SNR estimation based on amplitude modulation analysis with applications to noise suppression

Article *in* IEEE Transactions on Speech and Audio Processing · June 2003

DOI: 10.1109/TSA.2003.811542 · Source: IEEE Xplore

2 authors, including:

Jürgen Tchorz
Technische Hochschule Lübeck
39 PUBLICATIONS **386** CITATIONS

SNR Estimation Based on Amplitude Modulation Analysis With Applications to Noise Suppression

Jürgen Tchorz and Birger Kollmeier

***Abstract*—A single-microphone noise suppression algorithm is described that is based on a novel approach for the estimation of the signal-to-noise ratio (SNR) in different frequency channels: 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 nonspeech background noise. For noise suppression, the frequency bands are attenuated according to the estimated present SNR using a Wiener filter approach. Objective speech quality measures, informal listening tests, and the results of automatic speech recognition experiments indicate a substantial benefit from AMS-based noise suppression, in comparison to unprocessed noisy speech.**

***Index Terms*—Amplitude modulation processing, noise suppression, SNR estimation.**

I. 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 [1]. In prototype hearing instruments, binaural algorithms exploit phase and level differences or correlations between the two sides of the head for spatial filtering [2].

Single-microphone noise suppression algorithms, in contrast, try to separate speech from noise when only one microphone is available, i.e., without spatial information. Separation between speech and noise then requires a noise estimate. This can be obtained by detecting speech pauses. In speech pauses, the spectrum of the signal is measured and provides an estimate of the present noise floor. Spectral subtraction [3] and related schemes can then be used for noise suppression. There are two main prerequisites for noise estimation in speech pauses: i) the speech pause detector has to work properly (if speech parts are mistakenly labeled as "speech pause," the precision of the noise estimate decreases, and hence the quality of the processed signal can be severely degraded) and ii) the background noise is assumed to be relatively stationary between speech pauses, as the noise estimate cannot be updated while speech is active. In practice, however, these two prerequisites are often not met.

Other approaches have been described which do not require explicit speech pause detection for noise level (or SNR) estimation. Hirsch and Ehrlicher [4] proposed an algorithm which is based on the statistical analysis of the spectral energy envelope. Histograms of energy values are built for different frequency bands on signal segments of several hundred milliseconds. These histograms contain basically two modes: i) a low energy mode related to the contribution of (possibly noisy) speech pause frames) and ii) a high energy mode related to the contribution of (possibly noisy) speech frames. A noise level estimate is computed from these two histograms. An adaptive speech enhancement using SNR estimates based on this approach was proposed by [5]. The authors reported a noticeable suppression of the perceived noise with sometimes disturbing residual noise in informal listening experiments.

Martin [6] proposed a noise level estimator which is based on automatically tracking the low energy envelope of the signal within frequency bands. The average value of these minima is used as an estimate of the noise floor in the respective frequency band. The approach is based on the assumption that the noise is relatively stationary. In clean speech, it tends to overestimate the noise level by tracking soft speech portions. A detailed review on these two methods and related schemes, and quantitative comparisons of their performance can be found in [7].

The SNR estimation algorithm proposed in this paper also does not require explicit detection of speech pauses. 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 [8]) 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

Manuscript received January 16, 2001; revised October 15, 2002. This work was supported in part by the European Union (TIDE/SPACE) and BMBF. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hynek Hermansky.

The authors are with AG Medizinische Physik, Universität Oldenburg, 26111 Oldenburg, Germany (e-mail: juergen.tchorz@phonak.ch; birger.kollmeier@uni-oldenburg.de; http://medi.uni-oldenburg.de).

Digital Object Identifier 10.1109/TSA.2003.811542

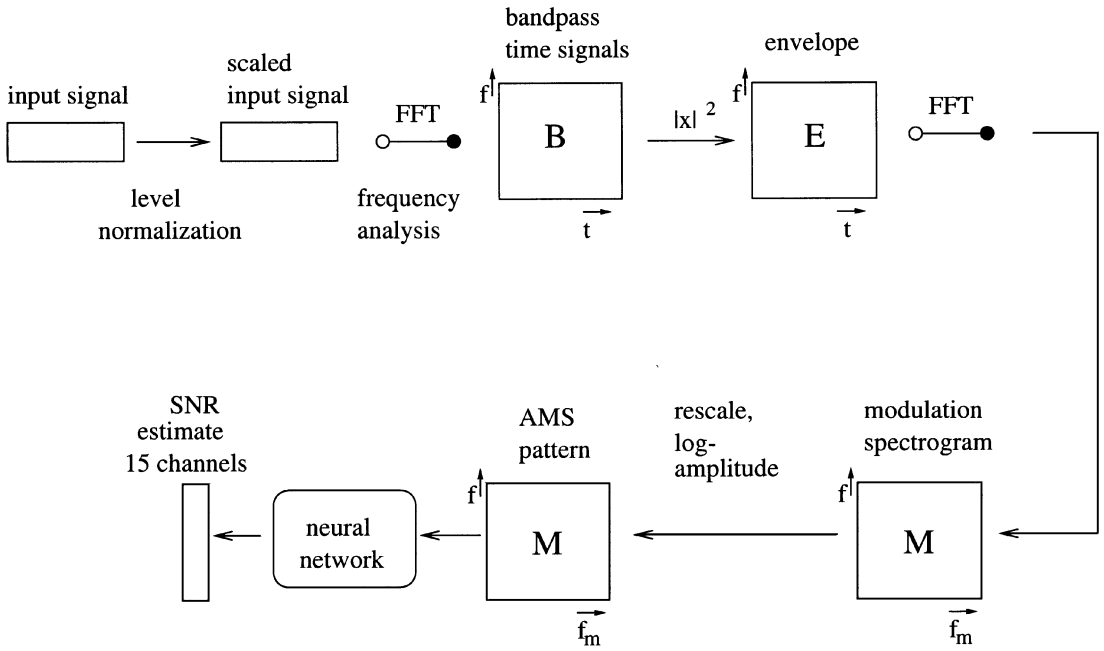

Fig. 1. Processing stages of AMS-based SNR estimation.

is different from the representation of most types of noise. The differences in the respective representations can be exploited by a neural network pattern recognition.

In Section II of this paper, the SNR estimation approach based on AMS patterns is described, and quantitative estimation results are presented. A comparison with SNR estimation based on voice activity detection is outlined in Section III. The noise suppression stage is described in Section IV.

II. 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. 1.

A. Feature Extraction

For SNR estimation, the input waveform is transformed into so-called amplitude modulation spectrograms (AMS), see [8]. These patterns are motivated from neurophysiological findings on amplitude modulation processing in higher stages of the auditory system in mammals. Langner and Schreiner [9], 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.* [10] observed periodotopical gradients in the human auditory cortex by means of magnetoenzephalography (MEG). Psychoacoustical evidence for modulation analysis in each frequency band is provided by Dau *et al.* [11], [12]

In the field of digital signal processing, Kollmeier and Koch [8] 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. 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 [13] and speech enhancement [14]. The application of AMS patterns on broadband SNR estimation is described in detail in [15].

First, the input signal which was digitized with 16 kHz sampling rate is long-term level adjusted. This 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. To avoid divisions by zero, the normalization function is limited by a lower threshold.

In a following processing step, the level-adjusted signal is subdivided into overlapping segments of 4.0 ms duration (64 samples) with a progression of 0.25 ms (four samples) 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 an 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. The frequency axis is transformed to a Bark scale with 15 channels by adding the magnitudes of neighboring FFT sub-bands, with center frequencies from 100 to 7300 Hz. Their respective envelopes are extracted by computing the square of the absolute values.

This envelope signal is again segmented into overlapping segments of 128 samples (32 ms) 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. The modulation frequency spectrum is scaled logarithmically, which is motivated by psychoacoustical findings on the shape of auditory modulation filters [16]. The modulation frequency range from 0 to 2000 Hz is restricted to the range between 50–400 Hz and 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 [10].

Very low modulation frequencies from articulator movement, which are characteristic for speech and which play an important role for speech intelligibility are 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*, but on the *detection* of speech and noise, and SNR estimation in short analysis frames. These two tasks must not be confused. Daily experience shows that short segments of speech which are too short to analyze low modulation frequencies around 4 Hz can be sufficient to identify them as "speech," without understanding the meaning (e.g., in a canteen situation).

The AMS representation is restricted to a 15×15 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. Bright and dark areas indicate high and low energies, respectively.

The AMS pattern on the top panel was generated from a voiced speech portion, uttered by a male speaker. The periodicity at the fundamental frequency (approximately 110 Hz) is represented in each center frequency band, i.e., the vertical bar which has the highest intensity (is brightest) at about 110 Hz modulation frequency. The second and third harmonics are represented by vertical bars centered at 220 and 330 Hz modulation frequency, respectively. 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 AMS pattern on the bottom panel was generated from speech simulating noise [17], 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 systematic structure across modulation frequencies such as harmonic peaks, and no obvious similarities between modulation spectra in different frequency channels, as in the upper panel.

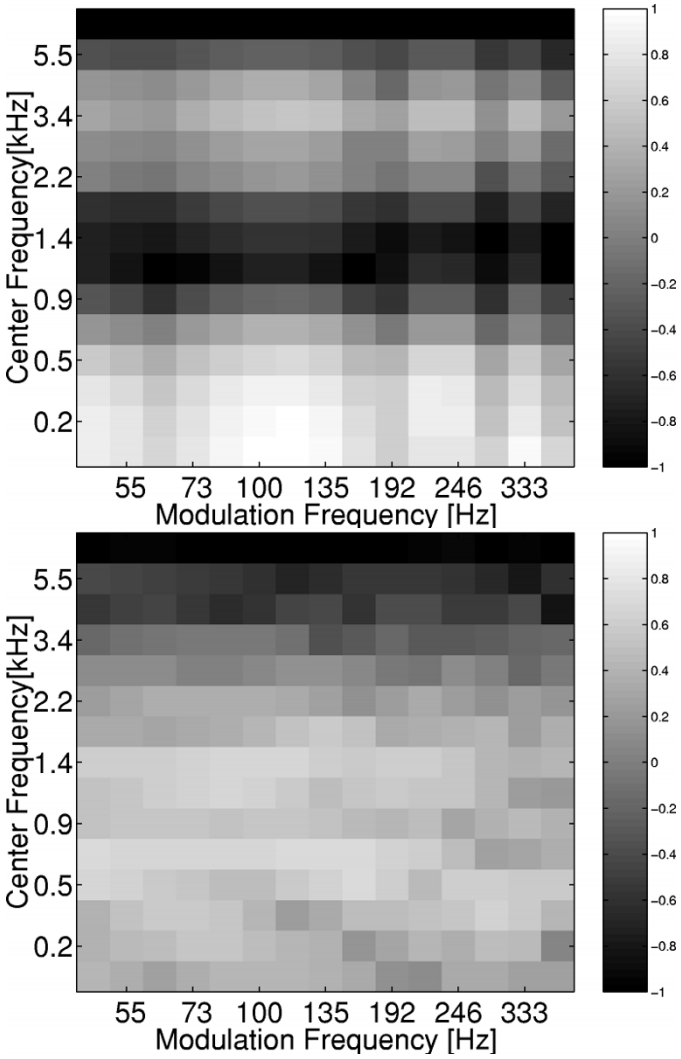

Fig. 2. AMS patterns generated from a voiced speech segment (top), and from speech simulating noise (bottom). Each AMS pattern represents a 32 ms portion of the input signal. Bright and dark areas indicate high and low energies, respectively.

B. 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. Artificial neural networks are widely used in a range of different pattern recognition tasks [18]. For SNR estimation based on AMS patterns, a standard feed-forward neural network is applied (SNNS, described in [19]). It consists of an input layer with 225 neurons (15×15, i.e., the resolution of AMS patterns, that are directly fed into the network), a hidden 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 narrowband 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.

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 [20]. 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 and serve as SNR estimates for the respective frequency channels in the present analysis frame.

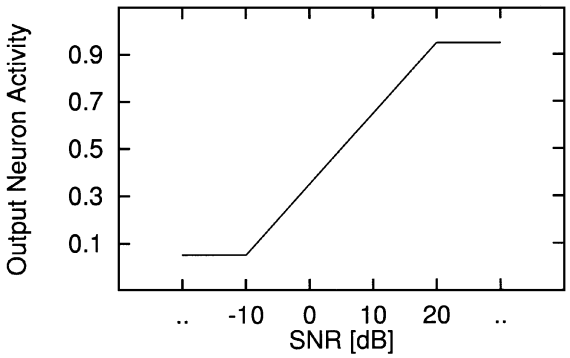

Fig. 3. Transformation function between SNR and output neuron activity for training and testing.

C. Speech and Noise 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 AMS patterns. The long-term, broadband 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 [21] and contained 2110 German sentences from 190 male and 210 female talkers. Forty-one 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 recordings for testing were not included in the training data. The network was trained with 100 cycles. The noise recordings for training and testing include a wide range of natural noise types, mostly traffic (inside and outside cars, trains, planes, boats, helicopters, etc.), machinery (engines, factories, construction sites, household, etc.) or social (restaurant, crowd in a sports stadium, school yard, etc.). Thus, many noisy situations of everyday life are covered, and the algorithm is not tuned to perform well in a specific situation (e.g., in a car for mobile communication applications). No artificially generated noise types (sine waves etc.) were used.

An example for the estimation of narrowband SNRs of noisy speech is illustrated in Fig. 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.

In signal portions with poor SNRs (i.e., in speech pauses or soft consonants), the estimator tends to overestimate the SNR, whereas in portions with very high SNRs, it is rather underestimated. This was also found for AMS-based broadband SNR estimation [15].

A quantitative measure of the estimation accuracy is obtained by computing the mean deviation D between the actual SNR a_i

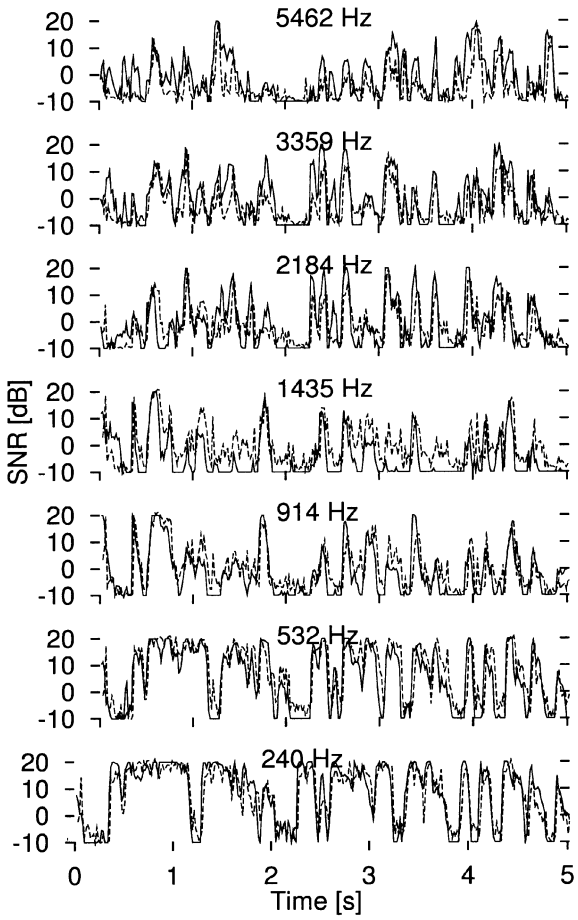

Fig. 4. Example for narrowband SNR estimation. Plotted are the measured (solid) and the estimated (dotted) SNRs as function of time for 7 out of 15 frequency channels.

and the estimated SNR e_i over N processed AMS patterns (with index i)

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

The mean estimation deviation D was calculated for all AMS analysis frames generated from the test data described in Section II-C, for all 15 frequency channels independently. The results are plotted in Fig. 5 (solid line). It can be seen that the estimation accuracy in the low- and mid frequency channels

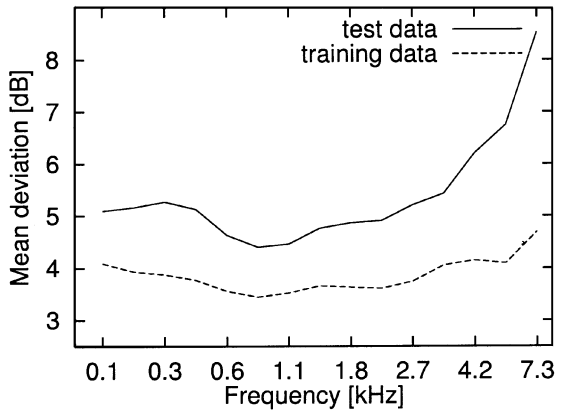

Fig. 5. Mean deviation between the estimated SNR and the "true" SNR which was measured prior to adding speech and noise as a function of the frequency channel for the test data (solid) and the training data (dotted).

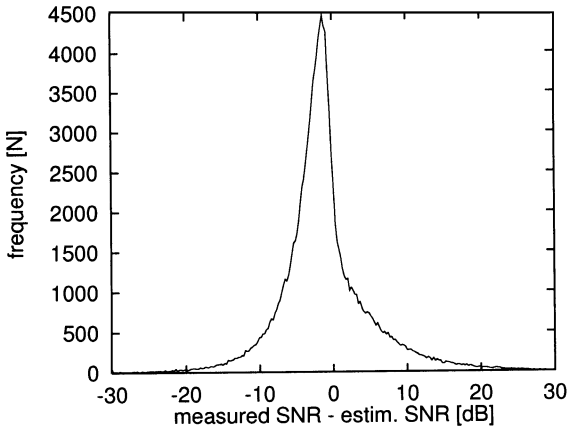

Fig. 6. Histogram of the differences $a_i - e_i$ between measured and estimated SNRs for the test data in the seventh frequency channel ($f_c = 1.1$ kHz).

is better compared to the high frequency region (which is also the case for the example plotted in Fig. 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 overtrained 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. 6. The maximum frequency is at about -1.3 dB, i.e., there is a slight estimation bias in this particular frequency channel toward worse SNRs than the actual ones. This bias varies from channel to channel, and there is no systematic error across all channels.

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.

A study on AMS-based broadband SNR estimation [15] quantitatively examined the most important cues that are necessary for reliable SNR estimation. It was shown that harmonicity is an important cue for analysis frames to be classified as "speech-like." To determine the influence of harmonicity on SNR estimation, 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 $f_{\text{shift}} = f + \text{rand}[-x \cdots x]$, where f is the frequency of the respective harmonic, and x is a frequency between 0 and 150 Hz. The highest output neuron activities (0.79) was reached with $f_{\text{shift}} = 0$, i.e., without disturbing the harmonic structure. With decreasing harmonicity, the output neuron activity decreased and indicated more and more noise-like signals.

The influence of 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. With a synthetic vowel as input, clearly higher average output neuron activities were reached (up to 0.95), compared to harmonic tone complexes. The highest output neuron activities were reached with fundamental frequencies between about 100 and 300 Hz, which is roughly the range of fundamental frequencies in human voices. The formant structure which is not given in pure tone complexes provides additional information and evidence for an analysis frame to be classified as "speech."

The performance of the algorithm in voiced and unvoiced speech was evaluated in an additional experiment. The average output neuron activity was measured for voiced and unvoiced phonemes extracted from 1350 phonetically labeled sentences from the PhonDat database [21]. For voiced phonemes ("n," "a," "i," "m," "l"), the average output neuron activity was 0.9. For unvoiced phonemes ("t," "s," "d," "f," "r," "k"), the average output neuron activity was 0.65, which was clearly higher than for most noise types that were tested (average: 0.25). Here, the level of the unvoiced phonemes after the long-term level normalization process (Section II) is softer, compared to the level of relatively stationary noise after level normalization. This difference is exploited by the neural network, as level is an important cue for SNR estimation [15].

Another set of experiments was conducted with reduced AMS patterns, i.e., only spectral or only temporal information was provided to the neural network, respectively. With these reduced patterns, SNR estimation was possible to some extend, but less accurate, compared to the full spectro-temporal joint representation in AMS patterns. When only temporal information was given, i.e., the modulation spectrum without any center frequency information, the mean deviation from the actual SNR was 6.6 dB. With a conventional spectrogram, the deviation was 7.6 dB. With the full AMS joint representation, 5.2 dB were reached. Thus, the conventional spectrogram representation was the least suited one for SNR estimation in these experiments, and temporal cues appeared to be more important.

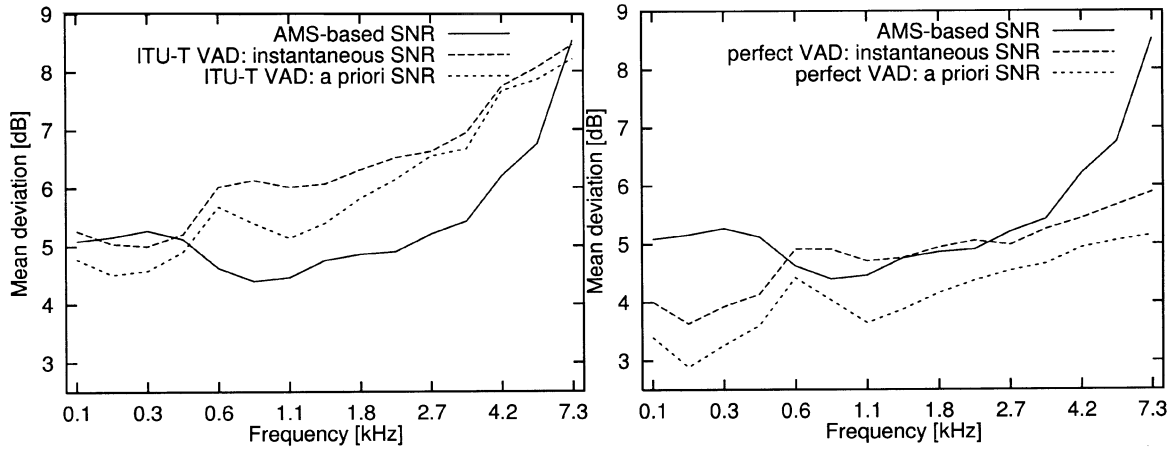

Fig. 7. Comparison between AMS-based (solid) and VAD-based (dotted) SNR estimation in 15 frequency channels. The left panel shows the results with a VAD standardized by ITU-T, on the right panel a "perfect" VAD was used.

Harmonicity appeared to be the most important cue for SNR estimation. In the full AMS pattern, however, spectral and level information also contribute to accuracy in SNR estimation.

III. Comparison With VAD-Based SNR Estimation

In common single-microphone 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. A VAD-based SNR estimation was chosen as reference condition, as direct SNR estimation algorithms which were described in the literature [4], [6], [7] typically require much longer analysis frames (at least 250 ms, better performance was reported using 500 ms and more) than the AMS-based approach, which analyzes 32 ms-frames.

Ris and Dupont [7] compared the accuracy of different direct SNR estimation algorithms. As reference, they used a speech/silence detector based on a forced HMM speech/silence alignment of a clean version of the speech data. The reference condition was shown to provide the most accurate SNR estimations. Thus, for comparison to the AMS-based approach, an SNR estimation scheme based on a high quality VAD was chosen, namely a VAD standardized by ITU-T [22]. 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 [23] was calculated for each spectral component

$$SNR_{[inst]} = 10\log(\gamma_k - 1) \tag{2}$$

with

$$\gamma_k = \frac{|R_k(l)|^2}{\lambda_d(k)} \tag{3}$$

where $R_k(l)$ is the modulus of the signal l 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 [3]. Several methods have been proposed to reduce musical noise. An approach which is widely used was introduced by Ephraim and Malah [23]. 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 Ephraim and Malah [23]. 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 II-C 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. 7 (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 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).

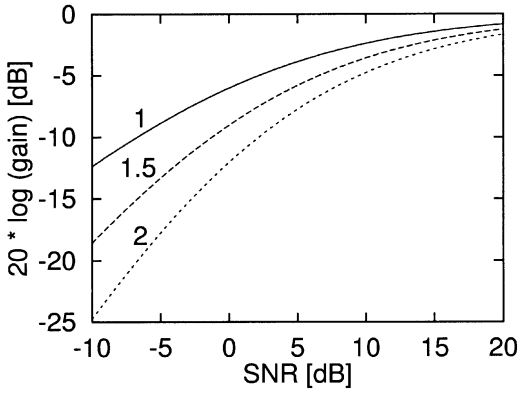

Fig. 8. Gain function for three different exponents x [see (4)].

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.

IV. 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

$$g_k = \left(\frac{SNR_k}{SNR_k + 1} \right)^x \tag{4}$$

where k denotes the frequency channel, SNR denotes 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. 8. 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. 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 linearly 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 smooths subsequent SNR estimates. 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 result in only moderate suppression of the noise (with a clearly audible noise floor).

Different recordings of processed 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 smooths 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.

Objective speech quality evaluations [24] with three different objective speech quality measures were conducted with the proposed noise suppression scheme. The measured improvement in speech quality was dependent on the type of background noise. A clear benefit was indicated in white Gaussian noise, whereas almost no differences between unprocessed and processed signals were measured in canteen babble noise.

The proposed noise suppression scheme was also evaluated in isolated-digit recognition experiments in different types of noise [25]. For comparison, recognition rates were measured with a standard noise suppression scheme consisting of a VAD-based SNR estimation and Spectral Subtraction including residual noise reduction. In all tested types of noise (stationary white noise, amplitude modulated speech simulating noise, and fast fluctuating printing room noise), the AMS-based approach allowed for higher recognition rates, compared to the VAD-based approach. This was particularly the case in fast fluctuating noise. With VAD-based noise suppression, an update of the noise estimate is not possible while speech is active, and the processed signal which is the input to the recognizer is distorted.

V. 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;
- single-microphone noise suppression based on 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 indicate that the analysis and representation of amplitude modulations play an important role in our auditory system. Technical sound signal processing, on the other hand, is commonly dominated by the analysis of *spectral* information, rather than modulation information. Spectral analysis in speech processing has a long history back to the invention of the spectrograph [26], 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 [8], [14] and automatic speech recognition [27]–[29]. 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.* [30] found that modulation frequencies up to 8 Hz are the most important ones in for speech intelligibility. Arai *et al.* [31] measured the intelligibility of syllables with temporally filtered cepstral trajectories. Their results suggest that intelligibility is not severely impaired as long as the filtered spectral components have a rate of change between 1 and 16 Hz. Shannon *et al*. [32] 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 broadband SNR estimation [15] it was shown that harmonicity appears to be an 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 representation 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 not linked with subjective listening tests, where the correlation between objective measures and subjective scores can be 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 Ephraim and Malah. 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%.

References

[1] W. Soede, A. J. Berkhout, and F. A. Bilsen, "Development of a directional hearing instrument based on array technology," *J. Acoust. Soc. Amer.*, vol. 94, no. 1, pp. 785–798, 1993.

[2] T. Wittkop, V. Hohmann, and B. Kollmeier, "Noise reduction strategies employing interaural parameters," *J. Acoust. Soc. Amer.*, vol. 105, no. 2, pp. 1211–1211, 1999.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[4] H. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Proc. Int. Conf. on Acoust., Speech and Signal Processing (ICASSP)*, 1995, pp. 153–156.

[5] C. Avendano, H. Hermansky, M. Vis, and A. Bayya, "Adaptive speech enhancement based on frequency-specific SNR estimation," in *IVTTA*. New York: IEEE, 1996, pp. 65–68.

[6] R. Martin, "An efficient algorithm to estimate the instantaneous SNR of speech signals," in *Proc. EUROSPEECH*, 1993, pp. 1093–1096.

[7] C. Ris and S. Dupont, "Assessing local noise level estimation methods: Applications to noise robust ASR," *Speech Commun.*, vol. 34, pp. 141–158, 2001.

[8] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *J. Acoust. Soc. Amer.*, vol. 95, no. 3, pp. 1593–1602, 1994.

[9] G. Langner and C. Schreiner, "Periodicity coding in the inferior colliculus of the cat. I. neuronal mechanisms," *J. Neurophysiol.*, vol. 60, pp. 1799–1822, 1988.

[10] G. Langner, M. Sams, P. Heil, and H. Schulze, "Frequency and periodicity are represented in orthogonal maps in the human auditory cortex: Evidence from magnetoencephalography," *J. Comp. Physiol. A*, vol. 181, pp. 665–676, 1997.
[11] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation: I. Modulation detection and masking with narrowband carriers," *J. Acoust. Soc. Amer.*, vol. 102, pp. 2892–2905, 1997.
[12] ——, "Modeling auditory processing of amplitude modulation: II. spectral and temporal integration," *J. Acoust. Soc. Amer.*, vol. 102, pp. 2906–2919, 1997.
[13] D. Yang, G. F. Meyer, and W. A. Ainsworth, "A neural model for auditory scene analysis," *J. Acoust. Soc. Amer.*, vol. 105, no. 2, p. 1092, 1999.
[14] H.-W. Strube and H. Wilmers, "Noise reduction for speech signals by operations on the modulation frequency spectrum," *J. Acoust. Soc. Amer.*, vol. 105, no. 2, p. 1092, 1999.
[15] J. Tchorz and B. Kollmeier, "Estimation of the signal-to-noise ratio with amplitude modulation spectrograms," Speech Commun., 2001, to be published.
[16] S. Ewert and T. Dau, "Frequency selectivity in amplitude-modulation processing," J. Acoust. Soc. Amer., 1999, submitted for publication.
[17] "Recommendation G.227," Comitée Consultatif Internationale de Télégraphique et Téléphonique CCITT, 1964.
[18] C. M. Bishop, *Neural Networks for Pattern Recognition*. New York: Oxford Univ. Press, 1995.
[19] A. Zell, *Simulation Neuronaler Netze*. Reading, MA: Addison-Wesley, 1994.
[20] D. Rumelhart, G. Hinton, and R. Williams, "Learning Internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, D. Rumelhart and J. McClell, Eds. Cambridge, MA: MIT Press, 1986, vol. 1, pp. 318–362.
[21] K. Kohler, G. Lex, M. Pätzold, M. Scheffers, A. Simpson, and W. Thon, "Handbuch zur Datenaufnahme und Transliteration in TP14 von VERBMOBIL-3.0," Verbmobil-Technischer, 1994.
[22] "Recommend. ITU-T G.729 Annex B," Int. Telecommun. Union, 1996.
[23] Y. Ephraïm and M. Malah, "Speech enhancement using a minimum mean-sqare error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, 1984.
[24] J. Tchorz, "Auditory-Based Signal Processing for Noise Suppression and Robust Speech Recognition,", BIS-Verlag, Oldenburg, Germany, 2000.
[25] J. Tchorz, M. Kleinschmidt, and B. Kollmeier, "Noise suppression based on neurophysiologically-motivated SNR estimation for robust speech recognition," in *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. Cambridge, MA: MIT Press, 2001, pp. 821–827.
[26] R. Koenig, H. Dunn, and L. Lacy, "The sound spectrograph," *J. Acoust. Soc. Amer.*, vol. 18, pp. 19–49, 1946.
[27] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 578–589, Oct. 1994.
[28] B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Commun.*, vol. 25, no. 1, pp. 117–132, 1998.
[29] J. Tchorz and B. Kollmeier, "A model of the auditory periphery as front end for automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 106, no. 4, pp. 2040–2050, 1999.
[30] R. Drullman, J. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, pp. 1053–1064, 1994.
[31] T. Arai, M. Pavel, H. Hermansky, and A. C. , "Syllable intelligibility for temporally filtered LPC cepstral trajectories," *J. Acoust. Soc. Amer.*, vol. 105, no. 5, pp. 2783–2791, 1999.
[32] R. Shannon, F.-G. Zeng, V. Kamath, J. Wygonsky, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, pp. 303–304, 1993.

Jürgen Tchorz studied physics in Oldenburg, Germany and Galway, Ireland. The work presented in this paper is part of his Ph.D. research conducted at the Universität Oldenburg.

His main research interests are auditory-based strategies for feature extraction in automatic speech recognition, and fast signal-to-noise ratio estimation for speech processing applications. Since 2001, he has been with a Swiss hearing aid manufacturer.

Birger Kollmeier received the Ph.D. degree in physics and the Ph.D. degree in medicine in Göttingen, Germany.

He was Assistant Professor (1986–1992) and, subsequent to his "Habilitation," Associate Professor (1992–1993) at the Drittes Physikalisches Institut, Göttingen. Since 1993, he has been a Full Professor in experimental and applied physics at the Universität Oldenburg, Head of the Medical Physics Group, Scientific Director of the Hörzentrum Oldenburg. Since 2000, he has been Chairman of the European Graduate School "Neurosensory Science and Systems," Speaker of the National Center of Excellence in biomedical engineering "hearing aid system technology (HörTech)." He supervised 29 Ph.D. dissertations and authored or co-authored more than 100 scientific papers and five books in various areas of hearing research, speech processing, auditory neuroscience, and audiology.

Dr. Kollmeier was awarded several fellowships and scientific prizes, including the Lothar-Cremer-Preis of the German Acoustical Society and the Alcatel-SEL Research Prize for technical communication. He is president of the German Audiological Society.