Sound Classification in Hearing Instruments

PETER NORDQVIST

Doctoral Thesis
Stockholm 2004
Akademisk avhandling som med tillstånd av Kungl. Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie doktorexamen torsdagen den 3 juni kl. 10.00 i Kollegiesalen, Administrationsbyggnaden, Kungl. Tekniska högskolan, Valhallavägen 79, Stockholm.

© Peter Nordqvist, juni 2004
Tryck: Universitetsservice US AB
Abstract

A variety of algorithms intended for the new generation of hearing aids is presented in this thesis. The main contribution of this work is the hidden Markov model (HMM) approach to classifying listening environments. This method is efficient and robust and well suited for hearing aid applications. This thesis shows that several advanced classification methods can be implemented in digital hearing aids with reasonable requirements on memory and calculation resources.

A method for analyzing complex hearing aid algorithms is presented. Data from each hearing aid and listening environment is displayed in three different forms: (1) Effective temporal characteristics (Gain-Time), (2) Effective compression characteristics (Input-Output), and (3) Effective frequency response (Insertion Gain). The method works as intended. Changes in the behavior of a hearing aid can be seen under realistic listening conditions. It is possible that the proposed method of analyzing hearing instruments generates too much information for the user.

An automatic gain controlled (AGC) hearing aid algorithm adapting to two sound sources in the listening environment is presented. The main idea of this algorithm is to: (1) adapt slowly (in approximately 10 seconds) to varying listening environments, e.g. when the user leaves a disciplined conference for a multi-babble coffee-break; (2) switch rapidly (in about 100 ms) between different dominant sound sources within one listening situation, such as the change from the user’s own voice to a distant speaker’s voice in a quiet conference room; (3) instantly reduce gain for strong transient sounds and then quickly return to the previous gain setting; and (4) not change the gain in silent pauses but instead keep the gain setting of the previous sound source. An acoustic evaluation shows that the algorithm works as intended.

A system for listening environment classification in hearing aids is also presented. The task is to automatically classify three different listening environments: ‘speech in quiet’, ‘speech in traffic’, and ‘speech in babble’. The study shows that the three listening environments can be robustly classified at a variety of signal-to-noise ratios with only a small set of pre-trained source HMMs. The measured classification hit rate was 96.7-99.5% when the classifier was tested with sounds representing one of the three environment categories included in the classifier. False alarm rates were 0.2-1.7% in these tests. The study also shows that the system can be implemented with the available resources in today’s digital hearing aids.

Another implementation of the classifier shows that it is possible to automatically detect when the person wearing the hearing aid uses the
telephone. It is demonstrated that future hearing aids may be able to
distinguish between the sound of a face-to-face conversation and a
telephone conversation, both in noisy and quiet surroundings. However,
this classification algorithm alone may not be fast enough to prevent initial
feedback problems when the user places the telephone handset at the ear.

A method using the classifier result for estimating signal and noise
spectra for different listening environments is presented. This evaluation
shows that it is possible to robustly estimate signal and noise spectra given
that the classifier has good performance.

An implementation and an evaluation of a single keyword recognizer for
a hearing instrument are presented. The performance for the best
parameter setting gives 7e-5 [1/s] in false alarm rate, i.e. one false alarm for
every four hours of continuous speech from the user, 100% hit rate for an
indoors quiet environment, 71% hit rate for an outdoors/traffic
environment and 50% hit rate for a babble noise environment. The
memory resource needed for the implemented system is estimated to 1820
words (16-bits). Optimization of the algorithm together with improved
technology will inevitably make it possible to implement the system in a
digital hearing aid within the next couple of years. A solution to extend the
number of keywords and integrate the system with a sound environment
classifier is also outlined.
Acknowledgements

I was introduced to the hearing aid area in 1996 by my supervisor Arne Leijon and since then we have worked together in several interesting projects. The time, energy, enthusiasm and theoretical clearness he has provided in this cooperation through the years are outstanding and much more than one can demand as a student. All the travels and events that have occurred during this period are memories for life.

This thesis would not have been possible without the professional relationship with the hearing aid company GN ReSound. They have provided knowledge and top of the line research equipment for testing and implementing new algorithms. Thanks especially to René Mortensen, Ole Dyrlund, Chas Pavlovic, Jim Kates, Brent Edwards, Bert de Vries, and Thomas Beierholm.

My upbringing as a researcher has also been formed by the people in the hearing group at KTH. The warm atmosphere and the social activities have been an excellent counterbalance to the hard work. I am particularly grateful to Karl-Erik Spens, Martin Dahlquist, Karolina Smeds, Anne-Marie Öster, Eva Agelfors, and Elisabeth Molin.

The Department of Speech, Music and Hearing has been an encouraging environment as a working place. The mix of different disciplines under the same roof generates a unique working environment.

Thanks to my friends for providing oases of relaxation.

I also want to thank my parents and brothers for not understanding anything about my research.

And last, thanks to my wife-to-be Eva-Lena for all the support she has given, with language and other more mysterious issues. I am also very grateful for the understanding that when I lack in interaction and response during sudden speech communication sessions it is only due to minor absent-mindedness and nothing else.
## Contents

I. Symbols 1

II. Abbreviations 2

III. Included papers 3

IV. Contributions 4

V. Introduction 5
   A. The peripheral auditory system 8
   B. Hearing loss 10
   C. The hearing aid 11
      1. Signal processing 12
      2. Fitting strategies 13
   D. Classification 14
      1. Bayesian classifier 16
      2. Hidden Markov models 18
      3. Vector quantizer 21
      4. Sound classification 21
      5. Keyword recognition 22

VI. Methods 23
   A. Platforms 23
   B. Signal processing 23
   C. Recordings 24
   D. Classifier implementation 26
      1. Feature extraction 27
      2. Sound feature classification 28
      3. Sound source classification 28
      4. Listening environment classification 30
      5. Feature extraction analysis 33
      6. Vector quantizer analysis after training 35
      7. HMM analysis after training 36

VII. Results and discussion 39
   A. Paper A 39
   B. Paper B 40
   C. Paper C and D 41
      1. Signal and noise spectrum estimation 41
   D. Paper E 44

VIII. Conclusion 45

IX. About the papers 47

X. Bibliography 49
I. Symbols

\( p^0 \) Vector are lower case and bold
\( \mathbf{A} \) Matrices are upper case and bold
\( \mathbf{A}_{i,j} \) Element in row \( i \) and column \( j \) of matrix \( \mathbf{A} \)
\( j, N \) Scalar variables are italic
\( Z \) Stochastic scalar variables are upper case and italic
\( \mathbf{X} \) Stochastic vectors are upper case and bold
\( N, Q \) State numbers
\( \mathbf{X} \) Observation vector
\( Z \) Observation index
\( M \) Number of observations
\( L \) Number of features
\( B \) Sample blocksize
\( \lambda \) Hidden Markov model
\( \mathbf{A} \) State transition matrix
\( \mathbf{B} \) Observation probability matrix
\( p^0 \) Initial state distribution
## II. Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>VQ</td>
<td>Vector Quantizer</td>
</tr>
<tr>
<td>BTE</td>
<td>Behind the Ear</td>
</tr>
<tr>
<td>ITC</td>
<td>In the Canal</td>
</tr>
<tr>
<td>CIC</td>
<td>Completely in the Canal</td>
</tr>
<tr>
<td>ITE</td>
<td>In the Ear</td>
</tr>
<tr>
<td>DSP</td>
<td>Digital Signal Processing</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-Noise Ratio</td>
</tr>
<tr>
<td>SPL</td>
<td>Sound Pressure Level, mean dB re. 20μPa</td>
</tr>
<tr>
<td>SII</td>
<td>Speech Intelligibility Index</td>
</tr>
<tr>
<td>MAP</td>
<td>Maximum a Posteriori Probability</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td>PCM</td>
<td>Pulse Code Modulated</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform</td>
</tr>
<tr>
<td>ADC</td>
<td>Analog to Digital Converter</td>
</tr>
<tr>
<td>DAC</td>
<td>Digital to Analog Converter</td>
</tr>
<tr>
<td>AGC</td>
<td>Automatic Gain Control</td>
</tr>
<tr>
<td>IHC</td>
<td>Inner Hair Cell</td>
</tr>
<tr>
<td>OHC</td>
<td>Outer Hair Cell</td>
</tr>
<tr>
<td>CT</td>
<td>Compression Threshold</td>
</tr>
<tr>
<td>CR</td>
<td>Compression Ratio</td>
</tr>
</tbody>
</table>
III. Included papers

(A) Nordqvist, P. (2000). “The behaviour of non-linear (WDRC) hearing instruments under realistic simulated listening conditions,” QPSR Vol 40, 65-68. (This work was also presented as a poster at the conference “Issues in advanced hearing aid research”, Lake Arrowhead, 1998.)

(B) Nordqvist, P. and Leijon, A. (2003). “Hearing-aid automatic gain control adapting to two sound sources in the environment, using three time constants,” submitted to JASA.


(D) Nordqvist, P. and Leijon, A. (2002). “Automatic classification of the telephone listening environment in a hearing aid,” QPSR, Vol 43, 45-49. (This work was also presented as a poster at the 141st Meeting Acoustic Society of America, Chicago, Vol. 109, No. 5, p. 2491, May 2001.)

IV. Contributions

This thesis shows that advanced classification methods can be implemented in digital hearing aids with reasonable requirements on memory and calculation resources. The author has implemented an algorithm for acoustic classification of listening environments. The method is based on hierarchical hidden Markov models that have made it possible to train and implement a robust classifier that works in listening environments with a large variety of signal-to-noise ratios, without having to train models for each listening environment. The presented algorithm has been evaluated with a large amount of sound material.

The author has also implemented a method for speech recognition in hearing instruments, making it possible for the user to control the hearing aid with spoken keywords. This method also works in different listening environments without the necessity for training models for each environment. The problem is to avoid false alarms while the hearing aid continuously listens for the keyword.

The main contribution of this thesis is the hidden Markov model approach for classifying listening environments. This method is efficient and robust and well suitable for hearing aid applications.
V. Introduction

The human population of the earth is growing and the age distribution is shifting toward higher ages in the developed countries. Around half a million people in Sweden have a moderate hearing loss. The number of people with mild hearing loss is 1.3 million. It is estimated that there are 560,000 people who would benefit from using a hearing aid, in addition to the 270,000 who already use hearing aids. These relatively high numbers are estimated in Sweden (SBU, 2003), a country with only nine million inhabitants. Clearly, hearing loss is a very common affliction among the population. Hearing loss and treatment of hearing loss are important research topics. Most of the people with hearing loss have mild or moderate hearing loss that can be treated with hearing aids. This category of people is also sensitive to sound quality and the behavior of the hearing aid.

Satisfaction with hearing aids has been examined by Kochkin in several studies, (Kochkin, 1993), (Kochkin, 2000), (Kochkin, 2001), (Kochkin, 2002), and (Kochkin, 2002). These studies showed the following: Only about 50-60 percent of the hearing aid users are satisfied with their hearing instruments. Better speech intelligibility in listening environments containing disturbing noise is considered by 95 percent of the hearing aid users as the most important area where improvements are necessary. More than 80 percent of the hearing aid users would like to have improved speech intelligibility for listening to speech in quiet and when using a telephone. The research for better and more advanced hearing aids is an important area that should increase in the future.

The digital hearing aid came out on the market in the middle of the nineties. Since then we have seen a rapid development towards smaller and more powerful signal processing units in the hearing aids. This development has made it possible to implement the first generation of advanced algorithms in the hearing aid, e.g. feedback suppression, noise reduction, basic listening environment classification, and directionality. The hardware development will continue and in the future it will be possible to implement the next generation of hearing aid algorithms, e.g. advanced listening environment classification, speech recognition and speech synthesis. In the future there will be fewer practical resource limitations (memory and calculation power) in the hearing aid and focus is going to change from hardware to software.

A comparison between modern analog hearing aids and today’s digital hearing aids shows that the difference in overall performance is remarkably small (Arlinger and Billermark, 1999; Arlinger, et al., 1998; Newman and
Sandridge, 2001; Walden, et al., 2000). Another study shows that the benefit of the new technology to the users in terms of speech perception is remarkably small (Bentler and Duve, 2000). Yet another investigation showed that different types of automatic gain control (AGC) systems implemented in hearing aids had similar speech perception performance (Stone, et al., 1999). Many investigations on noise reduction techniques show that the improvement in speech intelligibility is insignificant or very small (Dillon and Lovegrove, 1993; Elberling, et al., 1993; Ludvigsen, et al., 1993). The room left for improvements in the hearing aid seems to be very small when referring to speech intelligibility. This indicates the need for more research. Other aspects should be considered when designing new technology and circuitry for the hearing aids, e.g. sound quality, intelligent behavior, and adaptation to the user’s preferences.

This thesis is an attempt to develop more intelligent algorithms for the future hearing aid. The work is divided into four parts.

The first, paper A, is an analysis part, where the behavior of some of the current hearing aids on the market are analyzed. The goal of this first part is to determine what the hearing aids do, in order to get a starting point from which this thesis should continue. Since some of the hearing aids are multi-channel non-linear wide dynamic range hearing aids, it is not possible to use the traditional methods to measure the behavior of these hearing aids. Instead, an alternative method is proposed in order to determine the functionality of the hearing aids.

The second part, paper B, is a modification of an automatic gain controlled algorithm. This is a proposal on how to implement a more intelligent behavior in a hearing aid. The main idea of the algorithm is to include all the positive qualities of an automatic gain controlled hearing aid and at the same time avoid negative effects, e.g. so called pumping, in situations where the dominant sound alters between a strong and a weaker sound source.

The third part, paper C and D, is focused on listening environment classification in hearing aids. Listening environment classification is considered as the next generation of hearing aid algorithms, which makes the hearing aid “aware” of the surroundings. This feature can be used to control all other functions of a hearing aid. It is possible to implement sound environment classification with the technology available today, although sound environment classification is more complex to implement than traditional algorithms.

The fourth part of the thesis, paper E, introduces speech recognition in hearing aids. This is a future technology that allows the user to interact with the hearing aid, e.g. support an automatic listening environment classification algorithm or control other features in the hearing aid. This
technology is resource consuming and not yet possible to implement with the resources available in today’s hearing aids. However, it will soon be possible, and this part of the thesis discusses possibilities, problems, and limitations of this technology.
A. The peripheral auditory system

A healthy hearing system organ is an impressive sensory organ that can detect tones between 20 Hz and 20 kHz and has a dynamic range of about 100 dB between the hearing threshold and the level of discomfort. The hearing organ comprises the external ear, the middle ear and the inner ear (Figure 1). The external ear including the ear canal works as a passive acoustic amplifier. The sound pressure level at the eardrum is 5-20 dB (2 000-5 000 Hz) higher than the free field sound pressure level outside the outer ear (Shaw, 1974). Due to the shape of the outer ear, sounds coming approximately from in front of the head are amplified more than sounds coming from other directions.
The task of the middle ear is to work as an impedance converter and transfer the vibrations in the air to the liquid in the inner ear. The middle ear consists of three bones; malleus (which is attached to the eardrum), incus, and stapes (which is attached to the oval window). Two muscles, tensor tympani and stapedius (the smallest muscle in the body), are used to stabilize the bones. These two muscles can also contract and protect the inner ear when exposed to strong transient sounds.

The inner ear is a system called the osseus, or the bony labyrinth, with canals and cavities filled with liquid. From a hearing perspective, the most interesting part of the inner ear is the snail shell shaped cochlea, where the sound vibrations from the oval window are received and transmitted further into the neural system. The other parts of the inner ear are the semicircular canals and the vestibule. The cochlea contains about 12,000 outer hair cells (OHC) and about 3,500 inner hair cells (IHC) (Ulehlova, et al., 1987). The hair cells are placed along a 35 mm long area from the base of the cochlea to the apex. Each inner hair cell is connected to several neurons in the main auditory nerve. The vibrations in the fluid generated at the oval window causes the basilar membrane to move in a waveform pattern.

The electro-chemical state in the inner hair cells is correlated with the amplitude of the basilar membrane movement. Higher amplitude of the basilar membrane movement generates a higher firing rate in the neurons. High frequency sounds stimulate the basilar membrane more at the base while low frequency sounds stimulate the basilar membrane more at the apex. The movement characteristic of the basilar membrane is non-linear; it is more sensitive to weak sounds than to stronger sounds (Robles, et al., 1986). Tuning curves are narrower at low intensities than at high intensities. The non-linear characteristic of the basilar membrane is caused by the outer hair cells, which amplify small movements of the membrane (Ulfendahl, 1997).

Auditory neurons have a spontaneous nerve firing rate of three different types: low, medium, and high (Liberman, 1978). The different types of neurons and the OHC non-linearity make it possible to map a large dynamic range in sound pressure levels into firing rate, e.g. when a neuron with high spontaneous activity is saturated, a neuron with medium or low spontaneous activity in the same region is within its dynamic range and vice versa.
B. Hearing loss

There are two types of hearing loss: conductive and sensorineural. They can appear isolated or simultaneously. A problem that causes a hearing loss outside the cochlea is called a conductive hearing loss, and damage to the cochlea or the auditory nerve is referred to as a sensorineural hearing loss. Abnormalities at the eardrum, wax in the ear canal, injuries to bones in the middle ear, or inflammation in the middle ear can cause a conductive loss. A conductive loss causes a deteriorated impedance conversion between the eardrum and the oval window in the middle ear. This non-normal attenuation in the middle ear is linear and frequency dependent and can be compensated for with a linear hearing aid. Many conductive losses can be treated medically. There are also temporary hearing losses, which will heal automatically. For example, ear infections are the most common cause of temporary hearing loss in children.

A more problematic impairment is the sensorineural hearing loss. This includes damage to the inner and outer hair cells or abnormalities of the auditory nerve. Acoustic trauma, drugs or infections can cause a cochlear sensorineural hearing loss. A sensorineural hearing loss can also be congenital. Furthermore, it is usually permanent and cannot be treated. An outer hair cell dysfunction causes a reduced frequency selectivity and low sensitivity to weak sounds. Damage to the outer hair cells causes changes in the input/output characteristics of the basilar membrane movement resulting in a smaller dynamic range (Moore, 1996). This is the main reason for using automatic gain control in the hearing aid (Moore, et al., 1992).

Impairment due to ageing is the most common hearing loss and is called presbyacusis. Presbyacusis is a slowly growing permanent damage to hair cells. The National Institute of Deafness and Other Communication Disorders claims that about 30-35 percent of adults between the ages of 65 and 75 years have a hearing loss. It is estimated that 40-50 percent of people at age 75 and older have a hearing loss.

A hearing loss due to abnormalities in the hearing nerve is often caused by a tumor attached to the nerve.
C. The hearing aid

There are four common types of hearing aid models:

- in the canal (ITC)
- completely in the canal (CIC)
- in the ear (ITE)
- behind the ear (BTE).

The BTE hearing aid has the largest physical size. The CIC hearing aid and the ITC hearing aid are becoming more popular since they are small and can be hidden in the ear. A hearing aid comprises at least a microphone, an amplifier, a receiver, an ear mold, and a battery. Modern hearing aids are digital and also include an analog to digital converter (ADC), a digital signal processor (DSP), a digital to analog converter (DAC), and a memory (Figure 2). Larger models, ITE and BTE, also include a telecoil in order to work in locations with induction-loop support. Some hearing aids have directionality features, which mean that a directional microphone or several microphones are used.

![Figure 2 Overview of the components included in a modern BTE digital hearing aid.](image)
1. Signal processing

Various signal processing techniques are used in the hearing aids. Traditionally, the analogue hearing aid was linear or automatic gain controlled in one or a few frequency bands. Automatic gain control or compression is a central technique for compensating a sensorineural hearing loss. Damage to the outer hair cells reduces the dynamic range in the ear and a compressor can restore the dynamic range to some extent. A compressor is characterized mainly by its compression threshold (CT) (knee-point) and its compression ratio (CR). A signal with a sound pressure level below the threshold is linear amplified and a signal above is compressed. The compression ratio is the change in input to an aid versus the subsequent change in output.

When the digital hearing aid was introduced, more signal processing possibilities became available. Nowadays, hearing aids usually include feedback suppression, multi-band automatic gain control, beam forming, and transient reduction. Various noise suppression methods are also used in some of the hearing aids. The general opinion is that single channel noise reduction, where only one microphone is used, mainly improves sound quality, while effects on speech intelligibility are usually negative (Levitt, et al., 1993). A multi channel noise reduction approach, where two or more microphones are used, can improve both sound quality and speech intelligibility.

A new signal-processing scheme that is becoming popular is listening environment classification. An automatic listening environment controlled hearing aid can automatically switch between different behaviors for different listening environments according to the user’s preferences.

Command word recognition is a feature that will be used in future hearing aids. Instead of using a remote control or buttons, the hearing aid can be controlled with keywords from the user.

Another topic is communication through audio protocols, implemented with a small amount of resources, between hearing aids when used in pairs. The information from the two hearing aids can be combined to increase the directionality, synchronize the volume control, or to improve other features such as noise reduction.
2. Fitting strategies

A hearing aid must be fitted according to the hearing loss in order to maximize the benefit to the user. A fitting consists of two parts: prescription and fine-tuning. The prescription is usually based on very basic measurements from the patient, e.g. pure tone thresholds and uncomfortable levels. The measured data are used to calculate a preliminary gain prescription for the patient. The first prescription method used was the half gain rule (Berger, et al., 1980). The principle was simply to use the pure tone thresholds divided by two to determine the amount of amplification in the hearing aid. Modern prescription methods are more complicated and there are many differing opinions on how to design an optimal prescription method. The optimization criteria used in the calculation of the prescription method vary between the methods. One strategy is to optimize the prescription with respect to speech intelligibility. Speech intelligibility can be estimated by means of a speech test or through direct calculation of the speech intelligibility index (SII) (ANSI-S3.5, 1997).

The maximization curve with respect to speech intelligibility appears to be a very ‘flat’ function (Gabrielsson, et al., 1988) and many different gain settings may have similar speech intelligibility in noise (van Buuren, et al., 1995). It could be argued that it is not that important to search for the optimal prescription, since a patient may be satisfied with a broad range of different prescriptions. That is not true, however - even if many different prescriptions give the same speech intelligibility, subjects can clearly choose a favorite among them (Smeds, 2004).

The prescription is used as an initial setting of the hearing aid. The next stage of the fitting process is the fine-tuning. The fine-tuning is an iterative process based on purely subjective feedback from the user. The majority of the modern fitting methods consist of a combination of initial prescription and final fine-tuning.
D. Classification

A classification system consists of three components (Figure 3): a transducer that records and transforms a signal pattern, a feature extractor that removes irrelevant information in the signal pattern, and a classifier that takes a decision based on the extracted features. In sound classification, the transducer is a microphone that converts the variations in sound pressure into an electrical signal. The electrical signal is sampled and converted into a stream of integers. The integers are gathered into overlapping or non-overlapping frames, which are processed with a digital signal processor. Since the pure audio signal is computationally demanding to use directly, the relevant information is extracted from the signal.

Different features can be used in audio classification, (Scheirer and Slaney, 1997), in order to extract the important information in the signal. There are mainly two different classes of features: absolute features and relative features.

Absolute features, e.g. spectrum shape or sound pressure level, are often used in speech recognition where sound is recorded under relatively controlled conditions, e.g. where the distance between the speaker and the microphone is constant and the background noise is relatively low. A hearing aid microphone records sound under conditions that are not controlled. The distance to the sound source and the average long term spectrum shape may vary. Obstacles at the microphone inlet, changes in battery levels, or wind noise are also events that a hearing aid must be robust against. In listening environment classification, the distance to the microphone and the spectrum shape can vary for sounds that are subjectively classified into the same class, e.g. a babble noise with a slightly changed spectrum slope is still subjectively perceived as babble noise although the long-term spectrum has changed. One must therefore be very
careful when using absolute features in a hearing aid listening environment application.

A better solution is to use relative features that are more robust and that can handle any change in absolute long-term behavior.

A feature extractor module can use one type of features or a mix of different features. Some of the available features in sound classification are listed in Table 1.

<table>
<thead>
<tr>
<th>Features</th>
<th>Time Domain</th>
<th>Frequency Domain</th>
<th>Absolute Feature</th>
<th>Relative Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Zero Crossing</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>2. Sound Pressure Level</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>3. Spectral Centroid</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>4. Modulation Spectrum</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>5. Cepstrum</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>6. Delta Cepstrum</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>7. Delta-Delta Cepstrum</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

Table 1 Sound Classification Features

1. The zero crossing feature indicates how many times the signal crosses the signal level zero in a block of data. This is an estimate of the dominant frequency in the signal (Kedem, 1986).

2. The sound pressure level estimates the sound pressure level in the current signal block.

3. The spectral centroid estimates the “center of gravity” in the magnitude spectrum. The brightness of a sound is characterized by this feature. The spectral centroid is calculated as (Beauchamp, 1982):

\[
SP = \frac{\sum_{b=1}^{B-1} bX_b}{\sum_{b=1}^{B-1} X_b}
\]  

(1)

where \(X_b\) are the magnitude of the FFT bins calculated from a block of input sound samples.

4. The modulation spectrum is useful when discriminating speech from other sounds. The maximum energy modulation for speech is at about 4 Hz (Houtgast and Steeneken, 1973).
With the cepstrum class of features it is possible to describe the behavior of the log magnitude spectrum with just a few parameters (Mammone, et al., 1996). This is the most common class of features used in speech recognition and speaker verification. This is also the feature used in this thesis.

1. **Bayesian classifier**

The most basic classifier is the Bayesian classifier (Duda and Hart, 1973). A source has $N_S$ internal states, $S \in \{1, ..., N_S\}$. A transducer records the signal from the source and the recorded signal is processed through a feature extractor. The internal state of the source is unknown and the task is to guess the internal state based on the feature vector, $x = (x_1, ..., x_K)$. The classifier analyses the feature vector and takes a decision among $N_d$ different decisions. The feature vector is processed through $N_d$ different scalar-valued discriminant functions. The index, $d \in \{1, ..., N_d\}$, to the discriminant function with the largest value given the observation is generated as output.

A common special case is maximum a posteriori (MAP) decision where the task is to simply guess the source state with minimum error probability. The source has a priori source state distribution $P(S = j)$. This distribution is assumed to be known. The feature vector distribution is described by a conditional probability density function, $f_{X|S}(X = x | S = j)$, and these conditional distributions are also assumed to be known. In practice, both the a priori distributions and the conditional distributions are normally unknown and must be estimated from training data. The probability that the source is in state $S = j$ given the observation $x = (x_1, ..., x_K)$ can now be calculated by using Bayes’ rule as:

$$P_{S|X}(j|x) = \frac{f_{X|S}(x|j)P_S(j)}{f_X(x)}$$  \hfill (2)
The denominator in the equation is independent of the state and can be removed. The discriminant functions can be formulated as:

\[ g_j(x) = f_{x|s}(x|j)P_s(j) \]  

(3)

The decision function, the index of the discriminant function with the largest value given the observation, is:

\[ d(x) = \arg\max_j g_j(x) \]  

(4)

For a given observation sequence it was assumed here that the internal state in the source is fixed when generating the observed data sequence. In many processes the internal state in the source will change describing a discrete state sequence, \( S = (S(1),\ldots,S(2),\ldots,S(T)) \). In this case the observation sequence can have time-varying characteristics. Different states can have different probability density distributions for the output signal. Even though the Bayesian framework can be used when the internal state changes during the observation period, it is not the most efficient method. A better method, which includes a simple model of the dependencies between the current state and the previous state, is described in the next section.
Figure 4 Example of a simple hidden Markov model structure. Values close to arcs indicate state transition probabilities. Values inside circles represent observation probabilities and state numbers. The filled circle represents the start. For example in state $S=1$ the output may be $X=1$ with probability 0.8 and $X=2$ with probability 0.2. The probability for staying in state $S=1$ is 0.9 and the probability for entering state $S=2$ from state $S=1$ is 0.1.

2. Hidden Markov models

The Hidden Markov model is a statistical method, used in classification, that has been very successful in some areas, e.g. speech recognition (Rabiner, 1989), speaker verification, and hand writing recognition (Yasuda, et al., 2000). The hidden Markov model is the main classification method used in this thesis. A stochastic process $(X(1),...,X(t))$ is modeled as being generated by a hidden Markov model source with a number, $N$, of discrete states, $S(t) \in \{1,...,N\}$. Each state has an observation probability distribution, $b_j(x(t)) = P_{x_i|S}(x(t) = x_i|S(t) = j)$. This is the probability of an observation $x(t)$ given that the state is $S(t) = j$. Given state $S(t) = i$, the probability that the next state is $S(t+1) = j$ is described by a state transition probability $a_{ij} = P(S(t+1) = j|S(t) = i)$. The initial state distribution for a hidden Markov model is described by $p_0(j) = P(S(1) = j)$.

For hearing aid applications it is efficient to use discrete hidden Markov models. The only difference is that the multi-dimensional observation vector $x(t)$ is encoded into a discrete number $z(t)$; hence the observation
distribution at each state is described with a probability mass function, \( B_{j,m} = b_j(m) = P(Z(t) = m | S(t) = j) \). A hidden Markov model with discrete observation probabilities can be described in a compact form with a set of matrices, \( \lambda = \{ p_0, A, B \} \); Vector \( p_0 \) with elements \( p_0(j) \), \( A \) with elements \( a_{ij} \), and \( B \) with elements \( b_j(m) \).

An example of a hidden Markov model is illustrated in Figure 4 and described next. Consider a cat, a very unpredictable animal, which has two states: one state for “hungry”, \( S(t)=1 \), and one state for “satisfied”, \( S(t)=2 \). The observation is sampled at each time unit and \( Z(t) \) has two possible outcomes: the cat is scratching, \( Z(t)=1 \), or the cat is purring, \( Z(t)=2 \). It is very likely that a cat that is satisfied will turn into a hungry cat and it is very likely that a cat that is hungry will stay hungry. The initial condition of the cat is “satisfied”. A cat that is satisfied has a high probability for purring and a cat that is hungry has a high probability for scratching. The behavior of the cat may then be approximated with the following hidden Markov model.

\[
p_0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad A = \begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{pmatrix} \quad B = \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{pmatrix}
\]

The hidden Markov model described above is ergodic, i.e. all states in the model have transitions to all other states. This structure is suitable when the order of the states, as they are passed through, has no importance.

If the state order for the best matching state sequence given an observation sequence and model is important, e.g. classifying words or written letters. Then a left-right hidden Markov model structure is more suitable. In a left-right hidden Markov model structure a state only has transitions to itself and to states to the “right” with higher state numbers (illustrated in Figure 5).

The structure of a hidden Markov model can also be a mixture between ergodic and left-right structures.
Many stochastic processes can be easily modeled with hidden Markov models and this is the main reason for using it in sound environment classification. Three different interesting problems arise when working with HMMs:

1. What is the probability for an observation sequence given a model, $P(z(1),...,z(t)|\lambda)$.

2. What is the optimal corresponding state sequence given an observation sequence and a model:

$$\hat{i}(1),...,\hat{i}(t) = \arg \max_{i(1),...,i(t)} P(i(1),...,i(t-1), S(t) = i(t), z(1),...,z(t)|\lambda)$$

3. How should the model parameters $\lambda = \{p0, A, B\}$ be adjusted to maximize $P(z(1),...,z(t)|\lambda)$ for an observed “training” data sequence.

There are three algorithms that can be used to solve these three problems in an efficient way (Rabiner, 1989). The forward algorithm is used to solve the first problem. The Viterbi algorithm is used to solve the second problem and the Baum-Welch algorithm is used to solve the last problem. The Baum-Welch algorithm only guarantees to find a local maximum of $P(z(1),...,z(t)|\lambda)$. The initialization is therefore crucial and must be done carefully.

Practical experience with ergodic hidden Markov models has shown that the $p0$ and $A$ model parameters are less sensitive to different initializations than the observation probability matrix $B$. Different methods can be used to initialize the $B$ matrix, e.g. manual labeling of the observation sequence into states or automatic clustering methods of the observation sequence.
Left-right hidden Markov models are less sensitive to different initializations. It is possible to initialize all the model parameters uniformly and yet achieve good results in the classification after training.

3. Vector quantizer

The HMMs in this thesis are implemented as discrete models with a vector quantizer preceding the HMMs. A vector quantizer consists of $M$ codewords each associated with a codevector in the feature space. The vector quantizer is used to encode each multi-dimensional feature vector into a discrete number between 1 and $M$. The observation probability distribution for each state in the HMM is therefore estimated with a probability mass function instead of a probability density function. The centroids defined by the codevectors are placed in the feature space to minimize the distortion given the training data. Different distortion measures can be used; the one used in this thesis is the mean square distortion measure. The vector quantizer is trained with the generalized Lloyd algorithm (Linde, et al., 1980). A trained vector quantizer together with the feature space is illustrated in Figure 8.

The number of codewords used in the system is a design variable that must be considered when implementing the classifier. The quantization error introduced in the vector quantizer has an impact on the accuracy of the classifier. The difference in average error rate between discrete HMMs and continuous HMMs has been found to be on the order of 3.5 percent in recognition of spoken digits (Rabiner, 1989). The use of a vector quantizer together with a discrete HMM is computationally much less demanding compared to a continuous HMM. This advantage is important for implementations in digital hearing aids, although the continuous HMM has a slightly better performance.

4. Sound classification

Automatic sound classification is a technique that is important in many disciplines. Two obvious implementations are speech recognition and speaker identification where the task is to recognize speech from a user and to identify a user among several users respectively. Another usage of sound classification is automatic labeling of audio data to simplify searching in large databases with recorded sound material. An interesting implementation is automatic search after specific keywords, e.g. Alfa Romeo, in audio streams. It is then possible to measure how much a
company or a product is exposed in media after an advertisement campaign. Another application is to detect flying insects or larvae activities in grain silos to minimize the loss of grain (Pricipe, 1998). Yet another area is acoustic classification of sounds in water.

During one period in the middle of the nineties, the Swedish government suspected presence of foreign submarines in the archipelago inside the Swedish territory. The Swedish government claimed that they had recordings from foreign submarines. Later it was discovered that the classification results from the recordings were false alarms; some of the sounds on recording were generated from shoal of herring and minks.

Another implementation is to notify deaf people when important acoustic events occur. A microphone continuously records the sound environment in the home, and sound events, e.g. phone bell, door bell, post delivery, knocking on the door, dog barking, egg clock, fire alarm etc, are detected and reported to the user by light signals or mechanic vibrations. The usage of sound classification in hearing aids is a new area where this technique is useful.

5. **Keyword recognition**

Keyword recognition is a special case of sound classification and speech recognition where the task is to only recognize one or a few keywords. The keywords may exist isolated or embedded in continuous speech. The background noise in which the keywords are uttered may vary, e.g. quiet listening environments or babble noise environments. Keyword recognition exists in many applications, e.g. mobile phones, toys, automatic telephone services, etc. Currently, no keyword recognizers exist for hearing instruments. An early example of a possible implementation of a keyword recognizer system is presented in (Baker, 1975). Another system is presented in (Wilpon, et al., 1990). The performance of a keyword classifier can be estimated with two variables: false alarm rate, i.e. the frequency of detected keywords when no keywords are uttered, and hit rate, i.e. the ratio between the number of detected keywords and the number of uttered keywords. There is always a trade-off between these two performance variables when designing a keyword recognizer.
VI. Methods

A. Platforms

Various types of platforms have been used in order to implement the systems described in the thesis. The algorithms used in Paper A were implemented purely in Matlab. The algorithm described in Paper B was implemented in an experimental wearable digital hearing aid based on the Motorola DSP 56k family. The DSP software was developed in Assembler. The interface program was written in C++ using a software development kit supported by GN ReSound. The corresponding acoustic evaluation was implemented in Matlab. The algorithms used in Paper C were implemented in Matlab. Methods used in Paper D were implemented purely in Matlab. The algorithms in Paper E was implemented in C#(csharp) simulating a digital hearing aid in real-time. All methods and algorithms used in the thesis are implemented by the author, e.g. algorithms regarding hidden Markov models or vector quantizers.

B. Signal processing

Digital signal processing in hearing aids is often implemented blockwise, i.e. a block of $B$ sound samples from the microphone $x(t) = (x_n(t), n = 0, \ldots, B - 1)$ is processed for each time the main loop is traversed. This is due to the overhead needed anyway in the signal processing code being independent of the blocksize. Thus it is more power efficient to process more sound samples within the same main algorithm loop. On the other hand, the time delay in a hearing aid should not be greater than about 10 ms; otherwise the perceived subjective sound quality is reduced (Dillon, et al., 2003). The bandwidth of a hearing aid is about 7500 Hz, which requires a sampling frequency of about 16 kHz according to the Nyquist theorem. All the algorithms implemented in this thesis have 16 kHz sampling frequency and block sizes giving a time delay of 8 ms or less in the hearing aid. Almost all hearing aid algorithms are based on a fast Fourier transform (FFT) and the result from this transform are used to adaptively control the gain frequency response. Other algorithms, e.g. sound classification or keyword recognition, implemented in parallel with the basic filtering, should use the already existing result from the FFT in order to reduce power consumption. This strategy has been used in the implementations presented in this thesis.
C. Recordings

Most of the sound material used in the thesis has been recorded manually. Three different recording equipments have been used: a portable DAT player, a laptop computer and a portable MP3 player. An external electret microphone with nearly ideal frequency response was used. In all the recordings the microphone was placed behind the ear to simulate the position of a BTE microphone.

The distribution of the sound material used in the thesis is presented in Table 2, Table 3, and Table 4. The format of the recorded material is wave PCM with 16 bits resolution per sample. Some of the sound material was taken from hearing aid CDs available from hearing aid companies.

When developing a classification system it is important to separate material used for training from material used for evaluation. Otherwise the classification result may be too optimistic. This principle has been followed in this thesis.

<table>
<thead>
<tr>
<th>Total Duration (s)</th>
<th>Recordings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training Material</strong></td>
<td></td>
</tr>
<tr>
<td>Clean speech</td>
<td>347</td>
</tr>
<tr>
<td>Babble noise</td>
<td>233</td>
</tr>
<tr>
<td>Traffic noise</td>
<td>258</td>
</tr>
<tr>
<td><strong>Evaluation Material</strong></td>
<td></td>
</tr>
<tr>
<td>Clean speech</td>
<td>883</td>
</tr>
<tr>
<td>Babble noise</td>
<td>962</td>
</tr>
<tr>
<td>Traffic noise</td>
<td>474</td>
</tr>
<tr>
<td>Other noise</td>
<td>980</td>
</tr>
</tbody>
</table>

Table 2 Distribution of sound material used in paper C.
### Table 3 Distribution of sound material used in paper D.

<table>
<thead>
<tr>
<th></th>
<th>Total Duration (s)</th>
<th>Recordings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training Material</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clean speech</td>
<td>178</td>
<td>1</td>
</tr>
<tr>
<td>Phone speech</td>
<td>161</td>
<td>1</td>
</tr>
<tr>
<td>Traffic noise</td>
<td>95</td>
<td>1</td>
</tr>
<tr>
<td><strong>Evaluation Material</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Face-to-face conversation in quiet</td>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>Telephone conversation in quiet</td>
<td>23</td>
<td>1</td>
</tr>
<tr>
<td>Face-to-face conversation in traffic noise</td>
<td>64</td>
<td>2</td>
</tr>
<tr>
<td>Telephone conversation in traffic noise</td>
<td>63</td>
<td>2</td>
</tr>
</tbody>
</table>

### Table 4 Distribution of sound material used in paper E.

<table>
<thead>
<tr>
<th></th>
<th>Total Duration (s)</th>
<th>Recordings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training Material</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Keyword</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>Clean speech</td>
<td>409</td>
<td>1</td>
</tr>
<tr>
<td>Traffic noise</td>
<td>41</td>
<td>7</td>
</tr>
<tr>
<td><strong>Evaluation Material</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hit Rate Estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clean speech</td>
<td>612</td>
<td>1</td>
</tr>
<tr>
<td>Speech in babble noise</td>
<td>637</td>
<td>1</td>
</tr>
<tr>
<td>Speech in outdoor/traffic noise</td>
<td>2520</td>
<td>1</td>
</tr>
<tr>
<td>False Alarm Estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clean speech</td>
<td>14880</td>
<td>1</td>
</tr>
</tbody>
</table>
D. Classifier implementation

The main principle for the sound classification system used in this work is illustrated in Figure 6. The classifier is trained and designed to classify a number of different listening environments. A listening environment consists of one or several sound sources, e.g. the listening environment “speech in traffic” consists of the sound sources “traffic noise” and “clean speech”. Each sound source is modeled with a hidden Markov model. The output probabilities generated from the sound source models are analyzed with another hidden Markov model, called environment hidden Markov model. The environment HMM describes the allowed combinations of the sound sources. Each section in Figure 6 is described more in detail in the following sections.

Figure 6 Listening Environment Classification. The system consists of four layers: the feature extraction layer, the sound feature classification layer, the sound source classification layer, and the listening environment classification layer. The sound source classification uses a number of hidden Markov models (HMMs). Another HMM is used to determine the listening environment.
1. Feature extraction

One block of samples, \( x(t) = (x_n(t) \quad n = 0,\ldots, B-1) \), from the A/D-converter is the input to the feature extraction layer. The input block is multiplied with a Hamming window, \( w_n \), and \( L \) real cepstrum parameters, \( f_l(t) \), are calculated according to (Mammone, et al., 1996):

\[
f_l(t) = \frac{1}{B} \sum_{k=0}^{B-1} b_k^l \log \left( \sum_{n=0}^{B-1} w_n x_n(t) e^{-\frac{j2\pi kn}{B}} \right) \quad l = 0,\ldots, L - 1.
\]

(7)

where vector \( x(t) = (x_0(t),\ldots,x_{B-1}(t)) \) is a block of \( B \) samples from the transducer, vector \( w = (w_0,\ldots,w_{B-1}) \) is a window (e.g. Hamming), vector \( b^l = (b_0^l,\ldots,b_{B-1}^l) \) represents the basis function used in the cepstrum calculation and \( L \) is the number of cepstrum parameters.

The corresponding delta cepstrum parameter, \( \Delta f_l(t) \), is estimated by linear regression of a small buffer of stored cepstrum parameters (Young, et al., 2002):

\[
\Delta f_l(t) = \frac{\sum_{\theta=1}^{\Theta} \theta (f_l(t + \theta - \Theta) - f_l(t - \theta - \Theta))}{2 \sum_{\theta=1}^{\Theta} \theta^2}
\]

(8)

The distribution of the delta cepstrum features for various sound sources is illustrated in Figure 8. The delta-delta cepstrum parameter is estimated analogously with the delta cepstrum parameters used as input.
2. Sound feature classification

The stream of feature vectors is sent to the sound feature classification layer that consists of a vector quantizer (VQ) with $M$ codewords in the codebook $(c_1, \ldots, c_M)$. The feature vector

$$\Delta f(t) = (\Delta f_0(t), \ldots, \Delta f_{L-1}(t))^T$$

is encoded to the closest codeword in the VQ and the corresponding codeword index

$$z(t) = \arg \min_{i=1, \ldots, M} \|\Delta f(t) - c_i\|^2$$

is generated as output.

3. Sound source classification

This part of the classifier calculates the probabilities for each sound source category used in the classifier. The sound source classification layer consists of one HMM for each included sound source. Each HMM has $N$ internal states, not necessarily the same number of states for all sources. The current state at time $t$ is modeled as a stochastic variable $Q_{source}(t) \in \{1, \ldots, N\}$. Each sound model is specified as

$$\lambda_{source} = \{A_{source}, B_{source}\}$$

where $A_{source}$ is the state transition probability matrix with elements

$$a_{ij}^{source} = P(Q_{source}(t) = j | Q_{source}(t-1) = i)$$

and $B_{source}$ is the observation probability matrix with elements

$$B_{j,z(t)}^{source} = P(z(t) | Q_{source}(t) = j)$$

Each HMM is trained off-line with the Baum-Welch algorithm (Rabiner, 1989) on training data from the corresponding sound source. The HMM structure is ergodic, i.e. all states within one model are connected with each other. When the trained HMMs are included in the classifier, each source
model observes the stream of codeword indices, \((z(1),...,z(t))\), where \(z(t) \in \{1,\ldots,M\}\), coming from the VQ. The conditional state probability vector \(\hat{p}_{source}(t)\) is estimated with elements

\[
\hat{p}_{i_{source}}(t) = P\left(Q_{source}(t) = i | z(t),...,z(1), \lambda_{source}\right) \tag{12}
\]

These state probabilities are calculated with the forward algorithm (Rabiner, 1989):

\[
p_{source}(t) = \left(\left(A_{source}^{T} \hat{p}_{source}(t-1)\right) \circ B_{1,\ldots,N,z(t)}\right) \tag{13}
\]

Here, \(T\) indicates matrix transpose, and \(\circ\) denotes element-wise multiplication. \(p_{source}(t)\) is a non-normalized state likelihood vector with elements:

\[
p_{i_{source}}(t) = P\left(Q_{source}(t) = i, z(t),z(t-1),...,z(1), \lambda_{source}\right) \tag{14}
\]

The probability for the current observation given all previous observations and a source model can now be estimated as:

\[
\phi_{source}(t) = P\left(z(t)|z(t-1),...,z(1), \lambda_{source}\right) = \sum_{i=1}^{N} p_{i_{source}}(t) \tag{15}
\]

Normalization is used to avoid numerical problems:

\[
\hat{p}_{i_{source}}(t) = p_{i_{source}}(t) / \phi_{source}(t) \tag{16}
\]
Figure 7 Example of hierarchical environment hidden Markov model. Six source hidden Markov models are used to describe three listening environments. The transitions within environments and between environments determine the dynamic behavior of the system.

4. Listening environment classification

The sound source probabilities calculated in the previous section can be used directly to detect the sound sources. This might be useful if the task of the classification system is to detect isolated sound sources, e.g. estimate signal and noise spectrum. However, in this framework a listening environment is defined as a single sound source or a combination of two or more sound sources. Examples of sound sources are traffic noise, babble noise, clean speech, and telephone speech. The output data from the sound source models are therefore further processed by a final hierarchical HMM in order to determine the current listening environment. An example of a hierarchical HMM structure is illustrated in Figure 7. In speech recognition systems this layer of the classifier is often called the lexical layer where the rules for combining phonemes into words are defined.

The hierarchical HMM is an important part of the classifier. This part of the classifier models listening environments as combinations of sound
sources instead of modeling each listening environment with a separate HMM. The ability to model listening environments with different signal-to-noise ratios has been achieved by assuming that only one sound source in a listening environment is active at a time. This is a known limitation, as several sound sources are often active at the same time. The theoretically more correct solution would be to model each listening environment at several signal-to-noise ratios. However, this approach would require a very large number of models. The reduction in complexity is important when the algorithm is implemented in hearing instruments.

The final classifier block estimates for every frame the current probabilities for each environment category by observing the stream of sound source probability vectors from the previous block. The listening environment is represented as a discrete stochastic variable \( E(t) \), with outcomes coded as 1 for “environment #1”, 2 for “environment #2”, etc. The environment model consists of a number of states and a transition probability matrix \( A^\text{env} \). The current state in this HMM is modeled as a discrete stochastic variable \( S(t) \), with outcomes coded as 1 for “sound source #1”, 2 for “sound source #2”, etc. The hierarchical HMM observes the stream of vectors \( (u(t), \ldots, u(t)) \), where

\[
u(t) = \begin{pmatrix} \phi^\text{source#1}(t) & \phi^\text{source#2}(t) & \ldots \end{pmatrix}^T
\]

contains the estimated observation probabilities for each state. Notice that there exists no pre-trained observation probability matrix for the environment HMM. Instead, the observation probabilities are achieved directly from the sound source probabilities. The probability for being in a state given the current and all previous observations and given the hierarchical HMM,

\[
\hat{p}^\text{env}_i = P(S(t) = i | u(t), \ldots, u(1), A^\text{env})
\]

is calculated with the forward algorithm (Rabiner, 1989),

\[
p^\text{env}(t) = \left( (A^\text{env})^T \hat{p}^\text{env}(t-1) \right) \odot u(t)
\]
with elements \( p_i^{env} = P(S(t) = i, u(t)|u(t-1),...,u(1), A^{env}) \), and finally, with normalization:

\[
\hat{p}^{env}(t) = p^{env}(t) / \sum_i p_i^{env}(t) \tag{20}
\]

The probability for each listening environment, \( p^E(t) \), given all previous observations and given the hierarchical HMM, can now be calculated by summing the appropriate elements in \( \hat{p}^{env}(t) \), e.g. according to Figure 7.

\[
p^E(t) = \begin{pmatrix}
1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 \\
\end{pmatrix} \hat{p}^{env}(t) \tag{21}
\]
5. Feature extraction analysis

The features used in the system described in paper C are delta cepstrum features (equation 8). Delta cepstrum features are extracted and stored in a feature vector for every block of sound input samples coming from the microphone. The implementation, described in paper C, uses four differential cepstrum parameters. The features generated from the feature extraction layer (Figure 6) can be used directly in a classifier. If the features generated from each class are well separated in the feature space it is easy to classify each feature. This is often the solution in many practical classification systems.

A problem occurs if the observations in the feature space are not well separated; features generated from one class can be overlapped in the feature space with features generated from another class. This is the case for the feature space used in the implementation described in paper C. A
scatter plot of the 0th and 1st delta cepstrum features for three sound sources is displayed in Figure 8. The figure illustrates that the babble noise features are distributed within a smaller range compared with the range of the clean speech features. The traffic noise features are distributed within a smaller range compared with the range for the babble noise features. A feature vector within the range of the traffic noise features can apparently belong to any of the three categories.

A solution to this problem is to analyze a sequence of feature vectors instead of instant classification of a single feature vector. This is the reason for using hidden Markov models in the classifier. Hidden Markov models are used to model sequences of observations, e.g. phoneme sequences in speech recognition.

Figure 9 The $M=32$ delta log magnitude spectrum changes described by the vector quantizer. The delta cepstrum parameters are estimated within a section of 12 frames (96 ms).
6. Vector quantizer analysis after training

After the vector quantizer has been trained it is interesting to study how the code vectors are distributed. When using delta cepstrum features (equation 8) each codevector in the codebook describes a change of the log magnitude spectrum. The implementation in paper C uses \( M = 32 \) codewords. The delta cepstrum parameters are estimated within a block of frames (12 frames equal 96 ms). Each log magnitude spectrum change described by each code vector in the codebook is illustrated in Figure 9 and calculated as:

\[
\Delta \log |S|^m = \sum_{l=0}^{L-1} c_i^m b^l
\]  

where \( m \) is the number of the codevector, \( c_i^m \) element \( l \) in codevector \( #m \), \( L \) the number of features, and \( b^l \) the basis function used in the cepstrum calculation (equation 7). The codebook represents changes in log magnitude spectrum and it is possible to model a rich variation of log magnitude spectrum variations from one frame to another. It is illustrated that several possible increases and decreases in log magnitude spectrum from one frame to another are represented in the VQ.
Figure 10 Average magnitude spectrum change and average time duration for each state in a hidden Markov model. The model is trained on delta cepstrum features extracted from traffic noise. E.g. it is illustrated, in state 3 that the log magnitude spectrum slope grows for 25 ms, in state 4 that the log magnitude spectrum slope reduces for 26 ms.

7. HMM analysis after training

The hidden Markov models are used to model processes with time varying characteristics. In paper C the task is to model the sound sources traffic noise, babble noise, and clean speech. After the hidden Markov models are trained it is useful to analyze and study the models more closely. Only differential cepstrum features are used in the system (equation 2), and a vector quantizer precedes the hidden Markov models (equation 5).

Each state in each hidden Markov model has a probability, $b_j(m) = P(Z(t) = m | S(t) = j)$, for each log magnitude spectrum change, $\Delta \log |S|^m$. 


The mean of all possible log magnitude spectrum changes for each state can now be estimated as:

$$\Delta \log |S|_j = \sum_{m=1}^M b_j(m) \Delta \log |S|^m$$

(23)

where $M$ is the number of codewords.

Furthermore, the average time duration in each state $j$ can be calculated as (Rabiner, 1989):

$$d_j = \frac{1}{1 - a_{jj}} \cdot \frac{B}{f_s}$$

(24)

where $a_{jj} = P(S(t) = j | S(t-1) = j)$, $B$ is the block size, and $f_s$ the sampling frequency. The average log magnitude spectrum change for each state and the average time in each state for traffic noise is illustrated in Figure 10. E.g. it is illustrated in state 3 that the log magnitude spectrum slope grows for 25 ms, in state 4 that the log magnitude spectrum slope reduces for 26 ms. The hidden Markov model spends most of its time in state two with 28 ms closely followed by state 4 with 26 ms. The log magnitude spectrum changes described by the model is an estimate of the log magnitude spectrum variations represented in the training material, in this case traffic noise.
VII. Results and discussion

A. Paper A

The first part of the thesis investigated the behavior of non-linear hearing instruments under realistic simulated listening conditions. Three of the four hearing aids used in the investigation were digital. The result from the investigation showed that the new possibilities allowed by the digital technique were used very little in these early digital hearing aids. The only hearing aid that had a more complex behavior compared to the other hearing aids was the Widex Senso. Already at that time the Widex Senso hearing aid seemed to have a speech detection system, increasing the gain in the hearing aid when speech was detected in the background noise.

The presented method used to analyze the hearing aids worked as intended. It was possible to see changes in the behavior of a hearing aid under realistic listening conditions. Data from each hearing aid and listening environment was displayed in three different forms:

1. Effective temporal characteristics (Gain-Time)
2. Effective compression characteristics (Input-Output)
3. Effective frequency response (Insertion Gain).

It is possible that this method of analyzing hearing instruments generates too much information for the user. Future work should focus on how to reduce this information and still be able to describe the behavior of a modern complex hearing aid with just a few parameters. A method describing the behavior of the hearing aid with a hidden Markov model approach is presented in (Leijon and Nordqvist, 2000).
B. Paper B

The second part of the thesis presents an automatic gain control algorithm that uses a richer representation of the sound environment than previous algorithms. The main idea of this algorithm is to: (1) adapt slowly (in approximately 10 seconds) to varying listening environments, e.g. when the user leaves a disciplined conference for a multi-babble coffee-break; (2) switch rapidly (in about 100 ms) between different dominant sound sources within one listening situation, such as the change from the user’s own voice to a distant speaker’s voice in a quiet conference room; (3) instantly reduce gain for strong transient sounds and then quickly return to the previous gain setting; and (4) not change the gain in silent pauses but instead keep the gain setting of the previous sound source.

An acoustic evaluation showed that the algorithm worked as intended. This part of the thesis was also evaluated in parallel with a reference algorithm in a field test with nine test subjects. The average monosyllabic word recognition score in quiet with the algorithm was 68% for speech at 50 dB SPL RMS and 94% for speech at 80 dB SPL RMS, respectively. The corresponding scores for the reference algorithm were 60% and 82%. The average signal to noise threshold (for 40% recognition), with Hagerman’s sentences (Hagerman, 1982) at 70 dB SPL RMS in steady noise, was –4.6 dB for the algorithm and –3.8 dB for the reference algorithm, respectively. The signal to noise threshold with the modulated noise version of Hagerman’s sentences at 60 dB SPL RMS was –14.6 dB for the algorithm and –15.0 dB for the reference algorithm. Thus, in this test the reference system performed slightly better compared to the algorithm. Otherwise there was a tendency towards slightly better results with the proposed algorithm for other three tests. However, this tendency was statistically significant only for the monosyllabic words in quiet, at 80 dB SPL RMS presentation level.

The focus of the objective evaluation was speech intelligibility and this measure may not be optimal for evaluating new features in digital hearing aids.
C. Paper C and D

The third part of the thesis, consisting of paper C and D, is an attempt to introduce sound environment classification in the hearing aids. Two systems are presented. The first system, paper C, is an automatic listening environment classification algorithm. The task is to automatically classify three different listening environments: speech in quiet, speech in traffic, and speech in babble. The study showed that it is possible to robustly classify the listening environments speech in traffic noise, speech in babble, and clean speech at a variety of signal-to-noise ratios with only a small set of pre-trained source HMMs. The measured classification hit rate was 96.7-99.5% when the classifier was tested with sounds representing one of the three environment categories included in the classifier. False alarm rates were 0.2-1.7% in these tests. The study also showed that the system could be implemented with the available resources in today’s digital hearing aids.

The second system, paper D, is focused on the problem that occurs when using a hearing aid during phone calls. A system is proposed which automatically detects when the telephone is used. It was demonstrated that future hearing aids might be able to distinguish between the sound of a face-to-face conversation and a telephone conversation, both in noisy and quiet surroundings. The hearing aid can then automatically change its signal processing as needed for telephone conversation. However, this classifier alone may not be fast enough to prevent initial feedback problems when the user places the telephone handset at the ear.

1. Signal and noise spectrum estimation

Another use of the sound source probabilities (Figure 6) is to estimate a signal and noise spectrum for each listening environment. The current prescription method used in the hearing aid can be adapted to the current listening environment and signal-to-noise ratio.

The sound source probabilities $\phi_{\text{source}}$ are used to label each input block as either a signal block or a noise block. The current environment is determined as the maximum of the listening environment probabilities. Each listening environment has two low pass filters describing the signal and noise power spectrum, e.g. $S_{S}(t+1) = (1-\alpha)S_{S}(t) + \alpha S_{P}(t)$ for the signal spectrum and $S_{N}(t+1) = (1-\alpha)S_{N}(t) + \alpha S_{P}(t)$ for the noise spectrum. Depending on the sound source probabilities, the signal power spectrum or the noise power spectrum for the current listening
environment is updated with the power spectrum $S_p(t)$ calculated from the current input block.

A real-life recording of a conversation between two persons standing close to a road with heavy traffic is processed through the classifier. The true signal and noise spectrum is calculated as the average of the signal and noise spectrum for the complete duration of the file. The labeling was done manually by listening to the recording and labeling each frame of the material. The estimated signal and noise spectrum, generated from the classifier, and the true signal and noise spectrum, are presented in Figure 11. The corresponding result from a real-life recording of a conversation between two persons in babble noise environment is illustrated in Figure 12. Clearly it is possible to estimate the current signal and noise spectrum in the current listening environment. This is not the primary task for the classifier but may be useful in the hearing aid.

![Figure 11 Signal and noise spectrum estimation for speech in traffic noise. The true signal and noise spectrum calculated manually are illustrated with black lines. The estimated signal and noise spectra generated with the sound classifier are illustrated with grey lines.](image-url)
Figure 12 Signal and noise spectrum estimation for speech in babble noise. The true signal and noise spectrum calculated manually are illustrated with black lines. The estimated signal and noise spectra generated with the sound classifier are illustrated with grey lines.
D. Paper E

The fourth and last part of the thesis is an attempt to implement keyword recognition in hearing aids. An implementation and an evaluation of a single keyword recognizer for a hearing instrument are presented.

The performance for the best parameter setting was $7e-5$ [1/s] in false alarm rate, 100% hit rate for the indoors quiet environment, 71% hit rate for the outdoors/traffic environment and 50% hit rate for the babble noise environment. The false alarm rate corresponds to one false alarm for every four hours of continuous speech from the hearing aid user. Since the total duration of continuous speech from the hearing aid user per day is estimated to be less than four hours, it is reasonable to expect one false alarm per day or per two days. The hit rate for indoor quiet listening environments is excellent. The hit rate for outside/traffic noise may also be acceptable when considering that some of the keywords were uttered in high background noise levels. The hit rate for babble noise listening environments is low and is a problem that must be solved in a future application.

The memory resource needed for the implemented system is estimated to 1820 words (16-bits). Optimization of the algorithm together with improved technology will inevitably make it possible to implement the system in a digital hearing aid within the next couple of years. A solution to extend the number of keywords and integrate the system with a sound environment classifier is outlined.
VIII. Conclusion

A variety of algorithms intended for the new generation of hearing aids is presented in this thesis. The main contribution of this thesis is the hidden Markov model approach to classifying listening environments. This method is efficient and robust and well suited for hearing aid applications. The thesis also shows that the algorithms can be implemented with the memory and calculation resources available in the hearing aids.

A method for analyzing complex hearing aid algorithms has been presented. Data from each hearing aid and listening environment were displayed in three different forms: (1) Effective temporal characteristics (Gain-Time), (2) Effective compression characteristics (Input-Output), and (3) Effective frequency response (Insertion Gain). The method worked as intended. It was possible to see changes in behavior of a hearing aid under realistic listening conditions. The proposed method of analyzing hearing instruments may generate too much information for the user.

An automatic gain controlled hearing aid algorithm adapting to two sound sources in the listening environment has been presented. The main idea of this algorithm is to: (1) adapt slowly (in approximately 10 seconds) to varying listening environments, e.g. when the user leaves a disciplined conference for a multi-babble coffee-break; (2) switch rapidly (in about 100 ms) between different dominant sound sources within one listening situation, such as the change from the user's own voice to a distant speaker's voice in a quiet conference room; (3) instantly reduce gain for strong transient sounds and then quickly return to the previous gain setting; and (4) not change the gain in silent pauses but instead keep the gain setting of the previous sound source. An acoustic evaluation showed that the algorithm worked as intended.

A system for listening environment classification in hearing aids has also been presented. The task is to automatically classify three different listening environments: speech in quiet, speech in traffic, and speech in babble. The study showed that it is possible to robustly classify the listening environments speech in traffic noise, speech in babble, and clean speech at a variety of signal-to-noise ratios with only a small set of pre-trained source HMMs. The measured classification hit rate was 96.7-99.5% when the classifier was tested with sounds representing one of the three environment categories included in the classifier. False alarm rates were 0.2-1.7% in these tests. The study also showed that the system could be implemented with the available resources in today's digital hearing aids.

A separate implementation of the classifier showed that it was possible to automatically detect when the telephone was used. It was demonstrated
that future hearing aids might be able to distinguish between the sound of a face-to-face conversation and a telephone conversation, both in noisy and quiet surroundings. However, this classifier alone may not be fast enough to prevent initial feedback problems when the user places the telephone handset at the ear.

A method using the classifier result for estimating signal and noise spectra for different listening environments has been presented. This method showed that signal and noise spectra can be robustly estimated given that the classifier has good performance.

An implementation and an evaluation of a single keyword recognizer for a hearing instrument have been presented. The performance for the best parameter setting gives 7e-5 [1/s] in false alarm rate, i.e. one false alarm for every four hours of continuous speech from the user, 100% hit rate for an indoors quiet environment, 71% hit rate for an outdoors/traffic environment and 50% hit rate for a babble noise environment. The memory resource needed for the implemented system is estimated to 1 820 words (16-bits). Optimization of the algorithm together with improved technology will inevitably make it possible to implement the system in a digital hearing aid within the next couple of years.

A solution to extend the number of keywords and integrate the system with a sound environment classifier was also outlined.
IX. About the papers

(A) Nordqvist, P. (2000). "The behaviour of non-linear (WDRC) hearing instruments under realistic simulated listening conditions," QPSR Vol 40, 65-68 (This work was also presented as a poster at the conference “Issues in advanced hearing aid research”. Lake Arrowhead, 1998.)

Contribution: The methods were designed and implemented by Peter Nordqvist. The paper was written by Peter Nordqvist.


Contribution: The algorithm was designed and implemented by Peter Nordqvist with some support from Arne Leijon. The evaluation with speech recognition tests was done by Agneta Borryd and Karin Gustafsson. The paper was written by Peter Nordqvist.


Contribution: The algorithm was designed and implemented by Peter Nordqvist with support from Arne Leijon. The paper was written by Peter Nordqvist.

(D) Nordqvist, P. and Leijon, A. (2002). "Automatic classification of the telephone listening environment in a hearing aid," QPSR, Vol 43, 45-49 (This work was also presented as a poster at the 141st Meeting Acoustic Society of America, Chicago, Vol. 109, No. 5, p. 2491, May 2001.)

Contribution: The algorithm was designed and implemented by Peter Nordqvist with support from Arne Leijon. The paper was written by Peter Nordqvist.


Contribution: The algorithm was designed and implemented by Peter Nordqvist with support from Arne Leijon. The paper was written by Peter Nordqvist.
X. Bibliography


SBU, (The Swedish Council on Technology Assessment in Health Care, 2003).


Smeds, K. (2004). "Is normal or less than normal overall loudness preferred by first-time hearing aid users?," Ear Hear


"Comparison of benefits provided by different hearing aid technologies," J Am Acad Audiol 11, 540-560.

