also the variances of the features. The difference between the variances of the features will then deform the spherical form. This is performed by standardizing the distance between a sample and the mean of a class with the variances and covariances of the features. The resulting distance is called Mahalanobis distance; points with the same Mahalanobis distance to the mean of the class form then hyperellipsoids (ellipses in the two-dimensional case). The Mahalanobis distance is calculated by

$$d_j(\mathbf{x}) = \left(\mathbf{x} - \mu_j\right)^T \boldsymbol{K}_j^{-1} \left(\mathbf{x} - \mu_j\right) \tag{5.5}$$

where $\boldsymbol{K}_j$ is the covariance matrix of class ω_j, which is no longer diagonal. This classifier works best if the covariance matrices for all of the classes are identical. Otherwise, a pooled covariance matrix $\boldsymbol{K}_{\text{pool}}$ may be used instead of the $\boldsymbol{K}_j$, which is a weighted mean of all J covariance matrices $\boldsymbol{K}_j$. The weighting is done according to the observation index n_j and the total number of observations N, to ensure that the number of observations per class is not relevant:

$$\boldsymbol{K}_{\text{pool}} = \frac{1}{N} \sum_{j=1}^{J} n_j \boldsymbol{K}_j \tag{5.6}$$

The mean vectors μ_j and the covariance matrices $\boldsymbol{K}_j$ are not known and have to be estimated from the training data. Two possible methods for the estimation of these parameters are maximum likelihood and Bayesian interference. Bishop (1995) shows that these methods lead to the same values for a sufficiently large training set. The maximum likelihood estimation is:

$$\hat{\mu}_j = \frac{1}{N_j} \sum_{\mathbf{x} \in \omega_j} \mathbf{x}(n) \tag{5.7}$$

and

$$\hat{\boldsymbol{K}}_j = \frac{1}{N_j} \sum_{\mathbf{x} \in \omega_j} \left(\mathbf{x}(n) - \hat{\mu}_j\right)\left(\mathbf{x}(n) - \hat{\mu}_j\right)^T \tag{5.8}$$

where N_j is the number of pattern vectors from class ω_j and $\mathbf{x}(n)$ is the nth observation vector.

6.4.3 Bayes Classifier

The expression "Bayes classifier" mostly means that the classification is done with the help of histograms of the class-specific probabilities: The class-specific distribution $p(\mathbf{x} \mid \omega_j)$ is approximated with multidimensional histograms. For this purpose, each dimension in the F-

dimensional feature space is divided into M equidistant intervals, which divides the space itself into equal partitions. The histograms modeling the probabilities $p(\omega_j)$ are then calculated for each partition and each of the C classes from the occurrence of the training data. This corresponds to the Parzen window modeling with the special case of hypercubic volumina (see Duda and Hart, 1973; or Bishop, 1995, who uses the expression "kernel function"). Once the histograms have been constructed, the data can be discarded and only the information on the sizes and locations of the histogram bins need be retained. For classification, the most probable class is determined using the Bayes and the MAP rules (equations 5.2 and 5.3).

The number of probabilities is $C \cdot M^F$, which implies that the number of intervals and features should be kept low. Furthermore, too many intervals lead to bad generalization (the classical overfitting problem), and with too few intervals, the approximation of the distribution is not precise, which leads to bad classification. Apart from that, the number of intervals is limited by the amount of the training data (there should not be any empty intervals).

To decrease the need for memory with regard to the application in hearing instruments, a simplification can be accomplished. A joint probability can be calculated for the case of independent features:

$$p(\mathbf{x} \mid \omega_j) = p(x_1 \mid \omega_j) \cdot p(x_2 \mid \omega_j) \cdots p(x_F \mid \omega_j) \tag{5.9}$$

This means that it is sufficient to compute the one-dimensional histogram for each feature and class separately, which reduces the number of probabilities to $K \cdot M \cdot F$. However, the features are normally not independent, which leads to a loss of information, possibly decreasing the classification performance.

6.4.4 Multilayer Perceptron

Neural networks, and especially the multilayer perceptron (MLP), are universal approximators, that is, they allow to approximate any (discriminant) function to arbitrary accuracy. By training the classifier with labeled data, a function is determined that describes the a posteriori probability $P(\omega_j \mid \mathbf{x})$ for each class. The most probable class is then again obtained by the MAP rule (equation 5.3).

The MLP belongs to the group of feed-forward networks, which have two significant advantages: There exist good training algorithms and the computational burden in the classifying stage is moderate and deterministic (as opposed to the iterative determination of the solution for feedback networks).

The MLP used in this thesis is a two layer[6] perceptron as shown in Figure 6.5. The number of neurons in the input layer is the dimensionality F of the input feature vectors. The second layer consists of neurons with non-linear activation functions. It contains neither net inputs nor net outputs, and is therefore called hidden layer. The number of hidden neurons is not given and has to be adapted to the application; a good start is to have it in the range of the input and output nodes. The output layer consists of neurons with linear activation functions; there is one output neuron for each class.

[6] Sometimes the input layer is also counted; in this case, it would be a three layer perceptron.

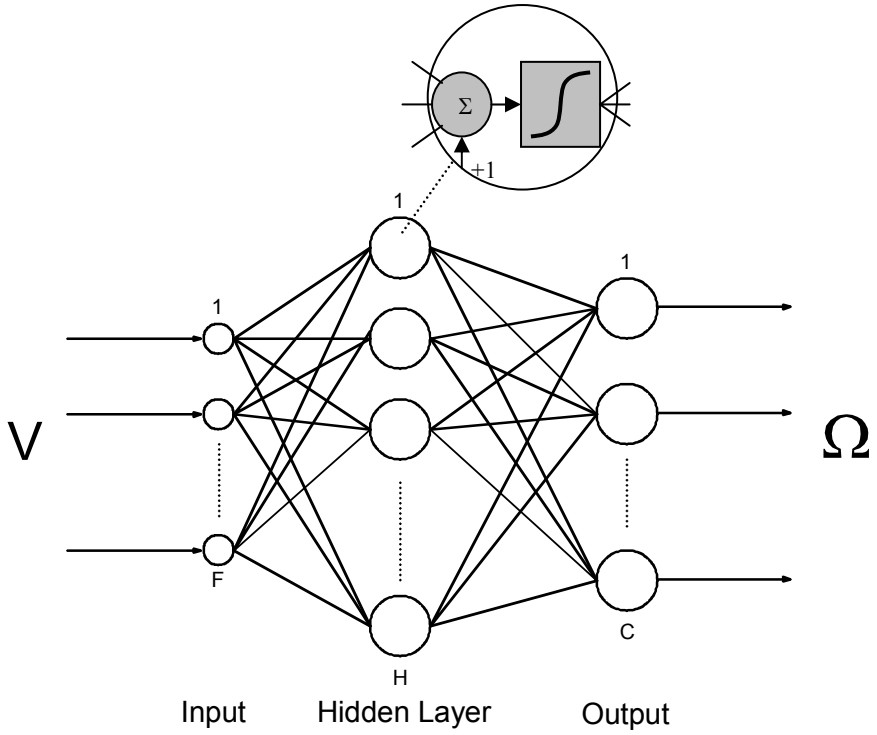


Figure 6.5: Structure of a two layer perceptron. The input layer consists just of distributing nodes; the hidden layer contains the non-linear neurons, whose outputs are combined again in the output layer.

The discriminant function for this two layer perceptron can be formulated as

$$\mathbf{d}(\mathbf{x}) = \mathbf{W}^{<2>}\sigma(\mathbf{W}^{<1>}\mathbf{x}) \tag{5.10}$$

The neural network computes all functions in parallel, which are combined in the vector $\mathbf{d}(\mathbf{x})$. The matrices $\mathbf{W}^{<1,2>}$ correspond to the weights of the hidden and the output layer, respectively. The activation function σ in the hidden layer can be a linear, step or sigmoid function. The latter is state of the art, as it is differentiable, which is required in the development of the training rule.

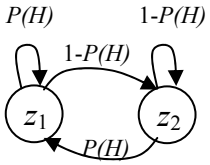
The goal of network training is not to memorize the data but to model the underlying generator of the data. The main problem in training a multilayer network lies in adjusting the weights in the hidden layer, because it is not known, what output of their nodes is best for classification. So, training has to start at the output layer, where the desired output of each node is known. This approach is called training by back propagation (see for example Gonzalez & Woods, 1993 or Bishop, 1995).

6.4.5 Hidden Markov Models

Hidden Markov models (HMMs) are a widely used statistical method of characterizing the spectral properties of the frames of a pattern; they are a powerful tool in speech recognition (see for example Rabiner and Juang, 1986/1993). Recently, a number of applications concerned also the recognition of some distinct noises, for example alarm signals, such as car horns and door bells (Oberle, 1999), or passing vehicles, such as cars, trucks and mopeds (Couvreur et al., 1998). One major advantage of HMMs is that they account for the temporal statistics of the occurrence of different states in the features.

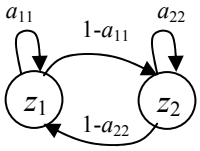
The idea of a HMM is to try to describe a number of observations (discrete or continuous time series) as a parametric random process. A model with a number of states is built, and based on the observations (the training data), the probability distributions in the states and the transition probabilities between the states are estimated. It is however not known a priori how many states the system has in reality (this is *hidden*), that is, the observations do not allow direct conclusions about the number of states. It is tried to choose a reasonable model according to the known properties of the signal and to common sense. The principle is explained with an example:

A number of observations of a tossed coin experiment is given: **O**=*(HHTTTHTHTT ...T)*, where **O** is the observation series, and *H* stands for heads and *T* for tails. The simplest explanation is that a single coin is tossed. The corresponding model has two states, z_1 = head and z_2 = tail:



Heads Tails

In this case, the model is observable, that is, not hidden. It might however be that more than one coin is involved, and somebody randomly chooses one of the coins each time (which we cannot see!). In the case of two coins, the states would be z_1 = coin 1 and z_2 = coin 2, and each state would be characterized by the probabilities of head and tail, $P_{1,2}(H)$ and $P_{1,2}(T)$ for coin 1 or 2, respectively:



$P(H) = P_1$ $P(H) = P_2$

$P(T) = 1\text{-}P_1$ $P(T) = 1\text{-}P_2$

The statistics of choosing coin 1 or coin 2 would then lie in the transition probabilities a_{ij}.

Given the observation sequence, there could also be three or more coins, which would result in a HMM of three or more states. The question is, which model best matches the actual observations. The more complex the model gets, the more parameters are unknown (the one-coin model only has one unknown parameter, the two-coin model four, and a three-coin model already nine). A fundamental question is whether the observation sequence is long and rich enough to be able to specify a complex model. If this is not the case, the model cannot be trained properly; in other words, the system is underspecified. This shows that it is crucial to choose a model as simple as possible, but as complex as necessary. In practice, a number of different models is implemented and compared.

Once the topology of the model has been chosen, the HMM parameter set λ has to be fixed:

$$\lambda = \{\mathbf{A}, \mathbf{B}, \pi\} \tag{5.11}$$

with

$\mathbf{A} = \{a_{ij}\}$ the state-transition probability distribution from state i to j

$\mathbf{B} = \{b_j(k)\}$ the observation probability distribution for state j and feature k

$\pi = \{\pi_i\}$ the initial state distribution of state i

In the training phase, the model parameters $\lambda = \{\mathbf{A}, \mathbf{B}, \pi\}$ are adjusted with an observation sequence $\mathbf{O}_{train}$ so that $P(\mathbf{O}_{train}|\lambda)$ is maximized; this is the so called HMM problem 3. The training is normally performed with the so called Baum-Welch algorithm. Then, an optimal state sequence **q** is searched for the model and a given observation $\mathbf{O}_{train}$. This means that the signal is segmented into states. The determination of the segment boundaries and the assignment of the model states to the segments is called the HMM problem 2. It is performed with the so called Viterbi algorithm.

In the testing phase, the probability $P(\mathbf{O}_{test}|\lambda)$ of the observation sequence $\mathbf{O}_{test}$ is computed for each HMM parameter set λ (one for each class), and the class corresponding to the HMM with the highest probability is chosen using the MAP rule (equation 5.3). The probability is calculated using the so called Forward-Backward algorithm (HMM problem 1).

For a detailed description of these algorithms, see for example Rabiner and Juang (1993), or Oberle (1999).

6.5 Discussion and Conclusions

A number of classifiers has been presented that may be used for hearing instrument applications because of the moderate need of computation power and memory. The requirement for computing time and memory are listed in the table below. If only little time and memory are available, the rule-based or minimum distance classifier can be a good choice. The need for memory in the Bayes classifier is drastically reduced especially for the case of many features if only one dimensional histograms are calculated. The number of hidden neurons is significant for the complexity of the MLP. For the HMM, the number of classes is crucial, because one model is needed per class.

Classifier	**Computing time**	**Memory**
Rule-Based	$O(CF)$	about CF
Minimum Distance, Euclidean	$O(CF)$	$C + CF$
Minimum Distance, Mahalanobis	$O(CF)$	$2CF$
Bayes, F-dim. Histogram	$O(1)$	I^F
Bayes, 1-dim. Histogram	$O(C(F-1))$	CIF
Multilayer Perceptron	$O(H(C+F))$ + nonlin.	$H(C+F)$
Hidden Markov Model	$O(CFT)$	$C(2FS+T)$

Meaning of the symbols	
F	Number of features
C	Number of classes
I	Number of intervals
H	Number of hidden neurons
S	Number of states
T	Number of transitions

One of the approaches, the HMM, can also account for the temporal statistics in the features; for the other approaches, some additional features have to be found that describe the temporal statistics of the given features if desired, and it has to be determined if such statistics are relevant at all for sound classification.

Instead of using a single, one-stage classifier, it is conceivable to combine two or more of the above approaches into a multi-stage classifier where different details of classification are extracted at different stages. A HMM could for example be used for the identification of coarse classes, and a rule-based stage could follow to further divide the coarse classes into subclasses.

The investigations in the next chapter will show which classifier gives the best classification, and which features are needed for this. In a second step, a simple multi-stage approach will be evaluated based on these results.

7 Evaluation of Different Classification Systems

7.1 Introduction

A number of features that could be used for sound classification have been presented in chapter 5. This was followed by an overview of pattern classifiers that shall make a classification based on the feature values that are extracted out of the signal, which is in this case the acoustic environment. In this chapter, it will be evaluated which feature set performs best with which classifier, for the classification of the four sound classes that have been considered as being most important in hearing instruments, that is 'speech', 'speech in noise', 'noise', and 'music'. The class 'silence' is left away, because its detection is easy and it would not make much sense to introduce test sounds that consist of quiet only. It will, however, be an important situation in a field trial.

The chapter starts with an overview of the system and the parameters that have to be set. It will be discussed how the soundset and the feature sets are chosen, and in which form the scores will be presented. After that, the results are presented and discussed for each of the classifiers, which includes a rule-based classifier, minimum-distance and Bayes classifier, a neural network and a hidden Markov model. Then the classifier performing best will be used in a simple multistage approach: The first stage will be the best classifier together with the best feature set, and the second stage a rule-based classifier together with a second feature set. The results will show that a hidden Markov model combined with a rule-based approach performs best, with an overall hit rate of over 90 %. The chapter concludes with a discussion about the different classifiers and feature sets, and about the limitations of the sound classification system.

7.2 Procedure

7.2.1 Overview

If we look again at the block diagram of the classification system in Figure 7.1, we see that a number of parameters can be set in each block:

- *Training and test set*: A soundset has to be defined for evaluation, and it has to be determined which part of this sound data is used for training, and which for testing.

- *Feature set*: A number of feature sets have to be selected for classification, to find the set that yields optimal performance.
- *Classifier parameters*: Depending on the classifier, a number of parameters can be set, for example the number of histogram intervals in a Bayes classifier, or the number of hidden nodes in a neural network, but also the kind of pre-processing that is applied to the feature vectors.
- *Smoothing constant*: In the post processing step, the transient behavior of the classification result is controlled, that is, it is determined how fast or sluggishly the system shall switch between classes.
- *Number of training and test cycles*: In the evaluation phase, the classification score is established only after a specific feature set has been processed with a number of different training and test sets.

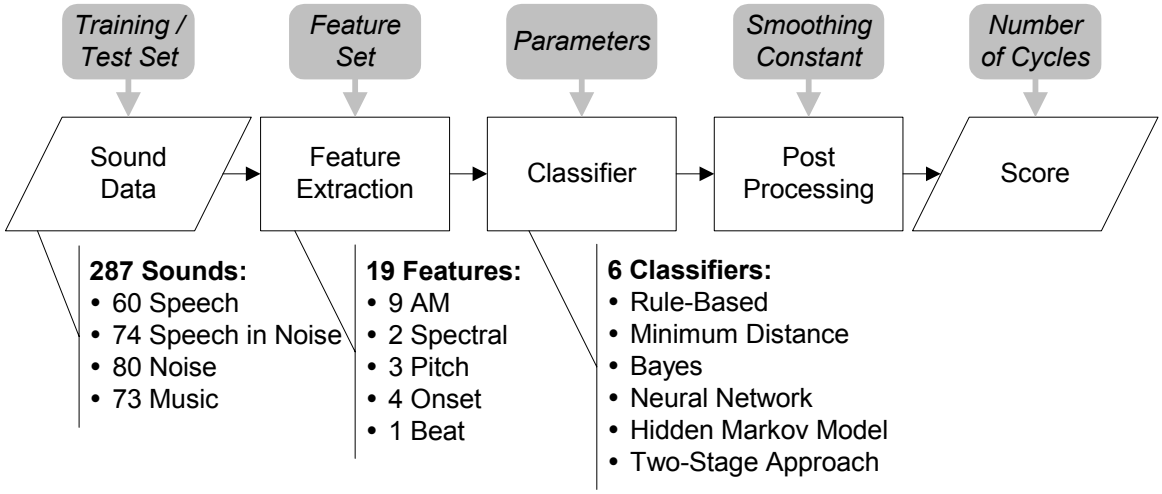


Figure 7.1: Block diagram of the classification system and the parameters that have to be set for evaluation.

The setting of these parameters will now be discussed for each of the blocks.

7.2.2 Sound Data

The soundset used for the evaluations contains 287 different sounds; it is described in detail in appendix A. Each sound belongs to one of the four classes and has a length of 30 seconds. About 5/6 of the sounds (83 %) is used for the training of the classifier, and 1/6 (17 %) for testing. Note that the trained classifier is not only tested with the test data, but also with the training data; the reason for this is explained in section 7.2.6 about the scores. The arrangement of the sounds for four classes is shown in the table below.

	Number of Sounds		
Class	**Training**	**Testing**	**Total**
'Speech'	50	10	60
'Speech in Noise'	62	12	74
'Noise'	67	13	80
'Music'	61	12	73
Total	**240**	**47**	**287**

If the soundset is divided into training and test sounds, one has to carefully select sounds within a class in order to ensure reliable results. If, for example, only speech without reverberation is used for training, then the reverberated speech sounds will probably all be misclassified in the testing phase, which will result in a very bad hit rate. In other words, both the training and the test set must be chosen in a way that they cover the whole range of each sound class homogeneously. However, it is not obvious how this choice shall be made. Therefore, the sounds for the training and test set are chosen at random, and this random choice is repeated several times. The "true" score is then the mean of the scores of these cycles.

The number of cycles has been set to 100. Preliminary experiments have shown that this leads to scores that differ not more than 3 % of the achieved scores of another series of 100 cycles. If a more accurate score is desired, the number of cycles has to be increased.

7.2.3 Feature Sets

As has been shown in chapter 5, some 20 features are extracted from the sound data. If all combinations of features should be evaluated, it would result in about 2^{20} different feature sets, which are of course too many to consider. Thus, an iterative strategy has been developed to find the best set. The recipe follows six steps[7]:

Step #	Recipe for Feature Combination
1	The pitch feature *Tonality* is used together with features describing the amplitude modulations (AM), that is *MLFS*, *M1*, *M2*, *M3*, *Width*, and *Skewness*.
2	The best AM set of step 1 is used without the *Tonality*, but together with the other pitch features, *Pitchvar* and *Deltapitch*.
3	The best set of step 2 is enriched with spectral features, *CGAV* and *CGFS*.
4	Onset features are added to the best set of step 3 (*Onsetm*, *Onsetv*, *Onsetc*, *Onseth*).
5	The best set of step 4 is reduced in succession by the AM feature(s) and by the spectral feature(s) of steps 1 and 3.
6	The beat feature *Beat* is added to the best set of step 4 and 5.

In addition to this iterative approach, the classification was performed with all of the above features: *MLFS*, *M1*, *M2*, *M3*, *Width*, *Skewness*, *Tonality*, *Pitchvar*, *Deltapitch*, *CGAV*, *CGFS*, *Onsetm*, *Onsetv*, *Onsetc*, *Onseth*, *Beat*.

This approach results in about 30 feature sets that have to be processed for each classifier in order to find the optimal combination.

7.2.4 Pattern Classifier

A number of classifier specific parameters will be described in the results section. However, it can generally be stated that the feature values are processed in parallel by the trainable classifiers. That is, each feature has the same weight when input to the classifier, and the

[7] This recipe has not been used for the rule-based classifier, whose structure is explicitly defined by the features that are chosen; for details, see the results section below.

classifier itself has to decide which features to consider most important, based on the training data.

The rule-based classifier, from this point of view, is a different approach. The features are investigated sequentially, and the importance of the features is determined by the order in which they are checked. This weighting allows to incorporate a priori information into the structure of the classifier.

7.2.5 Post Processing

It was mentioned in chapter 5, that one value per second is obtained for each feature to be extracted[8]. Thus, the classifier gets a new feature vector every second. The classifier outputs the most probable class for each feature vector, which results in a series of 30 (equal or different) class memberships for each sound of 30 seconds length.

In order to determine a classification score for the sound set, one single class membership has to be determined for each sound of 30 seconds length. The simplest way to do this is to take the class that occurs most frequently in the 30 class memberships of a sound.

Note that this method is chosen for evaluation purposes only; if the classifier is exploited in a hearing instrument, a smoothing constant can be introduced as proposed in Figure 7.1. The determined class can then, for example, be equal to the one that occurs most often in the last 10 seconds of the signal. Alternatively, a more complicated procedure may be chosen to control the transient behavior of the classification system.

7.2.6 Scores

For each feature set and classifier, classification scores will be obtained. After the training of the classifier, it is not only tested with the test set, but the training set is again taken to obtain scores for this set. By comparing the test score with the training score, important information can be gained:

- **Ability to generalize**: If the two scores are in the same order, then the classifier is able to generalize well, because the performance for known data (training set) is equal to that for unknown data (test set).
- **Overfitting**: If the score for the training set is much better than the one for the test set, then the classifier is overfitted to the training data; it behaves well for the 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.

The state of the art for displaying the scores is to generate a confusion matrix that indicates for each class how many sounds are classified correctly (hit rate), and how many are misclassified into other classes. Another way that allows to display the scores in a more compact format is to show the hit rates and false alarm rates for each class. They are defined as follows:

Hit rate: $$HR = N_{corr} / N_{class} \quad (6.1)$$

[8] The *Beat* feature is an exception, because only one value is obtained per 30 seconds. To use it together with the other features, the same value is taken for each of these 30 seconds.

False alarm rate: $FA = N_{wrong} / (N_{total} - N_{class})$ (6.2)

with N_{corr} number of correctly classified sounds in this class
N_{class} number of sounds in this class
N_{wrong} number of sounds of other classes wrongly classified as this class
N_{total} total number of sounds

The hit rate *HR* is the relation of the correctly classified to the total number of sounds in a class. The false alarm rate *FA* indicates how many of the sounds of other classes are wrongly classified as the actual class. Note that the sum of the two rates is not 100 %. The overall hit rate *OH* is then the mean of the hit rates of all classes.

With some 30 feature sets and a number of different classifiers and parameters, too much space would be needed to print all confusion matrices. Thus, only the hit and the false alarm rates of the best three scores per classifier will be shown throughout this chapter. Additionally, the best scores will be displayed graphically in the form of a receiver operating (ROC) graph, where the hit rate of each class is represented in percent on the ordinate, and the corresponding false alarm rates on the abscissa. The confusions of the best set per classifier will be discussed in detail.

Additionally, the difference of the hit rates of the training and the test set ΔH is calculated to provide an indicator for overfitting.

The abbreviations that will be used in the following are listed here:

Symbol	Meaning
HR	Hit Rate [%]
FA	False Alarm Rate [%]
OH	Overall Hit Rate [%]
ΔH	Difference of Training and Test Hit Rates [%]
Spe	Class 'Speech'
Noi	Class 'Noise'
SpN	Class 'Speech in Noise'
Mus	Class 'Music'

7.3 Results

7.3.1 Rule-Based Classifier

The rule-based classifier is a straightforward approach to define boundaries for every feature itself. Some rules are settled based on the training data and on the a priori knowledge. The procedure is to sequentially check feature after feature.

7.3.1.1 Parameters

A number of features have to be selected empirically for this classifier, and it has to be determined in which order the features are checked.

The structure that has been chosen here is shown in Figure 7.2. First, the modulation features *M1*, *M2*, *M3* are checked to determine whether the signal is 'speech' or something else. The feature *Tonality* then tells us if the signal is harmonic or not. A high tonality indicates music or harmonic noise, a medium tonality 'speech in noise', and a low tonality 'noise'. Finally, the features *Pitchvar* and *Deltapitch* shall distinguish between 'music' and harmonic noise.

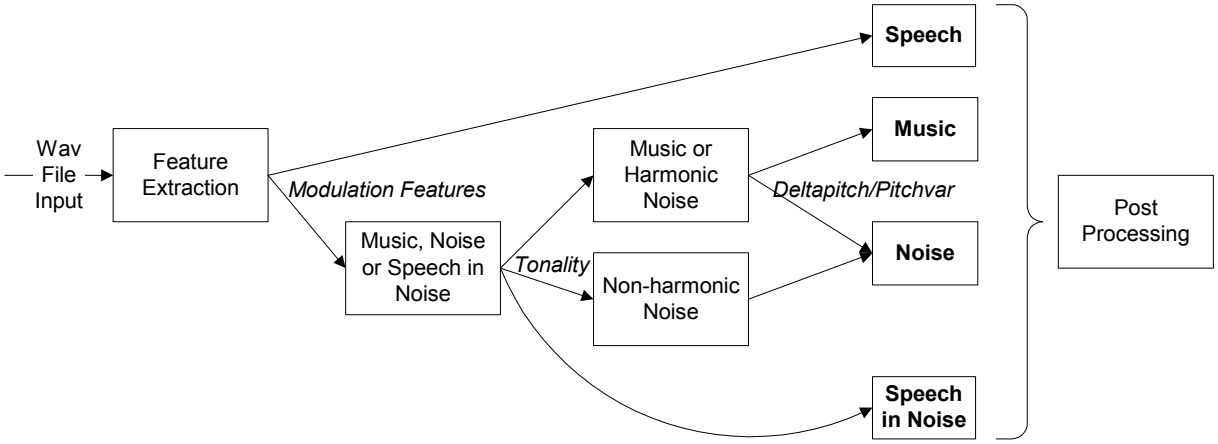


Figure 7.2: Structure of the rule-based classifier using AM and pitch features.

7.3.1.2 Scores

As no training set is needed for this kind of classifier (it is trained heuristically, that is manually), only test set scores are printed in the table below. Figure 7.3 shows the corresponding ROC graph.

Scores for Rule-Based Classifier	**Test**								
	OH	Spe		SpN		Noi		Mus	
		HR	*FA*	*HR*	*FA*	*HR*	*FA*	*HR*	*FA*
Tonality, Pitchvar, Deltapitch, M1, M2, M3	**77.9**	79.3	1.3	66.5	9.9	87.9	11.8	77.5	7.0

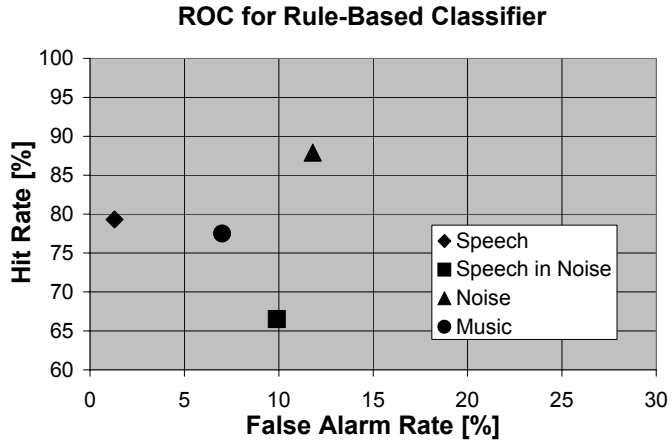


Figure 7.3: ROC for the rule-based classifier. The hit rate of each class is represented in percent on the ordinate, the corresponding false alarm rates are shown on the abscissa. The hit rate of the class 'speech in noise' is quite poor, and many sounds were misclassified as 'noise', resulting in a high false alarm rate.

Comments:

- The achieved hit rate is not really convincing, but only AM and pitch features were needed to get this score.
- Especially for the class 'speech in noise', additional features are required for better identification. For the *Tonality* feature, the values of 'speech in noise' lie quite close to those of 'noise', and it is difficult to set the boundaries. Thus, the false alarm rates of 'noise' and also of 'speech in noise' become quite high.
- Another reason for the high false alarm rate of 'noise' is, that a number of 'music' sounds were regarded as harmonic noise, if only the pitch features *Pitchvar* and *Deltapitch* were checked. On the other hand, 'noise' was never misclassified as 'music' or 'speech', only as 'speech in noise'.
- Most reverberated or compressed speech files were classified as 'music' or 'speech in noise'. The modulation features alone are not sufficient for identifying this sort of speech.
- If more decision rules are inserted into the classifier, it may become quite complex and the danger of unforeseen behavior for certain sounds increases. This can be regarded as some sort of overfitting.

7.3.1.3 Summary for the Rule-Based Classifier

A hit rate of 78 % is achieved with a simple rule-based classifier using 6 features. Many 'speech in noise' sounds, 'music', and reverberated or compressed speech sounds were misclassified. If more decision rules and features are introduced, it gets difficult to handle all possible cases.

7.3.2 Minimum Distance Classifier

The idea of the minimum distance classifier is to measure the distance of an observation to all classes and to choose the class with the shortest distance. Either the Euclidean distance or the Mahalanobis distance can be considered for this. The former considers only the mean of the features, the latter also their variances (see chapter 6.4.2).

7.3.2.1 Parameters

The only parameter is the type of distance, Euclidean or Mahalanobis distance.

7.3.2.2 Scores

The three best sets for both the Euclidean and Mahalanobis classifier are shown in the tables below. The best scores of each are also displayed as ROC graph in Figure 7.4.

Best Feature Sets for Euclidean Minimum-Distance	**Train**	**Test**									**ΔH**
	OH	*OH*	Spe		SpN		Noi		Mus		
			HR	*FA*	*HR*	*FA*	*HR*	*FA*	*HR*	*FA*	
Tonality, Pitchvar, M1, M2, CGFS, Onsetv, Beat	84.1	**83.3**	85.7	3.5	83.1	11.2	80.4	6.7	84.7	1.0	0.8
Tonality, M1, M2, CGFS, Onsetv, Beat	82.9	**81.9**	83.6	2.7	81.6	12.0	76.1	6.6	87.0	3.1	1.0
Tonality, M1, M2, CGFS, Onsetv	81.4	**81.4**	84.3	2.8	79.2	12.3	85.7	7.1	76.6	2.7	0.0

Best Feature Sets for Mahalanobis Minimum-Distance	Train	Test								ΔH
	OH	OH	Spe		SpN		Noi		Mus		
			HR	FA	HR	FA	HR	FA	HR	FA	
Tonality, Pitchvar, M1, CGFS, Onsetc	82.9	**82.9**	86.9	4.0	84.2	13.2	79.2	4.8	82.5	0.8	0.0
Tonality, M1, CGFS, Onsetm, Onsetc	82.8	**82.3**	80.0	3.5	84.4	14.6	78.5	4.0	86.0	1.6	0.5
Tonality, M1, CGFS, Onsetm, Onsetv, Onsetc	82.8	**81.9**	77.6	2.7	85.1	15.2	82.0	4.9	82.0	1.5	0.9

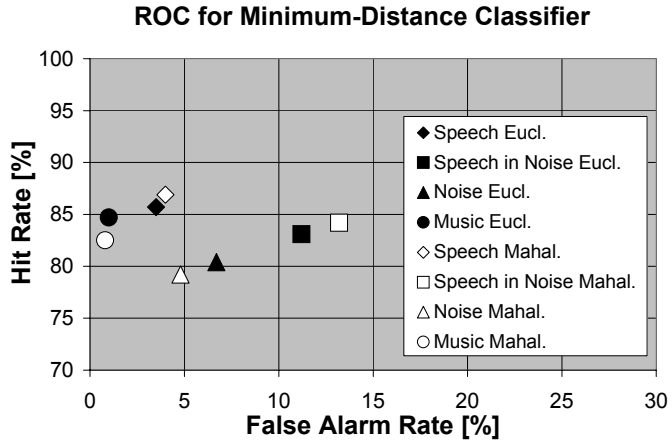


Figure 7.4: ROC for the minimum-distance classifiers. The performance is similar for both approaches. Many sounds were misclassified as 'speech in noise', resulting in a high false alarm rate.

Comments:

- The best set differs for the two classifiers. It contains features which describe the pitch, AM, the spectral form, onsets and for the Euclidean approach also the beat.
- Similar results were achieved with the Euclidean and the Mahalanobis distance.
- There is not much danger of overfitting (the difference ΔH of the hit rates of the training and the test set remains quite small), because it is not possible to divide the feature space in a complex way with this sort of classifier. If too many features are considered, both the training and test hit rates will decrease.
- 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. On the other hand, the files that were misclassified as 'noise' are mostly from the class 'speech in noise'; those misclassified as 'speech' originate from all of the three other classes.
- The beat feature helps to separate rhythmic noises and pop music from the class 'speech in noise'. This worked partly well for the Euclidean classifier; the Mahalanobis score however could not be improved with this feature, and many pop music sounds and rhythmic noises were classified as 'speech in noise'.

7.3.2.3 Summary for the Minimum-Distance Classifier

The best score is 83 % using 7 features. The Euclidean and the Mahalanobis distance gave similar results. 'Speech in noise' and 'noise' could not be well separated.

7.3.3 Bayes Classifier

The Bayes classifier does the classification with the help of histograms of the class-specific probabilities: The class-specific distribution $p(\mathbf{x}|\omega_j)$ is approximated with multidimensional histograms. If the features are assumed to be independent, it is sufficient to compute the one-dimensional histogram for each feature and class separately, which reduces the computational effort drastically. However, the features are normally not independent, which possibly decreases the classification.

7.3.3.1 Parameters

Either multidimensional histograms are calculated (the order of the dimension corresponds to the number of features), or one-dimensional histograms are calculated for each feature and class separately as an approximation. For both approaches, the optimal number of intervals into which the histogram is divided, has to be determined.

7.3.3.2 Scores

Preliminary experiments have shown that not more than 6 intervals can be taken for the multidimensional approach, otherwise the computational effort is too high. However, the scores that were achieved with the simpler one-dimensional approach were always better; thus, only these results are presented in the table below and in Figure 7.5.

Best Feature Sets for Bayes Classifier		**Train**	**Test**								**Δ*H***
Features	# of Inter-vals	*OH*	*OH*	Spe		SpN		Noi		Mus		
				HR	*FA*	*HR*	*FA*	*HR*	*FA*	*HR*	*FA*	
Tonality, Pitchvar, M1, M2, M3, CGFS, Onsetm, Onsetc	15	86.1	**84.3**	89.5	4.3	82.8	10.1	83.9	5.0	81.9	1.5	1.8
Tonality, Pitchvar, Deltapitch, M1, M2, M3, CGFS, Onsetv, Onsetc	20	86.9	**84.3**	90.4	5.5	79.6	8.3	86.4	5.7	81.8	1.4	2.6
Tonality, M1, M2, CGFS, Onsetv	25	84.0	**83.3**	93.4	4.9	82.7	12.2	83.5	4.2	75.4	0.9	0.7

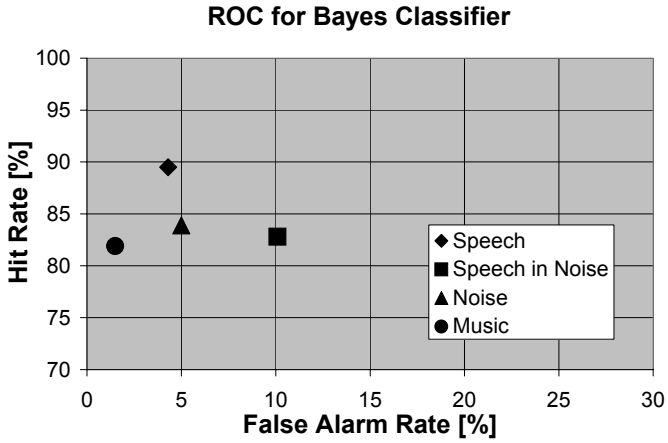


Figure 7.5: ROC for the Bayes classifier. 'Speech' achieves a high hit rate, but 'speech in noise' still has quite a high false alarm rate.

Comments:

- The best set contains pitch, AM, spectral and onset features.
- The more features are taken, the bigger is the danger of overfitting. A tradeoff between a high hit rate and a low ΔH has to be made.
- Overfitting can also occur if many histogram intervals are taken. Figure 7.6 shows how the optimum is reached for about 15 to 20 intervals. With more intervals, the training score remains equal or increases, but the test score may decrease.
- The false alarm rates show that the misclassified sounds were mostly regarded as 'speech in noise'. These sounds include reverberated speech, some cafeteria noises and other fluctuating noises, and pop music. Tonal noises, such as a vacuum cleaner, may be classified as 'music'. The files misclassified as 'noise' are all from the class 'speech in noise'; those misclassified as 'speech' originate from all of the three other classes, but especially from 'music'.
- Using the beat feature decreased the hit rates. This is probably because sounds from any of the classes may contain rhythmic parts, and the Bayes classifier can obviously not handle this.
- The last feature set in the table is a good example to show that the overall hit rate can be quite high, but the hit rate of certain classes (here it is 'music') may be poor. In this example, all pop music sounds were classified as 'speech in noise', which is also why the false alarm rate of 'speech in noise' is so high.

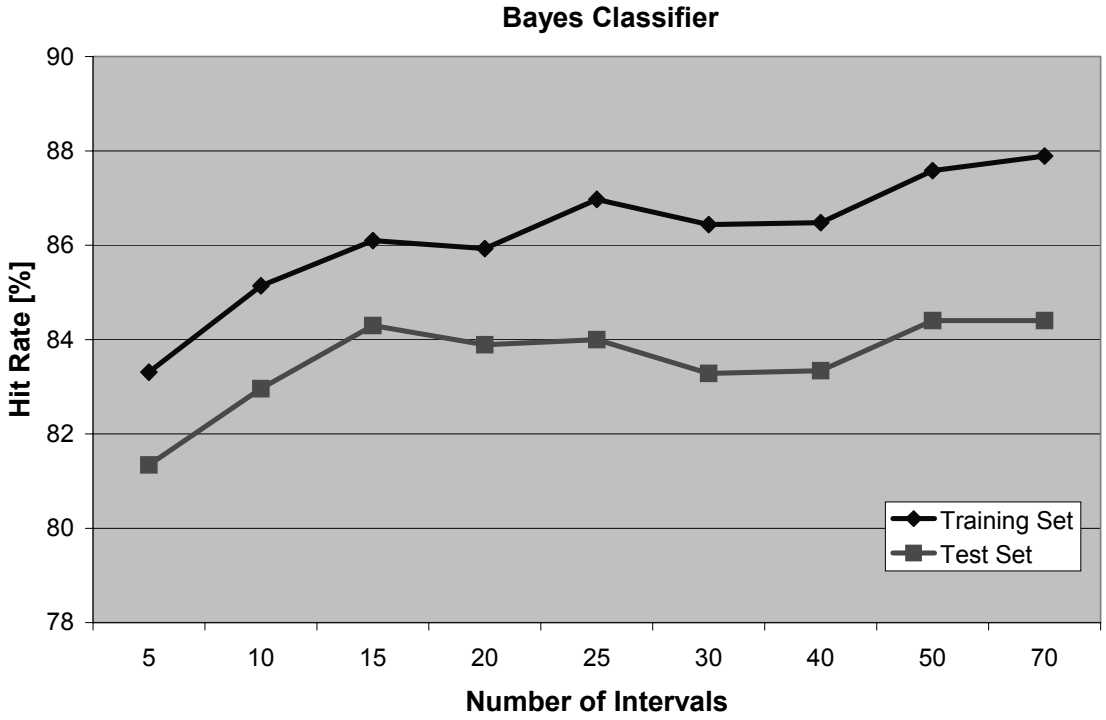


Figure 7.6: Hit rates for different number of intervals and feature set: Tonality, Pitchvar, M1, M2, M3, CGFS, Onsetm, Onsetc. The optimum lies at 15 to 20 intervals.

7.3.3.3 Summary for the Bayes Classifier

The best score is 84 % using 8 features. The approximation with one-dimensional histograms gave better results than with multidimensional ones, and it was even easier to compute. The optimal number of histogram intervals was 15 to 20. Overfitting could occur when many features and intervals were taken for training. Reverberated speech, fluctuating or tonal noises and pop music were often misclassified.

7.3.4 Multilayer Perceptron

The multilayer perceptron – a sort of neural network – allows to approximate any discriminant function to arbitrary accuracy. By training the classifier, a function is determined that describes the a posteriori probability $P(\omega_j \mid \mathbf{x})$ for each class.

7.3.4.1 Parameters

When using neural networks, it is important to normalize the input data so that all features have the same mean and standard deviation (see chapter 6.2), otherwise the weights cannot be calculated correctly. This is performed by applying the z-normalization.

A multilayer perceptron with one hidden layer was taken for all tests. The number of hidden neurons ranged from 2 to 12.

Two activation functions were evaluated: A tangens-sigmoid function and a linear function with saturation.

7.3.4.2 Scores

Preliminary experiments showed that the two activation functions lead to similar results. Thus, only the results for the tangens-sigmoid function are given in the table below and in the ROC graph in Figure 7.7.

Best Feature Sets for Neural Network		**Train**	**Test**								**Δ*H***
Features	# of hidden Nodes	*OH*	*OH*	Spe		SpN		Noi		Mus		
				HR	*FA*	*HR*	*FA*	*HR*	*FA*	*HR*	*FA*	
Tonality, Width, CGAV, CGFS, Onsetc, Beat	8	88.9	**87.1**	86.3	1.7	86.2	7.0	88.8	5.9	87.0	2.8	1.8
Tonality, Width, CGAV, CGFS, Onsetc, Beat	6	88.2	**86.7**	87.9	2.7	81.1	6.1	88.5	6.9	89.2	2.3	1.5
Tonality, Width, CGAV, CGFS, Onsetc	8	86.7	**85.9**	88.0	3.3	83.3	7.4	91.5	6.3	80.7	1.9	0.8

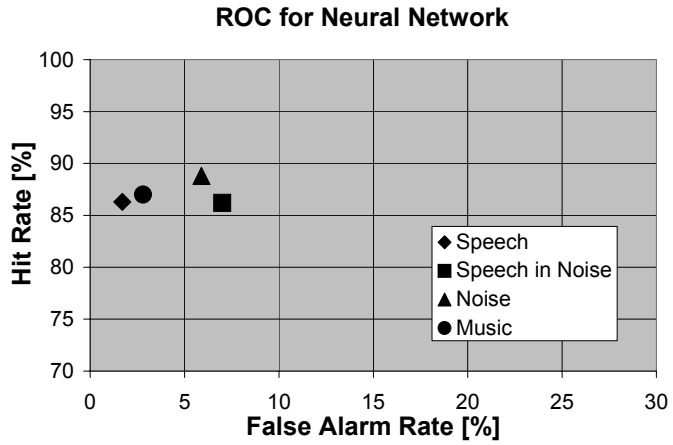


Figure 7.7: ROC for the multilayer perceptron. All hit rates are above 86 %, and the false alarm rate of 'speech in noise' has further improved.

Comments:

- The best feature set consists of pitch, AM, spectral, onset and beat features. If the beat feature is left away, the danger of overfitting is reduced (smaller Δ*H*).
- It has been stated above that the multilayer perceptron allows to shape any discriminant function. During the training, only the information that is relevant for the discrimination of the classes is picked from the features, which means that also features may be valuable that did not look very promising at first glance. Examples are the *CGAV* and *Beat* features, which differ for many sounds independently of the class.
- The optimal number of hidden nodes lies at 6 to 8 nodes, as shown in Figure 7.8. With more hidden nodes, there is the danger of overfitting, and the scores decrease.
- Most of the confusions occurred between the classes 'noise' and 'speech in noise'. This includes fluctuating noises like cafeteria noise, a passing train or a weaving machine, and 'speech in noise' with poor SNR. In addition to that, a few of the reverberated speech files were classified as 'speech in noise', and a few of the pop music sounds as 'noise',

especially if they were compressed (that is, recorded from the radio). Finally, a few tonal noises were regarded as 'music', and two files with singing as 'speech'.

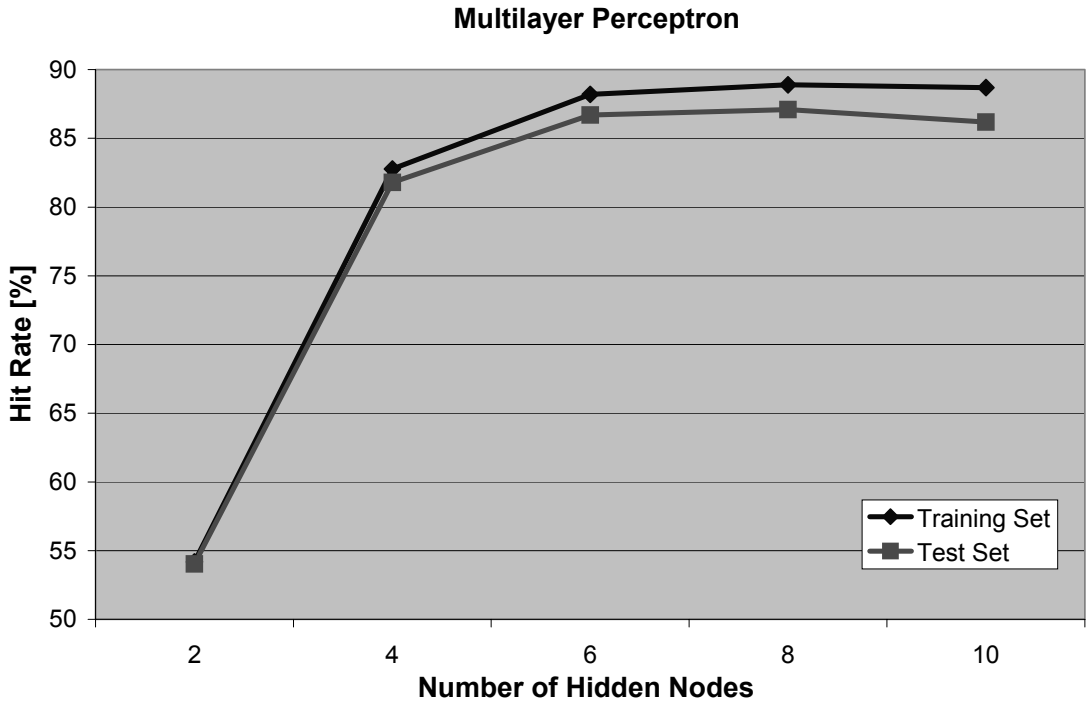


Figure 7.8: Hit rates for different number of hidden nodes and feature set: Tonality, Width, CGAV, CGFS, Onsetc, Beat. The optimum lies around 6 to 8 hidden nodes.

7.3.4.3 Summary for the Multilayer Perceptron

The best score is 87 % with 6 features and 8 hidden nodes. More hidden nodes caused overfitting. The choice of the activation function, sigmoid or step, was not crucial; similar results were achieved with either. Most confusions concerned 'noise' and 'speech in noise', but reverberated speech and pop music sounds could sometimes also be misclassified. However, the false alarm rates are significantly lower than for the other classifiers discussed so far.

7.3.5 Hidden Markov Model

Hidden Markov models (HMMs) are a widely used statistical method especially for speech recognition. One major advantage of HMMs is that they account for the temporal statistics of the occurrence of different states in the features. The idea of a HMM is to try to describe a number of observations as a parametric random process. A model with a number of states is built, and based on the observations (the training data), the probability distributions in the states and the transition probabilities between the states are estimated.

7.3.5.1 Parameters

The topology of the HMM is determined by the number of states and the number of transitions in the model. In speech recognition, left-to-right topologies are often used, to set the order in which the states have to appear. This way, given sequences of syllables or words

are modeled. For sound classification, there are no such sequences: Different states may occur in random order, as for example speech and pause segments in the class 'speech'. Thus, a so called ergodic HMM topology is selected, where each state can be reached from each other state.

The number of states is limited by the available training data. With the soundset used in this thesis, 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. This is very different for example from speech recognition, where each word is modeled by one single HMM.

The number of transitions in an ergodic HMM with two states is four.

7.3.5.2 Scores

The table shows the three best scores for the two-state ergodic HMM, and the best score is again displayed as a ROC graph in Figure 7.9.

Best Feature Sets for Ergodic HMM with 2 States	**Train**	**Test**								**Δ*H***
	OH	*OH*	Spe		SpN		Noi		Mus		
			HR	*FA*	*HR*	*FA*	*HR*	*FA*	*HR*	*FA*	
Tonality, Width, CGAV, CGFS, Onsetc, Onsetm	89.8	**87.6**	92.4	2.2	84.4	7.0	83.9	5.3	90.7	2.2	2.2
Tonality, Pitchvar, Width, CGAV, CGFS, Onsetc, Onsetm	89.5	**87.2**	92.4	2.7	74.3	3.5	94.3	10.0	88.2	1.0	2.3
Tonality, Pitchvar, Width, CGFS, Onsetc, Onsetm	87.8	**86.8**	93.0	2.2	75.2	3.9	92.9	10.2	86.7	1.6	1.0

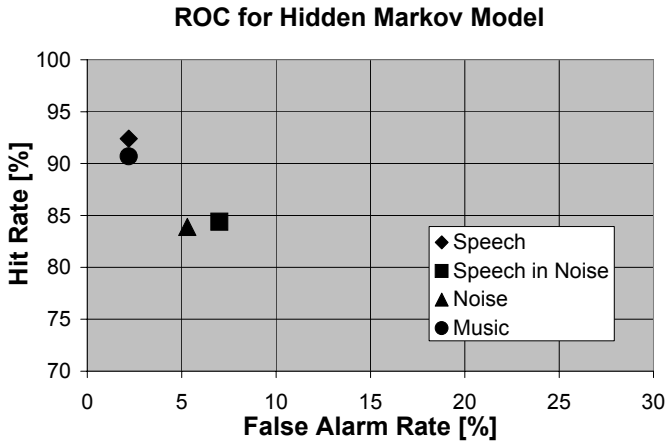


Figure 7.9: ROC for the HMM. 'Speech' and 'music' perform excellently, but a number of confusions occurred between 'speech in noise' and 'noise'.

Comments:

- The best feature set consists of pitch, AM, spectral and onset features. The beat feature only lead to overfitting ($\Delta H > 5$ %).
- There is obviously not very much temporal information in the features that could be exploited for sound classification (the hit rates are only slightly better compared to the

neural network). Maybe not all temporal information that lies in the *signal* is described by the features that are used.

- The more features are used, the more overfitting occurs. Again, a trade-off has to be made between the test hit rate and the ΔH.
- Most confusions occurred in the classes 'speech in noise' and 'noise'. Fluctuating noises were often regarded as 'speech in noise', and 'speech in noise' with poor SNR as 'noise'. Reverberated and compressed speech (from the radio) was mostly classified as 'speech in noise'. The same could happen to compressed pop music. In addition to that, few fluctuating noises and a drums sound were misclassified as 'speech', and a few tonal noises and a child speaking as 'music'.
- HMMs are commonly used for speech recognition, where cepstral coefficients are used as features. To compare the scores, the soundset was also classified using these features. However, cepstral coefficients gave only poor results; the overall hit rates were below 70 %.

7.3.5.3 Summary for the HMM

The best score is 88 % using 6 features and a two state ergodic HMM. Training was not possible with more states. Using state-of-the-art cepstral coefficients only achieved below 70 %. Most confusions concerned 'noise' and 'speech in noise', but compressed speech and pop music sounds could sometimes also be misclassified.

7.3.6 A Simple Multistage Strategy

7.3.6.1 Concept

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. If, for example, the spectral center of gravity *CGAV* of the signal is at high frequencies, then the experience tells us that this is mostly 'noise'. If the signal has been classified differently, it can be assumed that some other features led to a misclassification, which should be corrected considering the clear message of the *CGAV*. The correction of the classifier output should however be made conservatively, to prevent that a good solution achieved by a classifier is worsened.

In the following, the results of a first simple multistage approach are presented. The HMM classifier is used as first stage together with a first feature set, as shown in Figure 7.10. The output of the HMM is verified with a rule-based classifier and a second feature set. This second stage could also be regarded as a special form of post processing.

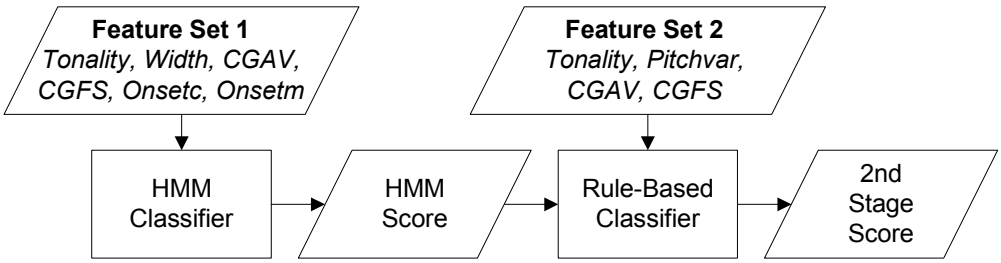


Figure 7.10: Simple multistage approach using a HMM as first stage and a rule-based classifier as second stage to verify the classification performed by the HMM.

The rule-based classifier contains only a few simple heuristic decisions that are based on four features. Different decisions are made depending on the class that is recognized in the first stage.

The *Tonality* feature is used in each class to adjust the class if the feature values are completely in the wrong range for the class determined by the HMM. The tonality is expected to be high for 'music', medium for 'speech', a bit lower for 'speech in noise' and low for 'noise'. If, for example, a sound is classified as 'speech' by the HMM, then it is expected that some features indicated to the HMM that it is a strongly fluctuating signal. If, on the other hand, the tonality is very low for that signal, it will probably not be 'speech', but 'speech in noise'.

Similar considerations can be made for the other three features, the variance of the pitch *Pitchvar*, the spectral center of gravity *CGAV*, and its fluctuations *CGFS*. This leads to the following "correction table":

Class after HMM	**Condition**	**New class**
'Speech'	if *Tonality* low	'Speech in noise'
	else if *CGFS* high	'Music
	else if *CGAV* high	'Noise
'Speech in noise'	if *Tonality* high	'Speech'
	else if *Tonality* low or *CGAV* high	'Noise'
'Noise'	if *Tonality* high	'Music'
'Music'	if *Tonality* low or *Pitchvar* low or *CGAV* high	'Noise'

It is perhaps astonishing that almost the same features are used for both classifier stages. Originally, it was thought that additional features would be used in the second stage. However, it turned out that especially the *Tonality* feature is suited for correcting the errors that the HMM had made. Thus, in the rule-based approach, the *Tonality* feature is considered to be most important. In other words, the HMM did not weight the features in the same manner as is done heuristically.

7.3.6.2 Scores

Only the test set has been used for evaluation of the second stage (the rule-based classifier was trained heuristically). The same feature set which achieved the best score was taken for the HMM which achieved the best score in section 7.3.5 (the small difference in the HMM results is due to a new set of 100 randomly chosen training and test files, as described in section 7.2.2). The table below and Figure 7.11 show the scores before and after the second stage.

Scores Before and After the Second Stage	**Test**								
	OH	Spe		SpN		Noi		Mus	
		HR	*FA*	*HR*	*FA*	*HR*	*FA*	*HR*	*FA*
Before (HMM only)	**87.3**	91.6	2.2	86.5	7.7	82.9	4.6	89.4	2.5
After (HMM & Rule-Based)	**90.5**	91.6	1.4	86.5	5.3	91.4	3.8	92.6	2.3

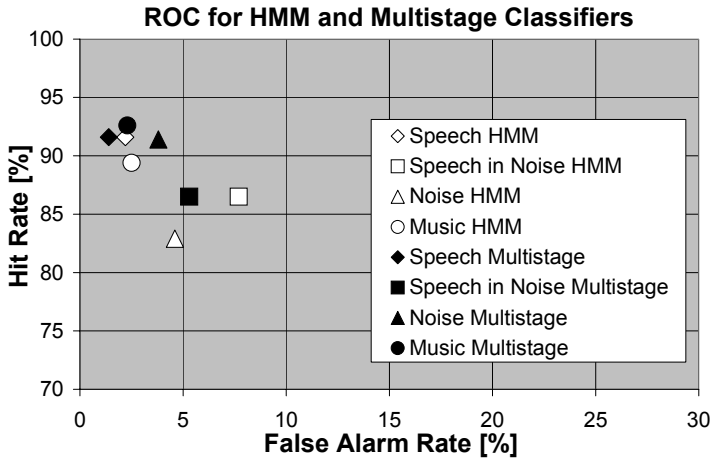


Figure 7.11: ROC graph for the HMM and the multistage classifier. Especially the hit rate of 'noise' and the false alarm rate of 'speech in noise' have improved after the second stage.

Comments:

- The overall hit rate is increased by some 3 %. The number of correctly classified 'speech' and 'speech in noise' sounds remained the same as before the second stage, but their false alarm rates were reduced. This is because a couple of 'noise' and 'music' files were moved to the correct class.
- The highest improvement of almost 9 % was achieved for the class 'noise', because the HMM did not perform very well there: Many fluctuating noises were regarded as being 'speech in noise', but taking the tonality into consideration was sufficient to reveal this error.
- The feature values for compressed and reverberated speech can be so close to those for 'speech in noise', that it was not possible to tell them apart. If a manual correction had been tried here, many true 'speech in noise' sounds would have been moved to the class 'speech'.
- It was not possible to prevent 'speech in noise' with very low SNR to be classified as 'noise', or with very high SNR as 'speech'. This, however, is also not necessarily desired. It shows again how the boundaries between 'speech', 'speech in noise', and 'noise' are somewhat fuzzy.

7.3.6.3 Summary for the multistage approach

The rather simple second stage following the HMM classifier increased the overall hit rate by about 3 % to 91 %. Especially the hit rate for the class 'noise' has improved, because many fluctuating noises were no longer misclassified as 'speech in noise'. For compressed and reverberated speech, and for 'speech in noise' with very high or very low SNR, no improvement could be achieved.

7.4 Discussion

Different feature sets and classifiers have been compared with each other. The results show that although not every classifier requires exactly the same feature set for optimal performance, the sets are quite similar: The pitch feature *Tonality*, which describes the harmonicity of the signal, together with one or more of the amplitude modulation (AM) features are the most effective. Further features that are always used are spectral and onset

features. The feature *TotPowdB*, which describes the signal level, has been omitted in the evaluation phase, because the sounds of the soundset were not recorded with their true signal level. In a real-time implementation (in a hearing instrument), a class 'silence' may be introduced that is determined by the *TotPowdB* feature. The optimal number of features depends on the classifier and lies between five and eight features.

It is obviously not reasonable to take as many features as possible. It is more efficient to find a few features that are as orthogonal as possible with regard to each other. The pitch and AM features are a good example for this approach; they alone gave already a hit rate of about 80 %, as the rule-based classifier shows. The addition of other features increased this score only by a few percent, which shows that the information lying in these features overlaps with the information that is contained in the pitch and AM features. Furthermore, the training algorithms of the classifiers did not perform optimally with too many features, or they resulted in undesired overfitting.

The single stage approach that performed best was a HMM classifier, achieving a hit rate of about 88 %. The neural network performed only slightly worse with 87 %. It should be considered that the HMM and the neural network scores are so close together that they lie within the predicted error range, which is 3 % of the achieved scores. Thus, it can be stated that neural network and HMM gave similar performance. 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. From this point of view, the good results that are achieved with the heuristic approach proposed by Zhang and Kuo (2001, see chapter 2.3) could possibly be improved if a more intelligent classifier was used. 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 post processing, 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 a HMM. HMMs are commonly more used for the identification of transient sounds, where the HMM states model the onset, the stationary and the offset part of a sound (see for example Oberle, 1999, or Zhang and Kuo, 2001). In continuous sounds, as they occur in our sound classes, the states represent different stationary parts that occur in random order, as for example parts with speech and parts with silence, or parts with speech and parts with noise. However, the problem of our sound classes is that the sounds within a class can differ very much (for example 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 a HMM. This would be different if optimal features were found; optimal in the sense that their temporal structure is the same for all sounds within a class (and different for all other sounds). It seems however unlikely that such features can be found for our sound classes. Nevertheless, it could be beneficial to look again for the common aspects of the sounds in each class, and to analyze the temporal behavior of the described features. The feature plots in chapter 5.2 do not show any temporal aspects, because only the averaged values over time are plotted. It is also possible that the features should be calculated more

often than every second to reveal more temporal information. However, some informal experiments showed that the best score for the current features and the HMM is achieved with one feature value per second.

The sounds that were most difficult to classify correctly are the following:

- Compressed (from the radio) and strongly reverberated speech was mostly classified as 'speech in noise'
- 'Speech in noise' with low SNR was regarded as 'noise', with high SNR as 'speech'
- A few fluctuating and tonal noises were classified as 'speech in noise'
- Compressed pop music (from the radio) was interpreted as being 'speech in noise'

These misclassified sounds were mostly the same for all feature sets and classifiers. On the other hand, there are some sounds in the four classes which were recognized very robustly. These include:

- Clean and slightly reverberated speech
- 'Speech in noise' with moderate SNRs (around 0.. 4 dB)
- Traffic and social noise
- Classical music, single instruments, and singing

At this moment, it is not clear how the performance for the misclassified signals can be reliably improved. Maybe a finer subdivision of the classes to be distinguished will simplify the task through a reduction of the intraclass variance. Possible subclasses are the classes 'clean speech' and 'reverberated or compressed speech', different noise classes like 'social', 'traffic' and other, or music classes like 'pop', 'classical' and 'singing', following the "wish" classes of chapter 1.2. However, for a subdivision of the classes, some specific additional features will have to be found. More reflections about further features will follow in the next chapter concerning future work.

In this context, it is also important to see that the scores are mainly a result of the soundset that was used. It was our intention 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, as for example compressed and strongly reverberated speech. How many of these sounds and how many clean speech sounds shall be put into the soundset? 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 soundset. 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 labeling 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 (for example a dialogue with interfering speakers). Again, the outcome of the classifier in such ambiguous situations depends on the labeling 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.

Maybe the purpose of sound classification should once more be considered regarding this context: The hearing instrument should switch to the optimal program depending on the quality of the acoustic environment. However, this quality could also be described in a more general way. For example, instead of saying that a signal is pop music, we could say that it is harmonic, fluctuating and rhythmic. Further characteristics could be 'high frequent' or 'low frequent', or more subjectively 'pleasant', 'metallic', 'scratching' etc. First approaches in this direction have been presented by Prante & Koop (2000), who classified a number of natural sounds depending on such general factors, or by Quast (2001), who tried to recognize nonverbal information in speech signals. Nevertheless, the optimum of the information for setting the parameters of the hearing instrument has to be specified in more detail first. And one should also be aware of the fact that the sound classes will most often occur as a mixture in everyday situations, for example "a bit noise, a bit silence, sometimes a bit music". Therefore, instead of switching between discrete programs, the hearing instrument could also change the parameters in a more dynamic way.

7.5 Conclusions

The following conclusions can be drawn from the evaluation phase:

- The most important part in a classification system are good features. In the described system, the pitch and AM features alone were sufficient to get about 80 % hit rate, but every additional percent needed much more effort.
- It is advantageous to utilize a multistage system, where the first stage produces a hypothesis about the class membership, which is then verified in a second stage. Here, the best system consists of a HMM followed by a second, rule-based stage which is able to correct a few of the misclassifications that the HMM has made. This approach achieved about 91 % hit rate.
- However, other classifiers, such as a neural network, achieved similar scores. This means that there is not much temporal information in the current features that could be modeled with a HMM. Anyway, it is worthwhile to use a classifier whose discriminant functions are trainable in a flexible way, rather than choosing a heuristic approach.
- 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'. This shows on the one hand that the current features are not adequate for all situations, and on the other hand that the sounds in the soundset can partly be ambiguous; for example, strongly reverberated speech may be perceived as being 'speech in noise' by some people, or pop music as being 'noise'. Thus, the ideas and preferences of the listener have to be considered already at the time of labeling the sound data.

Thus, the block diagram of the sound classification system could be extended with additional paths as shown in Figure 7.12. Feedback from the output of the pattern classifier or post processing block to the feature extraction or to the pattern classifiers shall indicate that the system makes use of a first feature set to draw a hypothesis, which is then verified depending on the first output and using a second feature set and/or different pattern classifier parameters.

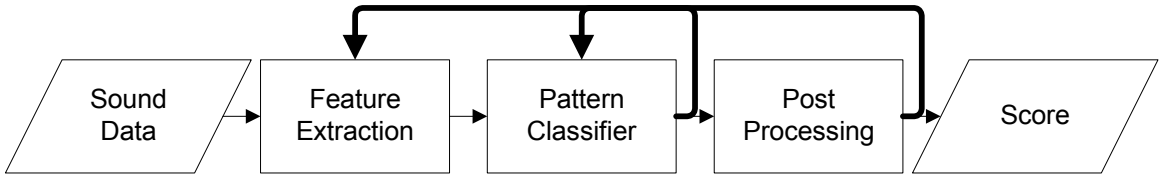


Figure 7.12: Sound classification system with additional feedback paths that indicate that the feature extraction and/or the pattern classification is adapted depending on the output, verifying the classification.

However, even a very sophisticated classification system will neither be able to look into the future to see how long a new acoustic situation will last, nor will it always know which sounds are regarded as desired signal and which as noise by the hearing instrument user.

8 Summary and Conclusions

8.1 Summary of Achievements and Conclusions

In this study, a sound classification system for the application in hearing instruments was developed. Motivated by the concept of Auditory Scene Analysis, auditory features were implemented that allow to classify the acoustic environment into the four classes 'speech', 'speech in noise', 'noise', and 'music'. Using a large sound database, different feature sets were evaluated together with several pattern classifiers. The combination of auditory features and an intelligent pattern classifier, consisting of a hidden Markov model and a second rule-based stage, achieved a high hit rate of 91 %. Compared to the existing system that was used in a field study (Phonak, 1999), improvements were obtained on two levels: The classification was implemented for four instead of two classes, and the performance for each class was significantly enhanced.

This work has been motivated by the fact that modern hearing instruments allow to switch between several hearing programs for different acoustic environments, such as for example single talker, speech in noise, music, etc. (chapter 1). These hearing programs can be activated by means of a switch at the hearing instrument or with a remote control. However, for the hearing instrument user it would be more convenient if the program selection were carried out automatically by the hearing instrument.

This assumption was confirmed by a field study. In this study with hearing impaired subjects, the usefulness and acceptance of an automatic program selection mode in the hearing instrument was investigated from the point of view of the user (chapter 3). It was shown that the automatic switching mode of the test instrument was deemed useful by a majority of test subjects, even if its performance was not perfect. These results were a strong motivation for the research described in this thesis. Furthermore, the need for a refinement of the classification into at least the classes 'speech', 'speech in noise', 'noise' and 'music' was clearly shown.

Thus, the goal of this study was on the one hand to build a system that robustly discriminates between these four classes. On the other hand, it was intended to provide fundamental knowledge as a basis for a future classification into more detailed subclasses, as for example clean and reverberated speech, social and traffic noise, classical and pop music, etc. (the "wish" classes in chapter 1).

A review of the literature showed that there are currently three algorithms that are exploited in hearing instruments (chapter 2). These algorithms allow a robust classification of clean

speech signals from other signals. Music however cannot be distinguished, and it is only partly possible to separate noise from speech in noise. Further algorithms that are designed for hearing instruments are neither able to classify music accurately, and are only partly designed to detect speech in noise. However, as stated above, the classes 'music' and 'speech in noise' are judged to be important to be recognized by the hearing instrument; for the class 'music', the sound quality should be optimized, for 'speech in noise', the intelligibility.

To consider how the auditory system performs sound classification, an overview of Auditory Scene Analysis was presented (chapter 4). By taking into account the mechanisms of Auditory Scene Analysis as well as the state of the art in sound classification, it was stated that a good sound classification system starts with a good feature extracting block. Without good features, a sophisticated pattern classifier is of little use. Thus, the main goal in this thesis was to find appropriate features before considering different pattern classifier architectures.

A number of adequate auditory features have been modeled (chapter 5). These features have partly been used in other applications, mainly for source tracking and source separation, rather than for sound classification, except for the amplitude modulation features, which have been employed in existing hearing instruments. The modeled auditory features were the following:

- The *amplitude variations* in the signal can be described in several ways: The amplitude modulations can directly be calculated for different modulation frequencies, or a feature describing the mean level fluctuations over time can be used, or the information can be gained by investigating the amplitude histogram, which also shows the mean level fluctuations. All three ways lead to features that allow to distinguish robustly between the class 'speech' and other classes.
- To distinguish between musical and non-musical sounds, the *harmonicity* of the signal was investigated. A feature that indicates how many tonal and non-tonal parts occur in the sound allows an effective recognition of musical sounds. However, the same feature is also suited for the separation of the classes 'speech', 'speech in noise' and 'noise', because the harmonicity decreases from class to class. The variance of the pitch is an additional feature for discriminating music and harmonic noises.
- The *spectral profile* was modeled in a rudimentary way by means of two features. The spectral center of gravity is a static characterization of the spectral profile, whereas the fluctuations of the spectral center of gravity describe dynamic properties of the spectral profile. These features supplement the amplitude variation and pitch features in classification; low and high frequency signals are well recognized with the help of the spectral center of gravity (these signals are mostly noises), and signals which are stationary in frequency are separated from non-stationary ones by the fluctuations of the spectral center of gravity (stationary signals are mostly noises or music).
- The onset features describe different aspects of the *amplitude onsets* in the signal. The mean onsets over time, the variance of the onsets over time, and the number of common onsets over frequency provide some specific information to identify particular sounds, such as in-the-car noise. In addition, the class 'speech' can also be separated from the other classes.
- A feature describing the *rhythm* in the signal by checking the amplitude onsets helps to distinguish between randomly fluctuating signals and signals that fluctuate rhythmically.

The latter can be music with a strong beat, such as pop music, or also noises that originate from machines with heavy motors.

Different pattern classifiers have been presented and evaluated together with the described features (chapters 6 and 7). For the application in hearing instruments, where computational speed and memory are limited, simple approaches (rule-based and minimum-distance classifier) have been compared with more complex ones (Bayes classifier, neural network, hidden Markov model, and a multistage approach). It was shown that a hit rate of about 80 % can be achieved with the simpler classifiers, which can be increased up to some 90 % when a more complex classifier is used. However, both the computing time and memory requirements are about four times bigger with the more complex than with the simpler approaches. Thus, a further increase of the classification score should be attempted by improving the features rather than the classifier. If good features can be found that are orthogonal with respect to each other, a simple classifier may be satisfying.

The classification system works well for most sounds within the four classes. There are a number of sounds in each class that were recognized very robustly: Clean and slightly reverberated speech, speech in noise with moderate SNR, traffic and social noise, and classical music, single instruments and singing. However, some sounds were problematic and were mostly misclassified: Speech in noise with very low or very high SNR, was classified as 'noise' or 'speech', respectively; compressed and strongly reverberated speech, a few tonal and fluctuating noises, and compressed pop music, were all classified as 'speech in noise'.

It is obvious that some files of the classes 'speech' and 'music' may also be perceived by human beings as sounding like the class 'speech in noise'; reverberated speech or pop music are just two examples. Hence, the classifier may indeed not be so wrong by yielding these confusions. In the case of some misclassfied fluctuating noises, however, other factors must be taken into account, as they will not even be perceived as 'speech in noise' by human beings.

8.2 Future Work

During the implementation and evaluation of the sound classification system described in this thesis, insight was gained on how the system could be improved and further advanced in the future. The following list contains the most important points requiring improvements as well as a few directions for future work.

- So far, the performance of the sound classification system has only been tested on sounds from the soundset. One of the next steps should therefore be to evaluate the system in a field experiment to gain more practical experience. For this purpose, the actual Matlab implementation comprising feature extraction, classification, and post processing should be transferred to a portable system which will enable to carry out the evaluation of the sound classification system in real-time. This approach will most probably also provide new ideas about possible optimization strategies.
- Up to now, the classification is performed for the four "must" classes 'speech', 'speech in noise', 'noise', and 'music'. However, some concepts for a refinement of the classification into the "wish" classes (see chapter 1.2) were presented throughout this thesis. A finer subdivision of the classes might also simplify the task through a reduction of the intraclass variance of the sound's characteristics. The "wish" classes are a good start for progressing. In addition, the classes could also be refined using more acoustical criteria. This might

apply in particular for the sound class 'noise', where for example continuous and transient sounds, or tonal and non-tonal sounds could be distinguished.

- The sound classes will most often occur as a mixture in everyday situations, for example "a bit noise, a bit silence, sometimes a bit music". Therefore, instead of switching between discrete programs, the hearing instrument could also change the parameters in a more dynamic way. This, however, is part of the post processing step, which has not been exploited much yet; it will become more important when evaluating the system in a field trial.
- Taking into account further or ameliorate existing features will be an important aspect for improving and refining the classification. This includes:
  - 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 (1999) analyzed the spectral profile for the classification of some specific environmental sounds, although on a more detailed layer than desired here.
  - A beat feature that can distinguish between beats originating from music and from noise. It is however not clear at this moment how this can be performed.
  - A feature that describes the amount of reverberation in the signal. This could improve the classification of reverberated speech signals. Approaches to determine the reverberation in a signal have been presented by Shoda and Ando (1998) and Ando et al. (1999).
  - A feature that determines the SNR of the signal, for a more gradual classification of signals containing speech and/or noise. SNR determination is often used in noise suppression algorithms (see for example Marzinzik 2000); thus, similar concepts could be employed here.
  - Spatial features, to analyze where the signal and where the noise comes from, or to check both front and back signal 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 instrument could be set accordingly.
- Finally, the simple multistage classification approach could be extended. More a priori knowledge could be exploited; especially the use of context information would be beneficial for a system that tries to take into consideration human perception.

8.3 Outlook

With this thesis and other recent work on sound classification, progress is being made towards automatic and robust classification of the acoustic environment. Automatic sensing of the current acoustic situation and automatic switching to the best fitting program is, however, just one gadget in the hearing instrument that contributes to the comfort of the user. The exploitation of psychoacoustical knowledge in combination with digital signal processing will further enhance both sound quality and speech intelligibility. The goal is undistorted communication in every listening situation.

However, we are far from achieving similar performance in a hearing instrument 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, we are 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. Grouping and segregation have to be performed based on primitive organization as well as learned schemas; these tasks must be carried out within milliseconds. However, in contrast to a hearing instrument, the human auditory system has had a life-time 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.

Appendix A
Sound Database

The sound database consists of 287 files of the classes 'speech', 'speech in noise', 'noise', and 'music', each of 30 seconds length. Each sound file represents exactly one of the four sound classes, that is, no class changes occur within the file. Most files originate from CDs, with some exceptions:

- The traffic noises were recorded at different places in the city of Zurich.
- The in-car noises were recorded in several different cars of Phonak employees.
- The files of the class 'speech in noise' are 'speech' and 'noise' files mixed together at different SNRs. The mixing and the SNR estimation were performed in Matlab. An *Oldenburger* sentence test (Wagener et al., 1999) was performed with a few normal hearing subjects to find out the SNR for the different noises at which they would understand 50 %. The SNR for mixing was then chosen 5 dB higher to account for a moderate hearing impairment (the 50 % speech reception threshold is about 5 dB higher for subjects with a moderate hearing loss than for normal hearing persons; see Killion, 1997).
- The clean speech files were so dry that slight reverberation was added using Syntrillium's CoolEdit 2000 audio processing software and its "warm room reverberation" preset.
- The reverberation of the strongly reverberated speech files was performed either in Matlab with room impulse responses that were available from Phonak's Sound Database (Phonak, 2000), or with the help of Syntrillium's CoolEdit 2000 audio processing software.

All files are stored as mono files with 16 bit resolution and 22 kHz sampling rate. The table below lists the files and their description.

Class	Sound No.	Description
60 Speech Files		*Clean speech with normal room reverberation ($T_{60} \approx 500$ ms):*
	1-10	Female speakers
	11-20	Male speakers
	21-23	Fast male speakers
	24-25	Fast female speakers
	26, 30	Male speakers with raised voice
	27-29	Female speakers with raised voice
	31-35	Dialogues between a male and a female speaker
	36-40	Children
		Compressed and more reverberated speech:
	41-44	Compressed speech from radio

Class	Sound No.	Description
	45	Compressed speech from radio, report via telephone
	46-47	Slightly compressed but quite reverberated speech from TV
	48	Compressed and quite reverberated speech from TV
	49-50	Slightly compressed but quite reverberated speech from TV
		Stronger reverberated speech:
	51	Speaker no.2 in church (T_{60} ~ 3200 ms)
	52	Speaker no.16 in church (T_{60} ~ 3200 ms)
	53	Speaker no.2 in warm room (T_{60} ~ 1200 ms)
	54	Speaker no.24 in warm room (T_{60} ~ 1200 ms)
	55	Speaker no.10 in empty echoic room (T_{60} ~ 7000 ms)
	56	Speaker no.24 in empty echoic room (T_{60} ~ 7000 ms)
	57	Speaker no.10 in large empty hall (T_{60} ~ 5000 ms, low HF absorption)
	58	Speaker no.20 in large empty hall (T_{60} ~ 5000 ms, low HF absorption)
	59	Speaker no.16 in large occupied hall (T_{60} ~ 4000 ms, high HF absorption)
	60	Speaker no.20 in large occupied hall (T_{60} ~ 4000 ms, high HF absorption)
74 Speech in Noise Files		*Speech in social noise:*
	61	Speaker no.20 in speech babble no.141, SNR 0 dB
	62	Speaker no.16 in speech babble no.142, SNR 4 dB
	63	Speaker no.10 in cafeteria no.143, SNR 0 dB
	64	Speaker no.2 in cafeteria no.144, SNR 4 dB
	65	Speaker no.24 in cafeteria no.145, SNR 0 dB
	66	Speaker no.20 in cafeteria no.146, SNR 4 dB
	67	Speaker no.16 in cafeteria no.147, SNR 0 dB
	68	Speaker no.10 in cafeteria no.148, SNR 4 dB
	69	Speaker no.2 in cafeteria no.149, SNR 0 dB
	70	Speaker no.24 in cafeteria no.150, SNR 4 dB
	71	Speaker no.20 in cafeteria no.151, SNR 0 dB
	72	Speaker no.16 in exhibition hall no.152, SNR 4 dB
	73	Speaker no.10 in party no.153, SNR 0 dB
	74	Speaker no.2 in party no.154, SNR 4 dB
	75	Speaker no.24 in restaurant no.155, SNR 0 dB
	76	Speaker no.20 in restaurant no.156, SNR 4 dB
	77	Speaker no.16 in public festival no.157, SNR 0 dB
		Speech in the car:
	78	Speaker no.10 in car noise no.158, SNR –9 dB
	79	Speaker no.2 in car noise no.159, SNR –5 dB
	80	Speaker no.24 in car noise no.160, SNR –9 dB
	81	Speaker no.20 in car noise no.161, SNR –5 dB
	82	Speaker no.16 in car noise no.162, SNR –9 dB
	83	Speaker no.10 in car noise no.163, SNR –5 dB
	84	Speaker no.2 in car noise no.164, SNR –9 dB
		Speech in traffic noise:
	85	Speaker no.24 in tractor noise no.165, SNR –1 dB
	86	Speaker no.20 in traffic noise no.166, SNR –3 dB
	87	Speaker no.16 in traffic noise no.167, SNR –1 dB
	88	Speaker no.10 in traffic noise no.168, SNR –3 dB
	89	Speaker no.2 in traffic noise no.169, SNR –1 dB
	90	Speaker no.24 in traffic noise no.170, SNR –3 dB
	91	Speaker no.20 in traffic noise no.171, SNR –1 dB
	92	Speaker no.16 in traffic noise no.172, SNR –3 dB
	93	Speaker no.10 in traffic noise no.173, SNR –1 dB
	94	Speaker no.2 in traffic noise no.174, SNR –3 dB
	95	Speaker no.24 in traffic noise no.175, SNR –1 dB
	96	Speaker no.20 in traffic noise no.176, SNR –3 dB
	97	Speaker no.16 in train noise no.177, SNR –1 dB
	98	Speaker no.10 in train noise no.178, SNR –3 dB
		Speech in industrial noise:
	99	Speaker no.2 in chainsaw noise no.179, SNR –4 dB
	100	Speaker no.24 in construction noise no.180, SNR 0 dB
	101	Speaker no.20 in drilling noise no.181, SNR –4 dB
	102	Speaker no.16 in grinding noise no.182, SNR 0 dB
	103	Speaker no.10 in teleprinter noise no.183, SNR –4 dB
	104	Speaker no.2 in jackhammer noise no.184, SNR 0 dB

Class	Sound No.	Description
	105	Speaker no.24 in lawnmower noise no.185, SNR –4 dB
	106	Speaker no.20 in printer noise no.186, SNR 0 dB
	107	Speaker no.16 in machine noise no.187, SNR –4 dB
	108	Speaker no.10 in mixer noise no.188, SNR 0 dB
	109	Speaker no.2 in piledriver noise no.189, SNR –4 dB
	110	Speaker no.24 in press noise no.190, SNR 0 dB
	111	Speaker no.20 in saw noise no.191, SNR –4 dB
	112	Speaker no.16 in saw noise no.192, SNR 0 dB
	113	Speaker no.10 in steamshovel noise no.193, SNR –4 dB
	114	Speaker no.2 in weaving machine noise no.194, SNR 0 dB
		Speech in other noise:
	115	Speaker no.24 in applause no.195, SNR –2 dB
	116	Speaker no.20 in blender noise no.196, SNR 2 dB
	117	Speaker no.16 in bowling noise no.197, SNR –2 dB
	118	Speaker no.10 in football stadium no.198, SNR 2 dB
	119	Speaker no.2 in frying noise no.199, SNR –2 dB
	120	Speaker no.24 in frying noise no.200, SNR 2 dB
	121	Speaker no.20 in office noise no.201, SNR –2 dB
	122	Speaker no.16 in printer noise no.202, SNR 2 dB
	123	Speaker no.10 in rain noise no.203, SNR –2 dB
	124	Speaker no.2 in shaver noise no.204, SNR 2 dB
	125	Speaker no.24 in shopping mall no.205, SNR –2 dB
	126	Speaker no.20 in shower noise no.206, SNR 2 dB
	127	Speaker no.16 in sink noise no.207, SNR –2 dB
	128	Speaker no.10 in volleyball game no.208, SNR 2 dB
	129	Speaker no.2 in supermarket no.209, SNR –2 dB
	130	Speaker no.24 in train no.210, SNR 2 dB
	131	Speaker no.20 in train station no.211, SNR –2 dB
	132	Speaker no.16 in typewriter no.212, SNR 2 dB
	133	Speaker no.10 in vacuum cleaner no.213, SNR –2 dB
	134	Speaker no.2 in videogame noise no.214, SNR 2 dB
80 Noise Files		*Social noise:*
	135-140	Cafeteria
	141-142	Speech babble
	143-151	Cafeteria
	152	Exhibition hall
	153-154	Party
	155-156	Restaurant
	157	Public festival
		In-the-car noise, at different speed:
	158-164	In-the-car
		Traffic noise:
	165	Tractor
	166-176	Traffic: cars, trams and trucks passing by
	177-178	Train passing by
		Industrial noise:
	179	Chainsaw
	180	Construction site, machines
	181	Drilling
	182	Grinding
	183	Teleprinter, high frequencies
	184	Jackhammer
	185	Lawnmower
	186	Printing machine
	187	Machine with air valves
	188	Mixer
	189	Piledriver
	190	Press
	191	Electric saw
	192	Manual saw
	193	Steamshovel
	194	Weaving machine
		Other noise:

Class	Sound No.	Description
	195	Applause
	196	Blender
	197	Bowling
	198	Football stadium
	199-200	Frying
	201	Office
	202	Printer
	203	Rain
	204	Electric shaver
	205	Shopping mall
	206	Shower
	207	Water running into sink
	208	Volleyball game
	209	Supermarket
	210	In a train
	211	Train station
	212	Typewriter
	213	Vacuum cleaner
	214	Videogames
73 Music Files		*Classical music:*
	215-216	Brass orchestra
	217	Bach
	218	Mozart
	219	Beethoven
	220	Brahms
	221	Bach
	222	Schumann
	223	Dvorak
	224	Tschaikowsky
	225	Debussy
	226	Berlioz
	227	Wagner
	228	Brahms
	229	Strauss
	230	Haydn, compressed from radio
	231	Wagner
	232	Strauss
	233	Mozart
		Pop music:
	234	Suzanne Vega
	235	Pop from Phonak CD
	236	Dire Straits
	237	Alan Parsons
	238	Züri West
	239	Midnight To Six
	240	Span
	241	Polo Hofer
	242	Lloyd Cole
	243	Tom Waits
	244	Tanita Tikaram
	245-248	Compressed pop from radio
		Single instruments:
	249	Contrabass
	250	Bass drum
	251	Bassoon
	252	Flute
	253	Cello
	254	Clarinet
	255	Flute
	256	Harp
	257	Horn
	258	Oboe
	259	Organ

Class	Sound No.	Description
	260	Piano
	261	Piccolo
	262	Saxophone
	263	Transverse flute
	264	Trombone
	265	Trumpet
	266	Viola
	267	Xylophone
		Singing:
	268	Chorus
	269-270	Chorus, children
	271	Male singer
	272	Chorus
	273, 275	Male singer
	274, 276-277	Female singer
	278	Chorus, woman
	279-280	Chorus, children
		Other music:
	281-282	Jazz
	283	Rock'n'Roll
	284, 287	Jazz
	285-286	Folk

Bibliography

Ando, Y., Sato, S., Sakai, H. (**1998**). "Fundamental subjective attributes of sound fields based on the model of auditory-brain system," in Computational Acoustics in Architecture, edited by J.J. Sendra (WIT Press), 63-99.

ANSI, American National Standards Institute (**1960, 1994**). "Acoustical Terminology S1," Acoustical Society of America, New York.

Baumann, U. (**1995**). Ein Verfahren zur akustischen Erkennung und Trennung multipler akustischer Objekte. PhD Thesis (Technische Universität München).

von Békésy, G. (**1960**). Experiments in Hearing (McGraw-Hill, New York).

Bishop, C. M. (**1995**). Neural Networks for Pattern Recognition (Clarendos Press, Oxford).

Boretzki, M., Kiessling, J., Margolf-Hackl, S., Kühnel, V., Volpert, S. (**2001**). "Adaptive Richtcharakteristik eines Doppelmikrofons und automatische Programmwahl: Nutzen für den schwerhörigen Menschen," Proc. 4. Jahrestagung der Deutschen Gesellschaft für Audiologie, Aachen, Germany.

Bregman, A. S. (**1990**). Auditory Scene Analysis (MIT Press, Cambridge).

Bregman, A. S. (**1993**). "Auditory Scene Analysis: Hearing in Complex Environments," in Thinking in Sound, edited by McAdams, S. and Bigand, E. (Oxford University Press).

Brown, G. J., Cooke, M. (**1993**). "Physiologically-motivated signal representations for computational auditory scene analysis", in Visual Representations of Speech Signals, edited by M. Cooke, S. Beet, M. Crawford (John Wiley).

Brown, G. J., Cooke, M. (**1994**). "Computational auditory scene analysis", Computer, Speech and Language **8**, 297-336.

Büchler, M. (**2001**). "How good are automatic program selection features? A look at the usefulness and acceptance of an automatic program selection mode," Hear. Rev. 9/2001, 50-54, 84. *Also in German in* Hörakustik 10/2001, 64-71.

Byrne, D., Dillon, H. (**1986**). "The National Acoustic Laboratories (NAL) new procedure for selecting the gain and frequency response of a hearing aid," Ear Hear. **7**, 257-265.

Cherry, E.C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975-979.

Cooke, Martin (**1993**). Modeling auditory processing and organization. PhD Thesis (University of Sheffield).

Cornelisse, L.E., Seewald, R.C., and Jamieson, D.G. (**1995**). "The input/output (i/o) formula: A theoretical approach to the fitting of personal amplification devices," J. Acoust. Soc. Am. **97**(3), 1854-1864.

Couvreur, C., Fontaine, V., Gaunard, P., Mubikangiey, C. G. (**1998**). "Automatic Classification of Environmental Noise Events by Hidden Markov Models," J. Appl. Acoustics **54**, No. 3, 187-206.

Cramer, E. M., Huggins, W. H. (**1958**). "Creation of Pitch through Binaural Interaction," J. Acoust. Soc. Am. **30**, 412-417.

Darwin, C. J., Carlyon, R. (**1995**). "Auditory Grouping," in Handbook of Perception and Cognition. Hearing, edited by B. C. J. Moore (Academic Press, San Diego).

Dau T., (**1996**). Modeling auditory processing of amplitude modulation. PhD Thesis (Fachbereich Physik, Universität Oldenburg).

Deller, J. R., Proakis, J. G., and Hansen, J. H. L. (**1993**). Discrete-Time Processing of Speech Signals (Macmillan Publishing Company, Englewood Cliffs).

Duda, R. O. and Hart, P. E. (**1973**). Pattern Classification and Scene Analysis (John Wiley & Sons).

Ellis, D. P. W. (**1996**). Prediction-driven computational auditory scene analysis. PhD thesis (Massachusetts Institute of Technology).

Engel, K., Singer, W. (**1997**). "Neuronale Grundlagen der Gestaltwahrnehmung", Spektrum der Wissenschaft "Kopf oder Computer", Dossier **4**/1997, 66-73.

Fedtke, T., Fuder, G., Hamann, D., Haubold, J. (**1991**). "Natürliche Klangbilder," in Plath, P. (Hrsg.), Neue Technologien in der Hörgeräte-Akustik - Herausforderung an die Audiologie. Materialsammlung vom 5. Multidisziplinären Kolloquium der GEERS-Stiftung am 12. und 13. März 1990 in Bonn, 116-136.

Feldbusch, F. (**1998**). "Geräuscherkennung mittels Neuronaler Netze," Zeitschrift für Audiologie 1/1998, 30-36.

Feldbusch, F. (**2001**). "A Heuristic for Feature Selection for the Classification with Neural Nets," Joint 9th IFSA World Congress and 20th NAFIPS International Conference, Vancouver.

Gabriel, B. (**2001**). "Nutzen moderner Hörgeräte-Features für Hörgeräte-Träger am Beispiel eines speziellen Hörgeräte-Typs," Zeitschrift für Audiologie **40**, Nr. 1, 16-31.

Gaunard, P., Mubikangiey, C. G., Couvreur, C., Fontaine, V. (**1998**). "Automatic Classification of Environmental Noise Events by Hidden Markov Models," Proc. ICASSP 1998, 3609-3612.

Gold, B., Morgan, N. (**2000**). Speech and Audio Signal Processing: Processing and Perception of Speech and Music (John Wiley).

Goldhor, R. S. (**1993**). "Recognition of Environmental Sounds," Proc. ICASSP 1993, 149-152.

Gonzalez, R. C., and Woods, R. E. (**1993**). Digital Image Processing (Addison-Wesley).

Green, D. M. (**1993**). "Auditory Intensity Discrimination," in Human Psychophysics (Springer, New York).

Handel, S. (**1989**). Listening (MIT, Cambridge, MA).

Handel, S. (**1995**). "Timbre Perception and Auditory Object Identification," in Handbook of Perception and Cognition. Hearing, edited by B. C. J. Moore (Academic Press, San Diego).

Hartmann, W. M. (**1998**). Signals, Sound, and Sensation (Springer, New York).

Haubold, J., Geers, W. (**1993**). "A-Life - a modern hearing aid fitting procedure based on natural acoustic patterns," in Recent Developments in Hearing Instrument Technology, edited by Beilin, J., Jensen, G. R., 497-503.

Holube, I. (**1998**). Speech Sensitive Processing - Voice Activity Detection System.

Houtsma, A. J. M. (**1989**). "Timbre," in Structure and Perception of Electroacoustic Sounds and Music, edited by Nielzen, S. and Olsson, O. (Excerpta Medica, Amsterdam), 157-159.

Karjalainen, M., Tohonen, T. (**1999**). "Multi-Pitch and Periodicity Analysis Model for Sound Separation and Auditory Scene Analysis," Proc. ICASSP 1999, Phoenix, AZ, 929-932.

Kashino, K., Murase, H. (**1999**). "A sound source identification system for ensemble music based on template adaption and music stream extraction," Speech Communication **27**, 337-349.

Kates, J. M. (**1995**). "Classification of background noises for hearing-aid applications," J. Acoust. Soc. Am. **97**, 461-470.

Kil, D.H., Shin, F.B. (**1996**). Pattern Recognition and Prediction with Applications to Signal Characterization (American Institute of Physics, Woodbury).

Killion, M. (**1997**). "The SIN report: Circuits haven't solved the hearing-in-noise problem," Hearing Journal 50 (10), 28-30, 32.

Kollmeier, B., and Koch, R. (**1994**). "Speech Enhancement Based on Physiological and Psychacoustical Models of Modulation perception and Binaural Interaction," J. Acoust. Soc. Am. **95**, 1593-1602.

Korl, S. (**1999**). Automatische Geräuschklassifizierung zur Anwendung in Hörgeräten. Diplomarbeit (Labor für experimentelle Audiologie, ORL-Klinik, Universitätsspital Zürich).

Lambrou, T., Kudumakis, P., Speller, R., Sandler, M., Linney, A. (1998). "Classification of Audio Signals Using Statistical Features on Time and Wavelet Transform Domains," ICASSP 1998, Seattle.

Leber, J.-F. (**1993**). The Recognition of Acoustical Signals using Neural Networks and an Open Simulator. PhD Thesis ETH No. 10016, (Swiss Federal Institute of Technology, Zürich); *also appearing in* Series in Microelectronics Vol. 20 (Hartung-Gorre Verlag, Konstanz).

Ludvigsen, C. (**1993**). Schaltungsanordnung für eine automatische Regelung von Hörhilfsgeräten. Deutsches Patent Nr. DE 43 40 817 A1.

Lyon, R., Shamma, S. (**1996**). "Auditory Representations of Timbre and Pitch", in Auditory Computation, edited by H. L. Hawkins, T. A. McMullen, A. N. Popper and R. R. Fay (Springer, New York).

Martens, J. P., VanImmerseel, L. M. (**1992**). "Pitch and voiced/unvoiced determination with an auditory model," J. Acoust. Soc. Am. **91** (6), 3511-3526.

Martin, K. D. (**1998**). "Toward Automatic Sound Source Recognition: Identifying Musical Instruments," NATO Computational Hearing Advanced Study Institute, Il Ciocco, Italy.

Martin, K. D., Kim, Y. E. (**1998**). "Musical instrument identification: A pattern-recognition approach," Proc. of 136th ASA meeting.

Martin, K. D., Scheirer, E. D., Vercoe, B. L. (**1998**). "Music Content Analysis through Models of Audition," ACM Multimedia Workshop on Content Processing of Music for Multimedia Applications, Bristol UK.

Marzinzik, M. (**2000**). Noise reduction schemes for digital hearing aids and their use for the hearing impaired. PhD Thesis (Fachbereich Physik, Universität Oldenburg).

Meddis, R., Hewitt, M. (**1991**). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," J. Acoust. Soc. Am. **89** (6), 2866-2882.

Mellinger, D. K. (**1992**). Event Formation and Separation in Musical Sound. PhD Thesis (Stanford University).

Mellinger, D. K., Mont-Reynaud, B. M. (**1996**). "Scene Analysis," in Auditory Computation, edited by H. L. Hawkins, T. A. McMullen, A. N. Popper and R. R. Fay (Springer, New York).

Moore, B. C. J. (**1993**). "Frequency Analysis and Pitch Perception," in Human Psychophysics, edited by W. A. Yost, A. Popper and R. R. Fay (Springer, Berlin).

Moore, B. C. J. (**1997**). An Introduction to the Psychology of Hearing (Academic Press, London).

Nordqvist, P. (**2000**). "Automatic Classification of Different Listening Environments in a Generalized Adaptive Hearing Aid," International Hearing Aid Research Conference IHCON 2000, Lake Tahoe, CA.

Oberle, S. and Kaelin, A. (**1995**). "Recognition of Acoustical Alarm Signals for the Profoundly Deaf Using Hidden Markov Models," Proc. ISCAS '95, Vol. 3, Seattle 1995, 2285-2288.

Oberle, S. (**1999**). Detektion und Estimation von akustischen Signalen mit Hidden-Markov-Modellen. Series in adaptive signal processing; Vol. 4 (Hartung-Gorre Verlag, Konstanz).

Ostendorf, M., Hohmann, V., and Kollmeier, B. (**1997**). "Empirische Klassifizierung verschiedener akustischer Signale und Sprache mittels einer Modulationsfrequenzanalyse," in Fortschritte der Akustik, DAGA '97, edited by DPG (DPG-Verlag, Bad Honnef).

Ostendorf, M., Hohmann, V., and Kollmeier, B. (**1998**). "Klassifikation von akustischen Signalen basierend auf der Analyse von Modulationsspektren zur Anwendung in digitalen Hörgeräten," in Fortschritte der Akustik - DAGA '98, 402-403.

Phonak Hearing Systems (**1999**). "Claro AutoSelect", company brochure no. 028-0148-02.

Phonak Hearing Systems (**2000**). "Phonak Sound Database", internal report.

Prante, H.U. and Koop, L. (**2000**). "Classification of Sounds with Temporal Feature Maps," Proc. 8th Oldenburg Symposium on Psychological Acoustics, 191-199.

Quast, H. (**2001**). "Automatische Erkennung nonverbaler Sprache," Fortschritte der Akustik - DAGA '01, Hamburg.

Rabiner, L., Juang, B.H. (**1986**). "An Introduction to Hidden Markov Models," IEEE Trans. Ac. Sp. Sig. Proc. **3** (1), 4-16.

Rabiner, L., Juang, B.H. (**1993**). Fundamentals of Speech Recognition (Prentice Hall, Englewood Cliffs).

Ringdahl, A., Mangold, S., Lindkvist, A. (**1993**). "Does the hearing aid user take advantage of different settings in multiprogrammable hearing aids in acoustically various listening environments?," in Recent Developments in Hearing Instrument Technology: Proceedings of the 15th Danavox Symposium, edited by J. Beilin and G. Jensen (Stoudard Jensen, Copenhagen), 453-464.

Scheirer, E. D. (**1997**). "Pulse Tracking with a Pitch Tracker," Proc. 1997 Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk, NY.

Scheirer, E. and Slaney, M. (**1997**). "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator," ICASSP 1997, 1331-1334.

Scheirer, E. D. (**1998**). "Tempo and Beat Analysis of Acoustic Musical Signals," J. Acoust. Soc. Am. **103** (1), 588-601.

Schürmann, J. (**1996**). Pattern Classification: a unified view of statistical and neural approaches (John Wiley & Sons).

Shepard, R. N. (**1991**). Einsichten & Anblicke: Illusion und Wahrnehmungskonflikte in Zeichnungen (Spektrum der Wissenschaft, Heidelberg).

Shoda, T., Ando, Y. (**1999**). "Calculation of speech intellegibility using four orthogonal factors extracted from ACF of source and sound field signals," 16th ICA and 135th ASA, Seattle, 2163-2164.

Slaney, M., Lyon, R. (**1993**). "On the importance of time - a temporal representation of sound," in Visual Representations of Speech Signals (Wiley).

Soltau, H., Schultz, T., Westphal, M., Waibel, A. (**1998**). "Recognition of Music Types," ICASSP 1998, Seattle, 1137-1140.

Stevens, S. S. (**1935**). "The relation of pitch to intensity," J. Acoust. Soc. Am. **6**, 150-154.

Terhardt, E. (**1998**). Akustische Kommunikation (Springer, Berlin, Heidelberg).

Viemeister, Neal F. and Plack, Christopher J. (**1993**). "Time Analysis," in Human Psychophysics, edited by W. A. Yost, A. Popper and R. R. Fay (Springer, Berlin).

Wagener, K., Kühnel, V., Kollmeier, B. (**1999**). "Entwicklung und Evaluation eines Satztestes für die deutsche Sprache. Design des Oldenburger Satztests," Zeitschrift für Audiologie **1**/1999, 4-15.

Warren, R. M. (**1999**). Auditory Perception (Cambridge University Press).

White, G. D. (**1991**). The Auditory Dictionary (University of Washington Press).

Yost, A. W. (**1994**). Fundamentals of Hearings – An Introduction (Academic Press).

Yost, W. A. (**1991**). "Auditory image perception and analysis: the basis for hearing," Hear. Res. **56**, 8-18.

Yost, A. W., Sheft, S. (**1993**). "Auditory Perception," in Human Psychophysics, edited by W. A. Yost, A. Popper and R. R. Fay (Springer, Berlin).

Zhang, T. and Kuo, C.-C. J. (**1998**). "Content-Based Classification and Retrieval of Audio," SPIE Conference on Advanced Signal Processing Algorithms, Architectures and Implementations VIII, San Diego, 432-443.

Zhang, T. and Kuo, C.-C. J. (**1999**). "Hierarchical Classification of Audio Data for Archiving and Retrieving," ICASSP 1999, Phoenix.

Zhang, T. and Kuo, C.-C. J. (**2001**). "Audio Content Analysis for Online Audiovisual Data Segmentation and Classification," IEEE Trans. Speech and Audio Proc. **9**, No. 4, 441-457.

Zwicker, E., and Fastl, H. (**1990**). Psychoacoustics (Springer Verlag, Berlin Heidelberg).

Dankeswort

Die vorliegende Arbeit wäre ohne Hilfe von mancher Seite nicht denkbar gewesen.

Für die Betreuung der Dissertation möchte ich in erster Linie drei Personen danken. Silvia Allegro ist mir von Seiten der Firma Phonak AG mit unermüdlichem Einsatz zur Seite gestanden und war immer bereit, Fragen und Probleme zu diskutieren. Ihre kritische Unterstützung sowie diejenige der Korreferenten Stefan Launer von Phonak AG und Norbert Dillier von der ORL-Klinik des Universitätsspitals Zürich haben mir immer wieder den richtigen Weg gewiesen. Dabei stellte es sich für mich meist erst im Nachhinein heraus, wie vorteilhaft es war, dass meine Betreuer gedanklich oft schon drei Schritte weiter waren als die tatsächliche Arbeit ("... das ist sehr ordentlich, aber weisst du, was ich an deiner Stelle auch mal noch machen würde ...").

Für die Übernahme des Referats gebührt mein Dank Herrn Prof. Dr. Peter Niederer vom Institut für Biomedizinsche Technik und Medizinische Informatik der ETH und Universität Zürich. Er hat durch seine interdisziplinäre Sicht der Dinge wertvolle Anregungen eingebracht.

Norbert Dillier und seiner ganzen audiologischen Forschungsgruppe an der ORL-Klinik des Universitätsspitals Zürich möchte ich für die angenehme und freundschaftliche Atmosphäre danken, in der ich meine Arbeit verrichten durfte. Meine Büronachbarin Simone Volpert hatte immer ein aufmunterndes Wort bereit, wenn etwas nicht so rund lief wie gewünscht. Meine Mitdoktoranden Mattia Ferrazzini, Olegs Timms und Urban Willi sorgten immer wieder für angeregte Diskussionen auch ausserhalb des Themas Audiologie. Waikong Lai war stets hilfsbereit zur Stelle, wenn es um grafisches Design, englische Rechtschreibung oder auch um etwas ganz anderes ging. Um administrative Dinge kümmerte sich Belja Breginc, und sie besorgte auch den lebenswichtigen Nachschub an Tee und Kaffe. Alle genannten, aber auch Fränzi Conod, Felix Beerli, Hubert Hauschild, Herbert Jakits, Markus Schmid sowie unsere australischen Gäste Colin Irwin und Andrew Knox waren eine willkommene Bereicherung der Kaffeerunde.

Patrizia Rauso und Linda Travnicek, in den ORL-Sekretariaten tätig, sorgten ebenso für erfrischende Mittags- und Kaffeepausen.

Mit seiner Diplomarbeit in unserer Gruppe hat Sascha Korl grundlegende und wegweisende Forschung betrieben, auf der ich verlässlich aufbauen konnte. Meine Anerkennung gilt deshalb seinem grossen Einsatz und seiner hervorragenden Arbeit.

Der Firma Phonak AG möchte ich für die grosszügige finanzielle Unterstützung für die Durchführung der Dissertation danken. Mein Dank gebührt aber auch der ganzen Crew von

Phonak Stäfa, insbesondere Herbie Bächler fürs Initiieren der Dissertation, Volker Kühnel für manchen audiologischen Rat, Maurice Boonen für die Unterstützung bei der Erstellung der Klangdatenbank, Hansueli Roeck für die Beratung im Bereich der digitalen Signalverarbeitung, und Jürgen Tchorz für das professionelle Gestalten von Publikationen.

Nicht zuletzt möchte ich meiner Familie und all meinen Freundinnen und Freunden ausserhalb der Spitalmauern für ihre ermutigenden Worte und ihre Unterstützung während der drei Dissertationsjahre danken, insbesondere Nicole Kropf, die auch einen grossen Teil des Lektorats übernommen hat.

Curriculum Vitae

Michael Büchler
Citizen of Brugg, Aargau, Switzerland
Born on April 26, 1967 in Zurich, Switzerland

Education and Professional Experience

July 1998 – Dec. 2001	Doctor of Technical Sciences (PhD), Swiss Federal Institute of Technology (ETH), and Laboratory of Experimental Audiology, University Hospital, Zurich, Switzerland.
May 1997 – June 1998	Signal Processing Research Engineer, at Phonak AG, Staefa, Switzerland.
Oct. 1994 – Apr. 1997	Microcontroller Software Research Engineer and Project Leader, at Feller AG, Horgen, Switzerland.
1988 – 1994	Diploma in Electrical Engineering, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland. Diploma Thesis: "Digital Interface for Cochlea Implant Speech Processor."
1988	Military Service as Communication Engineer
1980 – 1987	High School, Matura Type B, Kantonsschule Hohe Promenade, Zurich, Switzerland.
1974 – 1980	Primary School, Meilen, Switzerland.