

# Automatic Sound Classification Inspired by Auditory Scene Analysis

Silvia Allegro, Michael Büchler<sup>+</sup>, Stefan Launer

Signal Processing Department, Phonak AG, Switzerland

<sup>+</sup>Department of Otorhinolaryngology, University Hospital Zurich, Switzerland

silvia.allegro@phonak.ch

## Abstract

*A sound classification system for the automatic recognition of the acoustic environment in a hearing instrument is discussed. The system distinguishes the four sound classes 'clean speech', 'speech in noise', 'noise', and 'music' and is based on auditory features and hidden Markov models. The employed features describe level fluctuations, the spectral form and harmonicity. Sounds from a large database are employed for both training and testing of the system. The achieved recognition rates are very high except for the class 'noise'. Problems arise in the classification of pop music, reverberated speech, tonal noises and singing.*

## 1. Introduction

Modern hearing instruments typically provide several hearing programs to account for different acoustic situations such as e.g. quiet environment, noisy environment, music, etc. These hearing programs can be activated either by means of a switch at the hearing instrument or with a remote control. The manual switching between different hearing programs is however annoying and might even be difficult for some of the hearing instrument users. It would therefore be more convenient for the user if the hearing instrument would analyze the acoustical situation internally and carry out the appropriate program selection automatically.

There exist already simple approaches to automatic sound classification in hearing instruments, and even though today we are only able to reliably distinguish clean speech signals from all other sound kinds, practical experiences have shown that an automatic program selection system in the hearing instrument is appreciated very much by the user [1].

Our goal is therefore to reliably and robustly classify the complete spectrum of the acoustic environment in order to allow for an automatic program selection by the hearing instrument itself. So far we aim at distinguishing four sound classes: 'clean speech', 'speech in noise', 'noise', and 'music'. Our approach to sound classification is inspired by the human auditory system in that we extract auditory features as known from auditory scene analysis from the input signal. The individual sound classes are modeled by means of HMMs.

The remainder of this paper is structured as follows. Section 2 gives an overview on existing work on sound classification with a particular focus on hearing instrument applications. In section 3 a short overview on auditory scene analysis is given, and then our sound classification system comprised of feature extraction and classifier is described in detail. The achieved results are presented and discussed in section 4, and the conclusions of the present work are drawn in section 5.

## 2. Prior Art

Nowadays there exist already a few commercial hearing instruments which make use of sound classification techniques. *Widex* exploits Ludvigsen's amplitude statistics [2] for the differentiation of impulse-like sounds from continuous sounds in a noise canceler. Classification of the acoustic environment is also used at *Siemens* in the context of noise canceling. There the three sound kinds 'clean speech', 'speech in noise', and 'noise without speech' are distinguished by means of modulation frequency analysis [3] and Bayes classification [4]. A similar approach which is based on modulation analysis as well is utilized by *ReSound* for the detection of speech in a multiband noise canceler [5]. Sound classification is also used at *Phonak*, however not for noise canceling but for enabling the automatic program selection based on the actual acoustic environment. For this purpose a modified version [6] of the analysis of the temporal fluctuations and the spectrum as proposed by Kates [7] is used. This allows to reliably distinguish clean speech signals from all other sound kinds.

In the hearing instrument research community two further sound classification systems are known, which however, to the best of our knowledge, have not (yet) been applied in commercial hearing aids. In the approach of Nordqvist the sound is classified into clean speech and different kinds of noises by means of LPC coefficients and HMMs [8]. Feldbusch identifies clean speech, speech babble, and traffic noise by means of various time- and frequency-domain features and a neural network [9]. All of the above mentioned approaches allow a robust separation of clean speech signals from other signals. Music however can not be distinguished at all, and it is only partly possible to separate noise from speech in noise.

Another application of sound classification which has recently gained importance is the automatic data segmentation and indexing in multimedia databases. A nice example is the system of Zhang and Kuo where the audio signal is segmented and classified into twelve essential scenes using four signal features and a rule-based heuristic procedure extended by a HMM stage [10].

Other typical sound classification systems operate usually on much less universal target signals than the above mentioned applications. Examples of such systems are the recognition of different music styles and the identification of different instruments, the differentiation of speech and music signals, or the classification of different noise and alarm types.

### 3. Sound Classification Inspired by Auditory Scene Analysis

The basic structure of each sound classification system is illustrated in figure 1. The classifier separates the desired classes based on the features extracted from the input signal. Postprocessing is employed to correct possible classification errors and to control the transient behavior of the sound classification system.



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

#### 3.1. Auditory Scene Analysis

Auditory scene analysis (ASA) [11] describes mechanisms and processing strategies that the auditory system relies on in the analysis of the acoustic environment. Although this whole process is not yet completely understood, it is known that the auditory system extracts characteristic features from the acoustic signals. The features are analyzed based on grouping rules and possibly also on prior knowledge and hypotheses to form acoustic events. These events are then combined respectively segregated into multiple sound sources.

The features which are known to play a key role in auditory grouping, the so called auditory features, are *spectral separation*, *spectral profile*, *harmonicity*, *onsets* and *offsets*, *coherent amplitude* and *frequency variations*, *spatial separation*, and *temporal separation*. For more details on auditory features in particular and auditory scene analysis in general the reader is referred to the literature ([12] [13]).

Note that the auditory system attempts to separate and identify the individual sound sources, whereas our primary goal in sound classification is the discrimination of different sources which does not necessarily require their identification.

#### 3.2. Features for Sound Classification

In our approach to sound classification our aim is to mimic the human auditory system at least partially by making use of auditory features as known from auditory scene analysis. We are using three auditory feature groups so far: amplitude variations, spectral profile, and harmonicity.

Our *amplitude variations* are characterized by the range of the occurring sound levels. For this purpose we model the amplitude histogram of the sounds by means of percentiles. The width of the amplitude histogram is then used to characterize the modulation depth in the signal. This concept is illustrated in figure 2. A similar kind of amplitude statistics was already used by Ludvigsen for the differentiation of impulse-like sounds from continuous sounds [2].

The *spectral profile* is modeled in a very rudimentary way by means of two features, the spectral center of gravity, and the fluctuations of the spectral center of gravity. 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. Frequency information is extracted from a 20-channel Bark spectrum. These two features have already been employed in

an earlier sound classifier of Phonak [6]. Figure 3 illustrates the approach.

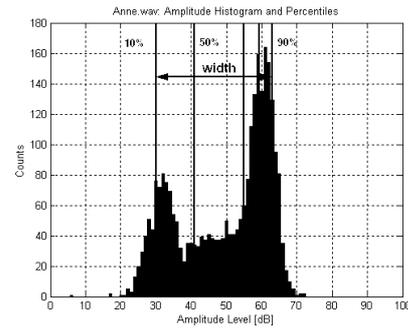


Figure 2: Characterization of the level fluctuations by means of amplitude statistics. The amplitude histogram is approximated by several percentiles. The width of the amplitude histogram describes the amplitude variations of the signal.

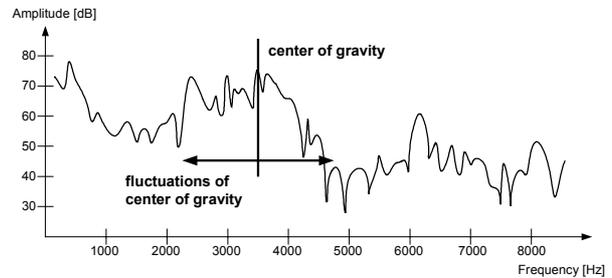


Figure 3: Spectral center of gravity and fluctuations of the spectral center of gravity as employed in the description of the spectral profile.

To describe *harmonicity* the pitch of the sound is usually employed. In our approach we characterize harmonicity by two features, the tonality of the sound on the one hand, and the pitch variance on the other hand. The tonality is defined as the ratio of harmonic to unharmonic parts in the sound over time. Note that the pitch value itself is not employed although it is used to extract both harmonicity features. Figure 4 illustrates the tonality and pitch variance of a sample sound.

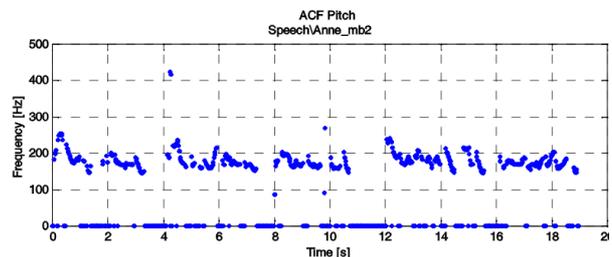


Figure 4: Pitch course of a speech sound. Where a pitch exists the sound is harmonic, otherwise, where no pitch can be extracted (indicated by 0 Hz) the sound is defined as unharmonic. Tonality is the ratio of harmonic to unharmonic periods over time. Pitch variance refers to the harmonic parts of the sound only.

### 3.3. Classifier

For the classifier we have evaluated two approaches: simple Bayes classification and HMM classification. The Bayes classifier is easy to train and very efficient, however, it can only account for static properties of the extracted features. The HMMs are more complex to train, but they allow to take into account to a certain extent also the dynamics of the features. In the following we concentrate on HMM classification since in our case it provides more robust results than the Bayes classification, especially for our sound class noise.

The determination of the parameters of the individual HMMs (“HMM training”) is carried out off-line by means of the freely available software tool HTK [14]. The sounds for training are taken from a large sound database collected at Phonak which comprises approximately 800 real-world sounds of 30 seconds length each and sampled at 22 kHz. All of the four sound classes that we wish to identify are represented with various examples. The sounds were either recorded in the real world (e.g. in a train station) or in our own sound-proof room, or taken from other media. The class ‘clean speech’ is comprised of different speakers speaking different languages, with different vocal effort, at different speed, and with different amounts of reverberation. The class ‘noise’ is the most widely varying sound class, comprising social noises, traffic noise, industrial noise and various other noises such as e.g. household and office noises. ‘Speech in noise’ sounds consist of speech signals mixed with noise signals at different SNRs. The class ‘music’, finally, comprises music styles from classics over pop and rock up to single instruments and singing. The original labeling of the sounds into the four sound classes has been carried out by hand based on listening. From this ample database, 287 selected sounds were employed for training.

The classification task itself is carried out by means of the Viterbi algorithm (see e.g. [15]). For a sound to be classified the likelihood of every model is computed, and the unknown sound is assigned the class of the model with the highest likelihood.

### 3.4. Postprocessing

As mentioned earlier, the purpose of postprocessing is to correct possible classification outliers and to control the transient behavior of the sound classification system. This is achieved in a very simple manner by observing the classification outcomes over a certain time and taking as a result the class which has occurred most often. By varying the length of the observation interval the transient behavior of the classification result is controlled.

## 4. Results and Discussion

The performance of the classification system is tested by classifying approximately 500 different real-world sounds from our sound database and comparing the resulting class to the known class of the sound. Note that only sounds which were not used for training are employed for testing. In this way we prevent a falsification of the result which could arise when the classifier “learns” particular sounds.

The performance is expressed by hit and false alarm rates and graphically represented in the form of receiver operating curves (ROC graphs). The hit rate is defined as the percent of correctly recognized sounds of a particular class, the corre-

sponding false alarm rate is the percentage of sounds which were falsely classified as this class.

The ROC graph of figure 5 shows the hit rate of our four classes on the ordinate and the corresponding false alarm rates on the abscissa. The results that we achieve so far are very promising. ‘Clean speech’ is identified with a hit rate of over 90% and a false alarm rate of less than 10%. Both ‘speech in noise’ and ‘music’ files were identified with hit rates of 80% and false alarm rates of approximately 7.8%. ‘Noise’ signals are identified with a hit rate of 65% and a false alarm rate of 10%.

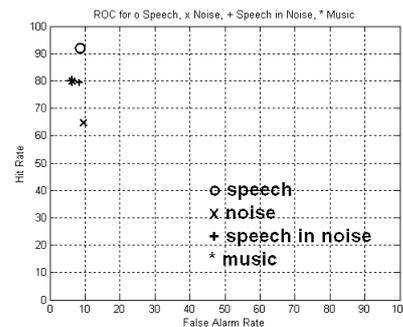


Figure 5: ROC graph for HMM classifier. The hit rate of each class is represented in percent on the ordinate, the corresponding false alarm rates are shown on the abscissa.

The relative low hit rate of only 65% for ‘noise’ signals is due to the wide variety of the different sounds comprised in the class ‘noise’, including signals such as e.g. “in a restaurant”, “in a car”, or “while shaving”. Nevertheless there is definitely room for improvement in the classification of ‘noise’ signals.

The following misclassifications occur repeatedly in our classification system. Pop music is often classified as ‘speech in noise’. Depending on the taste of the listener, this classification might however correspond to perception! Reverberated speech might also be classified as ‘speech in noise’. But also there, depending on the amount of reverberation the classifier might give a result similar to our perception. Other problematic sounds are tonal noises such as e.g. the sound of a shaver. Due to the high tonality of such signals they might be falsely classified as ‘music’. Singing, on the other hand, depending on the amount of pitch variation, is sometimes classified as ‘clean speech’ instead of ‘music’.

From the above discussion the fundamental limitations of any artificial sound classification system become obvious. On the one hand, the performance of a classifier is very much dependent on the employed training data. The training data must be chosen carefully to cover the whole range of each sound class homogeneously. This is however not always easy, in particular for classes covering a wide range of possible signals as is the case for our classes ‘noise’ and ‘music’. Also there are signals which are inherently ambiguous. Shall strongly reverberated speech, for example, be labeled as ‘clean speech’ or as ‘speech in noise’ when employed for training? These pre-labels influence 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 e.g. 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 training 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 noise (the listener is talking to somebody). No artificial classifier can read the listener’s mind, and therefore there exist always ambiguities in classification.

## 5. Summary and Conclusions

For the general classification of the acoustic environment in a hearing aid application we aim at distinguishing the four main sound classes *clean speech*, *speech in noise*, *noise*, and *music*. For this purpose we extract auditory features from the acoustic signal and classify them by means of HMMs. The employed features describe level fluctuations, the spectral form and harmonicity. Postprocessing allows to clean the classification result and to control the transient behavior of the classifier. Training of the HMMs is carried out by using sounds from an extensive sound database.

The results achieved so far are promising. All sound classes except the class noise are identified with hit rates over 80%. Only for noise signals the hit rate is with 65% considerably lower. The false alarm rates are below 10% for each class. The relative poor performance of the classifier on noise signals is due to the inhomogeneity of the class which comprises a wide range of very different signals. Common classification errors include misclassifications of pop music and reverberated speech into speech in noise, tonal noises into music, and singing into clean speech.

With this and other recent work on sound classification, progress is being made towards automatic and robust classification of the acoustic environment. However, we are far from achieving similar performance as our auditory system. Today’s limitations are – apart from our still lacking understanding of all of the auditory processes per se – the ambiguity and context-dependence of a large part of the acoustic situations.

## 6. References

- [1] Büchler, M., “Nützlichkeit und Akzeptanz einer automatischen Programmwahl in Hörgeräten”, *Focus 27*, Phonak AG, 2001.
- [2] Ludvigsen, C., “Schaltungsanordnung für die automatische Regelung von Hörhilfsgeräten”, *European Patent EP 0 732 036 B1*, 1997.
- [3] Ostendorf, M., Hohmann, V., Kollmeier, B., “Empirische Klassifizierung verschiedener akustischer Signale und Sprache mittels einer Modulationsfrequenzanalyse”, *DAGA '97*, 608-607, 1997.
- [4] Ostendorf, M., Hohmann, V., Kollmeier, B., “Klassifikation von akustischen Signalen basierend auf der Analyse von Modulationsspektren zur Anwendung in digitalen Hörgeräten”, *DAGA '98*, 402-403, 1998.
- [5] Edwards, B. W. et al., “Signal-Processing Algorithms for a New Software-Based Digital Hearing Device”, *The Hearing Journal*, vol. 51, 44-52, 1998.
- [6] Phonak AG, *Claro AutoSelect*, 28148(GB) / 0300, 1999.
- [7] Kates, J. M., “Classification of background noises for hearing-aid applications”, *Journal of the Acoustical Society of America*, vol. 97, 461-469, 1995.
- [8] Nordqvist, P., “Automatic classification of different listening environments in a generalized adaptive hearing aid”, *International Hearing Aid Research Conference*, Lake Tahoe, CA, 2000.
- [9] Feldbusch, F., “A Heuristic for Feature Selection for the Classification with Neural Nets”, *Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, Vancouver, 2001.
- [10] Zhang, T., Kuo, C.-C. J., “Audio Content Analysis for Online Audiovisual Data Segmentation and Classification”, *IEEE Trans. Speech and Audio Proc.*, vol. 9, no. 4, 441-457, 2001.
- [11] Bregman, A. S., *Auditory Scene Analysis*, MIT Press, Cambridge, 1990.
- [12] Mellinger, D. K., Mont-Reynaud, B. M., “Scene Analysis”, in *Auditory Computation*, edited by H. L. Hawkins et al., Springer, New York, 1996.
- [13] Yost, W. A., “Auditory Image Perception and Analysis: the Basis for Hearing”, *Hear. Res.*, vol. 56, 8-18, 1991.
- [14] Young, S. et al., *The HTK Handbook*, Entropic Ltd., version 2.2, 1999.
- [15] Rabiner, L. R., Juang, B. H., “An Introduction to Hidden Markov Models”, *IEEE ASSP Magazine*, January 1986.