Preface

Even in the more than 30-year-old science fiction movie "startrack" the main actors communicated with the board computer via natural speech - a fiction which has not yet been realized even after decades of intensive research in automatic speech recognition. An essential hurdle on the way to a natural-speech, acoustical man-machine communication is the lack of a suitable representation of speech for the computer. In other words: How should the speech signal be represented in the computer in such a way that the resulting patterns can be interpreted in a correct way? Another, related problem refers to the high susceptibility of automatic speech recognition systems towards interfering noise: Most automatic speech recognition systems operate very well in quiet, but completely fail as soon as little additional noise is added. What can be done to suppress non-speech background noise?

For these both fundamental problems in automatic speech recognition, Jürgen Tchorz has shown in this thesis the perhaps decisive solution: Why is the computer not able to understand and process speech in a way similar to our ear? Using this question as a starting point, he uses a pre-processing method for speech which is closely linked to a model of the "effective" auditory signal processing that was developed primarily in the Graduate School "Psychoakustik" in Oldenburg. With this auditory-based representation of speech Jürgen Tchorz can achieve a higher robustness of speech recognizers against interfering noise than with conventional systems. In addition, on the basis of new insights into modulation processing in our ear he has developed an astonishingly efficient noise suppression system which is able to separate speech from non-speech background noise. This noise-suppression procedure promises not only applications in man-machine communication, but rather it can be applied to telecommunications (noise suppression during telephone conversations) and to hearing-aid technology. The current work by Jürgen Tchorz therefore shows an innovative, new approach in automatic speech recognition, since the previous approaches known from the literature are more based on the physical properties of speech rather than the properties of our ear. Thus, this thesis is in line with early work of Manfred R. Schroeder

(the "scientific grandfather" of this thesis, since the author of this preface did his Ph.D. under his supervision) who first introduced linear predictive coding into speech coding and used principles of the ear to "hide" the residual quantization noise from the listener's ear by putting it into those spectral regions that are masked by the coded speech. This technology is nowadays used in each hand-held (cellular) mobile phone. Hopefully, the results of the current thesis will be implemented into each (mobile) computer in the future as well!

Jürgen Tchorz is the 20th Ph.D. candidate from the current author's group in Göttingen and (since 1993) in Oldenburg. In fact, he is the first graduate student from Oldenburg who worked on speech research and used the strong auditory-research-based background of the group only as a motivation, not as a "toolbox". In his unconventional, but always intriguing and efficient way he even helped to start up a whole group of speech researchers in Oldenburg and carried his knowledge into teaching of physics students, since he was the first Ph.D. student from our group to be awarded a faculty position for "support of scientific newcomers". Without knowing him in person, you cannot really assess, how much fun it was to work with him and to discover all his different superb talents (with creating public relations materials for the faculty of all kinds being just a little part of it), even though at first sight he always looks as if he has just woke up. Given this background (as well as the vast application possibilities for his work) it is not at all astonishing that shortly after completion of his Ph.D. several international companies in hearing-aid and speech business have been competing to recruit him as a staff member. If you want to find out why this is the case, there is a very simple way to do: Just read this thesis!

Birger Kollmeier, October 2000

Contents

Chapter 1

General Introduction

Computational speech processing has undergone a rapid development during recent years. Automatic speech recognition (ASR), for example, found its way out of the research laboratories into a wide range of practical applications, such as dictation, dialog systems for inquiries, or voice-driven banking. Computational speech processing is also employed in the field of human communication in a fast growing number of applications, such as mobile telephony, voice-over-IP, or in digital hearing instruments. In almost all of the applications listed above, however, background noise is a major problem. Hearing-impaired persons, for example, often complain that through the hearing aid noise becomes quite annoying which makes it exhausting or even impossible to understand a talker. Hence, they often prefer not to wear the hearing aid or to avoid noisy situations. The recognition rates of ASR systems, to give another example, typically drop significantly even in moderate background noise, which can make the usefulness of the whole system questionable. The current thesis is concerned with solutions for these problems by mimicking properties of the human auditory system to suppress unwanted noise and to increase the robustness of automatic speech recognizers.

Several approaches have been suggested to suppress disturbing background noise and to enhance speech recognition with hearing aids or to increase the robustness in ASR systems. Existing noise suppression

schemes can be grouped into two main categories. Directive algorithms perform the separation between the target and the noise signal by spatial filtering. A target signal (e.g. from the front direction) is passed through, and signals from other directions are suppressed. This can be realized by using directive microphones or microphone arrays (Soede *et al.*, 1993). In prototype hearing instruments, binaural algorithms exploit phase and level differences or correlations between the two sides of the head for spatial filtering (Wittkop *et al.*, 1997). Monaural noise suppression algorithms, in contrast, try to separate speech from noise when only one microphone is available, i.e. without spatial information. A monaural noise suppression approach which is widely used bases on Spectral Subtraction (Boll, 1979). The noise spectrum (which is measured and updated in speech pauses) is subtracted from the signal spectrum. After reconstruction, ideally, the signal is cleaned from noise. In ASR systems, higher robustness against noise can be achieved by single- or multi-channel noise suppression as described above, by model compensation (Siohan *et al.*, 1999), or by more noise-robust front ends, which are designed to extract vectors from the waveform that reflect distinctive features from speech but which are relatively insensitive against noise.

Despite the progress in noise suppression and more robust ASR, however, there is no speech recognition system available to-date with recognition performance even close to human speech intelligibility in noise. Most algorithms which are designed to reduce the impact of background noise in human communication or in ASR systems are "technical" approaches which do not or only little consider properties and characteristic features of auditory sound processing, even though the human auditory system can be regarded as a very robust "speech processor". We can detect and classify different sound sources, concentrate on one of them (e.g., a certain talker), and "fade out" the other sources from our focus, which allows us to understand speech even in very poor acoustical situations. These impressive skills are made possible by the interplay between the auditory "feature extraction", which detects, analyzes and sorts a range of different acoustic cues in the waveform, and the higher stages of the auditory system which perform their cognitive tasks basing on these cues.

While comparatively little is known on the complex processing in the higher stages, more insight has been gained on the details of the auditory periphery and the representation of signals in auditory stages behind the periphery, for example the analysis of amplitude modulations, or the mechanisms of spectral and temporal masking. These properties can already be observed in the first stages of the auditory system.

The current thesis therefore is concerned with the application of certain properties of the auditory system to computational speech processing. The goal is to reduce disturbing effects of background noise, with the underlying assumption that the biological model is better suited for the solution of the above described problems, compared to entirely "technical" approaches. A blind imitation of biological mechanisms, however, is not likely to yield an effective solution to the respective problem (most air planes have two wings and a tail unit, but they do not *flap* wings). Thus, it is important to determine and model the most effective and essential properties and characteristics of auditory processing.
Two major problems of computational speech processing are tackled in this thesis, namely the detection and suppression of noise in monaural input signals, and the extraction of noise-robust features in ASR systems. For noise detection and suppression, spectro-temporal patterns are generated from the waveform which reflect the representation of amplitude modulations in higher stages of the auditory system, and which allow for a distinction between speech and noise portions. For noise-robust ASR feature extraction, an effective psychoacoustical model of the auditory periphery is applied and investigated. Both algorithms are combined to further enhance the robustness in automatic speech recognition.

The thesis is structured as follows. In Chapter 2, an algorithm is presented which automatically detects the local acoustical situation in terms of the signal-to-noise ratio (SNR). The algorithm is motivated by psychoacoustical findings and neurophysiological experiments on the representation of amplitude modulations in the inferior colliculus and auditory cortex of mammals (Langner *et al.*, 1997;

Langner and Schreiner, 1988). These experiments revealed that, similar to center frequencies, modulation frequencies are analyzed and organized in "periodotopical" gradients, which were found to be almost orthogonal to the tonotopical gradients which respect to different center frequencies. Kollmeier and Koch (1994) applied these findings in the field of speech processing by introducing two-dimensional maps, so-called Amplitude Modulation Spectrograms (AMS), which contain information on both spectral and temporal characteristics of the input signal and which were applied in a binaural noise suppression scheme. In Chapter 2, AMS patterns and their contribution to reliable SNR prediction are studied in detail. In Chapter 3, the SNR prediction scheme is extended to frequency sub-bands and applied to noise suppression based on the SNR estimates. The local SNR is directly estimated in a range of frequency channels even if speech and noise are present at the same time, i.e., no explicit detection of speech pauses and no assumptions on noise stationarity during speech activity are necessary. The effects of "across-frequency" processing for SNR estimation are examined, and the results are compared with sub-band SNR estimation based on voice activity detection. Noise suppression is performed by attenuating different frequency channels according to their SNR in a following processing step. The quality of the novel noise suppression algorithm, compared to unprocessed speech, is evaluated with a range of objective speech quality measures.
In Chapter 4, the application of a model of the auditory periphery as front end for ASR is presented. The model which reflects both spectral and temporal properties of the auditory periphery was originally developed by Dau and others to predict human performance in typical psychoacoustical masking experiments (Dau *et al.*, 1996a; 1996b). The model provides feature vectors which are considered as an "internal representation" of sound. The auditory-based features serve as input for a HMM recognizer for digit recognition in noise, and the results are compared with the performance obtained with conventional mel-cepstral features. The different processing stages of the auditory model and their contribution to robust speech recognition are studied in detail, especially the role of adaptive amplitude compression and suppression of amplitude modulations outside the range of modulations originating from articulatory movements.

In Chapter 5, the noise suppression scheme which was developed in Chapter 3 is evaluated in ASR experiments. It is combined with the auditory front end to investigate whether it allows for further enhancement of robust digit recognition. The results are compared with Spectral Subtraction as a standard noise suppression approach.

Chapter 2

Estimation of the signal-to-noise ratio with amplitude modulation spectrograms [1]

Abstract

An algorithm is proposed which automatically estimates the local signal-to-noise ratio (SNR) between speech and noise. The feature extraction stage of the algorithm is motivated by neurophysiological findings on amplitude modulation processing in higher stages of the auditory system in mammals. It analyzes information on both center frequencies and amplitude modulations of the input signal. This information is represented in two-dimensional patterns, so-called Amplitude Modulation Spectrograms (AMS). A neural network is trained on a large number of AMS patterns generated from mixtures of speech and noise. After

[1] A modified version of this Chapter has been submitted to *Speech Communication*: Tchorz and Kollmeier (2000) "Estimation of the signal-to-noise ratio with amplitude modulation spectrograms".

training, the network supplies estimates of the local SNR when AMS patterns from "unknown" sound sources are presented. Classification experiments show a relatively accurate estimation of the present SNR in independent 32 ms analysis frames. Harmonicity appears to be the most important cue for analysis frames to be classified as "speech-like", but the spectro-temporal representation of sound in AMS patterns also allows for a reliable discrimination between unvoiced speech and noise.

2.1 Introduction

The automatic classification of the acoustical situation in terms of speech/non speech detection or signal-to-noise ratio (SNR) estimation is an important issue for various signal processing applications. In the field of mobile communication, accurate voice activity detection (VAD) is essential for silence compression. Digital processing in modern hearing instruments allows the implementation of a wide range of sound processing schemes which adapt on the present sound source. Finally, noise suppression for e.g. automatic speech recognition requires a fast and reliable estimate of the local noise level or signal-to-noise ratio (SNR). Typically, SNR estimation is realized by updating a measure of the background noise in speech pauses, which are detected by a VAD. For VAD-based SNR estimation, stationarity of the noise has to be assumed while speech is active. Furthermore, portions detected as speech pauses must not contain voice to allow for correct noise measurement, but at the same time all actual speech pauses should be detected for a fast update of the noise measurement. In reality, unfortunately, the combination of these two requirements is not often met. A range of different VAD algorithms are described in the literature which use different sets of feature parameters that are extracted from the waveform, and different types of classification paradigms which compute a speech/non-speech decision. The VAD standardized by the European Telecommunications Standards Institute (ETSI) for the Full Rate GSM codec (ETSI, 1996) is basically an energy detector. The applied threshold is constantly adapted in speech pauses, where pitch detection is used to prevent voiced speech portions from being classified as speech pauses. In a VAD standardized by the International Telecom-

munication Union (ITU, 1996), information on energy, zero-crossing rate, and spectral distortions is utilized for VAD, and thresholds are defined for classification.
Some direct SNR estimation schemes were developed which do not require explicit speech pause detection. An iterative approach described by Hirsch and Ehrlicher (1995) bases on the statistical analysis of a segment of the magnitude spectral envelope. Histograms of past values are build taking into account values below a dynamically updated threshold. The noise level is estimated as the smoothed maximum of this distribution. Being based on relative energy levels, however, the algorithm cannot distinguish between rising noise energy and the presence of speech. Furthermore, an accurate estimation of the noise energy requires analysis frames which include speech pauses or closures, typically more than 0.5 s. Thus, the noise estimate is rather "sluggish" and cannot follow rapid changes. Martin (1993) proposed a spectral analysis method which also requires a long segment of the input signal (about 0.6 s). The algorithm is based on the observation that a noise power estimate can be obtained by using minimum values of a smoothed power estimate. This approach implies that the estimate is biased when no speech is present. Dupont and Ris (1999) proposed a method which requires shorter analysis frames (about 0.3 s) by taking advantage of the fact that the spectral energy in valleys between the harmonics is close to the noise floor. A lower energy envelope follower is used for noise estimation. They quantitatively compared their SNR estimation approach with others, including (Hirsch and Ehrlicher, 1995) and a VAD-based scheme in different types of noise. In most situations, the VAD-based SNR estimator yielded the best results.
A fast SNR estimation scheme based on higher order statistics was introduced by Nemer et al. (1988). It analyses the kurtosis of noisy speech and uses a sinusoidal model for speech and a Gaussian assumption for noise. The authors report a fast and accurate estimation of the local SNR when these assumptions are met, which is the case for most mobile communication situations.

While technical VAD algorithms often fail to robustly detect speech pauses (especially in situations with low SNR), humans can easily de-

tect and classify different sound sources, and separate between speech and noise without problems. This is made possible by the interplay between the internal representation of sounds in the auditory system, and the higher processing stages in the brain which perform classification, recognition, and understanding basing on this internal representation. It is still unclear which are the most important features and cues within the acoustical waveform that allow for such impressive skills. Besides the well-known analysis and tonotopical representation of different center frequencies in the auditory system (e.g., on the basilar membrane), the analysis of amplitude modulations is assumed to provide further important information for human speech processing. Low modulation frequencies, for example, are known to play an important role for speech intelligibility. Drullman et al. (1994) found that modulation frequencies up to 8 Hz are the most important ones in for speech intelligibility. Shannon et al. (1993) conducted an impressive study on the importance of temporal amplitude modulations for speech intelligibility and observed nearly perfect speech recognition under conditions of highly reduced spectral information.

However, there is a difference between *understanding* speech and *detecting* speech. In a noisy canteen environment, for example, we can classify a very short prominent segment of speech as "human speaking" (and not, for example, "dog barking", or "cup being smashed"), even if we do not understand the meaning. In this case, low modulation frequencies in speech which are important for speech intelligibility probably play a minor role only. Thus, it is important to notice the difference between speech detection (or, in a wider sense, detection of acoustical objects), and speech intelligibility. Higher modulation frequencies which represent pitch information or harmonicity are likely to be more important for speech detection and sound classification.

During recent years, more insight has been gained about the coding of amplitude modulations in the auditory system. In psychoacoustical experiments, the auditory system's frequency selectivity for amplitude modulations were specified (Bacon and Grantham, 1989; Houtgast, 1989). Dau et al. (1997a; 1997b) showed that a separation of envelope fluctuations into different modulation frequency bands ("modulation filterbank") provides an adequate prediction of various psychoa-

coustical experiments. Ewert and Dau (1999) measured the shape of the "critical bands" in the envelope-frequency domain, or modulation filters, for target-modulation frequencies between 4 and 256 Hz with a noise carrier ranging from 1 to 4 kHz. Their results show that for low modulation frequencies the shapes of the modulation filters are reasonably symmetric on the logarithmic envelope-frequency scale, with almost constant quality factor (i.e., they become wider with increasing modulation frequency and have a more or less "form-invariant" modulation-filter shape on a logarithmic envelope-frequency scale). At higher test-modulation frequencies, the pattern broadens and becomes slightly asymmetric.
In neurophysiological experiments, Langner and Schreiner (1988), among others, found neurons in the inferior colliculus and auditory cortex of mammals which were tuned to certain modulation frequencies. The "periodotopical" organization of these neurons with respect to different best modulation frequencies was found to be almost orthogonal to the tonotopical organization of neurons with respect to center frequencies. Thus, a two-dimensional "feature set" represents both spectral and temporal properties of the acoustic signal. Recently, Langner et al. (1997) observed periodotopical gradients in the human auditory cortex by means of magnetoenzephalography (MEG). As stimuli, they used pure tones between 50 Hz and 1.6 kHz, and harmonic sounds which were composed of harmonics of 50-400 Hz and thus eclicted a pitch corresponding to these fundamental frequencies. All harmonic sounds had an upper cut-off frequency of 5 kHz, and the lower cut-off frequency was either 400 Hz or 800 Hz. Thus, both frequency range and pitch of their stimuli were in the range which is important and characteristic for human speech.
Kollmeier and Koch (1994) applied these psychoacoustical and neurophysiological findings in the field of digital signal processing and introduced two-dimensional patterns, so-called Amplitude Modulation Spectrograms (AMS) which contain information on both center frequencies and modulation frequencies for a binaural noise suppression scheme. They reported a small but stable improvement in terms of speech intelligibility, compared to unprocessed speech. Recently, similar kinds of feature patterns were applied to vowel segregation (Yang *et al.*, 1999), speech enhancement (Strube and Wilmers, 1999), and sound signal classification (Tchorz and Kollmeier, 1999a).

The SNR estimation algorithm which is outlined in this paper is also based on AMS patterns. In contrast to common VAD algorithms, it does not provide a binary speech/non-speech decision, but also covers the range in-between speech and noise by directly predicting the local SNR of the signal at every instant. In contrast to most other direct SNR estimation schemes, the proposed algorithm does not require relatively long segments of the input signal which should contain speech pauses or closures, but estimates the SNR from short analysis frames (typically 32 ms). Thus, the algorithm can almost instantaneously follow rapid changes in the acoustical situation.
The remainder of this paper is structured as follows. In Section 2.2, the algorithm and its processing steps are outlined. In Section 2.3, the SNR prediction experiments and their results are described. In Section 2.4, a comparison between the proposed SNR estimation algorithm with a VAD-based estimator is given. The question which of the extracted features contribute most to reliable SNR prediction is dealt with in Section 2.5. Section 2.6 examines the possibilities to enhance the accuracy of SNR prediction by extending the period of time which is considered to provide an estimate of the SNR. A discussion can be found in Section 2.7.

2.2 Classification Algorithm

The general idea of the classification algorithm described in this paper is to transform the incoming waveform into a series of neurophysiologically-motivated AMS patterns (Amplitude Modulation Spectrograms) which are assumed to carry sufficient information for speech/noise detection and SNR estimation. An artificial neural network is trained on a large number of AMS patterns which are generated from mixtures of speech and noise under defined conditions. After training, the response of the network when presenting AMS pattern from “unknown” sound samples serves as estimate of the local SNR.

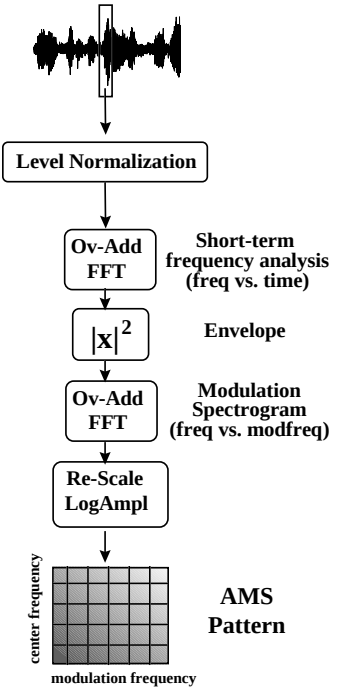


Figure 2.1: Signal processing steps for AMS pattern generation

2.2.1 Signal Processing

Figure 2.1 shows the processing steps which are performed to generate AMS patterns.

First, the input signal which was digitized with 16 kHz sampling rate is long-term level adjusted, i.e., changes in the overall level are compensated for, whereas short-term level differences (e.g., those between

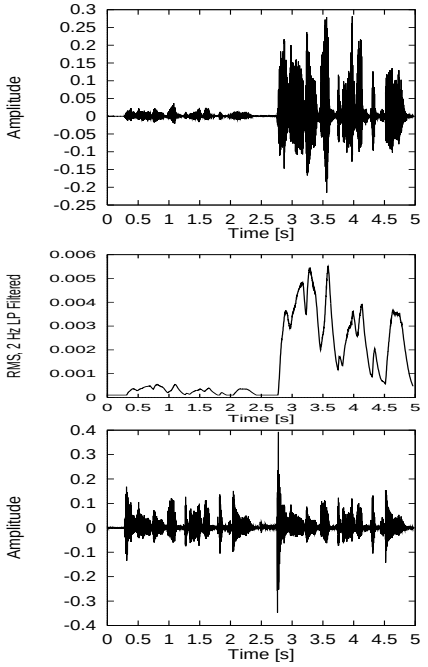


Figure 2.2: Level normalization scheme: Original input signal (top panel), level normalization function (middle), and normalized signal (bottom).

successive phonemes) are maintained to serve as additional cues for classification. This level adjustment is realized by dividing the input signal by its low pass filtered root-mean-square (rms) function which was calculated from 32 ms frames, with an overlap of 16 ms. The cut-off frequency of the low pass filter is 2 Hz. One example for long-term level adjustment is shown in Fig. 2.2. The input signal (top panel) is a concatenation of two identical sentences, the first sentence having a much smaller amplitude than the second one. The second panel shows the corresponding level normalization function. For normalization, the input signal is divided by this function. To avoid divisions by zero, the normalization function is limited by a lower threshold. The bottom panel shows the normalized signal. The overall level is equal for both sentences, but the local level fluctuates due to the amplitude variations between syllables and words. A short peak occurs at the onset of the second sentence in

the bottom panel, which is due to the normalization process. At the onset of the second sentence, the nominator is large, but the denominator is still quite small, until it adapts after a few ms.
In a following processing step, the level-adjusted signal is subdivided into overlapping segments of 4.0 ms duration with a progression of 0.25 ms for each new segment. Each segment is multiplied with a Hanning window and padded with zeros to obtain a frame of 128 samples which is transformed with a FFT into a complex spectrum, with a spectral resolution of 125 Hz. The resulting 64 complex samples are considered as a function of time, i.e., as a band pass filtered complex time signal. Their respective envelopes are extracted by squaring. This envelope signal is again segmented into overlapping segments of 128 samples (32ms) with an overlap of 64 samples. Each segment is multiplied with a Hanning window and padded with zeros to obtain a frame of 256 samples. A further FFT is computed and supplies a modulation spectrum in each frequency channel, with a modulation frequency resolution of 15.6 Hz. By an appropriate summation of neighbouring FFT bins the frequency axis is transformed to a Bark scale with 15 channels, with center frequencies from 100-7300 Hz. The modulation frequency spectrum is scaled logarithmically by appropriate summation, which is motivated by psychoacoustical findings on the shape of auditory modulation filters (Ewert and Dau, 1999). The modulation frequency range from 0-2000 Hz is restricted to the range between 50-400 Hz and has a resolution of 15 channels. Thus, the fundamental frequency of typical voiced speech is represented in the modulation spectrum. The chosen range corresponds to the fundamental frequencies which were used by Langner et al. in their neurophysiological experiments on amplitude modulation representation in the human auditory cortex (Langner *et al.*, 1997). Informal experiments showed that higher modulation frequencies do not contribute additional information for the task of speech/noise detection. Very low modulation frequencies from articulatory movements, which are characteristic for speech and which play an important role for speech intelligibility are also not taken into account, as they are not properly resolved due to the short analysis windows. Furthermore, the goal of the presented algorithm is not in the field of speech *intelligibility* or automatic speech recognition, but rather on speech / noise *detection*

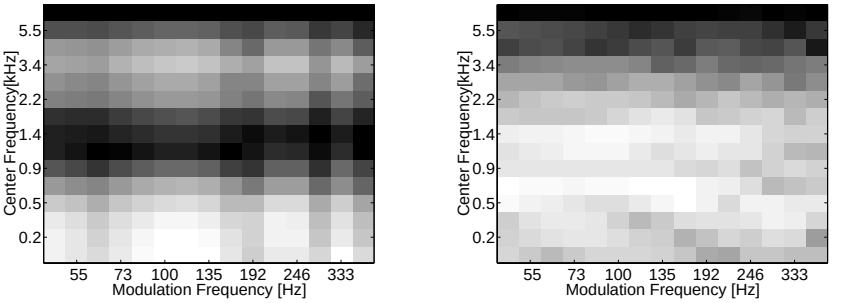


Figure 2.3: AMS patterns generated from a voiced speech segment (left), and from speech simulating noise (right). Each AMS pattern represents a 32 ms portion of the input signal. Bright and dark areas indicate high and low energies, respectively.

and SNR estimation in short analysis frames. The AMS representation is restricted to a 15 times 15 pattern to keep the amount of training data which is necessary to train a fully connected perceptron manageable, as this amount increases with the number of neurons in each layer.
In a last processing step, the amplitude range is log-compressed. Examples for AMS patterns can be seen in Fig. 2.3. Bright and dark areas indicate high and low energies, respectively.

The left AMS pattern was generated from a voiced speech portion, uttered by a male speaker. The periodicity at the fundamental frequency (approx. 110 Hz) is represented in each center frequency band, as well as the first and second harmonics. Due to the short length of the analysis frame (32 ms), the modulation frequency resolution is limited, and the peaks indicating the fundamental frequency are relatively broad. The right AMS pattern was generated from speech simulating noise (CCITT, 1964), i.e. noise with the same spectrum as the long-term spectrum of speech. The typical spectral tilt can be seen, which is due to less energy in higher frequency channels, but no structure across modulation frequencies such as harmonic peaks, and no similarities between modulation spectra in different frequency channels, as in the upper panel.

2.2.2 Neural Network Classification

A feed-forward neural network was used for the classification task (Zell *et al.*, 1995; SNNS, 1995). It consists of an input layer with 225 neurons (15 times 15, the resolution of AMS patterns, which are directly fed into the network), a hidden layer with 40 neurons, and an output layer with just one output neuron. The network was trained with 100 cycles using the Backpropagation-Momentum algorithm (Rumelhart *et al.*, 1986).

For training, a set of AMS patterns generated from noisy speech is presented to the network. The signal-to-noise ratio within each 32 ms AMS analysis frame is measured prior to adding speech and noise following the equation $\mathrm{SNR}_{[dB]} = 10\log(S^2/N^2)$, where S and N are the rms values of the speech and the noise signal in the respective analysis frame. The mixtures of speech and noise were generated artificially to allow for SNR control. Typical noisy speech effects such as Lombard speech are thus not taken into account. The local SNR which is measured within the analysis frame of an AMS pattern determines the target activity for the output neuron during training. A high AMS pattern SNR results in a target output neuron activity close to one, a low SNR in a target activity close to zero. The local SNR values which are considered range from -10 dB to 20 dB. This range is linearly transformed to output neuron activities from 0.05 to 0.95. The transformation function between the measured SNR and the target activity is plotted in Fig. 2.4. After training, the output neuron activity which occurs when presenting an AMS pattern generated from an "unknown" sound source is transformed using the function plotted in Fig. 2.4 and supplies an estimate of the local SNR.

Thus, the algorithm provides an estimation of the SNR in independent 32 ms frames. The SNR is directly predicted even if speech and noise are present at the same time, which is in contrast to "indirect" SNR estimation, where the accurate detection of speech pauses is necessary for noise energy measurement, and stationarity of noise is assumed during speech activity.
Currently, the SNR estimation algorithm is implemented on a SGI O2 R5000 (200MHz) workstation and requires about 8-fold real time for

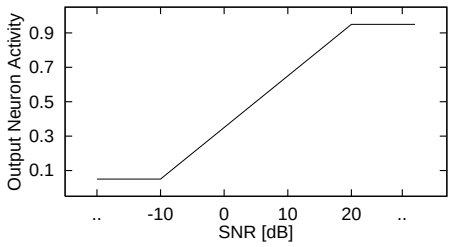


Figure 2.4: The transformation function which maps the local SNR onto output neuron target activity.

processing.

2.3 SNR estimation experiments

2.3.1 Setup

In the training phase, the neural network "learns" the characteristics of AMS patterns in different SNRs. During training, the whole range of possible SNRs should be covered with a sufficient number of representations. In total, 72 min of noisy speech with an overall SNR of 5 dB were transformed into 270000 AMS patterns, which were then presented to the network for training. The speech samples were taken from the "PhonDat" database (Kohler *et al.*, 1994) and contained 2110 German sentences from 190 male and 210 female talkers. The speech data contained only short segments of silence between the sentences. 41 types of natural noise were taken for training from various data bases. The network was trained with 100 cycles. For testing, a 36-min mixture of speech (200 speakers, PhonDat) and 54 noise types with an overall SNR of 5 dB was taken. The talkers and noise types for testing were not included in the training data. The local SNRs of the mixtures of speech and noise exhibited strong fluctuations. Histograms of the relative frequencies of the local SNRs of the training and the test material are plotted in Fig. 2.5.

The distributions of both data sets are very similar and have their maximum frequency at about 5 dB SNR.

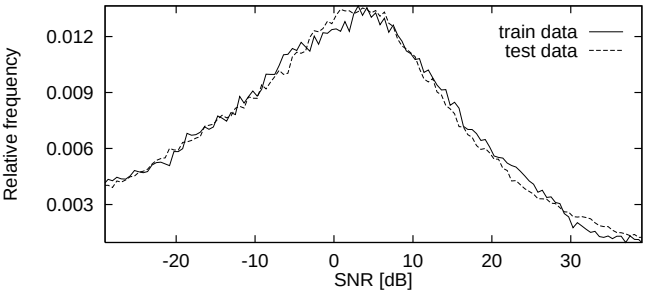


Figure 2.5: Histograms of the relative frequencies of the local SNRs of the mixtures of speech and noise for training and testing.

2.3.2 Results

An example of SNR estimation with the classification algorithm after it has been trained is illustrated in Fig. 2.6. It shows the actual local SNR as measured prior to adding speech and noise (solid), and the estimated SNR (dotted) for speech in non-stationary printing machine noise, which was part of the test set. It can be seen that the estimated SNR corresponds well with the actual SNR (except in very low SNRs) and follows it almost instantaneously. The above described example provides a qualitative impression of the performance of the classification algorithm. A quantitative measurement of the estimation accuracy is obtained by computing the mean deviation D between the actual SNR a_i and the estimated SNR e_i over N processed AMS patterns (with index i):

$$D = \frac{1}{N} \sum_{i=1}^{N} |e_i - a_i| \tag{2.1}$$

The mean SNR deviations D measured in the experiments are shown in Tab. 2.1 (first row). For the test set (as described in Section 2.3.1), the mean deviation between the actual SNR and the estimated SNR was 5.2 dB. When estimating the SNR of the training material, the algorithm achieved a mean deviation of 4.1 dB. Thus, the network failed to perfectly reproduce the training data, but it generalizes to “unknown”

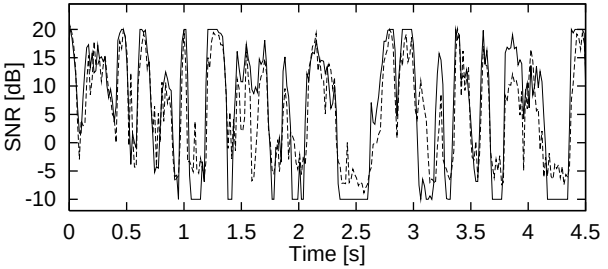


Figure 2.6: Example for SNR prediction for speech in non-stationary printing machine noise. The solid line shows the actual SNR as measured prior to adding speech and noise, the dotted line shows the estimated SNR provided by the classification algorithm.

	test set	train set
original algorithm	5.2	4.1
no level information	5.8	4.4
only spectral information	7.6	6.1
only modulation information	6.6	6.3
modulation + spectral information	5.8	4.8
16 ms analysis frames	5.8	4.5
64 ms analysis frames	5.3	4.3
128 ms analysis frames	4.6	3.3

Table 2.1: SNR prediction accuracy in terms of dB mean deviation (2.1) obtained with the classification algorithm and its modifications (Section 2.5).

sound sources, as the degradation of the performance for the test data is limited.

In Fig. 2.7, the SNR estimation accuracy as a function of the measured SNR for the test set is plotted. The solid line shows the absolute deviation from the measured SNR. For low input SNRs, the performance in general is worse than for high input SNRs. The dotted line shows the

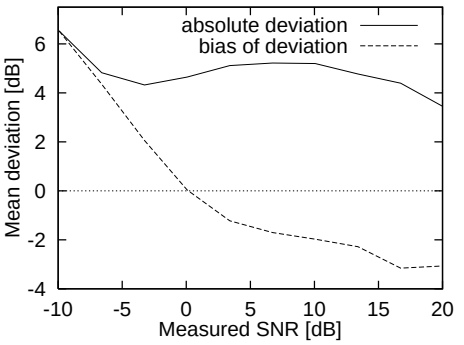


Figure 2.7: The solid line shows the absolute estimation error (mean deviation) as a function of the measured input SNR for the test set. The dotted line shows the bias of the SNR estimation depending on the input SNR. Low input SNRs in average are over-estimated, high input SNRs under-estimated.

bias of the SNR estimation depending on the input SNR, i.e., there are no absolute values computed in (2.1). For low input SNRs, there is a bias towards over-estimation of the SNR ($e_i - a_i$ is positive). Close to -10 dB SNR, almost all estimation errors are from over-estimation of the SNR. For high input SNRs, in contrast, the algorithm tends to under-estimate the SNR. The SNR range below -5 dB causes the highest estimation errors. In such low short-term SNRs speech is almost entirely masked by noise, and a difference between -5 and -10 dB in isolated frames is hardly audible, as informal listening experiments showed. If the SNR range below -5 dB is excluded from the evaluation, the overall mean deviation for the test set is 4.5 dB, compared to 5.2 dB including the very low SNRs.

2.4 Comparison with VAD-based SNR estimation

The performance of a voice-activity-detection (VAD) - based SNR estimation was compared to the SNR estimation approach outlined in this paper. For voice activity detection, a VAD standardized by ITU was used (ITU, 1996) that utilizes information on energy, zero-crossing rate, and spectral distortions. The noise energy estimate is updated in de-

tected speech pauses and low-pass filtered with a first-order FIR filter to avoid fast fluctuations of the estimate. The time constant of the low-pass filter was set to a value of 60 ms which was optimized for this experiment. The VAD-based instantaneous SNR was computed in 10 ms analysis frames:

$$\mathrm{SNR}_{[dB]} = 10 \log(\frac{R^2}{N^2} - 1), \tag{2.2}$$

where R is the rms of the signal plus noise in the analysis frame, and N is the present noise rms estimate. The mean deviation between the VAD-based estimate of the instantaneous SNR and the measured SNR was 5.4 dB on the test data described in Sec. 2.3.1. Thus, its accuracy is comparable to the AMS-based approach proposed in this paper (5.2 dB). The reliability of the VAD-based estimator of course strongly depends on the stationarity of the background noise. In constant noise, there is no advantage of the AMS-based approach to be expected. Fast fluctuations of the noise energy while speech is active, in contrast, can be followed by the AMS algorithm, but not by the VAD-based estimator. This is illustrated in Fig. 2.8 and 2.9. In Fig. 2.8, the input signal was a mixture between speech and stationary white noise, which are plotted separately on top. The VAD-based SNR estimator allows for almost perfect SNR prediction in this situation, which can be seen from the first panel. There is only very little difference between the measured (solid) and the estimated SNR (dotted). The AMS-based SNR estimator, in contrast, tends to over- and underestimate the SNR in very low and very high measured SNRs, respectively (second panel). In Fig. 2.9, the input signal was a mixture between speech and non-stationary construction site noise. In this situation, the VAD-based SNR estimator fails to update the noise measure. Thus, the noise burst which starts at $t = 0.6$ s leads to a large overestimation of the SNR, as the additional energy is regarded as "speech". The same holds for the smaller noise peaks at the end of the signal. Here, the AMS-based SNR estimator better tracks the acoustical situation as it is not dependent on explicit speech pause detection.

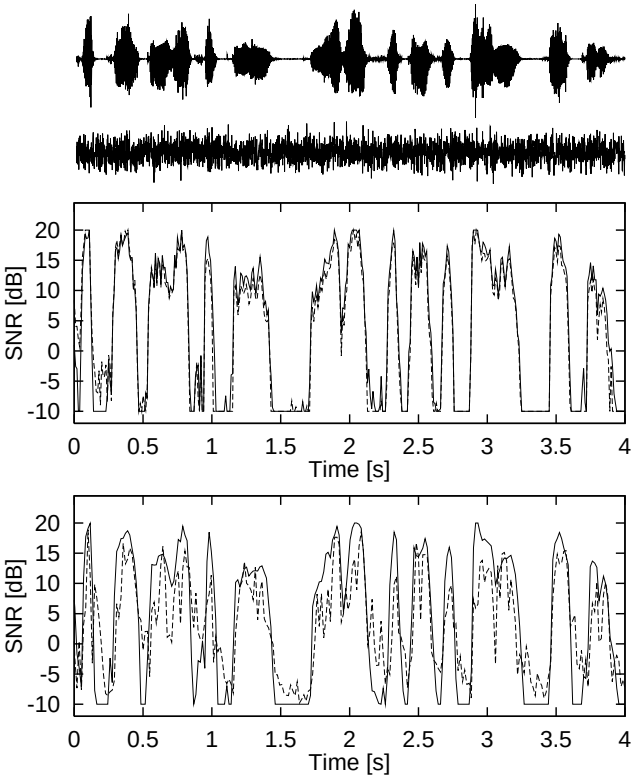


Figure 2.8: Comparison between VAD-based and AMS-based SNR estimation with a mixture of speech and stationary white noise (plotted separately on top). The solid and the dotted line show the measured and the estimated SNR, respectively. The VAD-based approach (first panel) yields almost perfect SNR estimation. The AMS-based estimator (second panel) tends to over- and underestimate the SNR in very low and very high measured SNRs, respectively.

2.5 Which features are important?

The algorithm based on AMS pattern recognition presented here was shown in the previous sections to provide a relatively accurate SNR

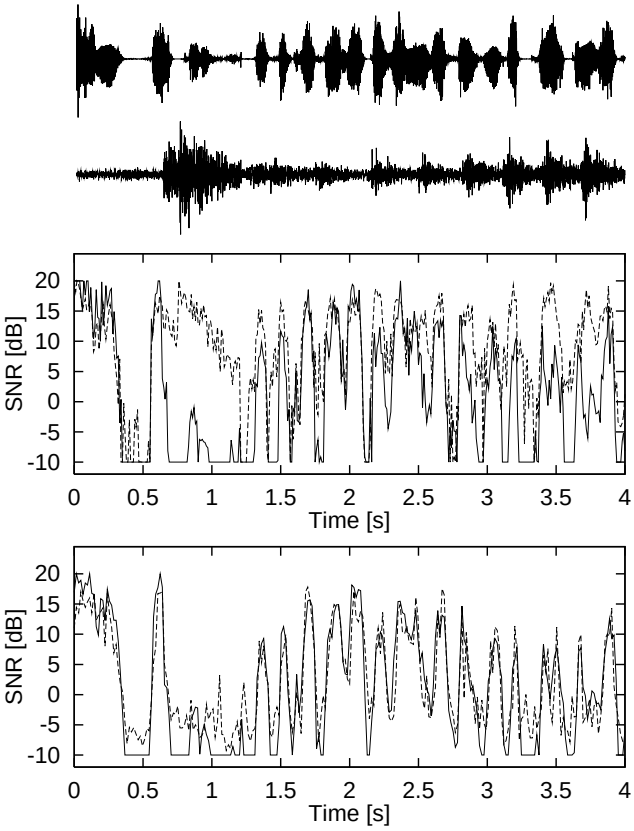


Figure 2.9: Comparison between VAD-based and AMS-based SNR estimation with a mixture of speech and non-stationary construction site noise (same notation as Fig. 2.8). Here, the VAD-based approach (first panel) cannot properly update the noise measure, in contrast to the the AMS-based estimator (second panel).

prediction for short analysis frames of unknown sound signals. This Section analyzes the features of the neurophysiologically-motivated AMS patterns that contribute most to successful classification of sounds and

reliable SNR estimation.

2.5.1 Modifications of AMS signal processing

There are three major dimensions of information encoded in AMS patterns: a) center frequency information, b) modulation frequency information, and c) local level information. SNR prediction experiments were carried out in which at a time one of these sources of information was eliminated in order to study its contribution to accurate SNR prediction.

2.5.1.1 Eliminating level information

The algorithm presented here performs an overall level compensation in the first signal processing step, i.e., the overall level is normalized, but local level differences of the input signal (e.g., between neighbouring phonemes) are maintained. As a consequence, AMS patterns generated from soft consonants, for example, exhibit smaller amplitudes than those computed from high-energy vowels. This energy information encoded in the AMS patterns might play a certain role for SNR prediction.
The elimination of level information was quantitatively explored in the following way: Instead of preserving local energy fluctuations as in the original algorithm described in Section 2.2.1, the energy of each 32 ms frame of the input signal (which is later transformed to an AMS pattern) was normalized to the same RMS value prior to AMS pattern generation. A SNR prediction experiment using this modified algorithm was performed with the same experimental setup as described in Section 2.3.1. The achieved SNR prediction accuracy in terms of mean SNR deviation (2.1) can be seen in Tab. 2.1 (second row). Without explicit level information, the mean deviation for the test set increased from 5.2 dB to 5.8 dB, which means a small degradation in prediction performance. The classification of the training data also was only slightly affected by the modification. Thus, implicit level information encoded in AMS patterns only provides a limited benefit for accurate SNR prediction since omitting this dimension of information has not a large impact.

2.5.1.2 Eliminating modulation information

In AMS patterns, modulation frequencies between about 50 and 400 Hz in different center frequency channels are encoded by the modulation spectra which are computed for each channel. Harmonicity in voiced speech, for example, is represented on the modulation axis by peaks at the fundamental frequency and its harmonics, which leads to characteristic AMS patterns for voiced speech. This dimension of information was removed in order to assess the importance of modulation information for SNR prediction. For this experiment, the long-term level normalized input signal was segmented into overlapping segments of 32 ms duration with a progression of 16 ms for each new frame. Each segment was multiplied with a Hanning window and transformed into its spectrum with a FFT. By appropriate summation of neighbouring FFT bins, the frequency axis was scaled logarithmically with a resolution of 15 channels with center frequencies ranging from 100-7300 Hz. The amplitude was log-compressed. These operations are similar to summing up the energy across all modulation frequencies for a given frequency channel in the two-dimensional representation given in Fig. 2.3 and hence converting it to a one-dimensional representation. Thus, the frequency resolution and level normalization of these new feature vectors is the same as in the original AMS patterns, but additional amplitude modulation information is missing. The neural network as described in Section 2.2.2 (with only 15 input neurons instead of 225) was trained and tested with the data described in Section 2.3.1. The SNR prediction accuracy obtained without modulation frequency information is given in Tab. 2.1 (3rd row). The estimation accuracy for the test set degraded from 5.2 dB to 7.6 dB. The mean deviation for training data classification increased from 4.1 dB to 6.1 dB. Thus, explicit analysis and representation of amplitude modulations appears to be a helpful dimension of information for accurate SNR prediction.

2.5.1.3 Eliminating spectral information

The third main dimension of information in AMS patterns besides level and modulation frequencies is the encoding of the signal spectrum (as in Fig. 2.3, where the typical spectral tilt of speech simulating noise can

be seen). This dimension of information was removed by modifying the original signal processing as follows: The long-term level normalized input signal was filtered using a 4th order band pass filter with cut-off frequencies of 100 and 7300 Hz in order to exploit the same overall frequency range of the signal as in the original algorithm. The envelope of the filtered signal was extracted by squaring and was segmented into 32 ms frames with a 16 ms shift. Subsequently it was Hanning-windowed and transformed into its modulation spectrum with a FFT. The modulation frequency axis was scaled logarithmically yielding a resolution of 15 channels with best modulation frequencies ranging from 50-400 Hz, and the amplitude was log-compressed. This operation is similar to summing up the energy across all frequency bands in the two-dimensional representation given in Fig. 2.3 and hence converting it into a one-dimensional representation of modulation frequencies. Thus, the amplitude modulation frequency resolution and level normalization of these new feature vectors are equal as in the original AMS patterns, but in this case, spectral information is missing. Again, a neural network was trained and tested on these reduced features, as described in Section 2.3.1. The results are shown in Tab. 2.1 (4th row). The classification accuracy for the test set degraded from 5.2 dB to 6.6 dB, and for the training data from 4.1 dB to 6.3 dB, respectively. Thus, the estimation of the local SNR from modulation cues only is more accurate than with spectral cues only, but still less reliable than with the joint representation as in the original algorithm.

2.5.1.4 Combination of spectral only and temporal only information

SNR estimation based on spectral or temporal cues alone yields a decrease in estimation accuracy in comparison to full AMS patterns. The "one-dimensional" patterns described in Sec. 2.5.1.2 and 2.5.1.3 were combined in a further experiment yielding patterns with both spectral only and temporal only information, but with only 30 dimensions instead of 225 dimensions as in the full AMS pattern, which reduces the computational load and storage requirements. The mean deviation in SNR estimation accuracy based on these reduced spectro-temporal features is given in Tab. 2.1 (5th row). Compared to the full joint representation,

the mean deviation increases from 5.2 dB to 5.8 dB for the test data and from 4.1 dB to 4.8 dB for the training data, respectively. Thus, the performance of the combined patterns is better than with spectral or temporal cues only, but still worse than with the full joint representation.

2.5.2 Artificial input signals

2.5.2.1 How important is harmonicity?

Harmonicity of sounds is well represented in AMS patterns, which can be seen in the two examples given in Fig. 2.3. For voiced speech, the fundamental frequency and its first two harmonics can easily be seen, in contrast to the pattern generated from (non-harmonic) speech simulating noise. Harmonicity is an important feature of voiced speech, and the question arises whether harmonicity is an important cue for SNR estimation based on AMS patterns. To determine the influence of harmonicity on the output neuron activity of the neural network (which serves as estimate for the SNR), artificial input signals with varying degrees of harmonicity were generated. The signals were composed of a fundamental frequency of 150 Hz and its harmonics up to 8 kHz, with all harmonics having the same amplitude. The frequencies of the harmonics were individually randomly shifted following the equation $f_{\text{shift}} = f + \text{rand}[-x..x]$, where f is the frequency of the respective harmonic, and x is a frequency between 0 and 150 Hz. For $x = 0$, the resulting signal is a tone complex with frequencies 150 Hz, 300 Hz, 450 Hz, etc. With increasing x harmonicity gets lost and the resulting sound becomes a random composition of sine waves. The output neuron activity for these artificial input signals as a function of x is plotted in Fig. 2.10. With increasing x and loss of harmonicity, the output neuron activity decreases until it reaches values which would indicate clear dominance of noise.

2.5.2.2 Variation of the fundamental frequency

The above described experiment demonstrated that harmonicity with a fundamental frequency typical for human speech is an important cue for an input signal to be classified as “speech-like”. The influence of

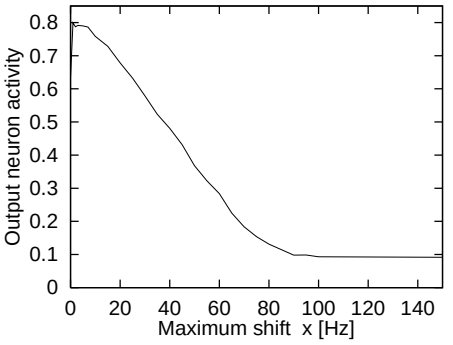


Figure 2.10: Output neuron activity for an artificial harmonic sound with increasing random shift x of the harmonic frequencies.

the fundamental frequency of harmonic sounds on the output neuron activity was determined in a further experiment, where a synthetically generated vowel ("a") with varying fundamental frequency served as input signal for the neural network. The resulting output neuron activity as a function of the fundamental frequency is plotted in Fig. 2.11. It can be seen that the maximum output neuron activities occurs at fundamental frequencies typical for human speech, but only slightly degrades for fundamental frequencies above this range. Note that the output neuron activity for a synthetic vowel with a fundamental frequency of 150 Hz is higher than for a tone complex with same fundamental frequency (Fig. 2.10), which indicates that harmonicity is not the only cue which is utilized for classification. In contrast to the tone complex, the synthetic vowel is also characterized by a speech-like formant structure and a spectral tilt.

2.5.2.3 The influence of additive noise

The impact of additive noise on the output neuron activity was evaluated in a further experiment. A synthetic vowel with fundamental frequency 150 Hz was distorted with additive speech simulating noise at different SNRs. The resulting output neuron activity as a function of the SNR is plotted in Fig. 2.12. Below -5 dB, the algorithm does not detect differences in the SNR. Above -5 dB, the activity increases monotonically with the SNR in a sigmoid-like curve. This corresponds well with the

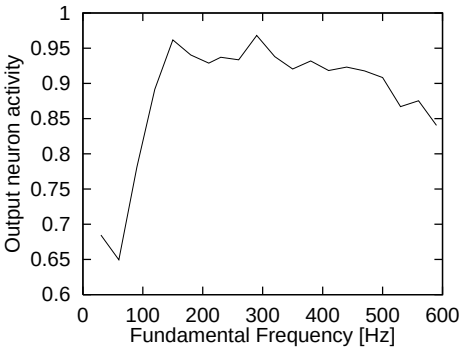


Figure 2.11: Output neuron activity for an synthetic vowel as a function of the fundamental frequency.

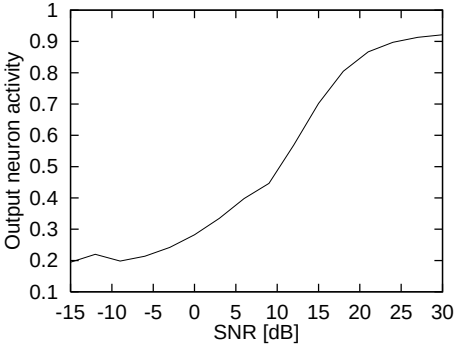


Figure 2.12: Output neuron activity for an synthetic vowel distorted with additive noise as a function of the SNR.

results from Sec. 2.3.2, where it was shown that the algorithm tends to over- and underestimate very low and very high SNRs, respectively.

2.5.3 Does the algorithm only track voiced speech?

The preceeding experiments revealed that harmonicity is an important cue for a signal to be classified as "speech". If harmonicity was the only cue, however, the computationally expensive effort for AMS signal processing and pattern recognition would not be necessary, as harmonicity can be tracked much easier. Furthermore, the algorithm could not detect unvoiced speech which does not exhibit harmonicity. Thus, the perfor-

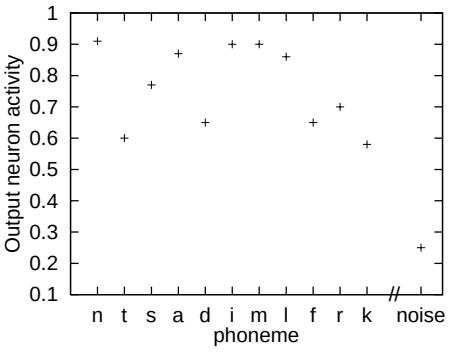


Figure 2.13: Average output neuron activity for different phonemes and noise.

mance of the algorithm in voiced and unvoiced speech was evaluated in an additional experiment. The average output neuron activity was measured depending on the present phoneme. In total, 1350 phonetically labelled sentences from the PhonDat database spoken by 45 speakers were processed (without adding noise) and classified by the network. The average output neuron activities for the most frequent phonemes are plotted in Fig. 2.13. For voiced phonemes, values around 0.9 are measured, whereas for unvoiced phonemes the average activity is between about 0.6 and 0.7. This is still well above the average activity for non-speech input, which was 0.25 for the noise data from the test set described in Sec. 2.3.1. These results indicate that harmonicity in voiced speech is probably the most important cue for classification and SNR estimation, but even without harmonicity, the spectral and temporal characteristics of speech and noise can be discriminated by the pattern recognizer from the joint representation in AMS patterns. Fig. 2.13 demonstrates that the proposed algorithm could be used as a binary speech/noise detector when an appropriate output neuron activity threshold is chosen.

2.5.4 Varying the analysis frame

The length of the analysis frame for AMS patterns is determined by the window length of the FFT which computes the modulation spectrum in each frequency channel, which was 32 ms for the above described

experiments. The influence of the frame length was determined in a further experiment. The analysis window was set to 16, 64 and 128 ms by variation of the FFT length. For all window sizes, logarithmically scaled modulation frequency bands were computed by averaging across neighbouring FFT-bins. The resolution of the modulation spectrum in each frequency band was kept constant using 15 bands ranging from 50-400 Hz, as in the original algorithm. Hence the AMS patterns for the different window sizes did not differ in the number of pixels evaluated by the neural net. Instead, they only differed with respect to the amount of input data that entered each AMS pattern. Neural networks were trained on the modified AMS patterns and the local SNR of the data was estimated. The mean deviations between the estimated and the measured SNR for the different analysis frames are shown in Tab. 2.1 (row 6-8). There is only a small difference in performance between the original algorithm (32 ms) and 64 ms analysis frames, whereas 16 ms and 128 ms frames reduced and enhanced the accuracy, respectively.

2.6 Low pass filtering of SNR trajectories

In the SNR prediction experiments described in the preceding sections, a SNR estimation was computed for each AMS pattern independently. The estimation based on the short-term analysis of single 32 ms frames of the input signal, without taking its temporal context into account. This allows for relatively fast SNR prediction, and the algorithm is able to quickly follow rapid changes of the sound situation. This has to be paid for by a limited SNR prediction accuracy, as the information available in isolated 32 ms frames is probably in principle not sufficient for very high estimation precision. Some possible applications of the algorithm, however, may not require a very fast update of the local SNR. In these cases it is possible to enhance the prediction accuracy of the algorithm by low pass filtering the time trajectory of successive SNR estimates. (A typical time trajectory of successive SNR estimates is shown in Fig. 2.6). By low pass filtering, "outliers" and prediction errors are smoothed. Figure 2.14 illustrates the gain in estimation accuracy by low pass filtering the time trajectories of the actual and the estimated SNR as a function of the filter's cutoff frequency, which varied from

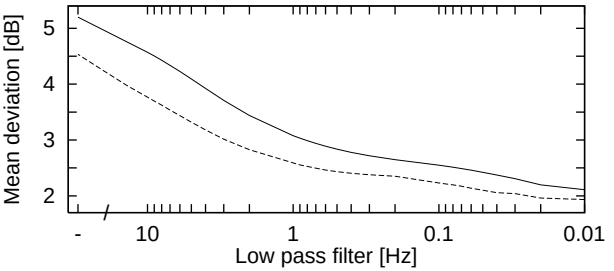


Figure 2.14: Effect of low pass filtering the time trajectories of the actual and the estimated SNRs for the test set. The mean deviation (solid line) decreases with decreasing cutoff frequency, as "outliers" are smoothed. The dotted line shows the mean deviation excluding the frames with a local SNR below -5 dB.

10 Hz down to 0.01 Hz. It can be seen how the mean deviation between the measured and the estimated SNR decreases with decreasing cutoff frequency, as short-term prediction errors are smoothed. On the other hand, of course, the sluggishness of the system increases, as the SNR prediction output does not instantaneously follow a new acoustical situation. Low pass filtering with 1 Hz, for example, lowers the mean deviation for the test set from 5.2 dB to 3.1 dB, but adaptation to a new acoustical situations takes a few hundred milliseconds then. The dotted line shows the mean deviation excluding the frames with a local SNR below -5 dB, where speech is almost entirely masked by noise and where the highest estimation errors occur (see Fig. 2.7).

2.7 Discussion

The main findings of this study can be summarized as follows:

- Neurophysiologically and psychoacoustically motivated Amplitude Modulation Spectrograms (AMS), in combination with artificial neural networks for pattern recognition, allow for automatic SNR estimation in a range of different acoustic situations.

- Three main dimensions of information are encoded in AMS patterns: information on center frequency, amplitude modulations, and local level fluctuations. All three dimensions contribute to the reliability of the classification algorithm.
- Harmonicity appears to be the most important cue for analysis frames to be classified as "speech-like", but the spectro-temporal representation of sound in AMS patterns also allows for reliable discrimination between unvoiced speech and noise.
- SNR estimation works for independent 32 ms analysis frames, but estimation accuracy can be enhanced by temporal smoothing of SNR estimates or enlarging the analysis frame.

In contrast to VAD algorithms which often focus on analyzing the spectrum of the signal, the sound classification approach presented in this paper explicitly utilizes temporal information by analyzing amplitude modulations between 50 Hz and 400 Hz. Providing this additional information in AMS patterns was found useful for automatic SNR estimation, as shown in Section 2.5.1.2. In the field of noise classification, similar findings were reported by Kates (1995). His results showed that envelope fluctuation features did add a significant amount of information to the noise classification.
In fact, the experiments from Section 2.5.1.3 indicate that signal classification and SNR prediction is possible to some extend *without* spectral information but only from the modulation spectrum.
The range of modulation frequencies which is considered in the presented approach is well above the modulation frequencies which are typical and characteristic for speech, namely those around 4 Hz. These slow modulations play an important role for speech intelligibility, but speech/noise classification basing on slow modulations requires long analysis frames (Ostendorf *et al.*, 1998) and cannot instantaneously detect a sudden change in the acoustical situation. Nevertheless, in applications such as hearing instruments, a slow adaptation rather than a fast one can be sufficient or even desired, as sudden changes of sound processing may irritate or annoy the user.
The experiments in Sec. 2.5 were intended to improve the insight into the mechanisms which are involved in the proposed SNR estimation

process (e.g., how important is harmonicity and its fundamental frequency? What is the impact of systematic variation of the noise level? How reliable is the detection of voiced speech vs. unvoiced speech?). These experiments revealed that harmonicity is the most important cue for speech detection, but harmonicity alone is not sufficient for accurate SNR estimation (Sec. 2.5.1.3), and not the only cue for speech detection (as unvoiced speech leads to much higher output neuron activities, compared to noise, Sec. 2.5.3). Thus, the spectro-temporal joint representation in AMS patterns cannot be replaced by a simple pitch detector (which would require less computational effort).

Our experiments demonstrate that amplitude modulation analysis which tries to mimic auditory modulation processing in a simple way is helpful for technical sound signal classification. However, the question whether auditory modulation analysis contributes to human sound detection and classification remains untouched by these experiments. It is still unclear which features of the acoustical waveform are considered by humans to perform auditory tasks such as signal classification, detection, and separation of different acoustical objects. In the field of Auditory Scene Analysis (Bregman, 1993; Unoki and Akagi, 1999), a couple of cues are proposed which allow for these skills, like common onset and offset, gradualness of change, harmonicity, and changes occurring in the acoustic event. However, the aim of this work was not to explore a wide range of possible helpful and important cues, but to concentrate on the contribution of amplitude modulation processing which was motivated by neurophysiological findings in the mammalian auditory system.

One potential advantage of the SNR estimation approach presented in this paper is its general structure which is not restricted to speech/noise detection and SNR prediction. No assumptions about specific characteristics of speech or noise are “hard wired” in the signal processing stage. It was simply assumed that speech and noise “look different” in the AMS pattern representation. Classification itself then is a matter of pattern recognition, which requires a sufficient amount of adequate training data. The application to other tasks in the field of sound classification and detection would not require a complete re-design of the algorithm, but just different training data and targets. On the other

hand, this potential advantage of the algorithm is at the same time one of its major disadvantages. There is no a priori knowledge about how to tell speech from noise implemented in the algorithm. The neural network learns the differences from a large amount of training data, but does hardly allow for direct and clear insight about its structure and dependencies. This might be unsatisfying from the scientific point of view, as it only allows for an indirect analysis of the features and their importance (as described in Section 2.5). However, we have to keep in mind that learning is also essential for all human cognitive skills such as understanding speech or recognizing a cat behind a tree from only its tail. Our biological "hardware" for analyzing physical information is fully developed within the first few months of life, but a successful exploitation of these streams of information for all the individual tasks that follow requires complex and (life-)long learning.

Further work will concentrate on extending the algorithm to sub-band estimation of the local SNR in different frequency channels. Reliable sub-band SNR prediction would allow for attenuation of noisy channels and thus enhancing the overall SNR of the signal. First experiments on predicting the SNR in 15 different frequency channels revealed promising results (Tchorz and Kollmeier, 1999c). Possible gains of such a noise suppression algorithm in terms of speech intelligibility, speech quality, and ease of listening will be investigated in further studies.

Acknowledgement

Part of this work was supported by the European Union (TIDE / SPACE). Many thanks to four anonymous reviewers. Their comments and suggestions helped a lot to improve the quality of the manuscript.

Chapter 3

Noise suppression based on amplitude modulation analysis

Abstract

This paper describes a monaural noise suppression algorithm. It bases on the estimation of the signal-to-noise ratio (SNR) in different frequency channels. For SNR estimation, the input signal is transformed into neurophysiologically-motivated spectro-temporal input features. These patterns are called Amplitude Modulation Spectrograms (AMS), as they contain information of both center frequencies and modulation frequencies within each 32 ms-analysis frame. The different representations of speech and noise in AMS patterns are detected by a neural network, which estimates the present SNR in each frequency channel. Quantitative experiments show a reliable estimation of the SNR for most types of background noise. "Across-frequency" processing enhances the SNR prediction accuracy, compared to independent SNR estimation in each frequency channel. For noise suppression, the frequency bands are attenuated according to the estimated present SNR using a Wiener filter

approach. Objective speech quality measures and informal listening tests indicate a benefit from AMS-based noise suppression, compared to unprocessed noisy speech.

3.1 Introduction

The suppression of noise is an important issue in a wide range of speech processing applications. In the field of automatic speech recognition, for example, background noise is a major problem which typically causes severe degradation of the recognition performance. In hearing instruments, noise suppression is desired to enhance speech intelligibility and speech quality in adverse environments. The same holds for mobile communication, such as hands-free telephony in cars.

Existing noise suppression approaches can be grouped into two main categories. Directive algorithms perform the separation between the target and the noise signal by spatial filtering. A target signal (e.g. from the front direction) is passed through, and signals from other directions are suppressed. This can be realized by using directive microphones or microphone arrays (Soede *et al.*, 1993). In prototype hearing instruments, binaural algorithms exploit phase and level differences or correlations between the two sides of the head for spatial filtering (Wittkop *et al.*, 1997).

Monaural noise suppression algorithms, in contrast, try to separate speech from noise when only one microphone is available, i.e. without spatial information. A monaural noise suppression approach which is widely used bases on Spectral Subtraction (Boll, 1979). The noise spectrum (which is measured and updated in speech pauses) is subtracted from the signal spectrum. After reconstruction, ideally, the signal is cleaned from noise. In practice, two major problems occur. First, if the speech pause detector classifies speech portions as "noise", the noise spectrum is wrongly updated which leads to distortions of the speech signal after spectral subtraction. Second, the noise spectrum is assumed to be stationary while speech is present. Frame-to-frame fluctuations of the noise lead to typical artifacts, known as "musical tones". Several methods have been proposed the reduce musical tones (Cappé, 1994; Linhard and Haulick, 1999; Seok and Bae, 1997).

The noise suppression algorithm presented in this paper does not require explicit detection of speech pauses, and no assumptions on noise stationarity are made while speech is active. It directly estimates the present SNR in different frequency channels with speech and noise being active at the same time. For SNR estimation, the input signal is transformed into neurophysiologically-motivated feature patterns. These patterns are called Amplitude Modulation Spectrograms (AMS), see (Kollmeier and Koch, 1994), as they contain information on both center frequencies and modulation frequencies within each analysis frame. It is shown that speech is represented in a characteristic way in AMS patterns, which is different from the representation of most types of noise. The differences in the respective representations can be exploited by neural network pattern recognition.
In Section 3.2 of this paper, the SNR estimation approach based on AMS patterns is described, and quantitative estimation results are presented. The influence of across-frequency processing for SNR estimation and a comparison with SNR estimation based on voice activity detection are outlined in Section 3.3 and 3.4, respectively. The noise suppression stage with informal listening results and objective quality measures is described in Section 3.5

3.2 SNR estimation

This Section outlines the processing steps which are applied to estimate the local SNR of noisy speech in different frequency channels. The SNR estimation process consists of two main parts: i) the feature extraction stage, where the incoming waveform is transformed into spectro-temporal feature patterns, and ii) a pattern recognition stage, where a neural network classifies the input features and estimates the SNR. A block diagram of the noise suppression algorithm including the SNR estimation stage is given in Fig. 3.1.

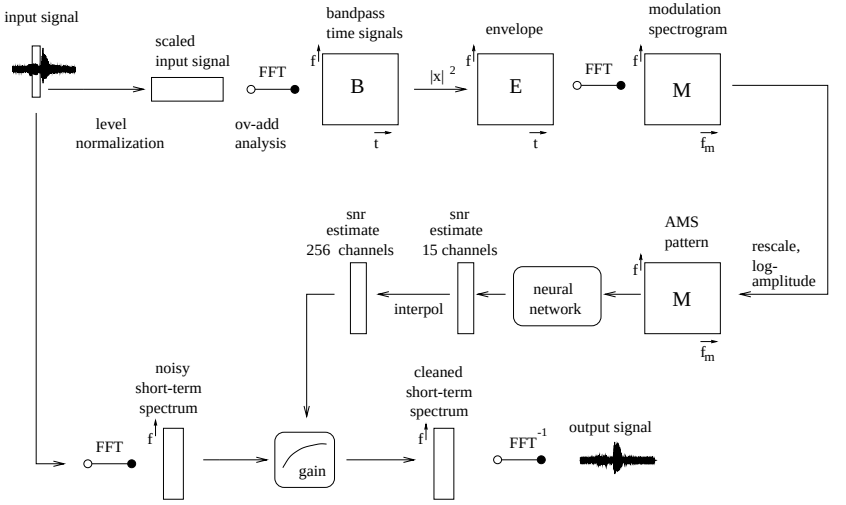


Figure 3.1: Processing stages of AMS-based noise suppression.

3.2.1 Feature extraction

For SNR estimation, the input waveform is transformed into so-called Amplitude Modulation Spectrograms (AMS), see (Kollmeier and Koch, 1994). These patterns are motivated from neurophysiological findings on amplitude modulation processing in higher stages of the auditory system in mammals. Langner and Schreiner (1988), among others, found neurons in the inferior colliculus and auditory cortex of mammals which were tuned to certain modulation frequencies. The "peridotopical" organization of these neurons with respect to different best modulation frequencies was found to be almost orthogonal to the tonotopical organization of neurons with respect to center frequencies. Thus, a two-dimensional "feature set" represents both spectral and temporal properties of the acoustical signal. More recently, Langner et al. (1997) observed periodotopical gradients in the human auditory cortex by means of magnetoencephalography (MEG). Psychoacoustical evidence for a modulation analysis in each frequency band is provided by Dau et al. (1997a;

1997b)
In the field of digital signal processing, Kollmeier and Koch (1994) applied these findings in a binaural noise suppression scheme and introduced two-dimensional AMS patterns, which contain information on both center frequencies and modulation frequencies. Using this algorithm, they reported a small but stable improvement in terms of speech intelligibility, compared to unprocessed speech. Recently, similar kinds of feature patterns were applied to vowel segregation (Yang *et al.*, 1999) and speech enhancement (Strube and Wilmers, 1999), The application of AMS patterns on broad-band SNR estimation is described in (Tchorz and Kollmeier, 2000) (Chapter 2 of this thesis).
For AMS pattern generation, the input signal is long-term level adjusted, i.e., changes in the overall level are compensated for, whereas short-term level differences (e.g., those between successive phonemes) are maintained to serve as additional cues for classification. This level adjustment is realized by dividing the input signal by its 2 Hz-low-pass filtered RMS function (which was calculated from 32 ms frames, with an overlap of 16 ms). In a following processing step, the level-adjusted signal is subdivided into overlapping segments of 4.0 ms duration with a progression of 0.25 ms for each new segment. Each segment is multiplied with a Hanning window and padded with zeros to obtain a frame of 128 samples which is transformed with a FFT into a complex spectrum. The resulting 64 complex samples are considered as a function of time, i.e., as band pass filtered complex time signal. Their respective envelopes are extracted by squaring. This envelope signal is again segmented into overlapping segments of 128 samples (32ms) with an overlap of 64 samples. A further FFT is computed and supplies a modulation spectrum in each frequency channel. By an appropriate summation of neighboring FFT bins both axes are scaled logarithmically with a resolution of 15 channels for center frequency (100-7300 Hz) and 15 channels for modulation frequency (50-400 Hz). In a last processing step, the amplitude range is log-compressed. Examples for AMS patterns can be seen in Fig. 3.2. Bright and dark areas indicate high and low energies, respectively. The left AMS pattern was generated from a voiced speech portion, uttered by a male speaker. The periodicity at the fundamental frequency (approx. 110 Hz) is represented in each center frequency

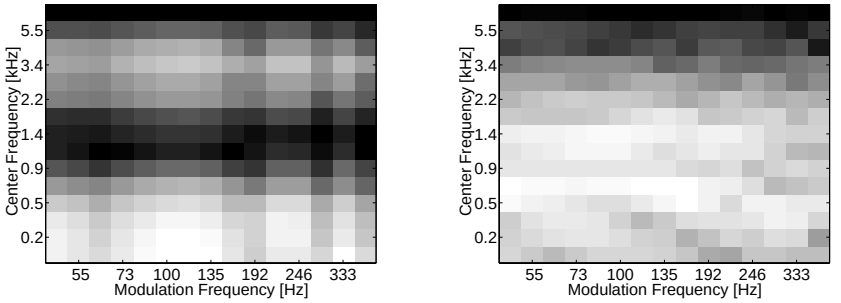


Figure 3.2: AMS patterns generated from a voiced speech segment (left), and from CCITT speech simulating noise (right). Each AMS pattern represents a 32 ms portion of the input signal. Bright and dark areas indicate high and low energies, respectively.

band, as well as the first and second harmonics at about 220 and 330 Hz, respectively. Due to the short length of the analysis frame (32 ms), the modulation frequency resolution is limited. Thus, the peaks indicating e.g. the fundamental frequency are relatively broad. The right AMS pattern was generated from speech simulating noise. The typical spectral tilt can be seen, but no structure across modulation frequencies.

3.2.2 Neural network classification

Amplitude Modulation Spectrograms are complex patterns which are assumed to carry important information to discriminate between speech and noise. The classification and SNR estimation task is considered as a pattern recognition problem (speech and noise obviously "look different" in the AMS representation in most cases). Artificial neural networks are widely used in a range of different pattern recognition tasks (Bishop, 1995). For SNR estimation based on AMS patterns, a standard feed-forward neural network is applied (SNNS, described in (Zell, 1994)). It consists of an input layer with 225 neurons (15 times 15, the resolution of AMS patterns, which are directly fed into the network), a hidden

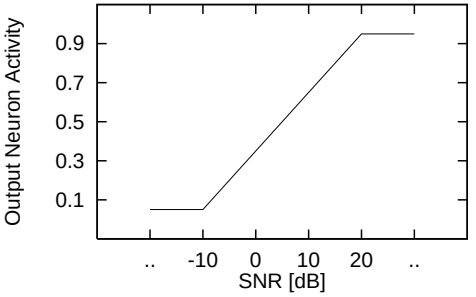


Figure 3.3: Transformation function between SNR and output neuron activity for training and testing.

layer with 160 neurons, and an output layer with 15 output neurons. The three layers are fully connected. Each output neuron represents one frequency channel. The activities of the output neurons indicate the respective SNR in the present analysis frame. For training of the neural network, mixtures of speech and noise were generated artificially to allow for SNR control. The narrow-band SNRs in 15 frequency channels (which were measured prior to adding speech and noise) are measured for each 32 ms AMS analysis frame of the training material. The measured SNR values are transformed to output neuron activities which serve as target activities for the output neurons during training. A high SNR results in a target output neuron activity close to one, a low SNR in a target activity close to zero, following the transformation function plotted in Fig. 3.3.

SNRs between -10 and 20 dB are linearly transformed to activities between 0.05 and 0.95. SNRs below -10 dB and above 20 dB are assigned to activities of 0.05 and 0.95, respectively. In the training phase, the neural network "learns" the characteristics of AMS patterns in different SNRs. The network is trained using the backpropagation-momentum algorithm (Rumelhart *et al.*, 1986). After training, AMS patterns generated from untrained sound material are presented to the network. The 15 output neuron activities that occur for each pattern are linearly re-transformed using the function shown in Fig. 3.3 and serve as SNR estimates for the respective frequency channels in the present analysis frame.

3.2.3 Speech material

For training of the neural network, a mixture of speech and noise with a total length of 72 min was processed and transformed into 270.000 AMS patterns. The long-term, broad-band SNR between speech and noise for the training data was 2.5 dB, but the *local* SNR in 32 ms analysis frames exhibited strong fluctuations (e.g., in speech pauses). The speech material for training was taken from the Phondat database (Kohler *et al.*, 1994) and contained 2110 German sentences from 190 male and 210 female talkers. 41 types of natural noise were taken for training from various data bases. For testing, a 36-min mixture of speech (200 speakers, Phondat) and 54 noise types was taken. The talkers and noise types for testing were not included in the training data. The network was trained with 100 cycles.

3.2.4 Results

An example for the estimation of narrow-band SNRs of noisy speech is illustrated in Fig. 3.4.

The input signal was a mixture of speech uttered by a male talker and power drill noise. The panels show the measured SNR (solid) and the estimated SNR (dotted) as a function of time in 7 out of 15 frequency channels. In the high-frequency bands (top), the SNR is relatively poor (due to the power drill noise, which is dominant in high frequencies). In general, the estimated SNR correlates with the measured SNR, but there are several prediction errors visible, especially in the high-frequency region. In low-frequency bands, there is a good correspondence between the measured and the estimated SNR.

A quantitative measure of the estimation accuracy is obtained by computing the mean deviation D between the actual SNR a_i and the estimated SNR e_i over N processed AMS patterns (with index i):

$$D = \frac{1}{N} \sum_{i=1}^{N} |a_i - e_i| \tag{3.1}$$

The mean estimation deviation D was calculated for all AMS analysis frames generated from the test data described in Section 3.2.3, for all

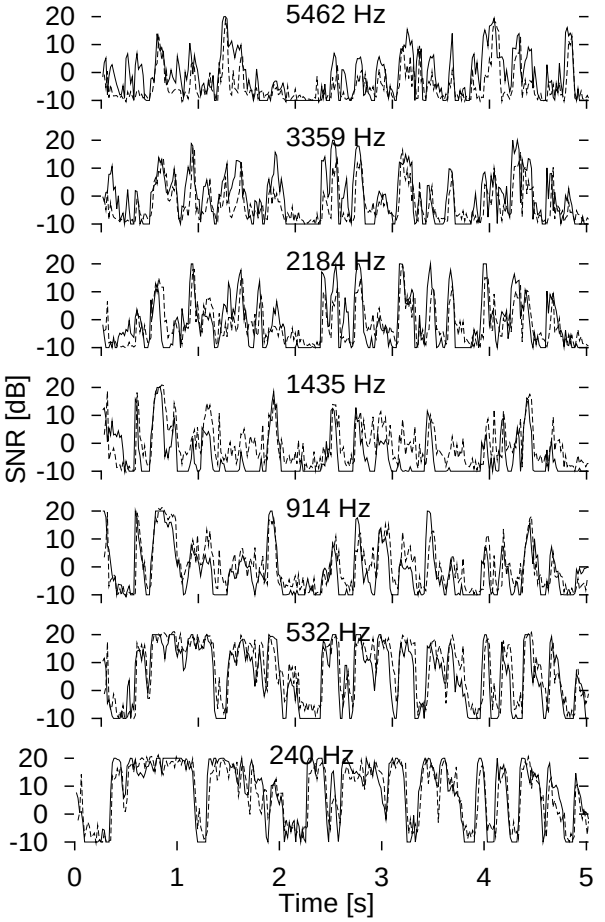


Figure 3.4: Example for narrow-band SNR estimation. Plotted are the measured (solid) and the estimated (dotted) SNRs as function of time for 7 out of 15 frequency channels.

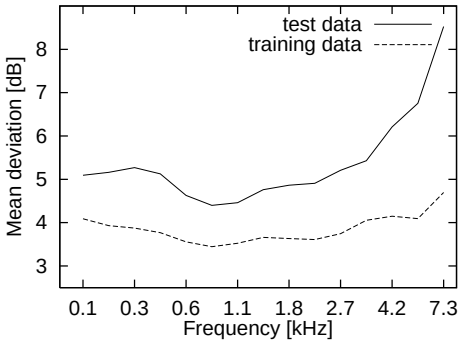


Figure 3.5: Mean deviation between the estimated SNR and the "true" SNR which was measured prior to adding speech and noise as a function of the frequency channel for the test data (solid) and the training data (dotted).

15 frequency channels independently. The results are plotted in Fig. 3.5 (solid line). It can be seen that the estimation accuracy in the low- and mid frequency channels is better compared to the high frequency region (which also is the case for the example plotted in Fig. 3.4). The average deviation between measured SNR and estimated SNR across all frequency channels is 5.4 dB. As expected, the estimation accuracy for the training data (dotted line) is better in all frequency channels. The difference between both data sets is not large, though, except for the highest frequency bands. This means that the network is not over-trained and generalizes to untrained test data to some extend. A histogram of the differences $a_i - e_i$ between measured and estimated SNRs for the test data in one exemplary frequency channel (f_c = 1.1 kHz) is plotted in Fig. 3.6. The maximum frequency is at about -1.3 dB, i.e. there is a slight estimation bias in this particular frequency channel towards worse SNRs than the actual ones. This bias varies from channel to channel, and there is no systematic error across all channels.

For some possible applications of the algorithm, a fast SNR estimation for independent 32ms-frames might not be necessary. Here, temporal smoothing of the estimates can enhance the accuracy of the prediction; "outliers" and short-term errors are attenuated. The dotted curves in Fig. 3.7 show the effect of low pass filtering the temporal trajectories of SNR measures and estimates prior to calculating the mean deviation

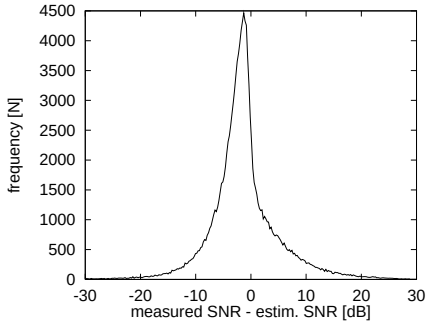


Figure 3.6: Histogram of the differences $a_i - e_i$ between measured and estimated SNRs for the test data in the 7th frequency channel ($f_c = 1.1$ kHz).

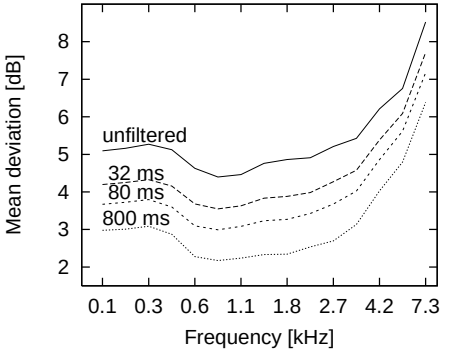


Figure 3.7: Mean deviations D between measured SNR and estimated SNR without and with temporal smoothing: Low pass filtering of successive SNR estimates enhances the estimation accuracy by smoothing short-term errors.

with a first-order low-pass filter. The mean deviation decreases with increasing time constant of the filter, but, of course, the "sluggishness" of the algorithm increases.

In AMS patterns, modulation frequencies between 50 and 400 Hz in different center frequency channels are encoded by the modulation spectra which are computed for each channel. Harmonicity in voiced speech, for example, is represented on the modulation axis by peaks at the fundamental frequency and its harmonics, which leads to characteristic AMS patterns for voiced speech. In a study on AMS-based broad-band SNR estimation (Tchorz and Kollmeier, 2000) (Chapter 2 of this thesis) which investigated the most important cues that are necessary for reliable SNR

estimation it was shown that harmonicity appears to be the most important cue for analysis frames to be classified as "speech-like", but the spectro-temporal representation of sound in AMS patterns also allows for reliable discrimination between unvoiced speech and noise, which would not be possible with a pitch detector.

3.3 Across-channel processing

In the presented algorithm, the signal-to-noise ratio estimation of a single frequency channel does not depend on information from this channel alone. The neural network which is used for classification is fully connected. Thus, information from the whole AMS pattern influences a single channel estimation to some extend. An example for this across-channel processing is illustrated in Fig. 3.8. The first panel shows the measured SNR in 15 frequency channels for a mixture of speech and CCITT speech simulating noise (CCITT, 1964) with an overall SNR of 15 dB. Bright and dark areas indicate high and low local SNRs, respectively. In the second panel, the estimated local SNRs basing on the AMS processing as described above is shown. Both the speaker and the noise type were not included in the training data. There is a good correspondence between the measured and the estimated SNRs across time and frequency channels. Nevertheless, obvious errors and deviations can be seen in the higher frequency channels, where in average the SNR is underestimated. The third panel shows the measured SNRs for the same input signal as in the above panels (speech + CCITT noise), but disturbed in addition with a sinusoid centered at 2.5 kHz. The sinusoid is represented by the vertical black bar across time. In this frequency region, speech is almost entirely masked by the sinusoid. The corresponding SNR estimation of this input signal is plotted in the forth panel. Two remarkable effects occur in this situation. First, the masking effect of the sinusoid in the respective frequency region is not recognized at all by the algorithm. Instated, it is simply "ignored". Obviously, the estimation of the SNR in this region is steered by neighboring frequency channels which were not disturbed by the sinusoid. Noise suppression basing on this SNR estimation would attenuate the CCITT noise, but not the sinusoid. The second effect is that the presence of the sinusoid *has* some impact on

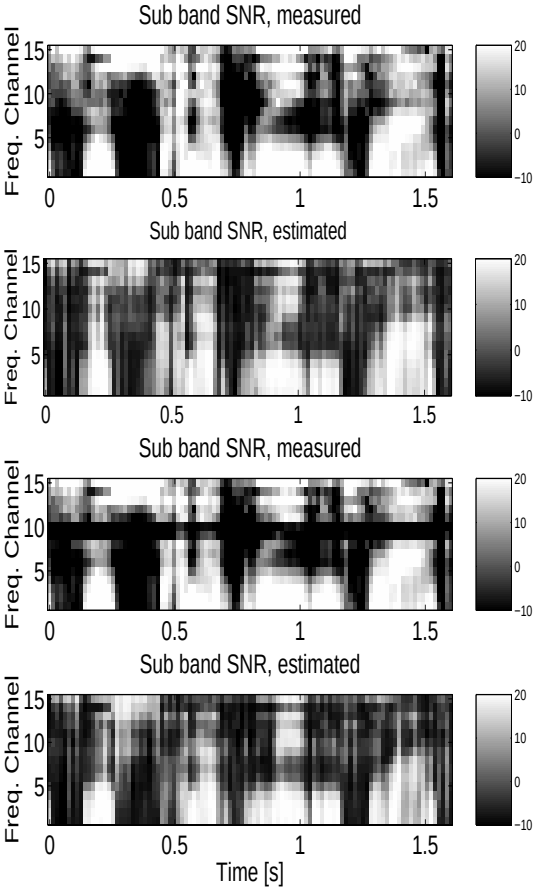


Figure 3.8: "Across frequency" processing in AMS-based SNR estimation. The first two panels show the measured and the estimated sub-band SNR for noisy speech. In the last two panels, the noisy speech was additionally distorted with a sinusoid, which can be seen from the measured SNR (3rd panel), but not from the estimated SNR (4th panel).

the SNR estimation for all other frequency channels, as there are small differences to the second panel, where the speech was only disturbed by CCITT noise. The inability of the algorithm to recognize and classify a sinusoid can be explained by the choice of the training data. All noises that were used for training were natural noises which occur in typical every-day environments. CCITT noise is not "natural", but its spectral and temporal properties are well represented in the training data. For the sinusoid, in contrast, there was nothing even similar contained in the training data, neither in the noise set, nor in the speech set. Thus, the neural network classification is somewhat "blind" for such an input signal. This illustrates the necessity of carefully choosing the training data according to the desired application, and to cover the range of potential input signals, if possible.

To quantitatively determine the influence of across-channel processing of the algorithm due to the full connection of the neural network, a modified, "isolated channel" classification scheme was implemented and tested. It does not use one single network for SNR estimation in all frequency channels, but 15 different networks which are trained on information from one single frequency channel and which estimate the SNR in this channel after training. Each neural network consists of an input layer with 15 neurons (the modulation spectrum of one frequency channel, i.e., one AMS pattern row), a hidden layer with 40 neurons, and an output layer with just one neuron. The target activity of the output neuron during training corresponds to the measured SNR in this channel. After training, the output neuron activity determines the estimated SNR in the particular frequency channel. Except from splitting up the network into 15 independent networks, the setup of the SNR prediction experiments is the same as described in Section 3.2. The resulting SNR prediction accuracy on the test data in terms of mean estimation deviation as a function of the frequency channel is plotted in Fig. 3.9 (solid curve). Compared to the original algorithm with one fully connected network (dotted curve), the mean deviation increased by 2–4 dB in all frequency channels. Nevertheless, these results indicate that it *is* possible to determine the SNR in a single frequency to some extend by just analyzing the modulations in that channel, as these results show. An increased accuracy, however, is made possible by across-channel pro-

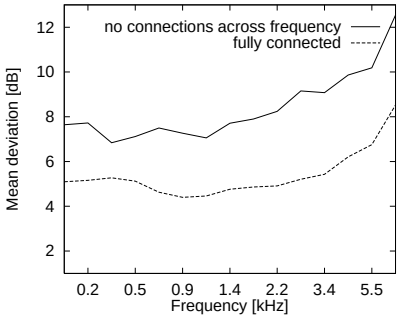


Figure 3.9: SNR estimation accuracy on the test data with no network connections across frequency (solid), and with fully connected network (dotted).

cessing, i.e. by exploiting information in neighboring frequency channels (provided the characteristics of the test data are covered by the training data).

3.4 Comparison with VAD-based SNR estimation

In common single-channel noise suppression algorithms, the noise spectrum estimate is updated in speech pauses using some voice activity detection (VAD). This allows for re-estimation of the clean speech signal from noisy speech under the assumption that the noise is sufficiently stationary during speech activity. Thus, an estimate of the SNR is provided for each analysis frame, in each frequency channel. The accuracy of a VAD-based SNR estimation was compared to the SNR estimation approach outlined in this paper. For voice activity detection, a VAD standardized by ITU-T was used (ITU, 1996). It utilizes information on energy, zero-crossing rate, and spectral distortions for voice activity detection. For this experiment, the FFT spectrum of the input signal was computed using 8 ms analysis frames and a shift of 4 ms. The noise spectrum estimate was updated in frames which were classified as speech pauses by the VAD. The "instantaneous SNR", as described in (Ephraïm and Malah, 1984) was calculated for each spectral component.

$$\mathrm{SNR}_{[inst]} = 10 \log(\gamma_k - 1), \tag{3.2}$$

with

$$\gamma_k = \frac{|R_k(l)|^2}{\lambda_d(k)}, \tag{3.3}$$

where $R_k(l)$ is the modulus of the signal plus noise resultant spectral component k, and $\lambda_d(k) = E\{|D_k|^2\}$ the variance of the kth spectral component of the noise. γ_k is interpreted as the *a posteriori* SNR. The instantaneous SNR typically fluctuates very fast, as the local noise energy in a certain frame can be quite different from the average noise spectrum estimate. These fluctuations cause the well-known "musical noise" which degrades the quality of speech enhanced by Spectral Subtraction (Boll, 1979). Several methods have been proposed to reduce musical noise. An approach which is widely used was introduced by Ephraïm and Malah (1984). In this approach, the gain function is determined by both the instantaneous SNR and the so-called *a priori* SNR, which is a weighted sum of the present instantaneous SNR and the recursively computed *a posteriori* SNR in the processed previous frame.

In our experiment, both the instantaneous SNR and the *a priori* SNR were calculated from the input signal, following Ephraïm and Malah (1984). To allow for direct comparisons with the AMS-based SNR estimation approach described in this paper, the time resolution of the instantaneous and *a priori* SNR estimates were reduced by taking the mean of eight successive frames, yielding 32 ms analysis frames with a shift of 16 ms, as in the AMS approach. By appropriate summation of neighboring FFT bins, a frequency resolution identical to the AMS approach was provided. The test material described in Section 3.2.3 was processed and the instantaneous and *a priori* SNR values were compared to the "true" SNR which was measured prior to mixing speech and noise. The achieved mean deviations in each frequency channel is plotted in Fig. 3.10 (left). When comparing the two VAD-based approaches, it can be seen that the *a priori* SNR provides a more reliable estimate of the present SNR than the instantaneous SNR. The accuracy of the AMS-based, direct SNR estimation approach, however, appears to be more accurate than the two VAD-based measures, especially in the mid-frequency region. In the lower frequency bands, the accuracy is comparable. The importance of a proper and reliable

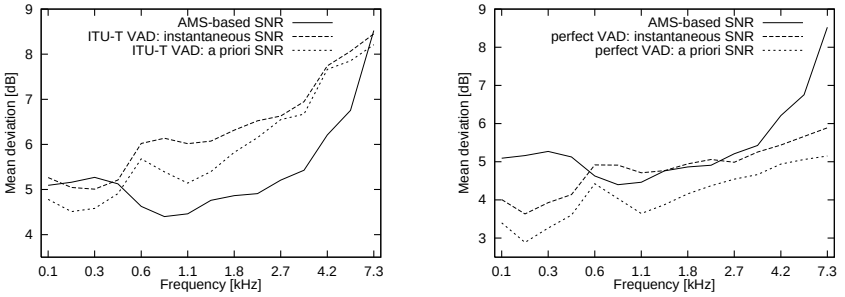


Figure 3.10: Comparison between AMS-based (solid) and VAD-based (dotted) SNR estimation in 15 frequency channels. The left panel shows the results with a VAD standardized by ITU-T, on the right panel a "perfect" VAD was used.

speech pause detection for the VAD-based approach is illustrated in the right panel. Here, the ITU-T VAD was replaced by a "perfect" VAD (the speech pauses were detected from the clean speech input with an energy criterion). Thus, there were no speech pauses missed and hence the noise estimate could be updated as often as possible. In addition, no speech portions were mistakenly classified as noise and distorted the noise measure. With perfect information on speech pauses, the VAD-based SNR estimation accuracy for the tested data was higher than with the direct AMS-based approach, especially in the lowest and highest frequency bands.

However, the VAD-based SNR estimation allows for estimation in narrow and independent frequency bins, and for short analysis frames. The AMS-based approach, in contrast, is restricted in both time and frequency resolution: Modulation analysis down to 50 Hz modulation frequency requires analysis frames of at least about 20 ms. In addition, increased center frequency resolution and hence SNR estimation in much more than 15 channels (as in the present AMS implementation) would require considerably higher costs in terms of necessary training data, processing time, and memory usage.

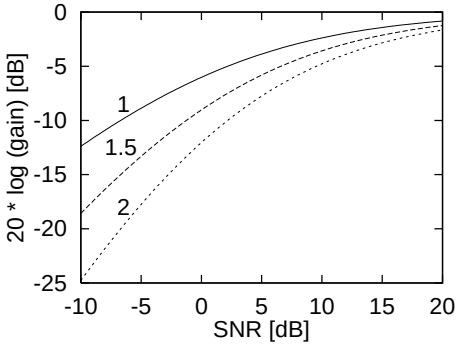


Figure 3.11: Gain function for three different exponents x (see Eqn.(3.4))

3.5 Noise suppression

Sub-band SNR estimates allow for noise suppression by attenuating frequency channels according to their local SNR. The gain function which is applied is given by:

$$g_k = \left(\frac{\mathrm{SNR}_k}{\mathrm{SNR}_k + 1} \right)^x, \tag{3.4}$$

where k denotes the frequency channel, SNR the signal-to-noise ratio on a linear scale, and x is an exponent which controls the strength of the attenuation. Note that for $x = 1$ the gain function is equivalent to a Wiener filter. The gain functions for the SNR range between -10 dB and 20 dB with three different exponents x are plotted in Fig. 3.11. The maximum attenuation with $x = 1$ is restricted to -12 dB, whereas choosing $x = 2$ allows for a maximum attenuation of -25 dB.

Noise suppression based on AMS-derived SNR estimations was performed in the frequency domain (see Fig. 3.1). The input signal is segmented into overlapping frames with a window length of 32 ms, and a shift of 16 ms is applied, i.e., each window corresponds to one AMS analysis frame. The FFT is computed in every window. The magnitude in each frequency bin is multiplied with the corresponding gain computed from the AMS-based SNR estimation. The gain in frequency bins which are not covered by the center frequencies from the SNR estimation is lin-

early interpolated from neighboring estimation frequencies. The phase of the noisy speech is extracted and applied to the attenuated magnitude spectrum. An inverse FFT is computed, and the enhanced speech is obtained by overlapping and adding.
A parameter of the proposed noise suppression approach is the cut-off frequency of the low pass filter which temporally smoothes subsequent SNR estimates (see Section 3.2.4). With filtering, prediction errors and thus incorrect attenuation are smoothed, but the adaptation to new acoustical situations gets slower. Another parameter is the attenuation exponent x. Values of 2 and higher result in a strong attenuation of the noise, but may also degrade the speech. Low values lead to only moderate suppression of the noise (with a clearly audible noise floor).

3.5.1 Informal listening results

A range of different samples of noisy speech were subject to informal listening tests. In general, a good quality of speech is maintained, and the background noise is clearly suppressed. There are no annoying "musical-noise"-like artifacts audible. The choice of the attenuation exponent x has only little impact on the quality of clean speech, which was well preserved for all speakers that were tested. With decreasing SNR, however, there is a tradeoff between the amount of noise suppression and distortions of the speech. A typical distortion of speech in poor signal-to-noise ratios is an unnatural spectral "coloring", rather than rough distortions.
Without temporal low-pass filtering of successive AMS-based SNR estimates, an independent adaptation to new acoustical situations is provided every 16 ms. Thus, estimation errors in single frames can cause unwanted fluctuations in the processed signal. Low-pass filtering of successive AMS-based SNR estimates with a cut-off frequency of about 2-4 Hz smoothes these fluctuations but still allows for quick adaptation to the present acoustical situation. With longer time constants for filtering, the noise slowly fades out in speech pauses. When speech commences, it takes some time until the gain increases again[1].

[1]Demonstrations for the proposed noise suppression scheme can be downloaded from

3.5.2 Objective speech quality evaluations

Besides the subjective validation of noise suppression algorithms by listeners, there is a range of objective measures available to evaluate the quality of such algorithms. An overview of objective speech quality measures can be found in (Quackenbush *et al.*, 1988). Three different objective measures were applied to noisy speech which was processed by the presented AMS-based noise suppression, namely the Log-Likelihood Ratio (LLR), the Log-Area Ratio (LAR), and a psychoacoustically-motivated speech quality measure. The first two methods (LLR and LAR) rely on mathematically based distance measures between the original and the degraded (noisy) speech signal, whereas the latter approach bases on PEMO, a model of the auditory periphery, which transforms the incoming waveform into an "internal representation" of the sound. The auditory model was originally developed to predict human performance in typical psychoacoustical masking experiments (Dau *et al.*, 1996a), but it was also applied in the field of speech processing, such as speech intelligibility prediction (Holube and Kollmeier, 1996) and feature extraction for automatic speech recognition (Tchorz and Kollmeier, 1999b) (Chapter 3 of this thesis). The application of PEMO to speech quality prediction is described by Hansen and Kollmeier (1997).
For the objective evaluation of the AMS-based noise suppression, speech uttered by different speakers was mixed with six different types of noise[2]. Noise was added with different SNRs ranging from -5 dB to 15 dB to explore dependencies on the overall SNR. The noisy speech was processed using an attenuation exponent $x = 1$, and no low pass filtering of SNR estimates. After noise suppression, the respective distances between clean and noisy speech, and between clean and processed noisy speech were computed as a function of the overall SNR. For LLR and LAR, the mean using the first 95% of all frames was calculated as overall performance measure, because a mean quality measure over all frames is typically biased by a few frames in the tails of the quality measure distribution (Hansen and Pellom, 1998). For these two measures, higher speech quality is indicated by a smaller distance to the clean speech. For

`http://medi.uni-oldenburg.de/members/juergen/ams.html`

[2]canteen babble, factory noise, white Gaussian noise, inside car, inside truck, traffic noise in town

PEMO, the overall correlation coefficient between the internal representations of clean speech and (processed) noisy speech served as quality measure. Thus, a higher correlation coefficient indicates better speech quality. The results of the objective measures averaged across noise conditions are plotted in Fig. 3.12. LLR and PEMO indicate a constant benefit of AMS noise suppression, independent of the overall SNR. In particular, these two measures do not detect degradation of processed speech in high SNRs. The LAR measure indicates a small benefit in poor SNRs, but a slight degradation of quality in high SNRs. The objective measures across all six tested types of noise has been averaged in Fig. 3.12, although the quality measures considerably varied between different noises. The largest benefit from processing was measured in white Gaussian noise, with all three measures. No benefit from processing was measured in babble noise. LLR and PEMO detected almost no differences between processed and unprocessed speech in this situation, and LAR indicated a decent degradation of speech quality due to processing. By informal listening, such a degradation could not be confirmed. However, the AMS-based noise suppression has only little effect in canteen noise, as lots of noise portions are classified as speech. From this, the results from the objective quality measures are not in contradiction to the informal listening results.

3.6 Discussion

The main findings of this study can be summarized as follows:

- Neurophysiologically-motivated Amplitude Modulation Spectrograms (AMS), in combination with artificial neural networks for pattern recognition, allow for automatic estimation of the present SNR in narrow frequency bands, even if both speech and noise are present at the same time

- SNR estimation is possible from modulation cues only, but estimation accuracy benefits from across channel processing

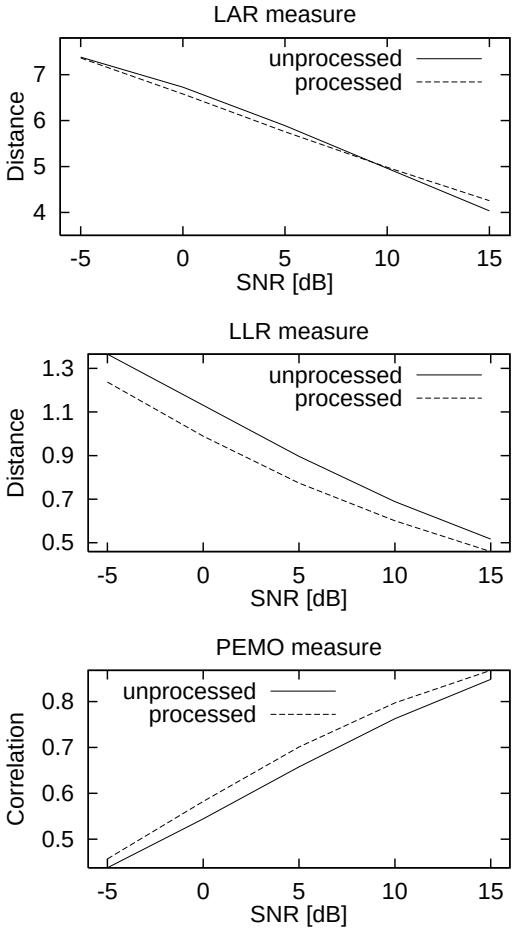


Figure 3.12: Three different objective measures of the speech quality for unprocessed noisy speech (solid) and noisy speech processed by AMS-based noise suppression (dotted).

- Monaural noise suppression from AMS-derived SNR estimates preserves the speech quality in SNRs which are not too poor and attenuates noise without musical noise-like artifacts.

Neurophysiological experiments on temporal processing clearly indicate that the analysis and representation of amplitude modulations play a central role in our auditory system. Technical sound signal processing, on the other hand, is commonly dominated by the the analysis of *spectral* information, rather than modulation information. Spectral analysis in speech processing has a a long history back to the invention of the spectrograph (Koenig *et al.*, 1946), and one is easily tended to take the importance of the frequency spectrum for granted. It was not before recent years that speech processing research focused on the analysis of *modulation* frequencies, especially in the field of noise reduction (Kollmeier and Koch, 1994; Strube and Wilmers, 1999) and automatic speech recognition (Hermansky and Morgan, 1994; Kingsbury *et al.*, 1998; Tchorz and Kollmeier, 1999b). In speech recognition, band pass filtering of low modulation frequencies of about 4 Hz attenuates the disturbing influence from background noise, which typically has a different modulation spectrum compared to speech. Low modulation frequencies also play an important role for speech intelligibility. Drullman et al. (1994) found that modulation frequencies up to 8 Hz are the most important ones in for speech intelligibility. Shannon et al. (1993) conducted an impressive study on the importance of temporal amplitude modulations for speech intelligibility and observed nearly perfect speech recognition under conditions of highly reduced spectral information.
However, it is important to notice the difference between speech intelligibility and speech detection (or, in a wider sense, detection of acoustical objects). Higher modulation frequencies which represent pitch information or harmonicity are likely to be more important for speech detection and sound classification. In a study on AMS-based broad-band SNR estimation (Tchorz and Kollmeier, 2000) (Chapter 2 of this thesis) it was shown that harmonicity appears to be the most important cue for analysis frames to be classified as "speech-like", but the spectro-temporal representation of sound in AMS patterns also allows for reliable discrimination between unvoiced speech and noise. Thus, the joint represen-

tation in AMS patterns cannot be replaced by a simple pitch detector (which would require less computational effort).
Amplitude Modulation Spectrograms for SNR estimation described in this paper do not allow for analysis of very low modulation frequencies, as the analysis windows have to be kept short for fast noise suppression. However, AMS processing can be regarded as a more general way of signal representation. The time constants and analysis frames are variable, and sub-band SNR prediction (in combination with a pattern recognizer) should be regarded as an example for a practical application of spectro-temporal feature extraction. The distinction between speech and noise is made possible by the choice of the training data, and no specific assumptions on speech or noise are "hard wired" in the algorithm. Thus, other applications such as classification of musical instruments or detection and suppression of certain types of noise are thinkable (but are not implemented to date).
A disadvantage of the proposed noise suppression scheme is the limited frequency resolution, as the SNR is estimated in only 15 channels. Hence, the suppression of noise types with sharp spectral peaks is not as efficient as in Spectral Subtraction or related algorithms. A smoother gain function across frequency, on the other hand, reduces annoying effects in the processed signal.
The objective speech quality measures indicate a benefit from AMS-based noise suppression. However, this finding is of limited evidence until the results are linked with subjective listening tests and the correlation between objective measures and subjective scores are determined. Thus, future work will include a more detailed evaluation of the proposed noise suppression algorithm with listening tests in normal-hearing and hearing-impaired persons, and comparisons with other monaural noise suppression algorithms such as Spectral Subtraction and the approach proposed by Ephraïm and Malah. Listening tests should not be restricted to speech intelligibility measurements in typically very poor signal-to-noise ratios. In addition, more "subjective" dimensions like ease of listening and overall sound quality should be covered, which are of great practical importance in SNR ranges where speech intelligibility is well above 50%.

Acknowledgements

Part of this work was supported by the European Union (TIDE/SPACE).

Chapter 4

A model of auditory perception as front end for automatic speech recognition [1]

Abstract

A front end for automatic speech recognizers is proposed and evaluated which is based on a quantitative model of the "effective" peripheral auditory processing. The model simulates both spectral and temporal properties of sound processing in the auditory system which were found in psychoacoustical and physiological experiments. The robustness of the auditory-based representation of speech was evaluated in speaker-independent, isolated word recognition experiments in different types of additive noise. The results show a higher robustness of the auditory front end in noise, compared to common mel-scale cepstral feature extraction. In a second set of experiments, different processing stages

[1] A slightly modified version of this Chapter appeared in *J. Acoust. Soc. Am.* (Tchorz and Kollmeier, 1999b).

of the auditory front end were modified to study their contribution to robust speech signal representation in detail. The adaptive compression stage which enhances temporal changes of the input signal appeared to be the most important processing stage towards robust speech representation in noise. Low pass filtering of the fast fluctuating envelope in each frequency band further reduces the influence of noise in the auditory-based representation of speech.

4.1 Introduction

Front ends for automatic speech recognition (ASR) systems are designed to transform the incoming speech signal into a representation which serves as input for later pattern recognition stages. The representation should extract and highlight important features from the speech signal which are relatively independent from speaker variability and channel conditions. It should suppress irrelevant redundancies contained in the speech waveform, thus reducing the data rate at subsequent processing stages of the recognition system. In addition, the representation of speech should be influenced as little as possible by both additive background noise and convolutive distortions (e.g., a change of the transmission channel) to allow for robust recognition in realistic environments outside the laboratory. Unfortunately, the desired robustness against noise is far away from being realized in present speech recognition systems. Even slightly disturbed speech often leads to a distinct decrease in the performance of ASR systems and makes the usefulness of the recognition system questionable.

The human auditory system, on the other hand, performs speech processing which is very robust against noise and allows us to understand speech even under poor conditions. Human speech recognition and understanding is made possible by the interplay between the auditory periphery, which transforms the incoming sound signal into its "internal representation", and the higher auditory processing stages in the brain, which performs the recognition task based on the internal representation (see Kollmeier (1990) for a review). While comparatively

little is known about the neural mechanisms of the central auditory processing stages in the brain, much more is known about the peripheral auditory processing stages. (Here we use the term periphery to characterize the first stages of auditory processing including filtering of the basilar membrane, half-wave rectification of the hair cells and encoding of hair cell movement into auditory nerve firing). The knowledge about these mechanisms is primarily based on physiological measurements and psychoacoustical experiments on frequency selectivity, loudness perception, short term adaptation, temporal masking, and other topics.

Despite the progress in understanding auditory processing mechanisms, only few aspects of sound processing in the auditory periphery are modeled and simulated in common front ends for ASR systems. One example is that most current front ends perform auditory frequency filtering proportional to a perceptually-based frequency scale rather than a linear scale, which in general improves recognition robustness in noise (Jankowski *et al.*, 1995). Another example is the use of the logarithm of certain speech features that approximate the nonlinear dynamic compression in the auditory system which allows us to cover the huge dynamical range between hearing threshold and uncomfortable loudness level.

Ideally, the automatic speech recognizer should operate as a pattern recognizer on the same input pattern that is available to the central auditory system after preprocessing in the peripheral auditory system, as the central auditory system can be regarded as a very powerful recognizer which bases on and makes use of the peripheral representation of sounds. This peripheral representation contains all information that is needed for robust speech recognition. In reality, of course, a technical pattern recognizer works different than our brain does, and it is questionable whether the technical recognizer can exploit auditory-like features for more robust recognition. Thus, the interactions between feature characteristics and recognizer properties have to be considered carefully. In addition, it is not clear which parts and details of peripheral auditory processing contribute to robust speech recognition (and thus might be worth to be simulated in a technical front end), and which parts are important for other skills of the auditory system (and thus would "whiten" the features with information that is not needed). A more detailed discussion on the

simulation of auditory processing in ASR is given by Hermansky (1998). Having the above mentioned restrictions in mind, an auditory-motivated representation of speech would include certain properties of auditory processing that have been characterized by physiological and psychoacoustical experiments and have not yet been incorporated in common ASR systems. One example of these properties is short term adaptation. With short term adaptation, we denote the combination of two effects: (i) Given a constant stimulus, the auditory nerve response decreases monotonically with increasing stimulus duration, i.e., it adapts and asymptotically approaches a steady-state rate (Kiang *et al.*, 1965). (ii) After stimulus offset, a period of recovery in auditory nerve activity with a firing rate below spontaneous emission in quiet and reduced response to a new stimulus can be observed (Smith, 1979; Delgutte and Kiang, 1984). Another example is the processing of amplitude modulations in the auditory system. During recent years, more insight has been gained about the coding of amplitude modulations in the auditory system and their contribution to pitch representation and speech perception. Langner and Schreiner (1988), for example, found neurons in the inferior colliculus of the cat that were tuned to certain modulation frequencies. Furthermore, a tonotopical organization of units with respect to their modulation tuning was found that appeared to be perpendicular to the tonotopical encoding of the carrier frequencies (Langner, 1992). An enhancement of speech intelligibility could be achieved when these properties were exploited (Kollmeier and Koch, 1994). The importance of temporal amplitude modulations for speech perception was demonstrated by Shannon et al. (1993). They observed nearly perfect speech recognition under conditions of greatly reduced spectral information. Their results indicate that amplitude modulations below 50 Hz in particular are important for speech perception.

Several researchers have proposed algorithms to model different psychoacoustical aspects or physiological processing stages of the auditory periphery. Only few of these models were tested in speech recognition systems, though. Ghitza (1988) introduced a model of temporal discharge patterns of auditory-nerve fibers (Ensemble Interval Histogram - EIH) as feature extraction for automatic speech recognition. In recognition experiments with the TIMIT database, EIH feature extraction

showed little robustness improvement when compared to common mel-scale cepstral coefficient front ends (Sandhu and Ghitza, 1995).
Seneff (1988) proposed a bank of filters model of auditory perception which simulates the transformation of basilar membrane motion into auditory nerve firing patterns. In a subsequent processing stage, a physiologically motivated "generalized synchrony detector" measures to which extend a frequency channels firing rate is periodic with the characteristic period $1/f_c$ of that certain frequency channel. Jankowski *et al.* (1995) evaluated the robustness of Seneffs and Ghitzas auditory front ends in additive and convolutive noise and compared them with a mel frequency filter bank (MFB) based cepstral front end as control feature extraction. In speaker-dependent, isolated word recognition experiments, almost no difference could be observed between the auditory models and the control front end when speech was degraded by convolutive distortions (e.g., by telephone filtering). In additive noise, the auditory-based front ends appeared to be slightly superior to the control front end. This small advantage had to be paid for with much higher computational effort compared to MFB cepstra. The authors emphasized the necessity of choosing an appropriate control front end when evaluating the robustness of auditory-based front ends. In their experiments, MBF cepstra significantly outperformed LPC-based cepstra used in other investigations which indicated a decent benefit from using auditory preprocessing in noisy environment.
Strope and Alwan (1997) presented a computationally efficient model of dynamic perception which augments common mel frequency cepstrum front ends. An additive logarithmic adaptation stage simulates short-term adaptation. The model's parameters were fitted to predict psychoacoustical forward masking experiments. A subsequent peak isolation mechanism is intended to further enhancing the dynamic spectral cues. First evaluations in speaker-independent, isolated digit recognition experiments in static additive noise yielded better results compared to common front ends.
An earlier approach to simulate short-term adaptation and to implement it in an ASR front end was presented by Cohen (1989). He combined critical band filtering with loudness scaling and a reservoir-type adaptation equation for simulating the hair cell action (Schroeder and Hall,

1974), which relates stimulus intensity to auditory-nerve firing rate. He reported lower error rates and reduced computational costs in large-vocabulary, connected word recognition experiments when compared to a standard bank of filters front end. Noise immunity was not addressed in his experiments.
In recent years, growing attention is paid to low frequency amplitude modulations. RASTA processing of speech (Hermansky and Morgan, 1994) filters the time trajectories of speech. The RASTA filter passes amplitude modulation components between about 1 and 12 Hz in a relatively flat pass band and with rather steep slopes. Ideally, by appropriate feature transformation, the disturbing components in the input signal should combine linearly with the components that origin from speech, so that these components can be separated by RASTA filtering. The log-RASTA approach uses static logarithmic compression prior to modulation filtering. It is primarily intended to remove convolutive distortions (e.g., due to changes in the transmission channel), as these distortions are additive in the log-domain. The adaptive J-RASTA approach is intended to suppress the influence of additive noise. Here, the compression prior to modulation filtering depends on an estimation of the present noise energy, which is measured in speech-free intervals. When applied as front end for ASR systems, considerable increase of robustness compared to standard front ends could be observed (see, e.g., Kasper *et al.* (1997)).
The outline of this paper is to describe the application of a psychoacoustically - motivated model of auditory perception to automatic speech recognition. A set of experiments in different types of noise evaluates the robustness of the auditory feature extraction compared to a standard front end. A subsequent analysis of single auditory model processing steps is intended to explore their contribution to robust recognition and to answer the question whether the parameters of the psychoacoustical model are optimal for the new task of ASR feature extraction.

4.2 Signal Processing

The intention of the quantitative model of auditory processing is to transform an incoming sound waveform into its “internal” representa-

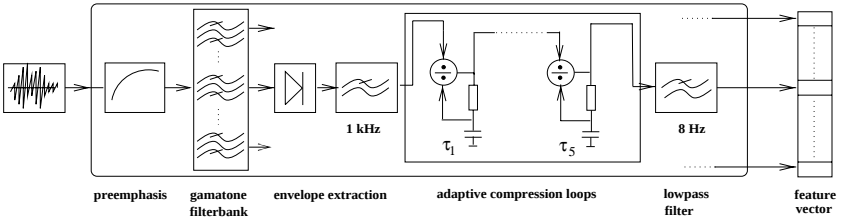


Figure 4.1: Processing stages of the auditory model.

tion. Rather than trying to model each physiological detail of auditory processing, the approach is to focus on the "effective" signal processing in the auditory system which uses as little physiological assumptions and physical parameters as necessary, but to predict as many psychoacoustical aspects and effects as possible.

The model was originally developed for describing human performance in typical psychoacoustical spectral and temporal masking experiments, e.g., predicting the thresholds in backward, simultaneous, and forward-masking experiments (Dau *et al.*, 1996a; 1996b). The parameters of the model were chosen to fit these experiments. Gap detection and modulation detection experiments were simulated with a combination of the model with a subsequent modulation filterbank (Dau *et al.*, 1997a; 1997b). In the field of speech processing, the auditory model was applied to objective speech quality measurement (Hansen and Kollmeier, 1997; 2000), speech intelligibility prediction in noise (Wesselkamp, 1994) and in hearing impaired listeners (Holube and Kollmeier, 1996).

4.2.1 Processing steps

A block diagram of the auditory model and its processing stages is shown in Fig. 4.1.

The first processing step of the auditory model is a preemphasis of the input signal with a first order differentiation. This flattens the typical spectral tilt of speech signals and reflects the transfer function of the outer ear (Djupesland and Zwislocki, 1972). The preemphasis was introduced for feature extraction in speech recognition systems and

was not used in previous applications of the auditory model. In the next processing stage, the preemphasized signal is filtered by a gammatone filterbank (Patterson *et al.*, 1987) using 19 frequency channels equally spaced on the ERB scale with center frequencies ranging from 300-4000 Hz. The impulse responses of the gammatone filterbank are similar to the impulse responses of the auditory system found in physiological measurements (Boer and Kruidenier, 1990). The implemented gammatone filterbank is linear. It does not consider nonlinear effects such as level-dependent upward spread of masking and combination tones. After gammatone filtering, each frequency channel is half wave-rectified and first order low pass filtered with a cutoff frequency of 1000 Hz for envelope extraction. Information about the fine structure of the signal at high frequencies gets lost; this reflects the limiting phase-locking for auditory nerve fibers above 1000 Hz. At this stage of processing, each frequency channel contains information about the amplitude magnitude of the input signal within the channel. Amplitude compression is performed in a following processing step. In contrast to conventional bank of filters front ends, the amplitude compression of the auditory model is not static (e.g., logarithmic) but adaptive, which is realized by an adaptation circuit consisting of five consecutive nonlinear adaptation loops (Püschel, 1988). Each of these loops consists of a divider and a RC-low pass filter with time constants $\tau_1 = 5$ ms, $\tau_2 = 50$ ms, $\tau_3 = 129$ ms, $\tau_4 = 253$ ms, and $\tau_5 = 500$ ms (Dau *et al.*, 1996a; 1997a). For each adaptation loop, the input signal is divided by the output signal of the low pass filter. Thus, a stationary input signal X is transformed to an output signal $Y = \sqrt{X}$. The output of five consecutive adaptation loops is then $Y = \sqrt[32]{X}$, which approximates the logarithm of the input signal. Fluctuations of the input signal that are very fast compared to the time constants of the adaptation loops, on the other hand, are transformed linearly because of time delayed denominator changes. Due to this transformation characteristic, changes in the input signal like onsets and offsets are emphasized, whereas steady-state portions are compressed. Thus, the adaptation loops can be regarded as an inherent memory of the system. Given an input signal, the output signal of the auditory model at time t does not only depend on the properties of the input signal within a narrow time frame around t, but

also on the characteristics of the input signal during a preceding time interval of about 200 ms. Due to the inherent memory of the auditory model, the dynamical structure of the input signal is taken into account over a relatively long period of time. Short term adaptation including enhancement of changes and temporal integration is simulated and allows a quantitative prediction of important temporal effects in auditory perception, such as backward- and forward masking (Dau *et al.*, 1996b). The last processing step of the auditory model is a first order low pass filter with a cutoff frequency of 8 Hz. The filter was introduced by Dau *et al.* (1996b) to optimize predictions of psychoacoustical masking experiments. It attenuates fast envelope fluctuations of the signal in each frequency channel.
The output of the auditory preprocessing is downsampled to a rate of 100 feature vectors per second to serve as input for the subsequent recognition device. The auditory model is implemented in C-code. Processing takes about one-third real time on a current standard personal computer.

4.2.2 Modulation filtering of the auditory model

Suppression of very slow envelope fluctuations by the adaptation loops and attenuation of fast fluctuations by the low pass filter results in a band pass characteristic of the amplitude modulation transfer function of the auditory model. We measured the modulation transfer function using a constant carrier (i.e., a sinusoid at 1 kHz) which was sinusoidal amplitude modulated at different modulation frequencies with a modulation depth of 20%. The attenuation of the modulated signal in the corresponding frequency channel after processing with the auditory model was measured for each modulation frequency. The resulting amplitude modulation transfer function is plotted in Fig. 4.2 (solid curve)[2].

The maximum amplitude modulation transmission of the model can be found at modulation frequencies around 6 Hz. There is a strong attenuation of modulation frequencies below 2 Hz due to the steady-state compression of the adaptation loops. In the high frequency part, the

[2]Note that due to the nonlinearity of the adaptation loops the effective modulation transfer function (MTF) obtained from the model depends strongly on the modulation spectrum of the input signal. Hence, the characteristic plotted in Fig. 4.2 is only an approximation of the MTF that may be obtained with speech signals.

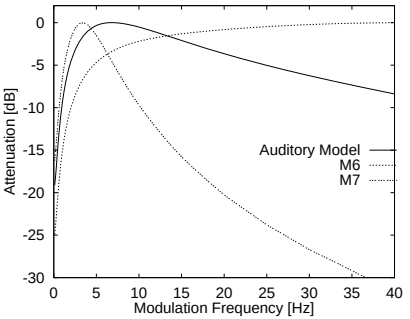


Figure 4.2: Amplitude modulation transfer function of the auditory model and of modifications M6 and M7 as described in section 4.4. The modulations in the output of the channel with a center frequency of 1 kHz were compared with the input modulations for a sinusoidally amplitude-modulated sinusoid at 1 kHz as a function of modulation frequency.

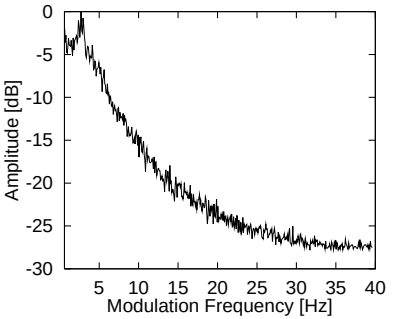


Figure 4.3: Broadband modulation spectrum of a speech sample.

transfer function declines with approximately 3 dB/octave.
A broadband modulation spectrum of speech is shown in Fig. 4.3. It was generated from 380 s of speech from 140 different American speakers from the TIMIT database (70 male, 70 female). The envelope was extracted from the waveform using half-wave rectification and low pass filtering using a 2nd-order Butterworth filter with a cutoff-frequency of 100 Hz. The peak around 3 Hz in the modulation spectrum is a characteristic feature of speech which origins from the average syllable rate (Houtgast

and Steeneken, 1985). It can be seen that the modulation transfer function of the auditory model resembles, but does not exactly fit the average modulation spectrum of speech in terms of most prominent modulation frequency and decline towards higher modulation frequencies.

4.2.3 Examples of sound and speech processing

As a first example, the response of the auditory preprocessing when stimulated with a pure tone is demonstrated. Figure 4.4 (first panel, left) shows the envelope of the stimulus (a 1000 Hz-tone with a duration of 500 ms).

In the second panel (left), the output of the corresponding frequency channel after processing is shown. The figure illustrates the simulation of short term adaptation, including initial linear response at signal onset, transition to steady-state compression, "undershoot" after signal offset due to discharged adaptation loops, and recovery time. An example of speech processing is shown in Fig. 4.5. The waveforms of an undisturbed word and that of the same word disturbed by additive speech simulating noise at 10 dB SNR are plotted on top. Their corresponding representations after processing with the auditory front end can be seen in the two panels in the second row. Peaks in the internal representations are indicated by bright spots, low feature values by dark areas. The two panels in the third row show one single frequency channel (with center frequency f_c=780 Hz) after processing for a more detailed study. The relatively stationary noise preceding and following the speech signal is suppressed due to the steady-state compression performed by the adaptation loops. The onsets and offsets of the speech signal are enhanced in both quiet and in noise. Due to the adaptive amplitude compression which emphasizes changes and suppresses constant portions, speech encoding of the model can be described as sparse and distinct.
To visualize the difference between adaptive compression and static compression, the two panels at the bottom of Fig. 4.5 show the representation for the same input signals and the same frequency channel with the adaptive compression stage replaced by a static log-compression. Here, the most prominent parts of the speech representation are maintained, but the "floor" of the representation during non-speech portions is shifted due to the background noise.

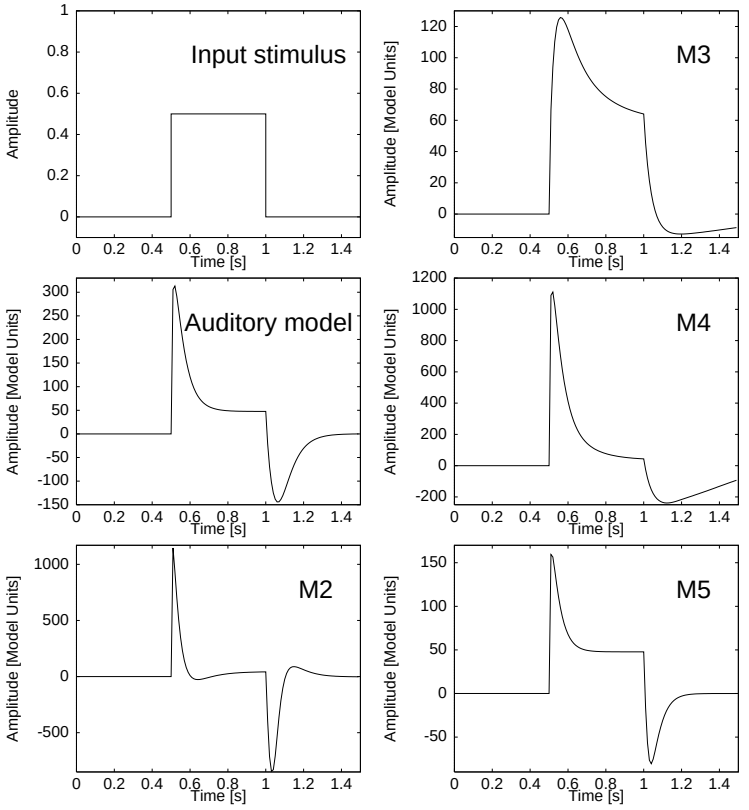


Figure 4.4: Simulation of short term adaptation in the auditory model. First panel (left): Envelope of the input stimulus, a 1000 Hz 500-ms tone pulse. Second panel (left): response of the auditory model in the frequency channel corresponding to the stimulus with initial linear response, transition to steady-state compression, and period of recovery after stimulus offset. Other panels: responses of modifications M2-M5 of the auditory model on the stimulus. See section 4.4.1 (Note that the scaling of the Y-axis is different for each figure).

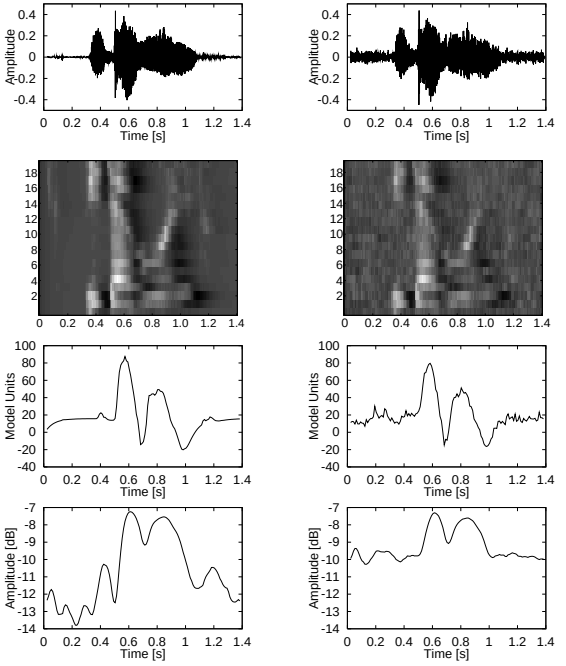


Figure 4.5: An example for speech processing performed by the auditory model. On top, the waveforms of an utterance of the German word "wiederholen" from a female speaker in quiet (left) and in speech simulating noise at 10 dB SNR (right) are shown. In the second row, the corresponding "internal" representations after preprocessing are visualized. High and low amplitudes are indicated by bright and dark areas, respectively. In the third row, the output of one single frequency channel (center frequency $c_f = 780$ Hz) is shown. It can be seen how onsets and offsets are contrasted by the adaptive compression stage. The relatively stationary background noise is compressed and causes only minor fluctuations in the representation. The last two pictures, in contrast, show the respective representations for the same input signals and the same frequency channel when the adaptive compression stage was replaced by a static log-compression. Here, the post prominent parts of the speech representation are well maintained, but the "floor" of the representation during non-speech portions is shifted due to the background noise.

4.3 Recognition Experiments

A number of speaker-independent, isolated digit recognition experiments in different types of additive noise were performed to evaluate the robustness of the auditory-based representation of speech quantitatively.

4.3.1 Experimental setup

The speech material for training of the word models and scoring was taken from the ZIFKOM database of Deutsche Telekom AG. Each German digit was spoken once by 200 different speakers (100 males, 100 females). The recording sessions took place in soundproof booths or quiet offices. The speech material was sampled at 16 kHz.
To cover a certain range of possible additive distortions of speech in actual ASR systems, three different types of noise were added to the speech material at different signal-to-noise ratios before feature extraction: white noise (WN), speech-simulating noise (SN), which was generated from a random superposition of words spoken by a male speaker, and background noise recorded on a construction site (CS). The first two noise types (WN and SN) are stationary, i.e., their spectral shape and energy do not change over time. The last noise (CS) exhibits fluctuations in both spectral shape and energy. To prepare the additive distortions, the RMS value of each word utterance including short pauses before and after the utterance was calculated separately. The background noises were scaled and added to the utterances with signal-to-noise ratios of 20, 15, 10, and 5 dB.
For training and testing, we used a standard continuous-density HMM recognizer with 5 Gaussian mixtures per state, diagonal covariance matrices and 6 emitting states per word model. The word models were trained with features from 100 undisturbed utterances of each digit. Features for testing were calculated from another 100 utterances of each digit which were distorted by additive noise before preprocessing. As control front end we used mel frequency cepstral coefficients (MFCC) (Davis and Mermelstein, 1980), which are widely used in common ASR systems. The coefficients were calculated from Hamming-windowed, preemphasized 32ms segments of the input signal with a frame period of 10ms. In our experiments, each mel cepstrum

feature vector contained 26 features (12 coefficients, log energy, and the respective first temporal derivatives as additional delta-features).

4.3.2 Results

The speaker-independent digit recognition rates in clean speech and in additive noise obtained with the auditory preprocessing and the control front end are shown in Fig. 4.6.

The recognition rates in per cent are plotted as a function of the signal-to-noise ratio in dB. The three panels show the results for white noise, speech simulating noise, and construction site noise, respectively. In undisturbed speech, the control front end yields a higher recognition rate (98.8%) than the auditory-based front end (97.1%). In additive noise, however, the auditory features are more robust than those of the control front end. Even in only slightly disturbed speech (20 dB SNR), the recognition rates obtained with the auditory model are significantly higher in all tested types of noise. The largest difference between the two front ends occur in white noise, where the error rate is decreased by a factor of 3, approximately. In construction site noise, small additive distortions lead to severely decreased recognition rates with the control front end, whereas the auditory model allows more robust recognition. In speech simulating noise, the difference in performance between the two front ends is smaller, but still distinct.

4.4 Recognition experiments with modifications of the auditory model

The preceding section showed that the auditory front end allows promising recognition rates in different types of noise. Two questions arise at this point: Firstly, how do the different signal processing stages of the auditory model contribute to robust recognition? Secondly, are the parameters of the model, which were previously determined to fit psychoacoustical experiments, optimal for the new task in the field of speech

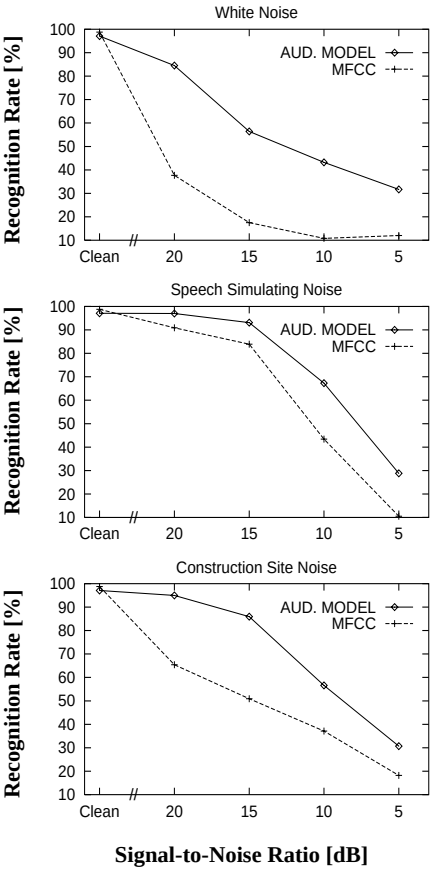


Figure 4.6: Recognition rates in % in different types of additive noise yielded with the auditory model and with the control front end (mel-scale cepstral coefficients, MFCC) as function of signal-to-noise ratio in dB.

processing? To answer these questions, we performed a number of recognition experiments with modifications of the auditory model.

4.4.1 Modifications

There are two processing steps which dominate the representation of speech performed by the auditory model compared to standard bank of filters front ends. Firstly, the nonlinear adaptation loops realize effects

of short term adaptation on the incoming signal. Steady-state portions are suppressed, whereas fast fluctuations are transformed without attenuation. Secondly, the 8 Hz low pass filter smoothes the fast fluctuating envelope and leads to a band pass modulation transfer function. These two processing steps were altered in different ways to analyze their respective contribution to robust representation of speech. In total, seven modified versions (M1-M7) of the auditory front end were implemented and tested.

In modifications M1-M4, the adaptive amplitude compression stage of the model was changed. To determine whether adaptive compression contributes to robust recognition at all, the five adaptation loops were replaced by static logarithmic compression (M1). Dynamic temporal properties of speech such as onsets and offsets are then no longer emphasized in the resulting feature vectors (see the bottom panels of Fig. 4.5).

In modification M2, the number of adaptation loops was increased from five to eight. The time constants $\tau_1 \dots \tau_8$ were linearly equispaced between 5 ms and 500 ms. Thus, temporal contrasts in the input signal are further enhanced in the representation. On the other hand, the amplitude X of a steady-state portion is compressed to $\sqrt[256]{X}$, i.e., there is almost no contribution of constant signal portions left in the representation.

For modification M3, the number of adaptation loops was decreased from 5 to 2, i.e., onsets and offsets are only moderately emphasized. Steady-state portions are compressed to $\sqrt[4]{x}$, the time constants are $\tau_1 = 5$ ms, and $\tau_2 = 500$ ms.

In modification M4, the number of adaptation loops was five, as in the original model. Their time constants were increased and set to $\tau_1 = 50$ ms, $\tau_2 = 300$ ms, $\tau_3 = 700$ ms, $\tau_4 = 1100$ ms, and $\tau_5 = 1600$ ms. This leads to a slower transition to steady-state compression. Only rather long steady-state portions are fully compressed to $\sqrt[32]{x}$, the attenuation of faster modulation frequencies is weaker than in the original model.

Shorter time constants as in the original model were chosen in modification M5. They were set to $\tau_1 = 2$ ms, $\tau_2 = 20$ ms, $\tau_3 = 50$ ms, $\tau_4 = 100$ ms, and $\tau_5 = 200$ ms. In this case, the transition to full steady-state compression is faster, and the peaks in the representation indicating onsets and offsets are narrower.

M1	logarithmic compression
M2	8 adaptation loops
M3	2 adaptation loops
M4	loops with longer time constants
M5	loops with shorter time constants
M6	no modulation low pass filtering
M7	enhanced modulation low pass filtering

Table 4.1: Modifications of the auditory model at a glance.

The different compression schemes of the modified auditory front ends M2-M5 are illustrated in Fig. 4.4, where their respective tone pulse responses are shown.
The modified versions M6 and M7 concern the 8 Hz low pass filter at the end of the auditory preprocessing. In modification M6, the filter was simply left out. This leads to a high pass amplitude modulation transfer function. High modulation frequencies are no longer attenuated, low modulation frequencies are damped by the steady-state compression of the adaptation loops. In M7, the 8 Hz low pass filter was replaced by a second-order low pass with a cutoff frequency of 4 Hz. Fast envelope fluctuations are almost fully compressed.
The modulation transfer functions of the two modified auditory front ends M6 and M7, compared to the original preprocessing, are shown in Fig. 4.2. Without low pass filtering of the envelope (M6), the modulation transfer function has a high pass characteristic due to the adaptive compression stage of the model. With the original 8 Hz 1st-order low pass filter, the decline in the high-frequency part is not 6dB/octave, as one could expect, but less, because the filter has to compensate for the increase in the transfer function towards high modulation frequencies without filtering. With increased low pass filtering of the envelope (M7), the decline in the modulation transfer function approximates the average modulation spectrum of speech in the low modulation frequency part (see Fig. 4.3). The modifications of the auditory model at a glance are listed in Table 4.1.

Speaker-independent, isolated-digit recognition experiments for clean

speech and in additive noise were performed with all modifications. The setup of the experiments was the same as described in Sec. 4.3.1

4.4.2 Results

The recognition rates obtained with modifications M1-M7 of the auditory model are plotted in Fig. 4.7. Replacing the adaptation loops by static logarithmic compression (M1) yields satisfactory results in clean speech, but the recognition rate drops very fast even if the signal-to-noise ratio is as high as 20 dB SNR. This indicates that adaptive compression of the envelope in each frequency channel plays an essential role for robust speech recognition with the auditory model. Changing the parameters of the adaptation loops appears to be disadvantageous in most situations. Taking eight adaptation loops instead of five (M2) and thus suppressing steady-state information almost always leads to increased error rates, especially in low signal-to-noise ratios. With only two adaptation loops (M3), performance in moderate noise is worse compared to the original model. In low signal-to-noise ratios, however, higher recognition rates can be observed in speech simulating noise and construction site noise. The effect of taking longer or shorter time constants for the low pass filters within the adaptation loops (M4 and M5, respectively) is limited. Longer time constants lead to slightly increased error rates in most situations, whereas shorter time constants yield similar results compared to the original model, except for the case of white noise, where the performance of the original model is a bit better in low signal-to-noise ratios.

In summary, the original parameters of the adaptive compression stage, which were optimized for predicting psychoacoustical experiments appear to be well suited for application in ASR systems. They seem to represent a broad optimum in terms of robust representation of speech in noise, and no distinct improvement of performance in all signal-to-noise ratios could be found when modifying them. The results indicate that there has to be some adaptive compression *at all* to improve robustness in noise, as can be seen from the performance of M1. Without adaptive compression, recognition rates are satisfactory in clean speech, but drop quite fast in noise.

Low pass filtering of the envelope after compression in each frequency

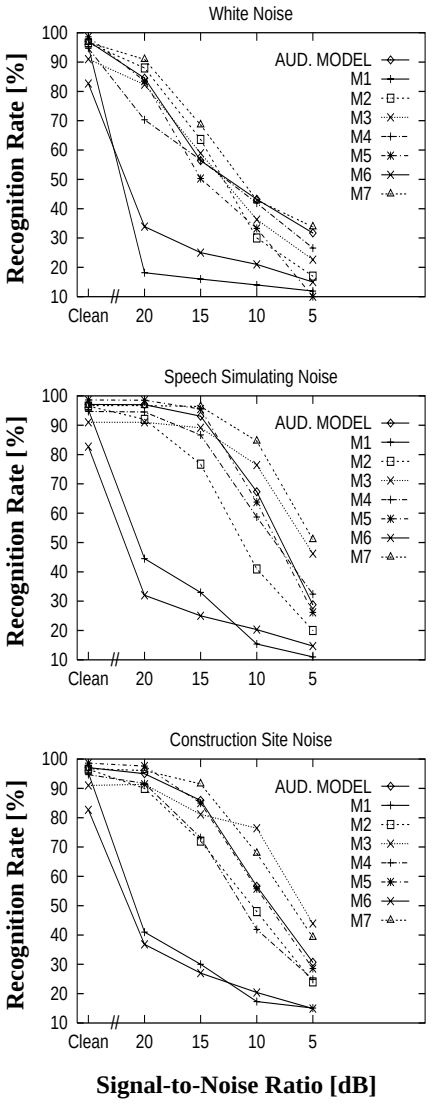


Figure 4.7: Recognition rates in % in different types of additive noise as function of signal-to-noise ratio in dB obtained with the original auditory model and with its modifications M1-M7 (see text).

channel also plays an important role for robust representation of speech, as the recognition rates obtained for modifications M6 and M7 of the auditory front end show. Without filtering (M6), only poor recognition rates in noise are yielded. Stronger attenuation of fast amplitude modulations (M7), on the other hand, leads to decent improvement of robustness in construction site noise and speech simulating noise. In clean speech, however, M7 results in a slightly degraded recognition rate (96.7%), compared to the original processing (97.1%). In contrast to the adaptive compression stage, the psychoacoustically-motivated low pass filter parameter at the end of the auditory preprocessing seems not to be optimal for speech processing. In our experiments, stronger attenuation of higher modulation frequencies (which better reflects the modulation spectrum of speech) enhanced robustness of speech recognition in noise, i.e., ASR performance benefits from variation of the original psychacoustically-motivated low pass filter.

4.5 Discussion

The main findings in this study can be summarized as follows:
(1) The presented auditory-based ASR front end allows more robust speaker-independent digit recognition compared to standard feature extraction in additive noise, even when the parameters of the psychoacoustically-motivated model were not refitted for the new task in the field of speech processing.
(2) An important processing step of the auditory model for robust representation of speech in additive noise is the adaptive compression stage, which encodes the dynamic evolution of the input signal and allows simulation of temporal aspects of processing found in the auditory system.
(3) Modulation band pass filtering centered at low modulation frequencies plays an important role for robust speech representation. Changing the original psychoacoustical filter parameters to better reflect the average modulation spectrum of speech further enhances the robustness of isolated digit recognition in noise.
Encoding the dynamic evolution of the input signal as realized by the auditory model emphasizes changes in the input signal relative to constant

portions. The signal representation of the auditory model is characterized by sparse and distinct peaks. These peaks are well maintained in additive noise and seem to serve as quite robust cues for the recognizer. A study on the interplay between auditory-based features and a recurrent neural network as recognizer supports this suggestion (Tchorz *et al.*, 1997). It was shown that the sparse and distinct peaks contain sufficient information for robust recognition with neural networks. Even if 80% of all feature values (i.e., those which did not exceed a certain threshold) were set to zero, a digit recognition rate of more than 90% was reached on these strongly reduced features.

Thus, the experiments presented here demonstrate that the consideration of short-term adaptation as in the auditory model might be fruitful for reducing the influence of additive noise in future ASR front ends. Findings from other researchers support this hypothesis (Strope and Alwan, 1997).

In human speech perception, analysis of low modulation frequencies plays an important role. In a study on the intelligibility of temporally-smeared speech, Drullman *et al.* (1994) found that modulation frequencies below 8 Hz are the most important ones for speech intelligibility. Components between 8 Hz and 16 Hz were found useful, too, whereas rates above 16 Hz are not required for speech intelligibility.

The 8 Hz low pass filter at the end of the auditory model preprocessing, which leads to a band pass modulation transfer function of the model, was introduced by Dau (1996b) to optimize predictions of psychoacoustical masking experiments. When applied as ASR front end, the band pass characteristic attenuates non-speech sources in disturbed speech. Further reduction of the filter's cut-off frequency allows for higher digit recognition rates in noise (which might be explained by the better correspondence with the average modulation spectrum of speech), but also slightly degraded the performance in clean speech.

Kanedera *et al.* (1997) measured the effects of band pass filtering of the time trajectories of spectral envelopes on speech recognition. Their results showed that most of the useful linguistic information for ASR is in modulation frequency components in the range from 1 Hz to 16 Hz, with the dominant component around 4 Hz. In noisy environment, the range below 2 Hz and above 16 Hz sometimes degraded the recognition

accuracy.
In isolated word recognition, there seems to be a tradeoff concerning the modulation frequencies above about 4 Hz. In clean speech, they carry useful information which can be exploited by the recognizer and thus should be passed through. In noise, however, the restriction to the most prominent modulation frequencies of speech helps to enhance robustness. In the field of sub-word unit recognition (e.g., phonemes), however, strong attenuation of higher modulation frequencies might lead to problems even in clean speech. Transitions between short segments are blurred by low pass filtering of the envelope. The feature representation of a single phoneme strongly depends on its temporal context then, and might make it difficult to train and recognize sub-word units.
The auditory model described in this paper exhibits similar features as RASTA processing of speech (Hermansky and Morgan, 1994) with respect to representation of temporal properties of the incoming signal. Both techniques perform some kind of envelope bandpass filtering around 4 Hz and hence take roughly 200 ms of "signal history" into account for feature calculation. This is in clear contrast to common short-term acoustic features, which represent independent 10-20 ms frames. Temporal processing in the auditory model is essential to quantitatively predict psychoacoustical masking experiments. While the auditory model used here was primarily developed to model a variety of temporal psychoacoustical effects, RASTA was also shown to model certain aspects of forward masking experiments (Hermansky and Pavel, 1998).
Kasper *et al.* (1997) compared the robustness of ASR systems with four different types of feature vectors, namely cepstral coefficients, features from the auditory model as described in this paper, and log-RASTA and J-RASTA coefficients (Hermansky and Morgan, 1994) in isolated-digit recognition experiments. In a first set of experiments, a HMM recognizer was used. The poorest recognition rates in noise (either additive noise or convolutive distortions of the test material) were measured with cepstral coefficients. Log-RASTA processing performed slightly better than the auditory model front end. Adaptive J-RASTA processing allowed for the highest recognition rates in almost all conditions. In a second set of experiments, a locally recurrent neural network was used as recognizer instead of the HMM recognizer (Kasper *et al.*, 1997). This lead to a

further decrease in robustness with cepstral coefficients as features. The results for RASTA processing were almost not affected by the different recognizer. The recognition rates rates for auditory model features, however, significantly increased with the neural network recognizer and were comparable to those obtained with adaptive J-RASTA processing (without requiring explicit speech-free portions for noise power estimation, as J-RASTA). Obviously, the neural network better exploits the characteristics of the auditory model features for robust recognition than the HMM recognizer does.
In a recent study, Kasper and Reininger (1999) further investigated these recognizer dependencies. They transformed the original feature vectors of the auditory model into their cepstra prior to HMM training and testing. With these modified features, the authors reported much improved recognition rates in both additive noise and convolutive distortions, compared to the original features. These results demonstrate that the original auditory model features are only partly suited for a HMM recognition framework with diagonal covariance matrices, as the feature values are partly correlated across frequency channels. For optimal performance in actual recognition systems, a careful adjustment between the front end and the recognizer is necessary. This point holds for most non-standard approaches to feature extraction and was discussed for different types of front ends by Bourlard *et al.* (1996).

4.6 Conclusion

The psychoacoustically-motivated auditory model which was originally developed to describe human performance in typical psychoacoustical spectral and temporal masking experiments yields promising results when applied as front end to ASR systems, especially in noisy environment. In the model, an appropriate temporal processing in each frequency channel of the auditory model plays an important role for robust representation of speech.
To further evaluate the potential of the auditory model in speech recognition systems, experiments with large word vocabulary as well as recognition experiments basing on sub-word units are necessary.

ACKNOWLEDGEMENTS

This work was supported by Deutsche Forschungsgemeinschaft (Ko 942). Many thanks to Torsten Dau, Martin Hansen, Matthias Wesselkamp, and all other members of the AG Medizinische Physik for support and fruitful discussions, and to Klaus Kasper and Herbert Reininger from Frankfurt University for close co-operatation and experiments with the auditory model. We also thank three anonymous reviewers for their comments and suggestions, which definitely helped to improve the paper.

Chapter 5

Noise suppression based on neurophysiologically-motivated SNR estimation for robust speech recognition [1]

Abstract

A novel noise suppression scheme is evaluated which is based on a neurophysiologically-motivated estimation of the local signal-to-noise ratio (SNR) in different frequency channels. For SNR-estimation, the input signal is transformed into so-called Amplitude Modulation Spectrograms (AMS), which represent both spectral and temporal characteristics of the respective analysis frame, and which imitate the representation of mod-

[1] This Chapter is an extended version of a paper which has been presented at the conference on Neural Information Processing Systems (NIPS) 2000: Tchorz, Kleinschmidt, and Kollmeier (2000) "Noise suppression based on neurophysiologically-motivated SNR estimation for robust speech recognition".

ulation frequencies in higher stages of the mammalian auditory system. A neural network is used to analyze AMS patterns generated from noisy speech and provides estimates of the local SNR. Noise suppression is achieved by attenuating frequency channels according to their SNR. The noise suppression algorithm is evaluated in speaker-independent digit recognition experiments and compared to noise suppression by Spectral Subtraction.
The results show that AMS-based noise suppression significantly improves digit recognition in noise, in comparison with unprocessed data and with Spectral Subtraction using noise measures based on voice activity detection. In stationary noise, perfect speech pause detection (which is not available in real systems) allows for a reliable estimation of the noise floor. In non-stationary noise, however, the AMS pattern-based signal classification and noise suppression is advantageous, as noise estimation which is restricted to speech pauses and which cannot be updated while speech is active was shown to even degrade the recognition rates.

5.1 Introduction

One of the major problems in automatic speech recognition (ASR) systems is their unsatisfactory robustness in noise, which severely degrades their usefulness in many practical applications. Several proposals have been made to increase the robustness of ASR systems, e.g. by model compensation or more noise-robust feature extraction (Hermansky and Morgan, 1994; Tchorz and Kollmeier, 1999b). Another method to overcome the lack of robustness of ASR systems is to suppress the background noise before feature extraction. Classical approaches for single-channel noise suppression are Spectral Subtraction (Boll, 1979) and related schemes, e.g. (Ephraïm and Malah, 1984), where the noise spectrum is usually estimated in detected speech pauses and subtracted from the signal. In these approaches, stationarity of the noise has to be assumed while speech is active. Furthermore, portions detected as speech pauses must not contain any speech in order to allow for correct noise measurement. At the same time all actual speech pauses should be detected for a fast update of the noise measurement. In reality, however, these partially conflicting requirements are often not met.

In order to overcome these problems, the noise suppression algorithm outlined in this work directly estimates the local SNR in a range of frequency channels even if speech and noise are present at the same time, i.e., no explicit detection of speech pauses and no assumptions on noise stationarity during speech activity are necessary. For SNR estimation, the input signal is transformed into spectro-temporal input features, which are neurophysiologically-motivated: experiments on amplitude modulation processing in higher stages of the auditory system in mammals show that modulations are represented in "periodotopical" gradients, which are almost orthogonal to the tonotopical organization of center frequencies (Langner *et al.*, 1997). Thus, both spectral and temporal information are represented in two-dimensional maps. These findings were applied to signal processing in a binaural noise suppression system (Kollmeier and Koch, 1994) with the introduction of so-called Amplitude Modulation Spectrograms (AMS), which contain information of both center frequencies and modulation frequencies. In the present study, the different representations of speech and noise in AMS patterns are detected by a neural network, which estimates the local SNR in each frequency channel. For noise suppression, the frequency bands are attenuated according to the estimated local SNR in the respective frequency channel.

The proposed noise suppression scheme is evaluated in isolated-digit recognition experiments. As recognizer, a combination of an auditory-based front end (Tchorz and Kollmeier, 1999b) (Chapter 4 of this thesis) and a locally-recurrent neural network (Kasper *et al.*, 1995) is used. In earlier studies, this combination was found to allow for more robust isolated-digit recognition rates, compared to a standard recognizer with mel-cepstral features and HMM modeling (Kasper *et al.*, 1997; Kleinschmidt *et al.*, 2000). Thus, the recognition experiments in this study are conducted with this particular combination to evaluate whether a further increase of robustness can be achieved with additional noise suppression.

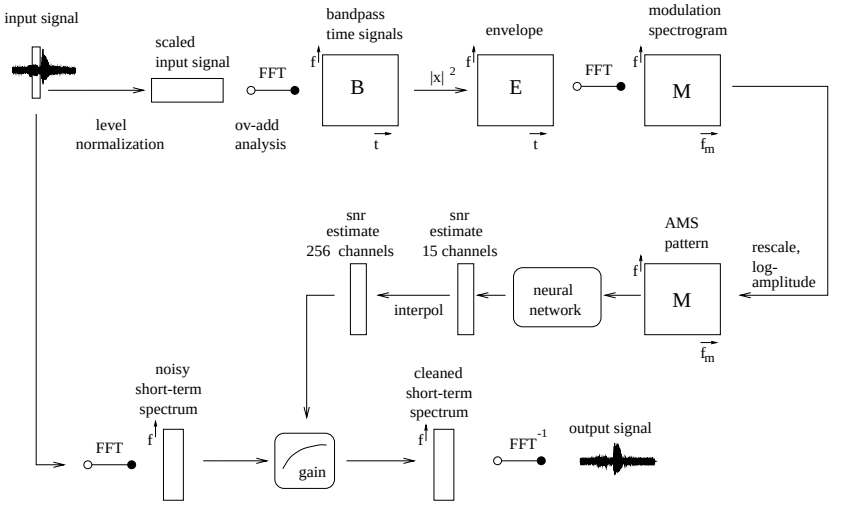


Figure 5.1: Processing stages of AMS-based noise suppression.

5.2 The recognition system

5.2.1 Noise suppression

Figure 5.1 shows the processing steps which are performed for noise suppression. To generate AMS patterns which are used for SNR estimation, the input signal (16 kHz sampling rate) is short-term level adjusted, i.e., each 32 ms segment which is later transformed into an AMS pattern is scaled to the same root-mean-square value. The level-adjusted signal is then subdivided into overlapping segments of 4.0 ms duration with a progression of 0.25 ms for each new segment. Each segment is multiplied by a Hanning window and padded with zeros to obtain a frame of 128 samples which is transformed with a FFT into a complex spectrum, with a spectral resolution of 125 Hz. The resulting 64 complex samples are considered as a function of time, i.e., as a band pass filtered complex time signal. Their respective envelopes are extracted by squaring. This envelope signal is again segmented into overlapping segments of 128 samples (32ms) with an overlap of 64 samples. Each segment is multiplied

with a Hanning window and padded with zeros to obtain a frame of 256 samples. A further FFT is computed and supplies a modulation spectrum in each frequency channel, with a modulation frequency resolution of 15.6 Hz. By an appropriate summation of neighbouring FFT bins, the frequency axis is transformed to a Bark scale with 15 channels, with center frequencies from 100-7300 Hz. The modulation frequency spectrum is scaled logarithmically by appropriate summation, which is motivated by psychoacoustical findings on the shape of auditory modulation filters (Ewert and Dau, 1999).
The modulation frequency spectrum is restricted to the range between 50-400 Hz and has a resolution of 15 channels. Thus, the fundamental frequency of typical voiced speech is represented in the modulation spectrum. The chosen range corresponds to the fundamental frequencies which were used by Langner et al. in their neurophysiological experiments on amplitude modulation representation in the human auditory cortex (Langner *et al.*, 1997). Informal experiments showed that higher modulation frequencies do not contribute additional information for SNR estimation. Very low modulation frequencies from articulatory movements (which are characteristic for speech and which play an important role for speech intelligibility) are also not taken into account, as they are not properly resolved due to the short analysis windows. In a last processing step, the amplitude range is log-compressed.
The AMS representation is restricted to a 15 times 15 pattern to limit the amount of training data which is necessary to train the fully connected perceptron. This is important since this amount increases with the number of neurons in each layer. Examples for AMS patterns can be seen in Fig. 5.2. The AMS pattern on the left side was generated from a voiced speech portion. The periodicity at the fundamental frequency (approx. 110 Hz) is represented in each center frequency band. The AMS pattern on the right side was generated from speech simulating noise. The typical spectral tilt can be seen, but no structure across modulation frequencies.

For classifying AMS patterns and estimating the narrow-band SNR of each AMS pattern, a feed-forward neural network is implemented. The net consists of 225 input neurons (15*15, the AMS resolution of center frequencies and modulation frequencies, respectively), a hidden layer

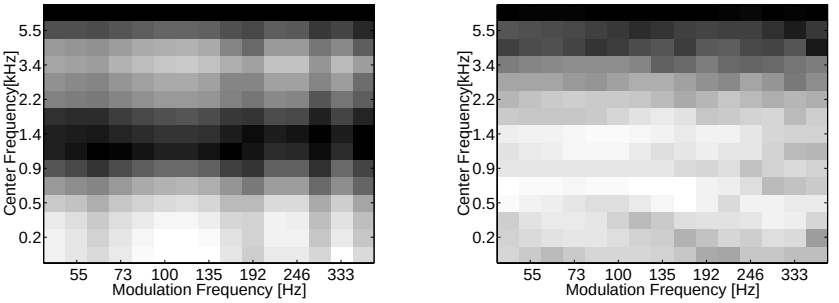


Figure 5.2: AMS patterns generated from a voiced speech segment (left), and from speech simulating noise (right). Each AMS pattern represents a 32 ms portion of the input signal. Bright and dark areas indicate high and low energies, respectively.

with 160 neurons, and an output layer with 15 neurons. The activity of each output neuron indicates the SNR in one of the 15 center frequency channels. For training, the narrow-band SNRs in 15 channels are measured for each AMS analysis frame of the training material prior to adding speech and noise. The neural network was trained with AMS patterns generated from 72 min of noisy speech from 400 talkers and 41 natural noise types. The measured SNR values are transformed to output neuron activities which serve as target activities during training (SNRs between -10 and 20 dB are linearly transformed to activities between 0.05 and 0.95. SNRs below -10 dB and above 30 dB are assigned to activities of 0.05 and 0.95, respectively). After training, AMS patterns generated from "unknown" sound material are presented to the network. The 15 output neuron activities that appear for each pattern serve as SNR estimates for the respective frequency channels. In a detailed study on AMS-based broad-band SNR estimation (Tchorz and Kollmeier, 2000) (Chapter 2 of this thesis) it was shown that harmonicity which is well represented in AMS patterns is the most important cue for the neural network to distinguish between speech and noise. However, harmonicity is not the only cue, as the algorithm allows for reliable discrimination between unvoiced speech and noise.

Sub-band SNR estimates are utilized for noise suppression by attenuating frequency channels according to their local SNR. The gain function which was applied is given by

$$g_k = (\mathrm{SNR}_k/(\mathrm{SNR}_k + 1))^x, \tag{5.1}$$

where k denotes the frequency band, SNR the signal-to-noise ratio on a linear scale, and x is an exponent which controls the strength of the attenuation. Note that for x=1 this is equivalent to a Wiener filter. Noise suppression based on AMS-derived SNR estimation is performed in the frequency domain. The input signal is segmented into overlapping frames with a window length of 32 ms, and a shift of 16 ms is applied, i.e., each window corresponds to one AMS analysis frame. The FFT is computed in every window. The magnitude in each frequency bin is multiplied by the corresponding gain computed from the AMS-based SNR estimation. The gain in frequency bins which are not covered by the center frequencies from the SNR estimation is linearly interpolated from neighboring estimation frequencies. The phase of the input signal is remained unchanged and applied to the attenuated magnitude spectrum. An inverse FFT is computed, and the enhanced speech is obtained by overlapping and adding.
A visualization for the proposed noise suppression is given in Fig. 5.3. It can be seen that the background noise is not completely attenuated, but a small noise floor is left.

5.2.2 Auditory-based ASR feature extraction

The front end which is used in the recognition system is based on a quantitative model of the "effective" peripheral auditory processing. The model simulates both spectral and temporal properties of sound processing in the auditory system which were found in psychoacoustical and physiological experiments. The model was originally developed for describing human performance in typical psychoacoustical spectral and temporal masking experiments, e.g., predicting the thresholds in backward, simultaneous, and forward-masking experiments (Dau *et al.*, 1996b; 1997a). The main processing stages of the auditory model are

Figure 5.3: Example for AMS-based noise suppression. Top: undisturbed waveform of the German digit *sieben*. Middle: utterance disturbed with white Gaussian noise at 5 dB SNR. Bottom: waveform after processing of the disturbed utterance

gammatone filtering, envelope extraction in each frequency channel, adaptive amplitude compression, and low pass filtering of the envelope in each band. The adaptive compression stage compresses steady-state portions of the input signal logarithmically. Changes like onsets or offsets, in contrast, are transformed linearly. A detailed description of the auditory-based front end is given in (Tchorz and Kollmeier, 1999b) (Chapter 3 of this thesis).

5.2.3 Neural network recognizer

For scoring of the input features, a locally recurrent neural network (LRNN) is employed with three layers of neurons (150 input, 289 hidden, and 10 output neurons). Hidden layer neurons have recurrent connections to their 24 nearest neighbours. The input matrix consists of 5 times the auditory model feature vector with 30 elements, glued together in order to allow the network to memorize a time sequence of input matrices. The network was trained using the Backpropagation-trough-time algorithm with 200 iterations (see (Kasper *et al.*, 1995) for a detailed description of the recognizer).

5.3 Recognition experiments

5.3.1 Setup

The speech material for training of the word models and scoring was taken from the ZIFKOM database of Deutsche Telekom AG. Each German digit was spoken once by 200 different speakers (100 males, 100 females). The recording sessions took place in soundproof booths or quiet offices. The speech material was sampled at 16 kHz.
Three different types of noise were added to the speech material at different signal-to-noise ratios before feature extraction: a) white Gaussian noise, b) speech-simulating noise which is characterized by a long-term speech spectrum and amplitude modulations which reflect an uncorrelated superposition of 6 speakers (ICRA, 1997)[2], and c) background noise recorded in a printing room which strongly fluctuates in both amplitude and spectral shape. The SNR for the speech in noise samples was adjusted by first computing the RMS value of each word utterance including short pauses before and after the utterance. Then the background noises were added to the utterances with signal-to-noise ratios ranging from 20 to -10 dB. The word models were trained with features from 100 undisturbed and unprocessed utterances of each digit. Features for testing were calculated from another 100 utterances of each digit which were distorted by additive noise before preprocessing. The recognition rates were measured without noise suppression and with noise suppression as described in Section 5.2.1.
Five different exponents x of the gain function 5.1 were applied, rang-

[2]The speech simulating noise is generated by splitting the signal into three bands with cross-over frequencies of 850 Hz and 2500 Hz using very steep filters (>100 dB/octave). In the next processing step, each of the three bands are processed according to Schroeder (1968), which means that with a probability of 50% the sign of each sample of the speech is at random either reversed or kept unaltered. Thus, each of the modified signals is completely unintelligible and has a flat white spectrum. Next, the Schroeder processed signals are again filtered with the same three filters as above and scaled to have the same RMS-value. The three bands are added and filtered with a male or female speech shaped filter in close accordance with LTASS (Byrne, 1994). In a last processing step, the phase is randomized. The resulting signals have a long-term spectrum according to LTASS and modulation characteristics like natural speech. The applied noise consist of an uncorrelated superposition of 6 persons babble (1f + 1m + 2f(-6 dB) + 2m(-6 dB))

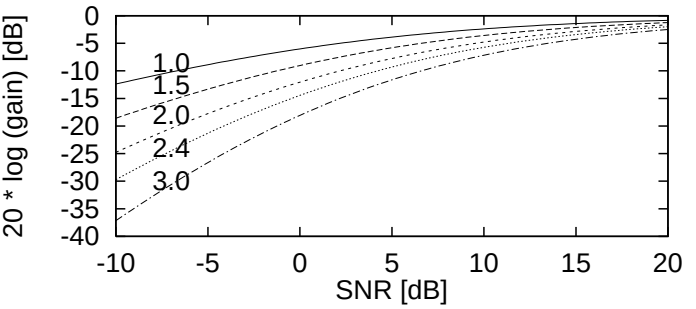
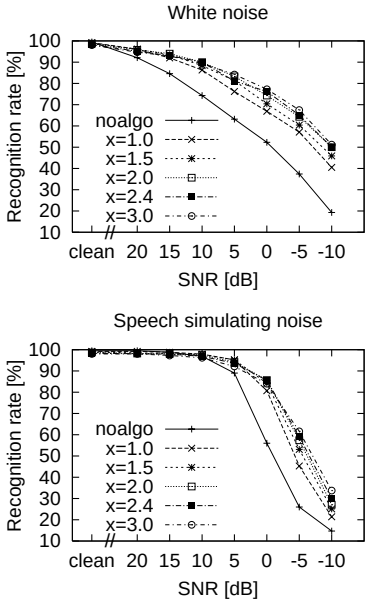


Figure 5.4: Five different gain functions were applied in the recognition experiments. The exponent x (Eqn. 5.1) ranges from 1-3.

ing from 1-3. The gain functions which were employed are plotted in Fig. 5.4.

With small x, the maximum attenuation of noise is limited, but speech quality in favourable SNRs is best preserved (as the impact of SNR estimation errors is limited). With high x, in contrast, there is a strong attenuation of noise, but the speech quality may be degraded due to estimation errors. It can be seen from Fig. 5.4 that all signal portions are attenuated (even those with an estimated SNR close to 20 dB, which is the upper limit in the estimation process). Thus, each speech file was scaled after noise suppression to have the same maximum peak amplitude as the noisy signal before processing.

For comparison, the recognition rates were measured with noise suppression based on Spectral Subtraction including residual noise reduction (Boll, 1979) before feature extraction. Two methods for noise estimation were applied. In the first method, speech pauses in the noisy signals were detected using the voice activity detector (VAD) standardized by the International Telecommunication Union (ITU, 1996), which utilizes information on energy, zero-crossing rate, and spectral distortions for voice activity detection. The noise measure was updated in speech pauses using a low pass filter with a time constant of 40 ms to temporally smooth the measure. In the second method, the noise spectrum was measured in speech pauses which were detected from the *clean*



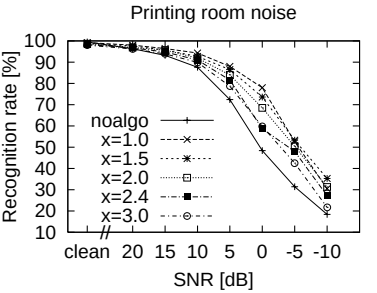


Figure 5.5: Speaker-independent, isolated digit recognition rates for three types of noise as a function of the SNR without noise suppression (noalgo) and with AMS-based noise suppression for different gain function exponents x ranging from 1.0-3.0.

utterances using an energy criterion (thus, perfect speech pause information is provided, which is not available in real applications). This control condition allows to determine the influence of accurate speech pause detection on Spectral Subtraction performance when applied in an ASR system.

5.3.2 Results

The speaker-independent, isolated-digit recognition rates which were obtained with AMS-based noise suppression in three types of background noise are plotted in Fig. 5.5.

In all types of noise and SNRs (except speech simulating noise in high SNRs), recognition rates increase with AMS-based noise suppression, compared to unprocessed data. In white noise, a strong attenuation of noisy parts (i.e., a high gain function exponent x) is advantageous. To a smaller extend, this also holds for speech simulating noise, which

un-	AMS-based					SpecSub	
processed	x=1	x=1.5	x=2	x=2.4	x=3	VAD	perf
99.5	99.2	98.9	98.7	98.3	98	99.1	99.1

Table 5.1: Recognition rates in % in clean speech for unprocessed data, after AMS-based noise suppression (with different gain function exponents x), and after Spectral Subtraction (with VAD-based and perfect speech pause detection).

has a constant spectrum but fluctuates in level. In printing office noise which fluctuates in both spectrum and level, the situation is reversed. Here, moderate noise suppression (x=1) yields the highest recognition rates. In this particular type of noise, the optimal value between the two conflicting goals "strong attenuation of noise" and "preservation of speech quality" is close to the latter one.
The recognition rates in clean speech for AMS-based noise suppression and the two variations of Spectral Subtraction are shown in Tab. 5.1. The highest recognition rate is obtained with unprocessed speech (99.5% on the test data). Any processing (either AMS-based or Spectral Subtraction) introduces artifacts and leads to higher error rates. For AMS-based processing, the amount of artifacts on clean speech depends on the choice of the gain function exponent x. With increasing exponent, the amount of artifacts is increased (but even with x=3, a speaker-independent digit recognition rate of 98% is obtained).

In Fig. 5.6, the recognition rates for AMS-based noise suppression (x=1.5) compared to Spectral Subtraction with VAD-based and perfect speech pause detection are plotted. In all tested noises, noise suppression with the proposed algorithm yields higher performance in comparison to Spectral Subtraction with VAD-based noise measurement. Spectral Subtraction with perfect speech pause detection allows for higher recognition rates than the AMS-based approach in stationary white noise. Here, the noise measure for Spectral Subtraction is very accurate during speech activity and allows for effective noise removal. AMS-based noise suppression estimates the SNR in every analysis frame, and no a priori information on speech-free segments is provided to the algorithm.

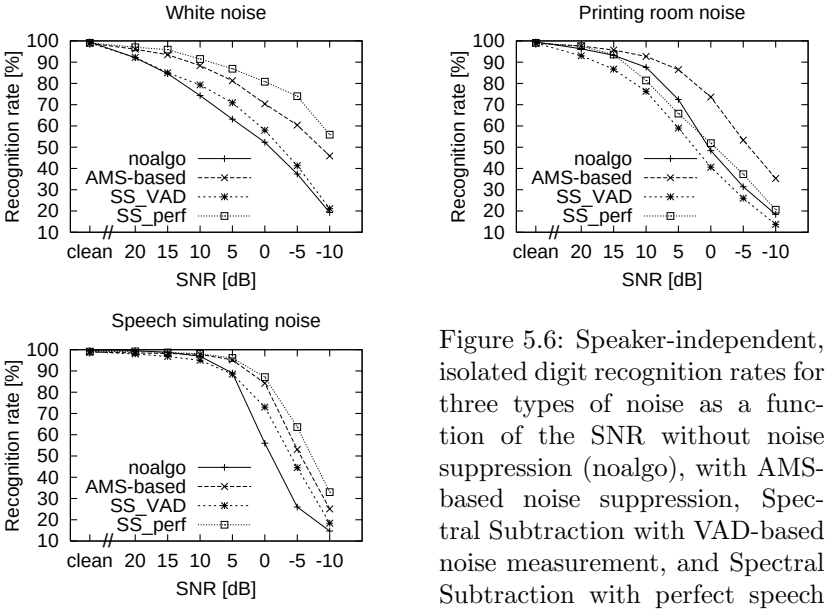


Figure 5.6: Speaker-independent, isolated digit recognition rates for three types of noise as a function of the SNR without noise suppression (noalgo), with AMS-based noise suppression, Spectral Subtraction with VAD-based noise measurement, and Spectral Subtraction with perfect speech pause information.

In speech simulating noise, which fluctuates in level but not in spectral shape, Spectral Subtraction with perfect speech pause detection works slightly better than AMS-based noise suppression. In printing room noise, which fluctuates in both level and spectrum, the AMS-based approach yields the best results. Here, Spectral Subtraction even degrades the recognition rates in some SNRs, compared to the unprocessed data. The noise measure from VAD-based or perfect speech pause detection cannot be updated while speech is active. Thus, an incorrect spectrum is subtracted and leads to artifacts and degraded recognition performance. The large discrepancy in recognition rates between the two versions of Spectral Subtraction (VAD-based speech pause detection, and perfect speech pause detection) can be explained by VAD detection errors. Not all speech pauses were detected by the VAD, and some speech portions were identified as "noise", resulting in a wrong noise spectrum being

subtracted from the signal. These effects are illustrated in Fig. 5.7. The input signal (second panel) was a mixture of speech (*"sechs"*, plotted separately in the first panel) and white Gaussian noise at 10 dB SNR. The two unvoiced consonants */s/* were in part mistakenly identified as speech pause by the VAD (indicated by the black bars below the third panel). Thus, the respective portions were suppressed by the Spectral Subtraction, and the resulting signal does almost not contain the unvoiced parts, but mostly the voiced part which was correctly identified as "speech". With a perfect speech pause detection (fourth panel), the correct spectrum can be subtracted, and the unvoiced parts of the speech signal are audible. AMS-based SNR estimation is to some extend capable to detect the unvoiced fricatives in white noise, which are partly maintained in the processed signal (bottom panel). SNR estimation errors are likely to be locally restricted, as the SNR is predicted continuously. VAD-based noise measurement errors, in contrast, may persist over relatively long segments when no correct speech pauses are detected.

5.4 Discussion

The proposed neurophysiologically-motivated noise suppression scheme was shown to significantly improve digit recognition in noise in comparison with unprocessed data and with Spectral Subtraction using VAD-based noise measures. A perfect speech pause detection (which is not available yet in real systems) allows for a reliable estimation of the noise floor in stationary noise. In non-stationary noise, however, the AMS pattern-based signal classification and noise suppression is advantageous, as it does not depend on speech pause detection and no assumption is necessary about the noise being stationary while speech is active.
Spectral Subtraction as described in (Boll, 1979) produces musical tones, i.e. fast fluctuating spectral peaks. In an earlier study (Tchorz *et al.*, 1997), the LRNN recognizer was shown to be relatively insensitive against fast random fluctuations, compared to a HMM recognizer, and benefits from this type of noise suppression. Human listeners, in contrast, tend to be annoyed by these artifacts. Hence, the assessment criteria for a noise suppression scheme strongly depend on the desired application. Parameter settings which are suited for automatic speech

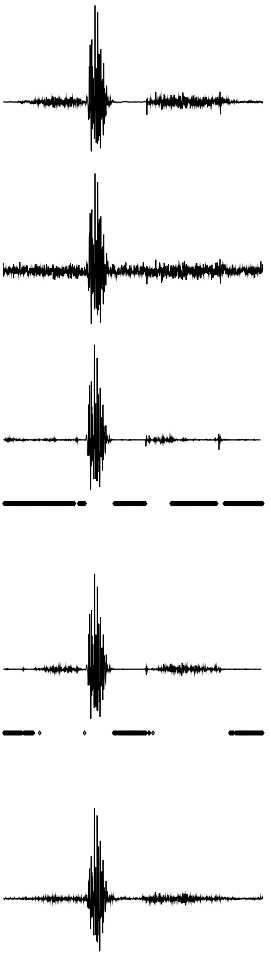


Figure 5.7: The top panel shows the waveform of an utterance of the German digit *sechs*, spoken by a male talker. The first consonant was pronounced voiceless (as in *sex*). The second panel shows the same utterance, disturbed with white Gaussian noise at 10 dB SNR. In the third panel, the noisy signal was processed by Spectral Subtraction using a VAD. The detected speech pauses are marked by the bars below. The fourth panel shows the signal being processed with Spectral Subtraction using perfect pause detection (bars below). In the bottom panel, the signal after AMS-based noise suppression is plotted.

recognition are not necessarily appropriate for human listeners. In a recent study conducted by Kleinschmidt et al. (2000) on the effects of different types of noise suppression in digit recognition it was shown that, in general, noise suppression schemes which strongly attenuate noise and which at the same time produce lots of artifacts performed best.
The neurophysiologically-based noise suppression scheme outlined in this paper does not produce fast fluctuating, musical noise-like artifacts. In general, a good quality of speech is maintained. The choice of the attenuation exponent x has only little impact on the quality of speech in favourable SNRs, which is well preserved. With decreasing SNR, however, there is a tradeoff between the amount of noise suppression and distortions of the speech. Large gain function exponents x which are optimal for suppression of stationary noise clearly degrade speech quality in poor SNRs, and cause annoying artifacts. A typical distortion of speech in poor signal-to-noise ratios is an unnatural spectral "coloring" (rather than fast fluctuating distortions). For an assessment of these effects, a further evaluation in human listeners is necessary, not only in terms of speech intelligibility, but also with respect to subjective speech quality and listening effort. The current implementation of the noise suppression algorithm, however, does not allow for an application in e.g. digital hearing instruments, as the minimum processing delay is 32 ms (the length of one AMS frame from which the SNR estimate is determined). Hence, procedures have to be developed and tested which predict the "real-time" SNR from past estimates, e.g. by utilizing them as *a priori* SNR measures, which were employed by Ephraïm and Malah (1984) to reduce annoying musical noise.

Acknowledgements

Many thanks to Klaus Kasper and Herbert Reininger from Institut für Angewandte Physik, Universität Frankfurt/M. for supplying us with their LRNN implementation.

Chapter 6

Summary and conclusion

Background noise is a major problem in a range of speech processing applications, in both communication between humans and speech interfaces to machines. The current thesis is concerned with the application of certain properties of the auditory system to computational speech processing, aiming at reducing the disturbing effects of background noise. Two main problems of computational speech processing are tackled, namely the detection and suppression of noise in monaural input signals, and the extraction of noise-robust features for ASR systems.

The first problem is dealt with in Chapter 2 and 3 of this thesis, where a noise suppression algorithm based on a novel, neurophysiologically-motivated SNR estimation is proposed. For SNR estimation, spectro-temporal patterns (so-called Amplitude Modulation Spectrograms, AMS) are extracted from the waveform. These patterns contain information on both center frequencies and modulation frequencies, and imitate the respective representation of sounds in higher stages of the auditory system in a simple way. In AMS patterns, differences between speech and noise are reflected in the spectro-temporal joint representation, which are exploited by a neural network pattern recognizer to automatically distinguish between speech and noise. In situations with speech and noise being present at the same time, a fast estimation of the local SNR is possible, with the highest accuracy in signal-to-noise

ratios between -5 and 15 dB.
Experiments on the mechanisms and the most important features for SNR estimation revealed that harmonicity appears to be the most important cue for a segment to be classified as "speech", but not the only one, as the algorithm is able to reliably separate between unvoiced speech and noise. For SNR estimation, the full joint representation of AMS patterns with both spectral and temporal information is necessary to yield high estimation accuracies, compared to reduced representations with only spectral or temporal information.
For narrow-band SNR estimation, which is developed in Chapter 3, across-frequency connections of the neural network play an important role for reliable performance. This might be explained by the fact that in many real situations, the evolution of the SNR is not independent in different frequency bands, but correlates at least with neighbouring channels.
The most important difference between the proposed algorithm and common SNR estimation methods is that it directly predicts the local SNR even if both speech and noise are present at the same time. Common approaches either have to rely on proper speech pause detection (with the assumption that the noise is constant while speech is active), or require relatively long signal segments for e.g. analysis of amplitude histograms or slow modulation frequencies.
Informal listening experiments revealed that noise suppression based on the proposed SNR estimation yields a clear suppression of most noise types, with only little annoying artifacts from degradation of the speech signal (at least in favourable to moderate SNRs). Musical noise-like, fast fluctuating artifacts as known from Spectral Subtraction do not occur. Objective speech quality measures indicate a benefit from AMS-based noise suppression in most situations.
The problem of additive noise in automatic speech recognition (ASR) systems is addressed in Chapter 4. One method amongst others to enhance the performance in unfavourable conditions is to increase the robustness of the feature extraction stage. A model of the auditory periphery which was originally developed to predict human performance in typical psychoacoustical masking experiments was applied as front end in an ASR system. Compared to a standard mel-cepstral front

end, it allows for more robust digit recognition in different types of background noise. The processing stages of the auditory model were studied in detail, and it was shown that the adaptive compression stage of the model, which linearly transforms fast changes in the input signal like onsets or offsets, but compresses steady-state portions is essential for robust recognition in noise. Combined with the final low pass filter, a modulation transfer function of the auditory model is achieved which reflects the average modulation spectrum of speech. Thus, very fast or very slow modulations, which are not likely to to origin from speech are attenuated, and noise portions are attenuated.
In Chapter 5, the noise suppression scheme which was developed in Chapter 2 and 3 of this thesis was evaluated in digit recognition experiments, where noisy speech was enhanced prior to feature extraction with the auditory model as front end, which was described in Chapter 4. The results show that a further enhancement of robustness is attained. In all tested types of noise, higher recognition rates were achieved with the proposed noise suppression scheme, compared to unprocessed digits and to digits which were processed by Spectral Subtraction with voice activity detection (VAD) - based speech pause tracking. Especially in non-stationary noise, direct estimation of the SNR in every analysis frame as performed by the proposed algorithm showed to be advantageous. With VAD-based SNR estimation, speech portions mistakenly identified as "noise" may lead to wrong noise estimates for a relatively long period of time and thus produce speech distortions after Spectral Subtraction.

In conclusion, findings from neurophysiology and psychoacoustics were successfully applied in two different fields of technical sound signal processing, namely noise suppression and speech recognition. Finally, both approaches could be combined in a beneficial way in order to allow for an even more noise robust automatic speech recognition system. These applications can be regarded as further examples for the observation that it might be very helpful to have a look at nature, and try to understand how it works. The attempt to imitate the "feature extraction" of the human auditory system, of course, has to be very low-level and simplistic, as it is extremely complex and only partly understood yet.

The cognitive skills *behind* auditory system feature extraction (which already allow small children to successfully manage the really impressive acoustic tasks of daily life), however, are far away from being understood or even imitated by machines.

Bibliography

Bacon, S.P. and Grantham, D.W. (**1989**). Modulation masking: Effects of modulation frequency, depth, and phase. *J. Acoust. Soc. Am.*, **85**:2575–2580.

Bishop, C. M. (**1995**). *Neural networks for pattern recognition*. Oxford University Press.

Boer, E. de and Kruidenier, C. (**1990**). On ringing limits of the auditory periphery. *Biol. Cybern.*, **63**:433–442.

Boll, S. (**1979**). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal Processing*, **27**(2):113–120.

Bourlard, H., Hermansky, H. and Morgan, N. (**1996**). Towards increasing speech recognition error rate. *Speech Communication*, **18**:205–231.

Bregman, A.S. (**1993**). Auditory scene analysis: Listening in complex environments. In *Thinking in sound*, pp. 10–36. Oxford University Press.

Byrne, D. et al. (**1994**). An international comparison of long-term average speech spectra. *J. Acoust. Soc. Am.*, **96**:2108–2120.

Cappé, O. (**1994**). Elimination of the musical noise phenomenon with the Ephraïm and Malah noise suppressor. *IEEE Trans. Speech Audio Processing*, **2**(2):345–349.

CCITT (**1964**). *Recommendation G.227*. Comitée Consultatif Internationale de Télégraphique et Téléphonique (CCITT).

Cohen, J.R. (**1989**). Application of an auditory model to speech recognition. *J. Acoust. Soc. Am.*, **58**:2623–2629.

Dau, T., Kollmeier, B. and Kohlrausch, A. (**1997a**). Modeling auditory processing of amplitude modulation: I. modulation detection and masking with narrowband carriers. *J. Acoust. Soc. Am.*, **102**:2892–2905.

Dau, T., Kollmeier, B. and Kohlrausch, A. (**1997b**). Modeling auditory processing of amplitude modulation: II. spectral and temporal integration. *J. Acoust. Soc. Am.*, **102**:2906–2919.

Dau, T., Püschel, D. and Kohlrausch, A. (**1996a**). A quantitative model of the "effective" signal processing in the auditory system: I. model structure. *J. Acoust. Soc. Am.*, **99**:3615–3622.

Dau, T., Püschel, D. and Kohlrausch, A. (**1996b**). A quantitative model of the "effective" signal processing in the auditory system: II. simulations and measurements. *J. Acoust. Soc. Am.*, **99**:3623–3631.

Davis, K. and Mermelstein, P. (**1980**). Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Processing*, **28**(4):357–366.

Delgutte, B. and Kiang, N.Y.S. (**1984**). Speech coding in the auditory nerve: IV. sounds with consonant-like dynamic characteristics. *J. Acoust. Soc. Am.*, **75**:897–907.

Djupesland, G. and Zwislocki, J.J. (**1972**). Sound pressure distribution in the outer ear. *Scand. Audiol.*, **1**:197–203.

Drullman, R., Festen, J.M. and Plomp, R. (**1994**). Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.*, **95**:1053–1064.

Dupont, S. and Ris, Ch. (**1999**). Assessing local noise level estimation methods. In *Proc. Workshop on robust methods for speech recognition in adverse environments*, pp. 115–118. Tampere, Finland.

Ephraïm, Y. and Malah, M. (**1984**). Speech enhancement using a minimum mean-sqare error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, **32**(6):1109–1121.

ETSI (**1996**). *Recommendation GSM 06.32 (ETS 300 580-6)*. European Telecommunications Standards Institute (ETSI).

Ewert, S. and Dau, T. (**1999**). Frequency selectivity in amplitude-modulation processing. *J. Acoust. Soc. Am.* (submitted).

Ghitza, O. (**1988**). Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment. *J. Phonetics*, **16**:109–123.

Hansen, J.H.L. and Pellom, B. (**1998**). An effective quality evaluation protocol for speech enhancement algorithms. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, pp. 2819–2822. Sydney, Australia.

Hansen, M. and Kollmeier, B. (**1997**). Using a quantitative psychoacoustical signal representation for objective speech quality measurement. In *Proc. Int. Conf. on Acoust., Speech and Signal Processing (ICASSP)*, pp. 1387–1391, IEEE. Munich.

Hansen, M. and Kollmeier, B. (**2000**). Objective modeling of speech quality with a psychoacoustically validated auditory model. *J. Audio Eng. Soc.* (in press).

Hermansky, H. (**1998**). Should recognizers have ears? *Speech Communication*, **25**:3–24.

Hermansky, H. and Morgan, N. (**1994**). RASTA processing of speech. *IEEE Trans. Speech Audio Processing*, **2**(4):578–589.

Hermansky, H. and Pavel, M. (**1998**). RASTA model and forward masking. In *Proc. NATO/ASI Conference on Computational Hearing*, pp. 157–162. Il Ciocco, Italy.

Hirsch, H.G. and Ehrlicher, C. (**1995**). Noise estimation techniques for robust speech recognition. In *Proc. Int. Conf. on Acoust., Speech and Signal Processing (ICASSP)*, pp. 153–156, IEEE.

Holube, I. and Kollmeier, B. (**1996**). Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. *J. Acoust. Soc. Am.*, **100**:1703–1716.

Houtgast, T. (**1989**). Frequency selectivity in amplitude-modulation detection. *J. Acoust. Soc. Am.*, **85**:1676–1680.

Houtgast, T. and Steeneken, H.J.M. (**1985**). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J. Acoust. Soc. Am.*, **77**:1069–1077.

ICRA (**1997**). *ICRA noise signals, Ver. 0.3*. International Collegium of Rehabilitative Audiology. Compact Disc, produced by Widex, Danmark.

ITU (**1996**). *Recommendation ITU-T G.729 Annex B*. International Telecommunication Union.

Jankowski, C.R., Vo, H-D.H. and Lippmann, R.P. (**1995**). A comparison of signal processing front ends for automatic word recognition. *IEEE Trans. Speech Audio Processing*, **3**:286–293.

Kanedera, N., Arai, T., Hermansky, H. and Pavel, M. (**1997**). On the importance of various modulation frequencies for speech recognition. In *Proc. EUROSPEECH*, pp. 1079–1082, ESCA. Rhodes, Greece.

Kasper, K. and Reininger, R. (**1999**). Evaluation of PEMO in robust speech recognition. *J. Acoust. Soc. Am.*, **105**(2):1157.

Kasper, K., Reininger, R. and Wolf, D. (**1997**). Exploiting the potential of auditory preprocessing for robust speech recognition by locally recurrent neural networks. In *Proc. Int. Conf. on Acoust., Speech and Signal Processing (ICASSP)*, pp. 1223–1227. Munich, Germany.

Kasper, K., Reininger, R., Wolf, D. and Wüst, H. (**1995**). A speech recognizer with low complexity based on rnn. In *Neural Networks*

for Signal Processing V, Proc. of the IEEE workshop, pp. 272–281. Cambridge (MA).

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

Kiang, N.Y.S., Watanabe, T., Thomas, E.C. and Clark, L.F. (**1965**). *Discharge patterns of single fibers in the cat's auditory nerve.* M.I.T. Press, Cambridge, MA.

Kingsbury, B., Morgan, N. and Greenberg, S. (**1998**). Robust speech recognition using the modulation spectrogram. *Speech Communication*, **25**(1):117–132.

Kleinschmidt, M., Tchorz, J. and Kollmeier, B. (**2000**). Combining speech enhancement and auditory feature extraction for robust speech recognition. *Speech Communication.* (accepted).

Koenig, R., Dunn, H. and Lacy, L. (**1946**). The sound spectrograph. *J. Acoust. Soc. Am.*, **18**:19–49.

Kohler, K., Lex, G., Pätzold, M., Scheffers, M., Simpson, A. and Thon, W. (**1994**). *Handbuch zur Datenaufnahme und Transliteration in TP14 von VERBMOBIL-3.0.* Technical Report: Verbmobil-Technischer Report.

Kollmeier, B. (**1990**). *Messmethodik, Modellierung und Verbesserung der Verständlichkeit von Sprache.* Habilitationsschrift, Universität Göttingen.

Kollmeier, B. and Koch, R. (**1994**). Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. *J. Acoust. Soc. Am.*, **95**(3):1593–1602.

Langner, G. (**1992**). Periodicity coding in the auditory system. *Hear. Res.*, **60**:115–142.

Langner, G., Sams, M., Heil, P. and Schulze, H. (**1997**). Frequency and periodicity are represented in orthogonal maps in the human auditory cortex: evidence from magnetoencephalography. *J. Comp. Physiol. A*, **181**:665–676.

Langner, G. and Schreiner, C.E. (**1988**). Periodicity coding in the inferior colliculus of the cat. I. neuronal mechanisms. *J. Neurophysiol.*, **60**:1799–1822.

Linhard, K. and Haulick, T. (**1999**). Noise subtraction with parametric recursive gain curves. In *Proc. EUROSPEECH*, Vol. **6**, pp. 2611–2614, ISCA. Budapest, Hungary.

Martin, R. (**1993**). An efficient algorithm to estimate the instantaneous SNR of speech signals. In *Proc. EUROSPEECH*, pp. 1093–1096, ESCA.

Nemer, E., Goubran, R. and Mahmoud, S. (**1988**). SNR estimation of speech signals using subbands and fourth-order statistics. *IEEE Signal Processing Letters*, **6**:1799–1822.

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

Patterson, R.D., Nimmo-Smith, J., Holdsworth, J. and Rice, P. (**1987**). *An efficient auditory filterbank based on the gammatone function.* Technical Report: paper presented at at a meeting of the IOC Speech Group on Auditory Modeling at RSRE, Dec. 14-15.

Püschel, D. (**1988**). *Prinzipien der zeitlichen Analyse beim Hören.* Doctoral thesis, Universität Göttingen.

Quackenbush, S.R., Barnwell, T.P. and Clements, M.A. (**1988**). *Objective measures of speech quality.* Prentice-Hall, NY.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (**1986**). Learning internal representations by error propagation. In D.E. Rumelhart and J.L. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. **1**, pp. 318–362. MIT Press, Cambridge, MA.

Sandhu, S. and Ghitza, O. (**1995**). A comparative study of mel cepstra and EIH for phone classification under adverse conditions. In *Proc.*

Int. Conf. on Acoust., Speech and Signal Processing (ICASSP), pp. 401–405, IEEE.

Schroeder, M.R. (**1968**). Reference signal for signal quality studies. *J. Acoust. Soc. Am.*, **44**:1735–1736.

Schroeder, M.R. and Hall, J.L. (**1974**). A model for mechanical to neural transduction in the auditory receptor. *J. Acoust. Soc. Am.*, **55**:1055–1060.

Seneff, S. (**1988**). A joint synchrony/mean-rate model of auditory speech processing. *J. Phonetics*, **16**:55–76.

Seok, J. W. and Bae, K. S. (**1997**). Speech enhancement with reduction of noise components in the wavelet domain. In *Proc. Int. Conf. on Acoust., Speech and Signal Processing (ICASSP)*, pp. 1323–1326, IEEE. Munich.

Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonsky, J. and Ekelid, M. (**1993**). Speech recognition with primarily temporal cues. *Science*, **270**:303–304.

Siohan, O., Chesta, C. and Lee, C.H. (**1999**). Hidden Markov Model adaptation using maximum a posteriori linear regression. In *Proc. Workshop on robust methods for speech recognition in adverse environments*, pp. 147–150. Tampere, Finland.

Smith, R.L. (**1979**). Adaptation, saturation, and physiological masking in single auditory-nerve fibers. *J. Acoust. Soc. Am.*, **65**:166–178.

SNNS (**1995**). *http://www-ra.informatik.uni-tuebingen.de/SNNS/.*

Soede, W., Berkhout, A. J. and Bilsen, F. A. (**1993**). Development of a directional hearing instrument based on array technology. *J. Acoust. Soc. Am.*, **94**(1):785–798.

Strope, B. and Alwan, A. (**1997**). A model of dynamic auditory perception and its application to robust word recognition. *IEEE Trans. Speech Audio Processing*, **5**(5):451–464.

Strube, H.-W. and Wilmers, H. (**1999**). Noise reduction for speech signals by operations on the modulation frequency spectrum. *J. Acoust. Soc. Am.*, **105**(2):1092.

Tchorz, J., Kasper, K., Reininger, H. and Kollmeier, B. (**1997**). On the interplay between auditory-based features and locally recurrent neural networks. In *Proc. EUROSPEECH '97*, pp. 2075–2078, ESCA. Rhodes, Greece.

Tchorz, J. and Kollmeier, B. (**1999a**). Automatic classification of the acoustical situation using amplitude modulation spectrograms. *J. Acoust. Soc. Am.*, **105**(2):1157.

Tchorz, J. and Kollmeier, B. (**1999b**). A model of the auditory periphery as front end for automatic speech recognition. *J. Acoust. Soc. Am.*, **106**(4):2040–2050.

Tchorz, J. and Kollmeier, B. (**1999c**). Speech detection and SNR prediction basing on amplitude modulation pattern recognition. In *Proc. EUROSPEECH*, pp. 2399 – 2404, ISCA. Budapest, Hungary.

Tchorz, J. and Kollmeier, B. (**2000**). Estimation of the signal-to-noise ratio with amplitude modulation spectrograms. *Speech Communication.* (submitted).

Unoki, M. and Akagi, M. (**1999**). A method of signal extraction from noisy signal based on auditory scene analysis. *Speech Communication*, **27**:261–279.

Wesselkamp, M. (**1994**). *Messung und Modellierung der Verständlichkeit von Sprache.* Doctoral thesis, Universität Göttingen.

Wittkop, T., Albani, S., Hohmann, V., Peissig, J., Woods, W. S. and Kollmeier, B. (**1997**). Speech processing for hearing aids: Noise reduction motivated by models of binaural interaction. *Acustica united with Acta Acustica*, **83**:684–699.

Yang, D., Meyer, G. F. and Ainsworth, W. A. (**1999**). A neural model for auditory scene analysis. *J. Acoust. Soc. Am.*, **105**(2):1092.

Zell, A. (**1994**). *Simulation Neuronaler Netze*. Addison-Wesley, Bonn; Paris; Reading, Mass.

Zell, A., Mamier, G., Vogt, M. and Mache, N. (**1995**). Der Stuttgarter Neuronale Netze Simulator. In G. Dorffner, K. Möller, G. Paaßand S. Vogel (Eds.), *Konnektionismus und Neuronale Netze*, Vol. **272**, pp. 335–350. GMD-Studien.