7. References

/1/ H. Hermansky, N. Morgan, A. Bayya, P. Kohn: Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA_PLP), Eurospeech, 1991, pp. 1367-1370

/2/ H.G. Hirsch, P. Meyer, H.W. Rühl: Improved speech recognition using high-pass filtering of subband envelopes, Eurospeech, 1991, pp. 413-416

/3/ S.F. Boll: Suppression of acoustic noise in speech using spectral subtraction, IEEE ASSP-28, No.2, 1979, pp.113-120

/4/ P. Vary: Noise suppression by spectral magnitude estimation- Mechanism and theoretical limits, Signal Processing, 1985, pp. 387-400

/5/ P. Lockwood, J. Boudy: Experiments with a nonlinear spectral subtractor, Hidden Markov Models and the projection, for robust speech recognition in cars, Speech Communication 11, 1992, pp. 215-228

/6/ H.G. Hirsch, H.W. Rühl: Automatic speech recognition in a noisy environment, Eurospeech, 1989, pp. 652-655

6. Conclusions

A method is presented in this report to estimate the noise spectrum of speech utterances which are disturbed by additive noise. One advantage is that no speech pause detection is required.

The processing is based on a calculation of the distribution density function of spectral magnitude values in a subband. The histogram for one subband is calculated for a past segment with a defined duration. Good results were obtained for a segment of 0,5 s. In this case the noise spectrum can also follow a slowly changing noise.

Two applications of this technique are described in this report. The first one is an estimation of the actual SNR of a speech segment. Good results were obtained for a wide range of SNRs and for different noise signals. The second one is the use for an enhancement of noisy speech. The enhancement techniques attempted were a new form of spectral subtraction, and a modified high-pass filtering of the spectral envelopes in subbands.

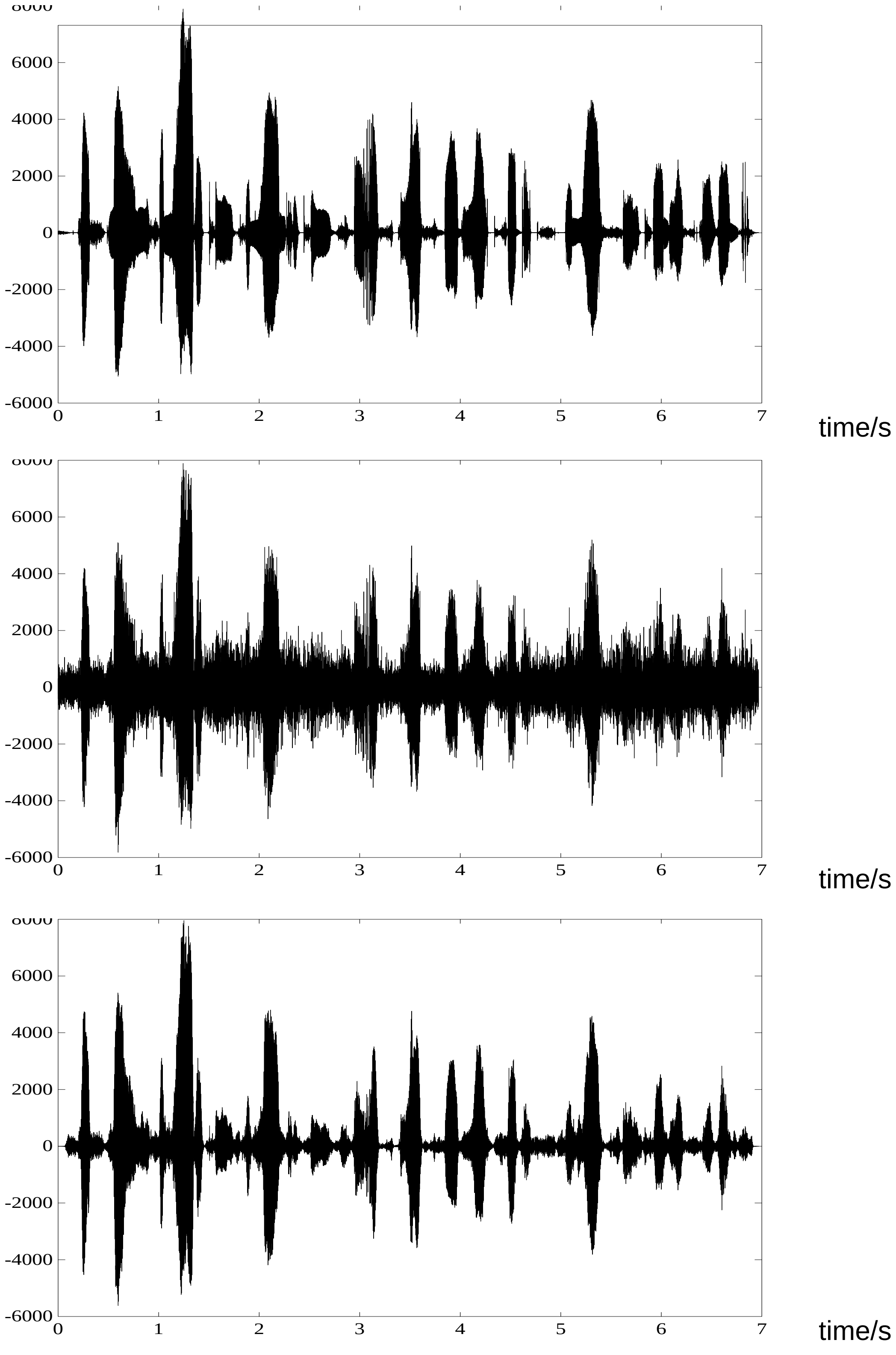

Figure 5.5: Time signals of clean and noisy speech and after processing with the modified high-pass filtering

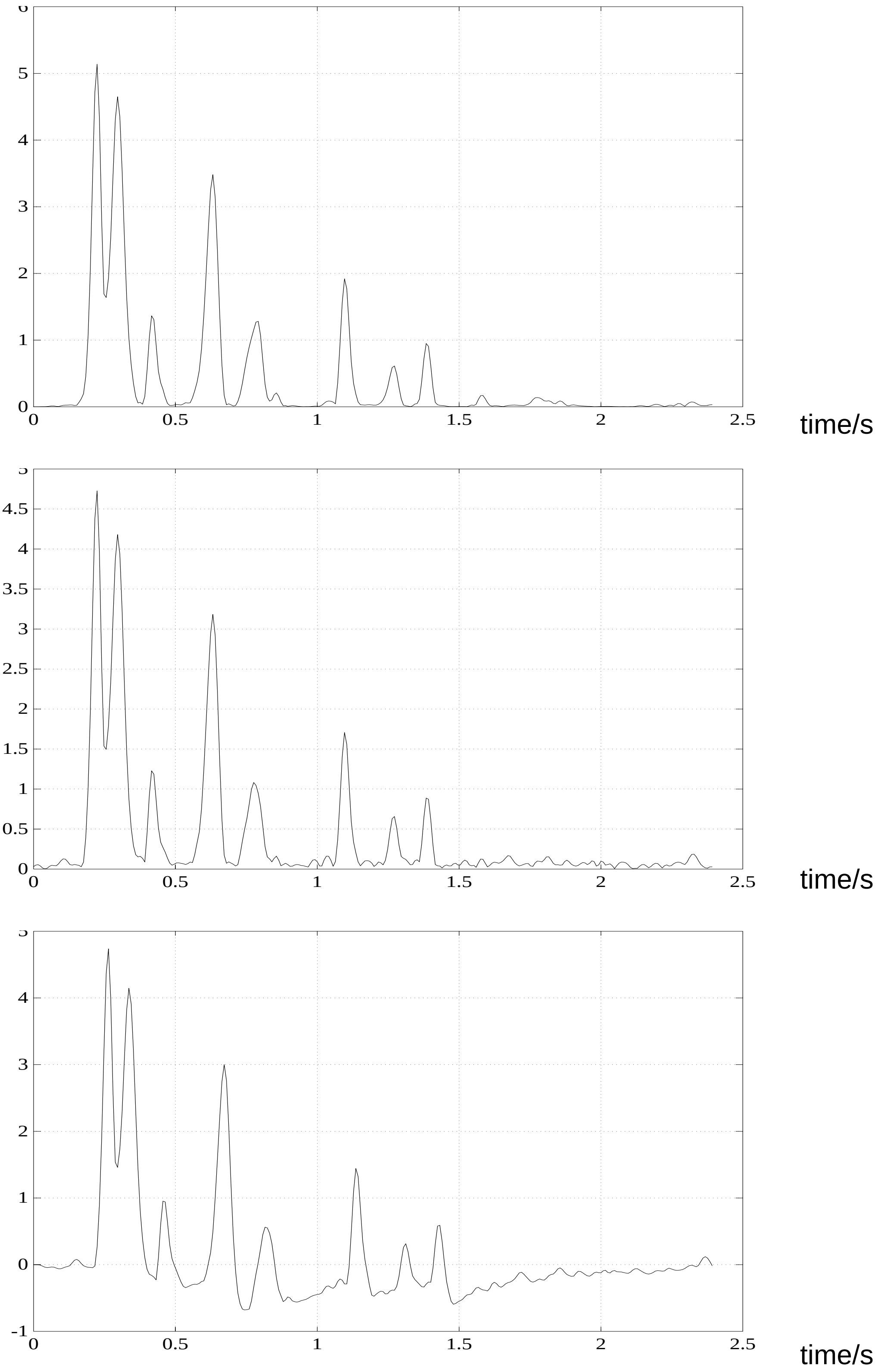

Figure 5.4: Spectral envelope of clean and noisy speech and after modified high-pass filtering

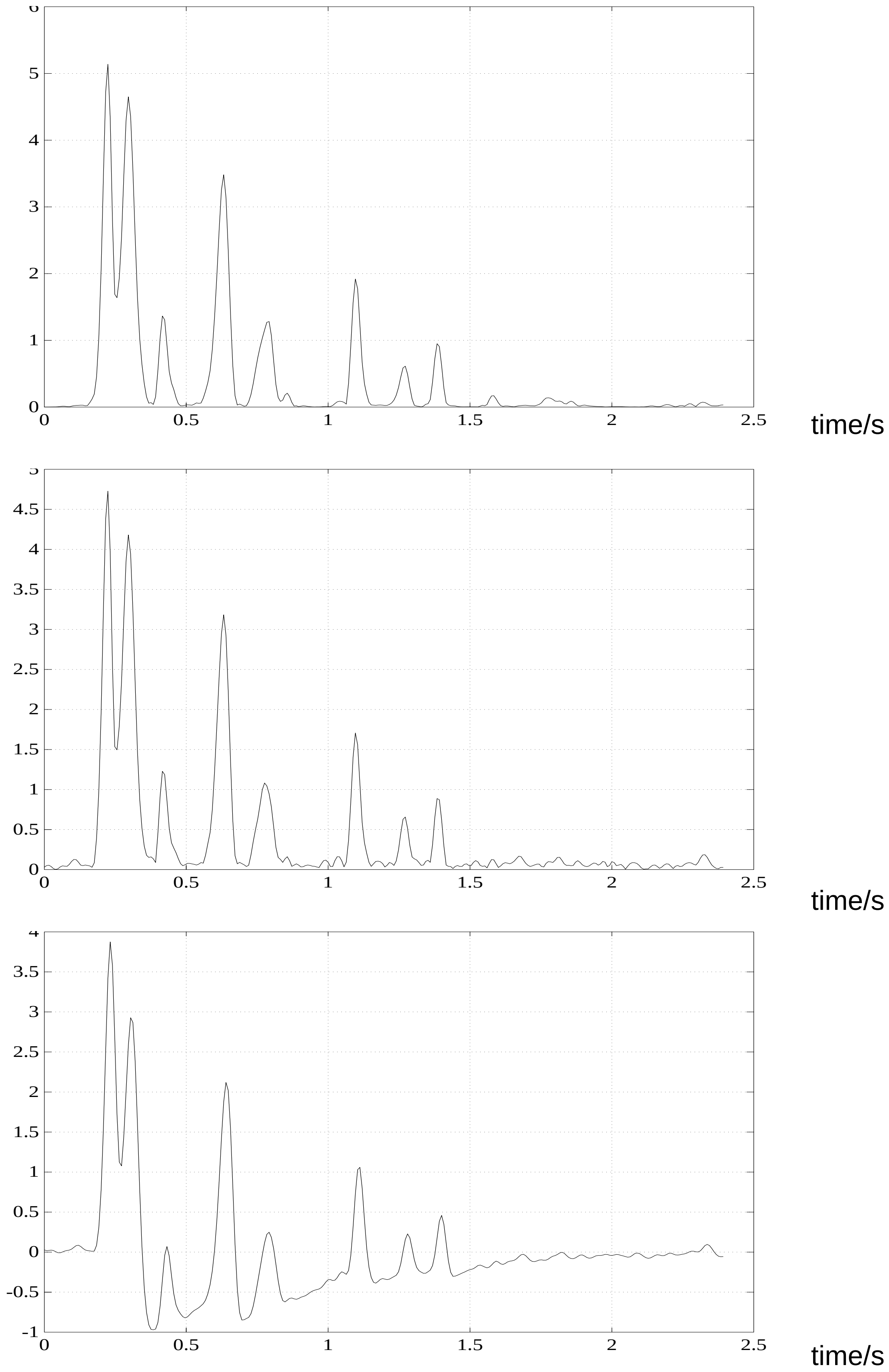

Figure 5.2: Spectral envelope of clean and noisy speech and after RASTA high-pass filtering

The second speech enhancement algorithm is based on a high-pass filtering of the spectral magnitude contours in subbands. The known filter function from /1/ is used in each subband. Applying this filtering directly to a spectral trajectory in one subband, many negative values occur as shown in figure 5.2.

The envelope of clean speech, of noisy speech and the filtered envelope in a subband with a centre frequency of 500 Hz are shown. Setting all negative values to zero the noise is considerably reduced but also certain parts of the speech are suppressed.

Because of this the filter structure is modified as shown in figure 5.3.

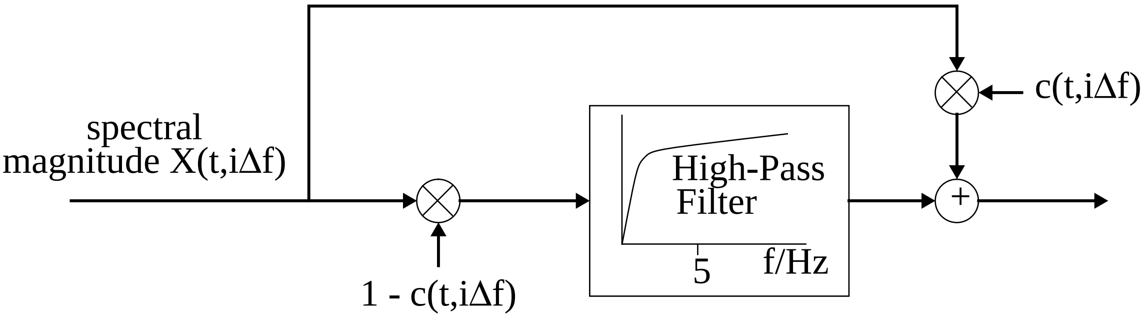

Figure 5.3: Modified high-pass filtering

The attenuation of the DC-component can be varied with this filter scheme with the factor c. In principal the attenuation should be high for noise segments and should be made dependent on the actual SNR for a speech segment. The weighting function 1 - N/X already used for spectral subtraction, is applied according to the additional weighting factor c.

The spectral trajectories of figure 5.2 are shown again in figure 5.4 but this time the processing is done with the modified filter structure.

From this experiment, it appears that the suppression of speech segments is much less of a problem than in the case of a static high-pass filter. The result of processing a whole noisy speech sentence is shown in figure 5.5.

The time signals of the clean, the noisy and the processed speech are plotted. A considerable improvement of the SNR can be obtained with this processing. The generation of musical tones seems to be a little bit less of a problem than in the case of spectral subtraction.

Some experiments with automatic recognition for a combination of this technique and Rasta-PLP are currently in progress and will be presented at ICASSP93 for signals corrupted by convolutional and additive noise.

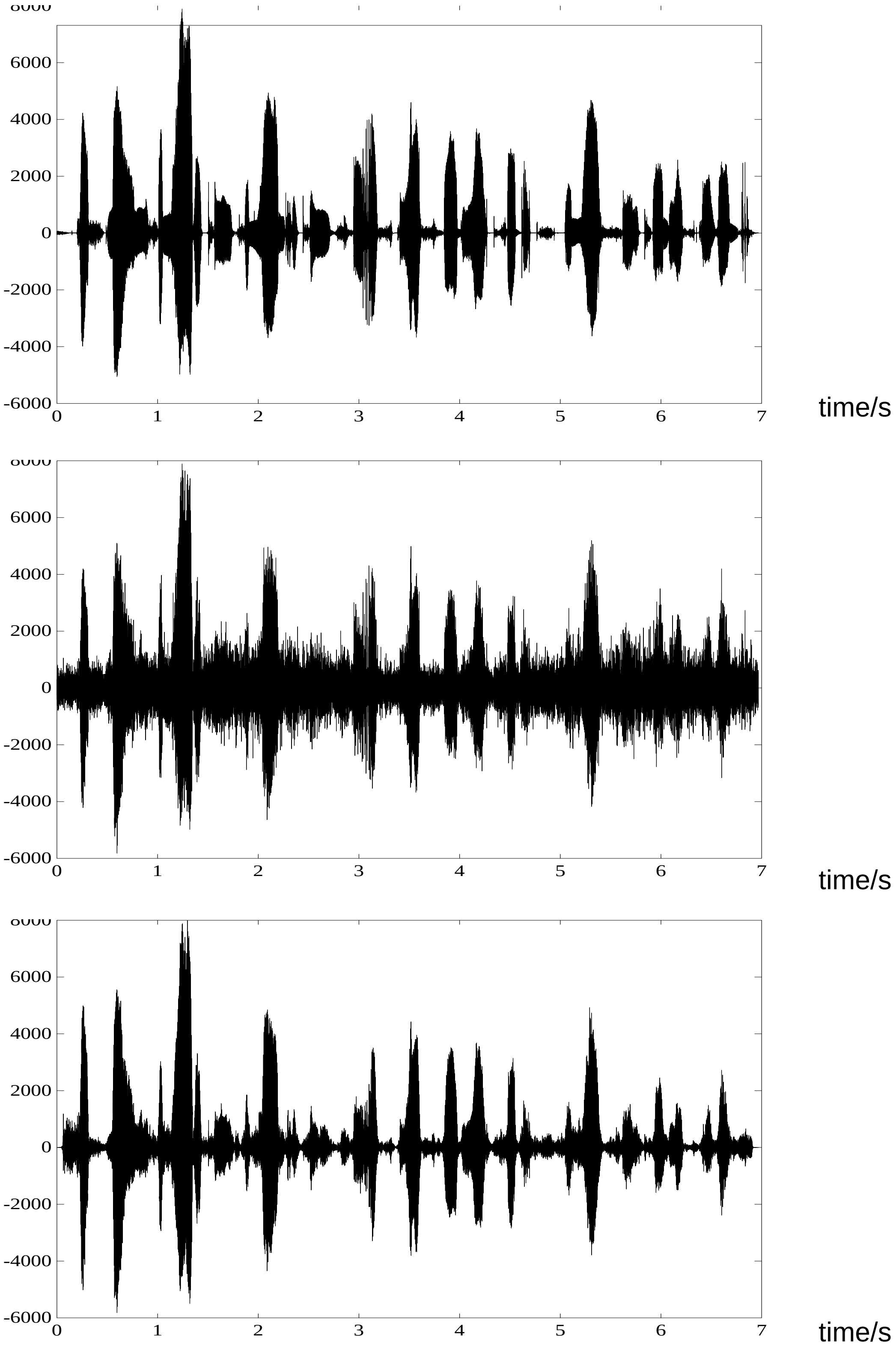

Figure 5.1: Time signals of clean and noisy speech and after processing with a spectral subtraction technique

5. Speech Enhancement

The estimation of noise spectra can also be used for speech enhancement. Two applications were considered in this study.One is the well-known spectral subtraction technique. The other one is based on a modified high-pass filtering of the spectral envelopes in subbands.

Only the magnitude spectral values are processed in both cases. The phase of the noisy speech is used for the resynthesis. An existing program called “synthese” was used for the resynthesis, in which time signals are generated from processed spectra with the overlap-add method.

The noise spectrum estimation described above (section 3) is used for the spectral subtraction. The subtraction is applied in 128 bands. The estimated magnitude noise spectral value N(iΔf) is used for an adaptive weighting of X(iΔf), the spectral magnitude of the noisy speech, in each subband with a centre frequency iΔf.

An estimation of the magnitude component of the speech is calculated as

$$\hat{S}(i\Delta f) = \left(1 - \frac{N(i\Delta f)}{X(i\Delta f)}\right) X(i\Delta f)$$

The contour of X(iΔf) is usually smoothed with an exponentially decaying weighting of past values.

Various modifications of the weighting function 1 - (N(iΔf)/X(iΔf)) are possible /4/, /5/, e.g. the realization as a Wiener filter.

It is possible that negative weighting factors occur. One solution is to set these factors to zero. The time signals of clean speech, of noisy speech where a car noise was added with a SNR of 5 dB and of the processed noisy speech are shown in figure 5.1.

A considerable improvement of the SNR can be seen. However, listening to the resynthesized speech one can hear a new artificial noise that is often referred to as ‘‘musical tones’’. This degradation, common to spectral subtraction-based enhancement techniques, significantly disturbs the subjective impression. However, earlier experiments /6/ had shown that this artificial noise has little influence on the recognition rates of an isolated word recognizer. Thus, recognition rates could be improved by the introduction of the spectral subtraction technique.

The delay is smaller for an analysis window of 250 ms but some errors occur for the noise level estimation. In the case of a 1 s window the curve doesnot fit the modulation characteristic as good as in the case of 500 ms. The length of 2 s is too high to follow the varying noise level.

The length of the analysis window should be chosen for the particular noisy situations to which the processing is applied. A length of 500 ms seems to be a good compromise for the cases we have examined.

Experiments with naturally recorded noisy speech signals have shown a good agreement of the estimated SNRs with the expected curves.

The results for different window lengths are shown in figure 4.10.

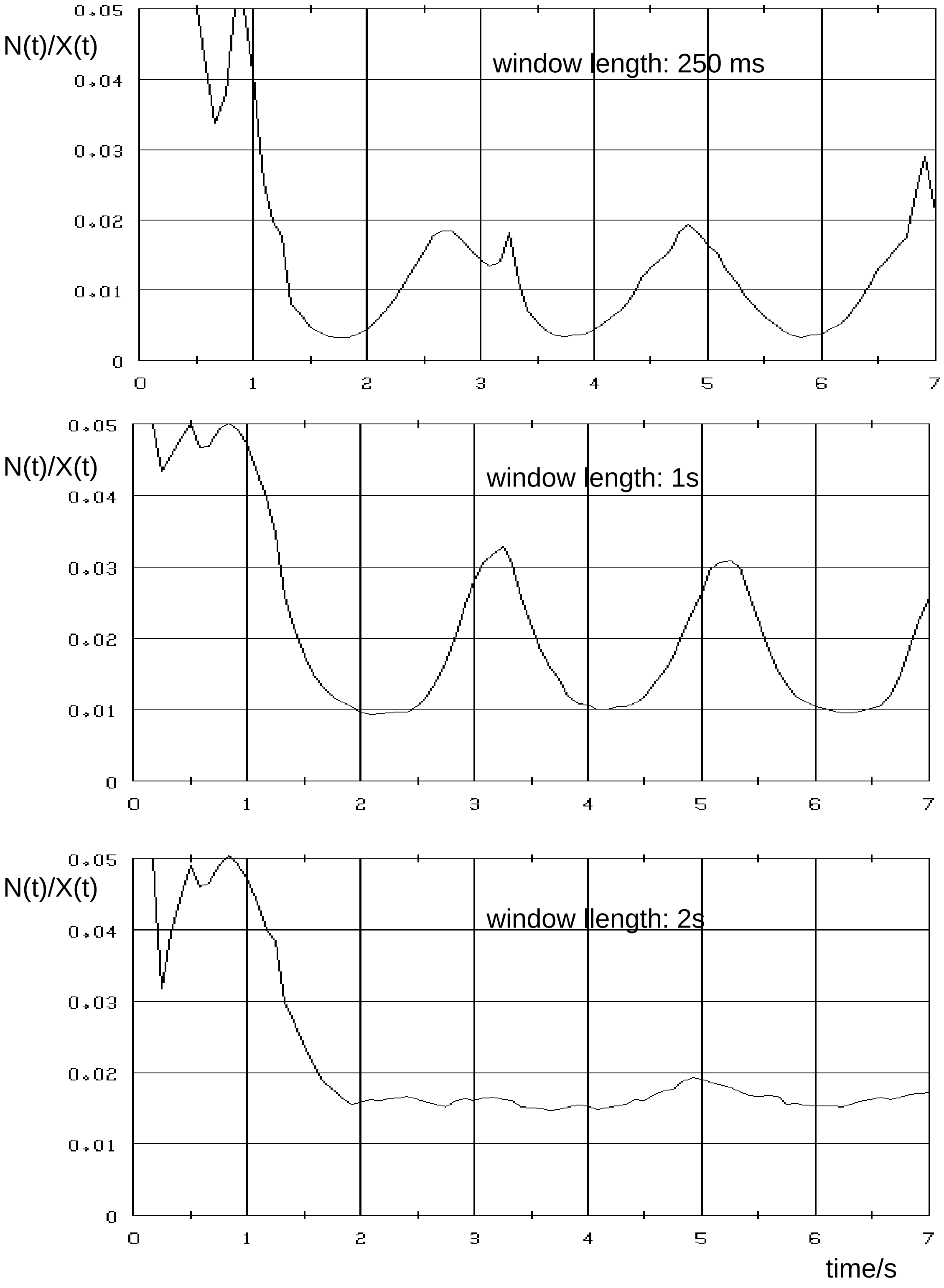

Figure 4.10: Estimated N(t)/X(t) for different window lengths

Some experiments were done adding noise with a varying SNR. A modulated Gaussian noise was added to the speech signal shown in figure 4.1 with an overall SNR of 10 dB. The modulation signal itself can be seen in figure 4.8. The result for the estimation of N(t)/X(t) is shown in figure 4.9 using an analysis window of 500 ms.

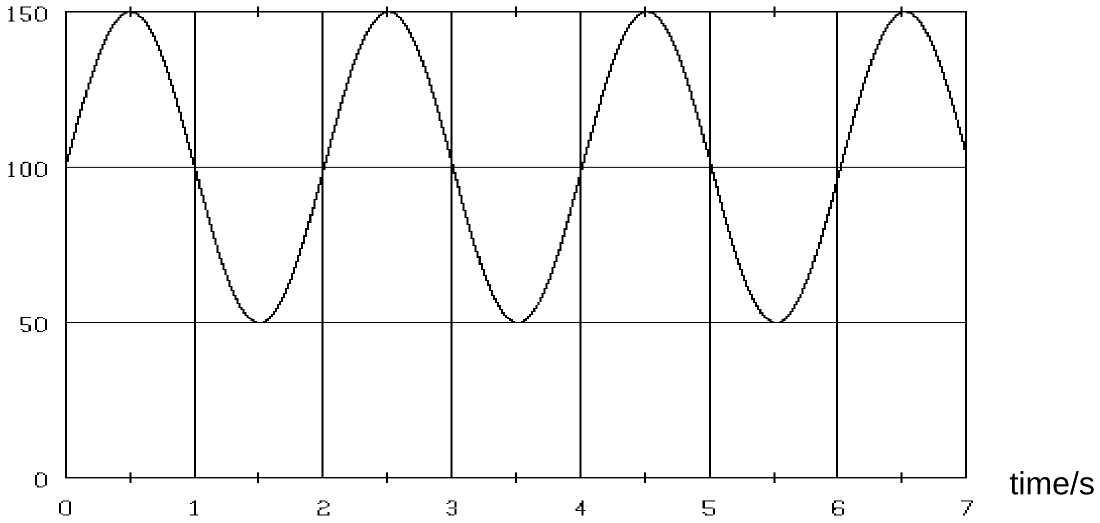

Figure 4.8: Time signal for the modulation of a Gaussian noise

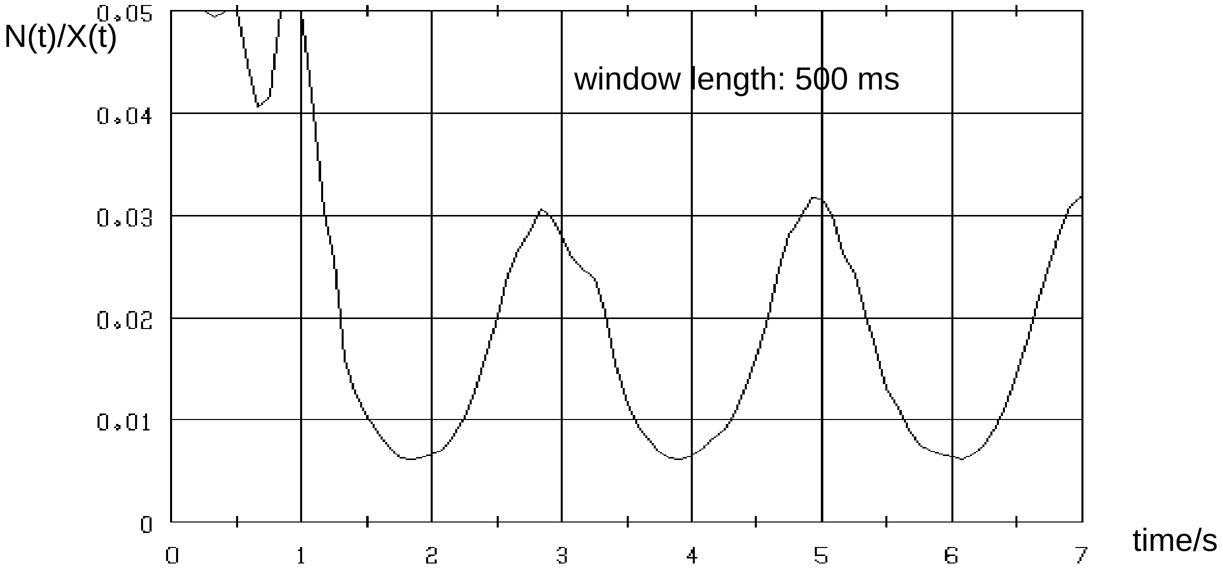

Figure 4.9: Estimated N(t)/X(t) for a signal disturbed by a modulated Gaussian noise

The estimation of N/X follows this artificial modulation characteristic quite good. A delay of about 500 ms can be considered because of the analysis window in the past.

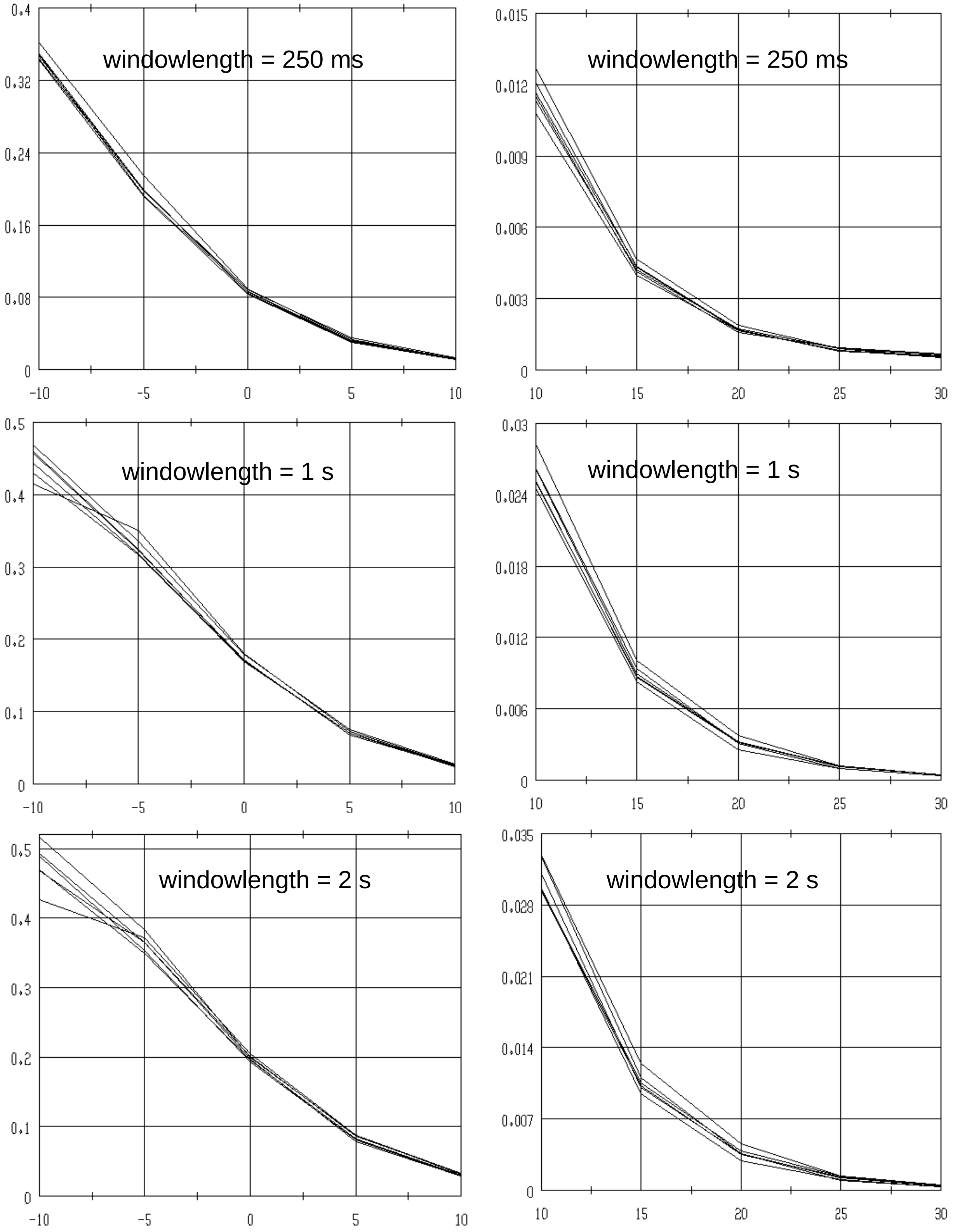

Figure 4.7: Average estimation of the relation N/X for different window lengths

An average of these estimated values is calculated for the last 4 s of the signal corresponding to the nearly constant part of the curve. The result of this averaging is shown in figures 4.4 and 4.5 for different SNRs and different noise signals where a window length of 500 ms was used for the noise estimation.

Stationary segments of two naturally recorded signals and four artificially generated signals were used as noise signals. The results are shown for a 40 dB range in dynamics. The noise signals were

1) car noise
2) computer room noise
3) white Gaussian noise within a bandwidth of 0 to 4 kHz
4) white Gaussian noise within a bandwidth of 0 to 0.5 kHz
5) white Gaussian noise within a bandwidth of 0 to1 kHz
6) white Gaussian noise within a bandwidth of 2 to 4 kHz

The accuracy of the estimation slightly decreases for high SNRs. But overall a high correlation can be seen for the different noise signals. The accuracy is in a range of about 1 to 2 dB. The results can be used to realize a mapping from the estimated N_{energy}/X_{energy} to the real SNR.

The results when using different window lengths for the noise level estimation are shown in figure 4.7.

The scaling of the ordinates is different for the different window lengths because of the calculation of X in the corresponding window. A much higher maximum of X occurs for the short window of 250 ms which is nearly comparable with the estimation of the energy of a vowel. No big difference can be seen for the different lengths of the window when comparing the correlations of SNR to the computed N/(S+N). However, these curves are the result of an averaging over 4 s so that the influence of the temporal fluctuations can not be seen.

part of the curve in figure 4.3 which it ideally should take for a SNR of 5 dB.
Some further results using the signal of figure 4.1 are shown in figure 4.4 for different SNRs.

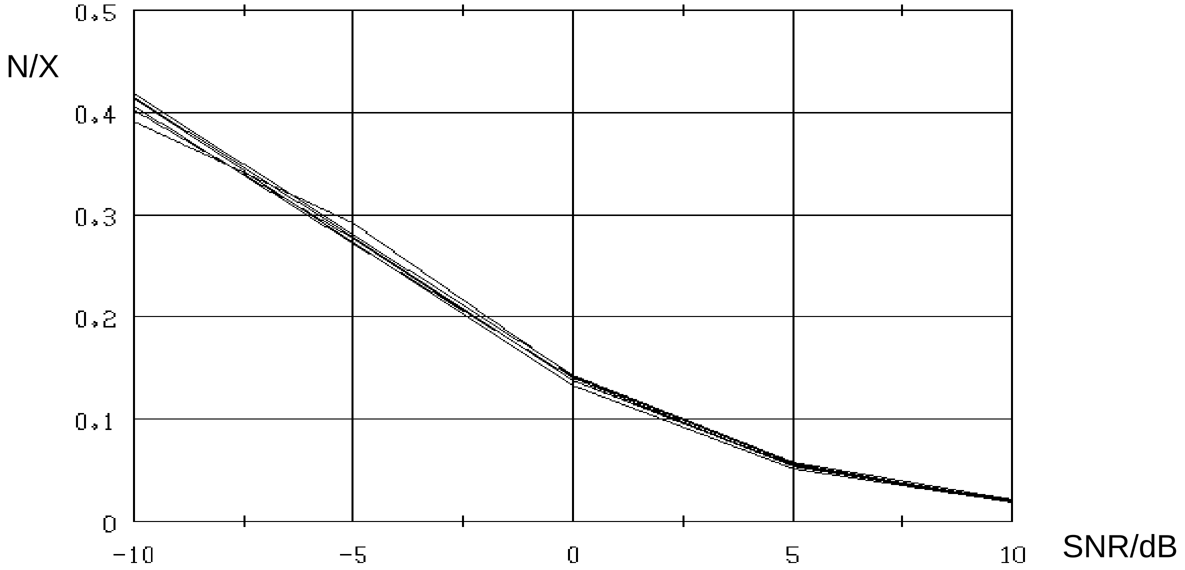

Figure 4.5: Average estimation of the relation N/X for different noise signals

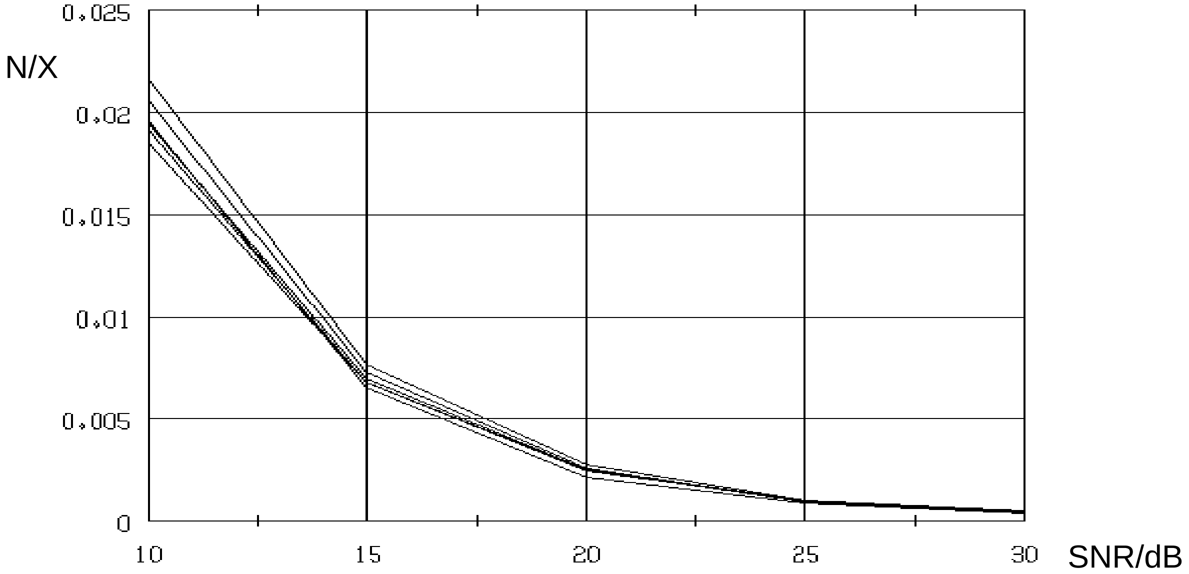

Figure 4.6: Average estimation of the relation N/X for different noise signals

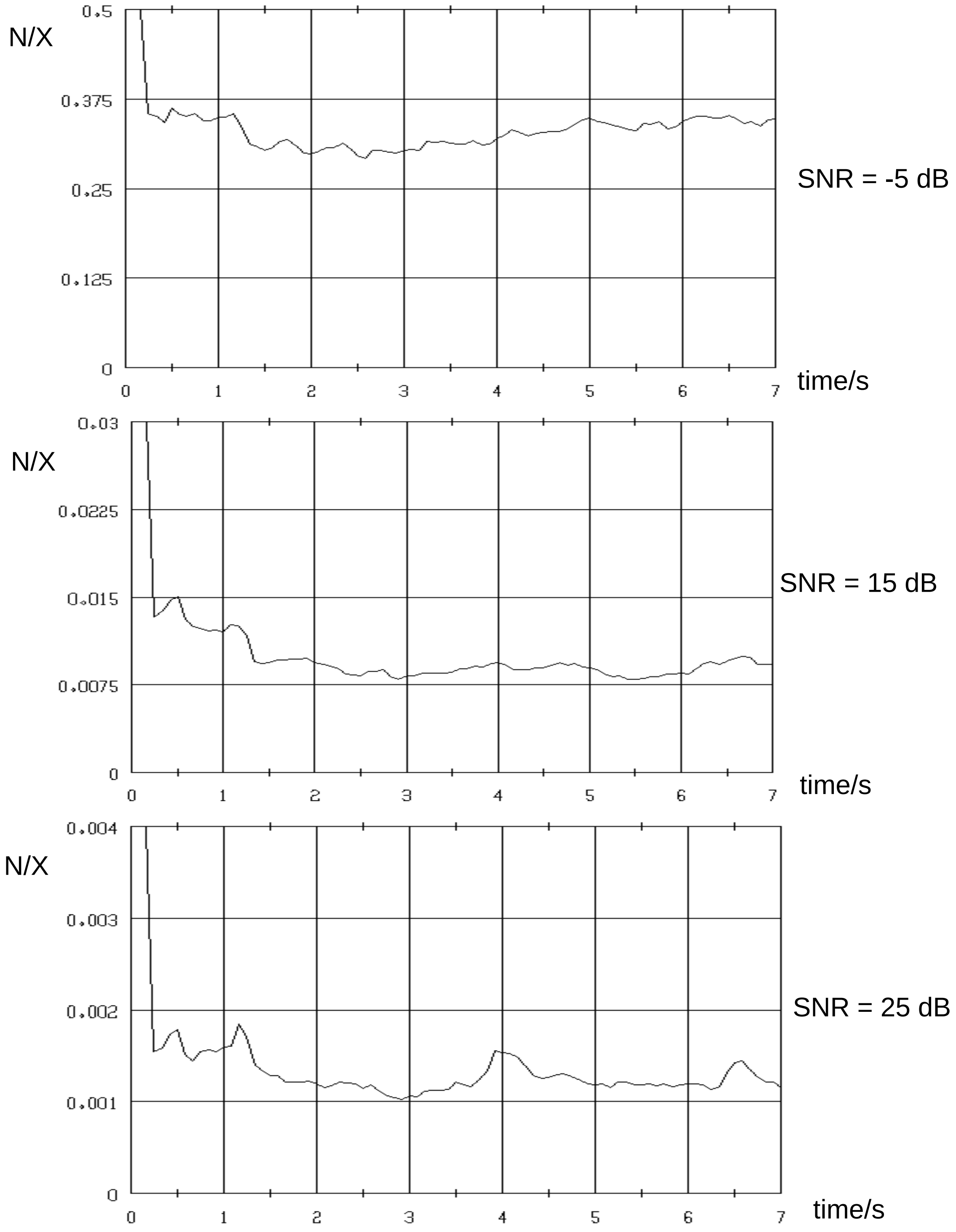

Figure 4.4: Estimation of the relation N(t)/X(t) for different SNRs

The calculation of the maximum of the average energy $X_{energy}(t)$ is actually not the same as computing the long term average of X. The value of $X_{energy}(t)$ is higher than a long term average. Because of this the relation doesn't take the value of 0.24 in the constant

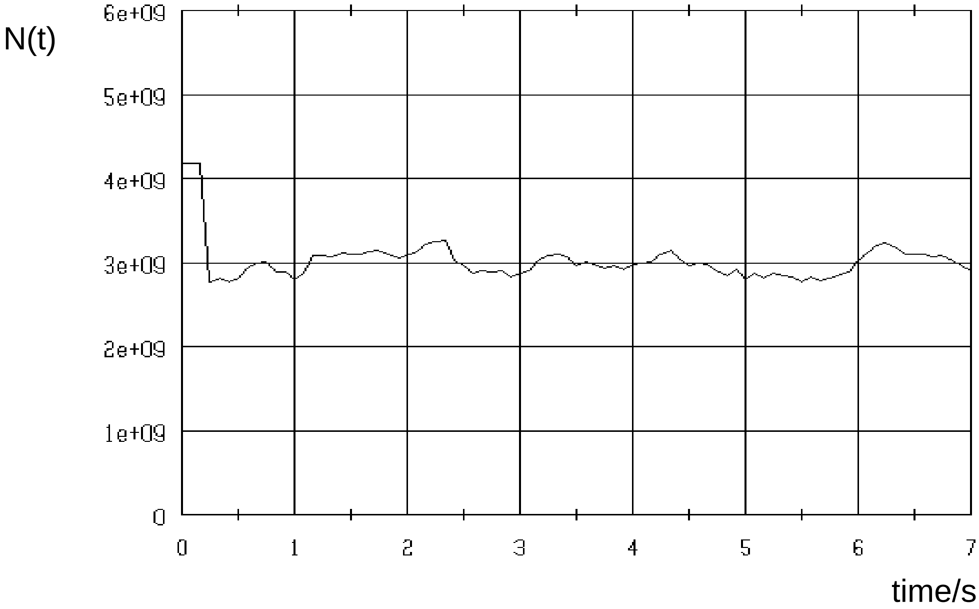

Figure 4.2: Estimation of the noise level N(t) for the signal of figure 4.1

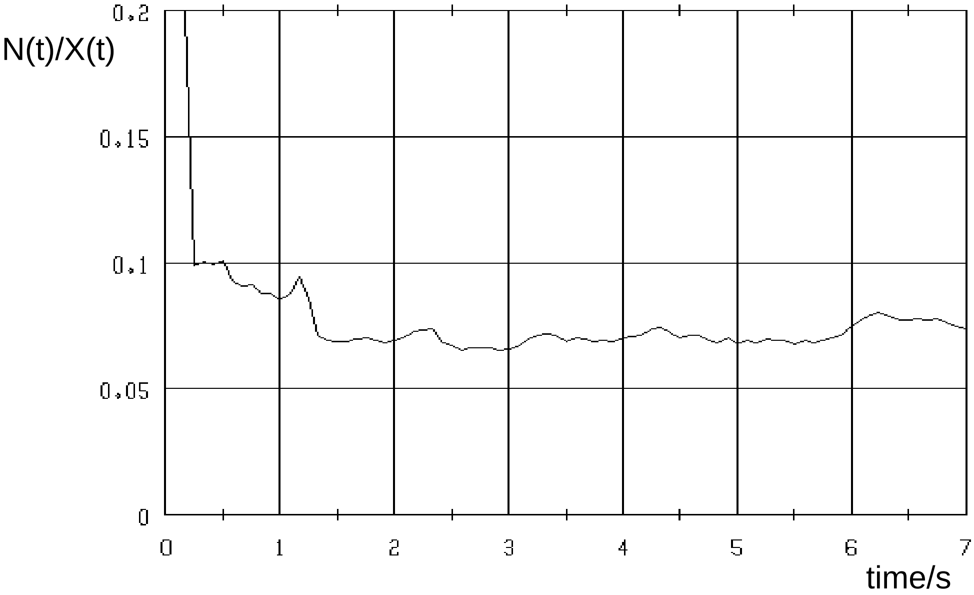

Figure 4.3: Estimation of the relation N(t)/X(t) for the signal of figure 4.1

The noise level estimation seems to be a little bit too high at the beginning of the speech. One reason for this is a noise inside the speech signal itself caused by breathing of the speaker. The relation N(t)/X(t) takes a high value of about 0.7 at the beginning because there is nearly only noise. Then the curve rapidly slopes when the first high values for X(t) are calculated at the beginning of speech activity. Later on the relation takes a nearly constant value of about 0.07.

the task precludes a long initial delay.

The average energy X(t) is calculated for a past segment where the length of this window corresponds to the window length for the calculation of the distribution density function for the noise level estimation.

If the actual value of this average energy X(t) would always be used, the result of the relation N(t)/X(t) would be a kind of segmental N/(S+N) ratio. Instead, the maximum of the energy X(t) up to this time is used. This maximum is slowly decreased with an exponential decay to adapt to an overall change of the signal level so as not to use a local peak value of X over a long time.

A result for the estimation of N(t)/X(t) can be seen in figure 4.3. A sentence with a duration of about 7s was used as a speech signal. A Gaussian noise was artificially added with a SNR of 5 dB. The time signal is shown in figure 4.1. The estimation of the noise level N(t) itself can be seen in figure 4.2. The noise estimation as well as the calculation of X(t) were done within a window of 1 s.

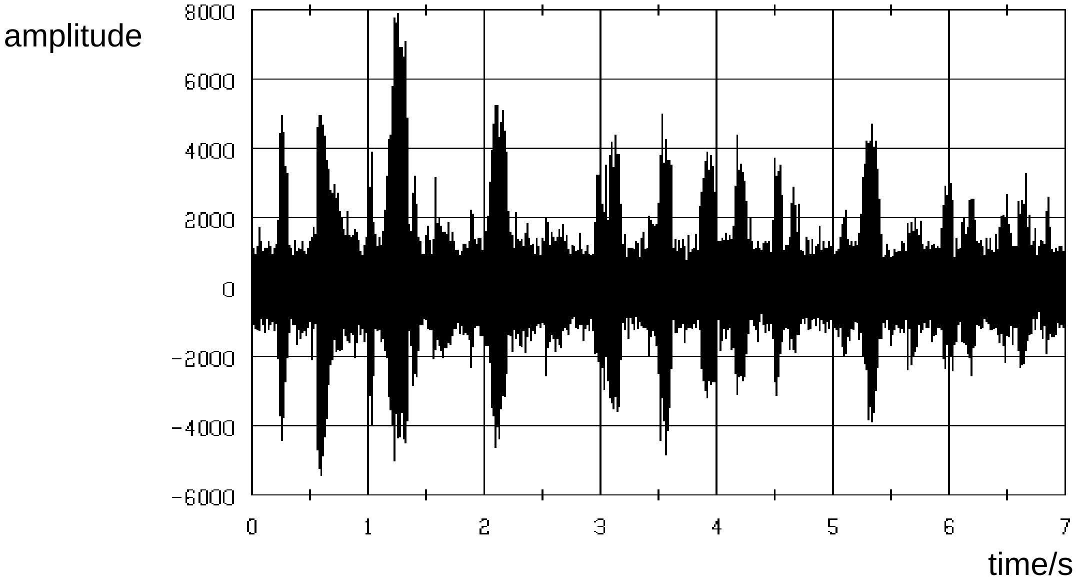

Figure 4.1: Time signal of a noisy sentence

4. SNR (signal-to-noise ratio) estimation

The estimation of the noise spectrum can be used for an estimation of the SNR (signal-to-noise ratio). Despite the well-known engineering meaning of the acronym, SNR must first be defined in further detail.

In the field of speech coding the term SNR is often used as a so called segmental SNR where the SNR is actually only a description of the ratio of speech energy to noise energy for a short time period, e.g. 20 ms. Often the expression is really interpreted as a long-term measurement, especially for stationary noise situations. In this case the average energy of the speech is usually calculated over a longer time period, at least some seconds and including speech pauses. The noise energy is calculated for the same time period assuming that the noise is nearly stationary during this time. This can be ideally done when artificially mixing a speech and a noise signal with an optional SNR.

In this application the term SNR can apply to any of a range of temporal scales.
On one hand it should be more related to the long-term SNR. But on the other hand it should be possible to follow slowly changing noise situations as they occur in real situations.

Given a real noisy signal it is difficult to directly estimate the SNR without any specific knowledge about the speech or noise energy. Because of this the noise to signal-plus-noise ratio N/(S+N) was considered instead of the signal to noise ratio itself.

An estimation of the short-term energy Nenergy of the noise at a specific time t is calculated with Parseval’s relation..

$$N_{energy}(t) = \frac{1}{nfft} \sum_{i=0}^{nfft} N_{spec}(i\Delta f, t)$$

where $N_{spec}(i\Delta f)$ is the estimated spectral magnitude of the noise in a subband with a centre frequency of $i\Delta f$ where Δf = <sampling frequency> divided by nfft and nfft = <FFT-length>.

The short term energy of the noisy signal x is calculated in the same way.

$$X_{energy}(t) = \frac{1}{nfft} \sum_{i=0}^{nfft} X_{spec}(i\Delta f, t)$$

where X_{spec} is the spectral magnitude of the noisy speech signal.

The average energy of the noisy signal can not be calculated over a longer period of speech, (e.g. several seconds), in this application because the interactive character of

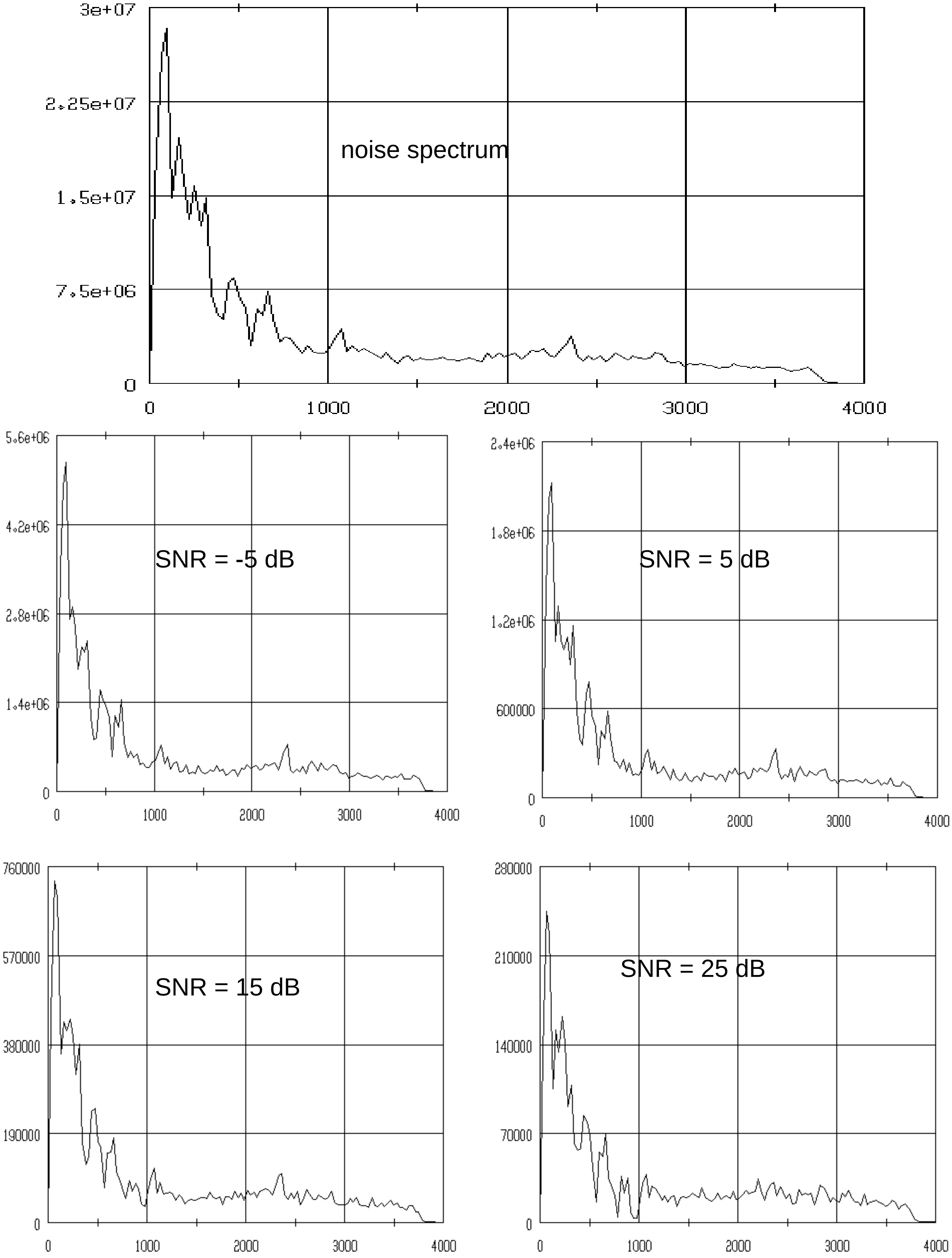

Figure 3.4: Average noise and estimated noise spectra for different SNRs

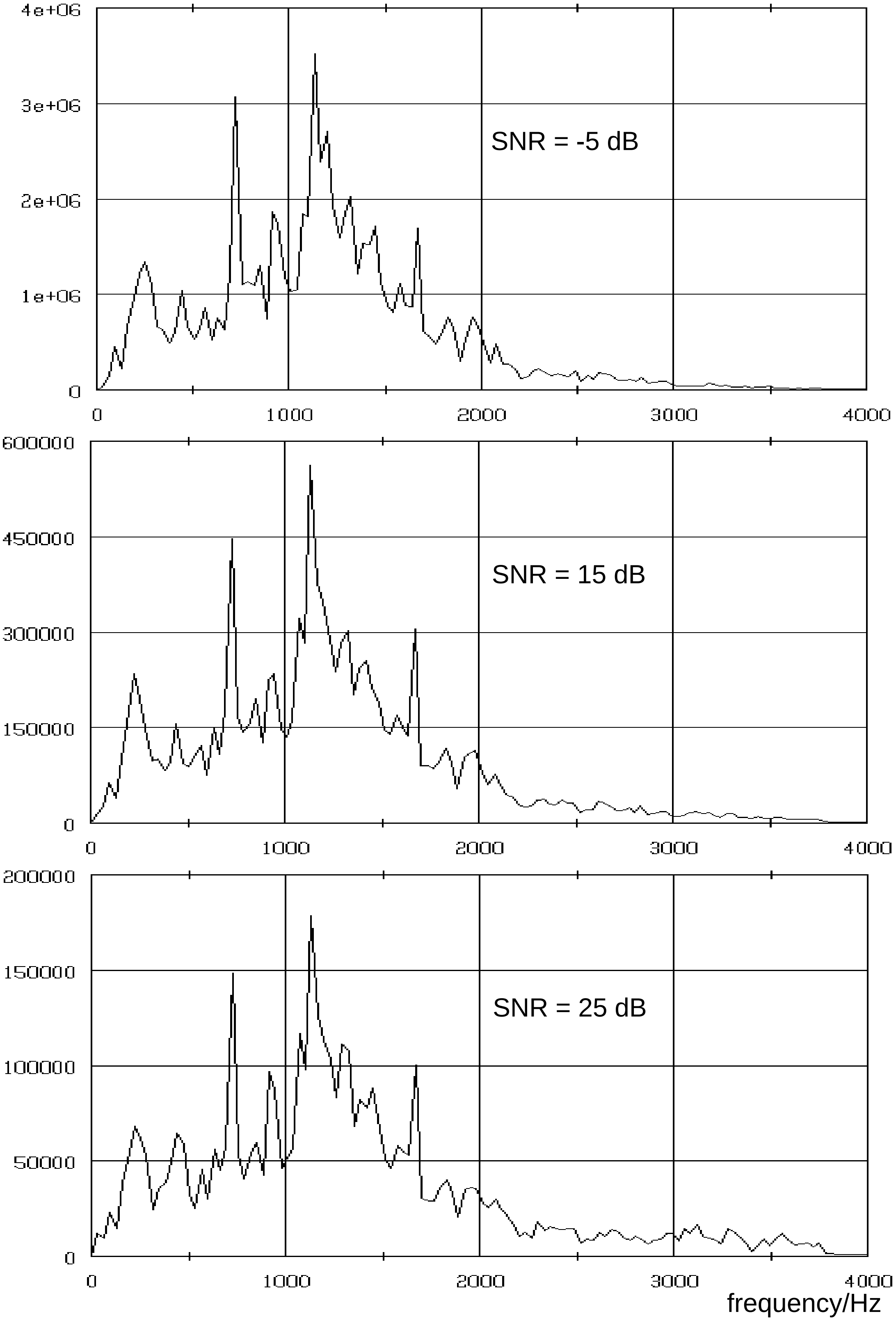

Figure 3.3: Estimated magnitude noise spectra for different SNRs

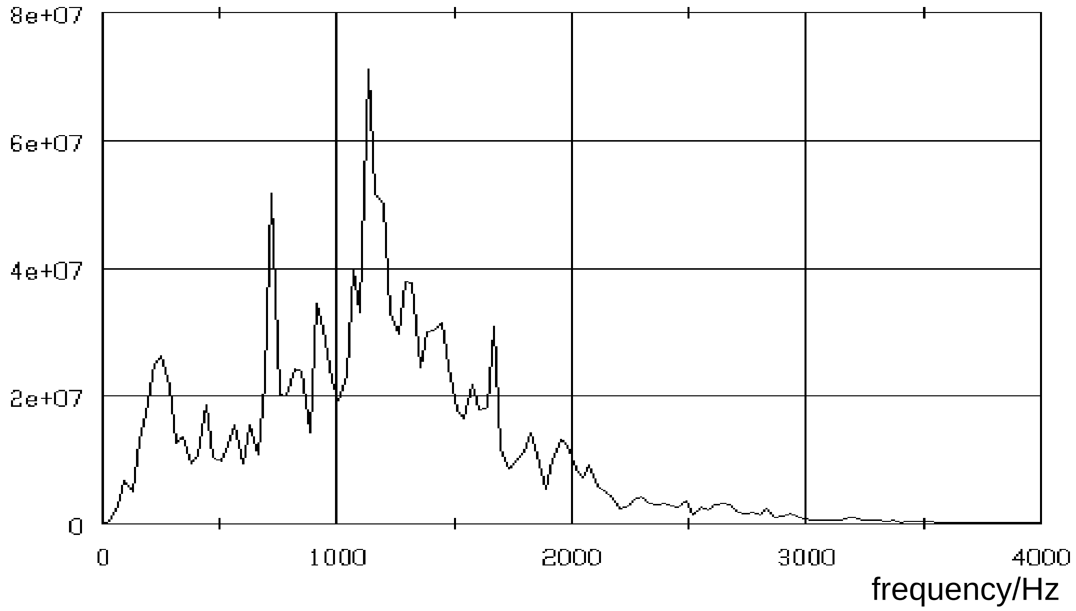

Figure 3.1: Average magnitude spectrum of a noise signal

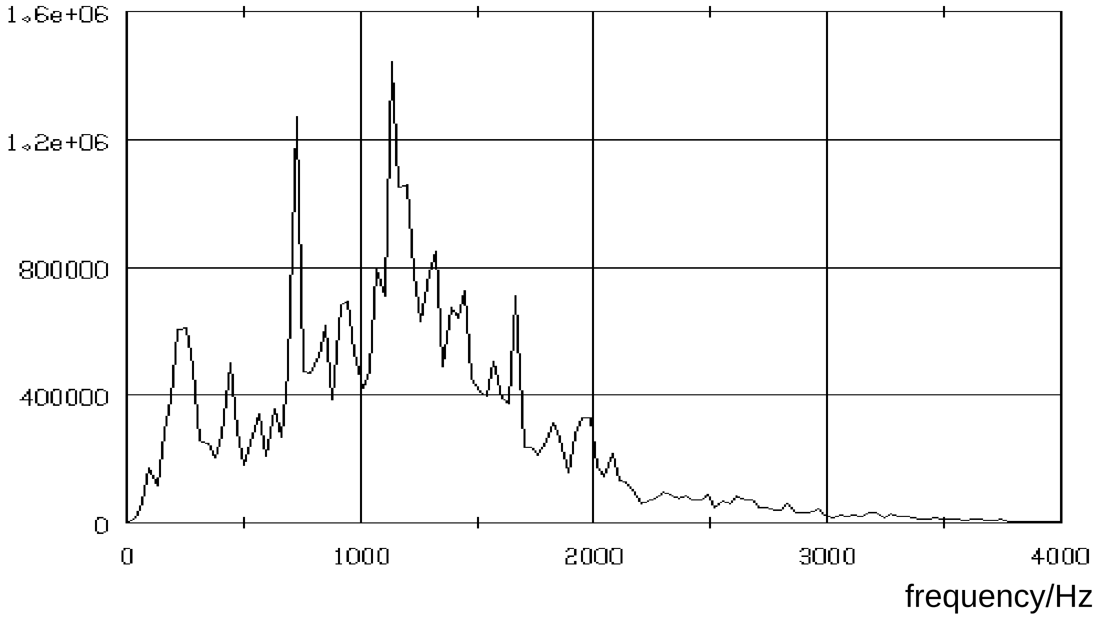

Figure 3.2: Average estimated magnitude noise spectrum of a noisy speech signal with a SNR of 5 dB

the noise level inside each of these bands is estimated by taking an average of a certain amount of the smallest spectral values of the last second. Then a distribution density function is calculated with a variable window length. Up to 50 values of the past are considered which take a value between 0 and the upper limit of the noise level. The maximum of the distribution density function is calculated as mentioned before.

The estimated magnitude spectral values are smoothed in all cases by computing a weighted sum of the actual estimated noise level and all estimated values of the past. The weighting is done with an exponentially decaying curve.

A result of this kind of noise estimation can be seen in figure 3.2. There is shown the average estimated spectrum of a noise signal. The estimation was done after adding the noise signal to a nearly clean speech signal. The length of the analysis window for the distribution density function was 500 ms. The average spectrum of the noise itself is shown in figure 3.1. The averaging was done for the whole noise signal.

The estimation of the noise spectrum appears to work well. However, some overestimation can be seen for a frequency component at about 750 Hz with high energy.

Some further estimated noise spectra are shown in figure 3.3 for additive noise resulting in different SNRs. The estimation seems to be nearly independent of the SNR. Some small differences can be seen for the case of a high SNR of 25 dB. The reason for this is the influence of the “clean” speech signal itself. The speech was recorded at a SNR of about 35 to 40 dB. This noise of the “clean” speech can already be seen for example in the region of higher frequencies.

A spectrum of another noise signal and some corresponding noise spectral estimates are shown in figure 3.4.

3. Practical realization

In a practical application the calculation must be done online with running speech. The length of a window has to be defined for the calculation of the distribution density function. On one hand, the length should be as high as possible to increase the accuracy of the noise level estimation. On the other hand, the length may not exceed a certain duration if a signal with a varying noise level should be analyzed. The spectral analysis is done with a universal program for a short-term spectral analysis.

The FFT length used was 256 so that the centre frequencies of the estimated spectrum have a distance of 31.25 Hz at a sampling frequency of 8000 Hz. The window for weighting the speech samples was a Kaiser window multiplied by a sinc function. The influence of different window types was not examined but it could be assumed that it is not very high. The spectrum is calculated every 8 ms so that the magnitude spectral values inside one band are given with a sampling frequency of 125 Hz.

The first estimation of the noise spectrum used the average of the first ten magnitude spectral values, calculated within each band. It is assumed that the first incoming speech samples of a recording are related to the noise. This noise estimation is used up to a time of 250 ms after starting the recording. Afterwards a window with an increasing length is used up to the final window length for the calculation of the distribution density function. We finally considered window lengths of 250ms, 500ms, 1s and 2s.

The distribution density function is calculated for the magnitude spectral values inside the window for each band. The function is computed in a range from 0 to the maximum spectral value which was found in this band up to this time. The accuracy is 0.25 per cent of the maximum which corresponds to dividing the whole range into 400 intervals. A search for the maximum peak value is done where at first an accuracy of 2 % for the distribution function is used. This is done by summing 8 neighboring values of the function, corresponding to a smoothing of the distribution density function.

If the maximum value is higher than 10 % of the maximum spectral value inside this band the estimated value is taken directly. If the detected maximum is anywhere in the range of 5 to 10 per cent an accuracy of 1 % is used, in the range of 2.5 to 5 % an accuracy of 0.5 % and under 2.5 % the highest available accuracy of 0.25 % is used for a more accurate noise level estimation. This kind of analysis is related to the fact that on one hand a more smoothed version of the distribution function should be used for the noise level estimation in cases of a low SNR inside a band. On the other hand the resolution has to be higher for high SNRs where you have only a small amount of noise in one channel. The fixing of the analysis intervals and the accuracy was done empirically.

The following processing is done to avoid a poor estimation of the noise level inside some low frequency bands where spectral magnitude values very often occur with a high value (already mentioned in the preceding section). The five channels with the highest amount of energy are calculated by looking back 1 s of time to the past. An upper limit for

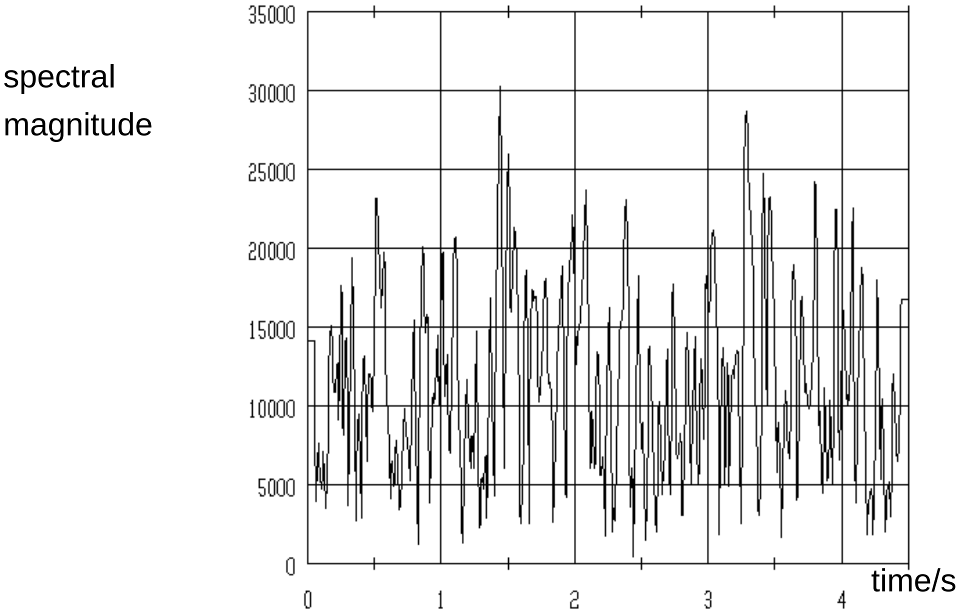

Figure 2.5: Spectral envelope in a subband with a centre frequency of 219 Hz

One problem for this envelope occurs with a statistical analysis in a short time window of e.g. 500 ms around the time of 2s. The signal takes nearly only high values inside this window so that it will be impossible to estimate a reasonable noise level. To avoid this problem we must detect channels with such a behaviour and to extend the analysis window in these cases. Usually these are only channels with a centre frequency less than 500 Hz and with high energy. Both these criteria are used for the detection of channels with a possible behaviour similar to the spectral envelopes of figure 2.5.

Some channels don’t show the good behaviour seen in figure 2.3. One result can be seen in figure 2.4 for a channel with a high centre frequency of about 2950 Hz. There is nearly no speech energy but a significant noise energy in this channel. The noise was a Gaussian noise in this case.

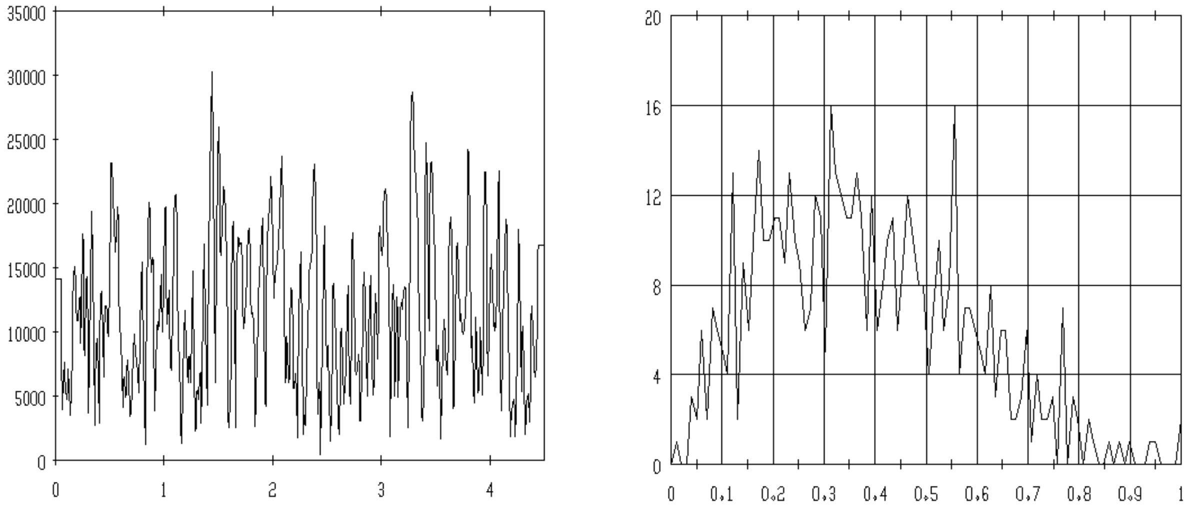

Figure 2.4: Spectral envelope and the corresponding distribution density function of a frequency band with a centre frequency of 2590 Hz

Sometimes a maximum respectively a noise level is calculated with an unrealistic high value in these cases. Because of this the possible estimated noise level is limited to the average spectral value. This is related to the fact that no noise energy can occur which is higher than the total amount of energy inside a band.

Another channel with a nonideal behaviour is shown in figure 2.5. This is a channel with a centre frequency of about 219 Hz where the spectral magnitude of the signal often takes a very high value.

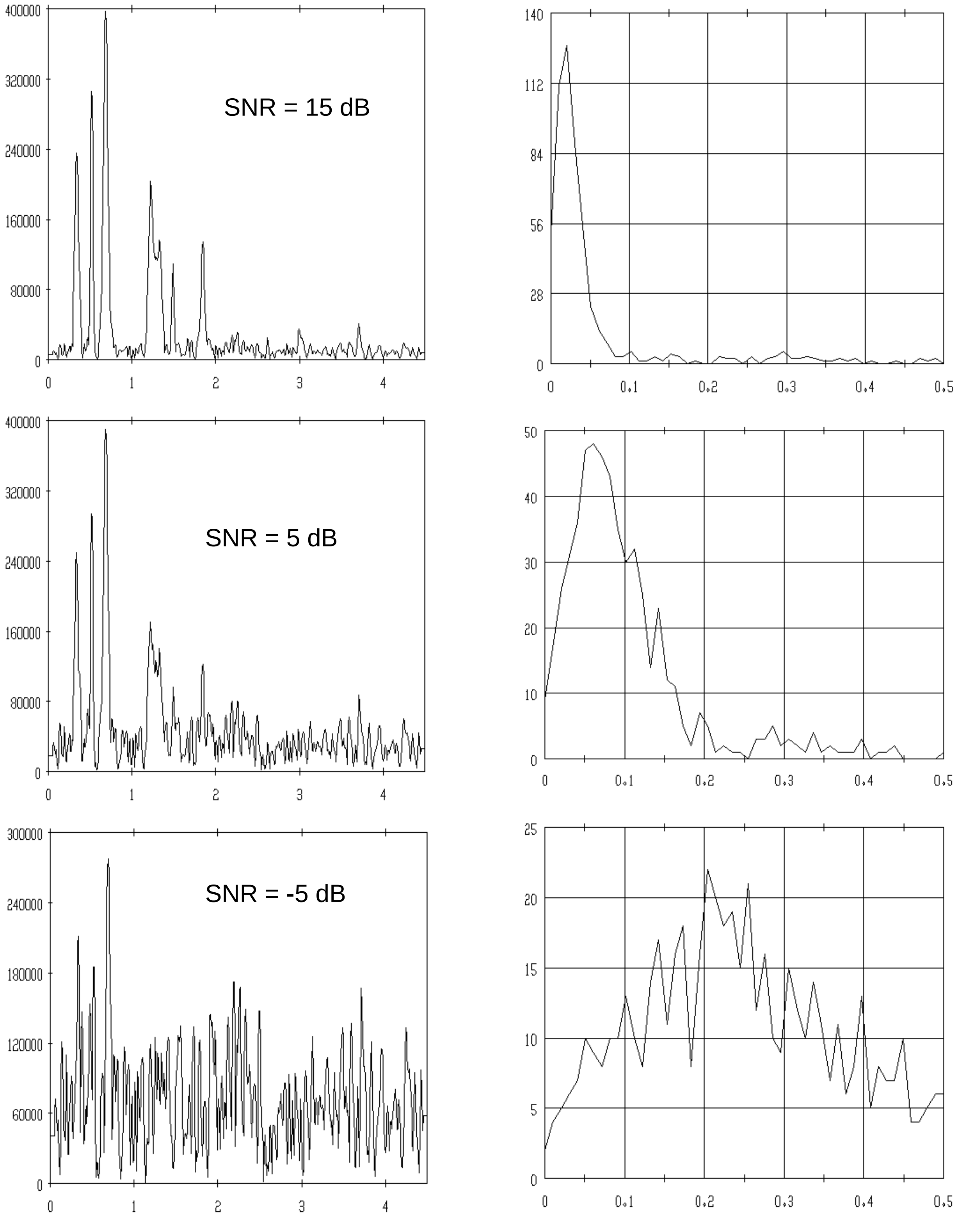

Figure 2.3: Spectral envelopes and distribution functions for different SNRs

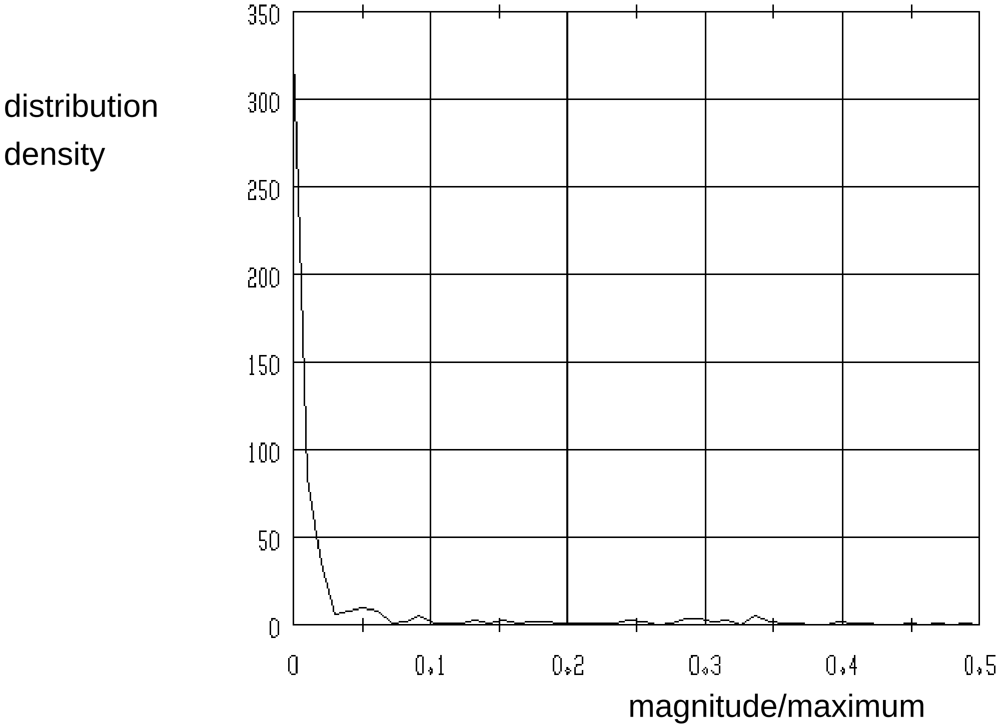

Figure 2.2: Distribution density function for the spectral envelope in figure 2.1

for channels with a low SNR. A reduction of the accuracy for low SNRs has the effect of smoothing the distribution function and improving the maximum detection. On the other hand the accuracy has to be high for channels with less noise to get a reasonable estimation for the amount of noise at all. Because of this the accuracy for the calculation of the distribution density function is made dependent on the actual SNR inside one channel. The accuracy is less for channels with a bad SNR and higher for channels with a better SNR.

2. Principal Idea

The principal idea to estimate the noise level in a certain subband is based on a statistical analysis of a segment of the magnitude spectral envelope.

Looking at such a spectral envelope in figure 2.1 and the corresponding distribution density function in figure 2.2 the most commonly occurring spectral magnitude value is zero. The spectral envelope was calculated for a clean speech signal with a duration of about 4.5 s and in a subband of about 500 Hz. The distribution density function was calculated for the whole duration of 4.5 s with an accuracy of about 1 percent in regard to the maximum spectral value inside this subband. The function is shown for the range of 0 to 50 percent of the maximum. Only a few values occur which are higher than 50 percent.

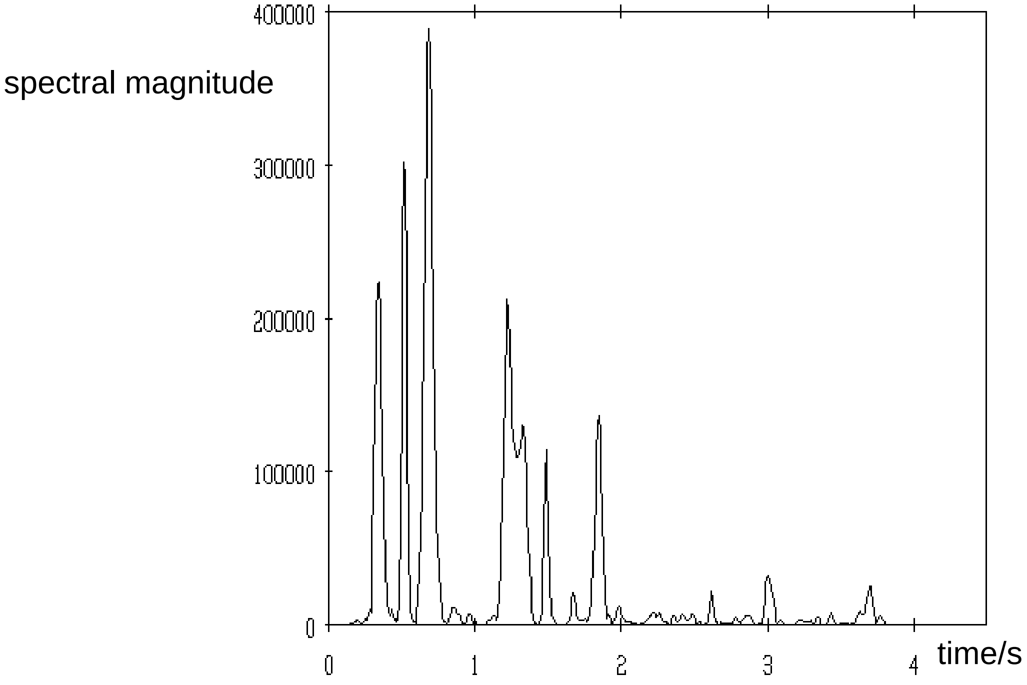

Figure 2.1: Spectral envelope in a band with a centre frequency of 500 Hz

Noise was added artificially to this speech signal to produce different SNRs. The results can be seen in figure 2.3. The noise was a bandpass limited Gaussian noise with a centre frequency of 500 Hz and a bandwidth of 200 Hz.

An increase of the maximum value in the distribution function can be observed for a decreasing SNR. This most frequently occurring value can be taken as an estimation for the noise level inside this band.
Also, an increasing variance of the spectral magnitude values of the noise can be seen for an decreasing SNR. Because of a broad distribution the estimation isn’t so accurate

realistic situations with a varying noise level. Another disadvantage is the fact that the algorithm cannot adapt to a varying noise level during segments of speech. The adaptive filtering is always based on the estimated noise spectrum of the preceding speech pause.

An improvement of the spectral subtraction technique would be an estimation of the noise spectrum without the necessity of a speech pause detection. A method is presented in this report to estimate the noise spectrum without a speech pause detection.

One application for this method presented in this report is the estimation of the actual SNR of a noisy signal. Furthermore the technique is applied to speech enhancement based on a spectral subtraction respectively on a nonlinear high-pass filtering of the spectral envelopes dependent on the actual SNR.

1. Introduction

At the time of this writing, some experiments in robust speech recognition had already been done at ICSI.
The original goal of this work was improved recognition of speech recorded with different microphones and transmitted over channels with different frequency characteristics. One practical application of this is the recognition of speech recorded via telephone lines where you have microphones and channels with different transmission characteristics. It could be shown that the recognition rates can be improved when introducing a high-pass filtering of the logarithmic spectral envelopes in subbands /1/.
This idea is based on the fact that a frequency characteristic corresponds to a multiplication of the speech spectrum with the frequency response of the transmission channel. The result would be a constant additive component in the logarithmic spectral envelopes in subbands (assuming a nearly constant transmission characteristic). Because of this, a high-pass filtering leads to a suppression of these constant components.

Another aspect is the superposition of noise in many applications of speech recognizers in real environments, e.g. voice dialing in a car or serving any kind of machines on the street or in workshops. This noise would result in a nearly constant additive component to the magnitude spectral envelopes in subbands (assuming a nearly stationary noise) . It could be shown that recognition rates can be improved by high-pass filtering the magnitude spectral envelopes /2/.

Additive noise as well as a certain frequency characteristic are present in many real situations. One way to handle both effects could be to use a combination of processing in the magnitude as well as in the logarithmic spectral domain. Another possibility could be a processing anywhere between the magnitude and the logarithmic domain dependent of the amount of noise in the specific situation. This would presuppose an estimation of the signal-to-noise ratio (SNR).

Looking at the first possibility several processing techniques are well known to reduce the noise in the magnitude spectral domain. One could be the already mentioned high-pass filtering. A disadvantage of this method is the suppression of certain spectral features in speech segments. Introducing a high-pass filter with a total suppression of the DC component, not only the constant noise components are suppressed but also the constant component of the speech. Because of this just the spectral features of the phonemes with less energy are reduced in the case of a preceding phoneme with higher energy and spectral components in the same subbands.

One solution could be a kind of nonlinear filtering with the goal of preserving the spectral features of the phonemes with less energy on one hand but suppressing the noise components on the other hand. Another method to reduce the noise is the well known spectral subtraction technique /3/,/4/. This technique is based on the estimation of the noise spectrum during speech pauses and an adaptive filtering with the estimated noise spectrum. A major disadvantage is the necessity of the detection of speech pauses to estimate the noise spectrum. This is a very difficult and ultimately unsolved problem for

Estimation of Noise Spectra and its Application to SNR-Estimation and Speech Enhancement

H. Günter Hirsch

Contents

Estimation of Noise Spectrum and its Application to SNR-Estimation and Speech Enhancement

H.Günter Hirsch

Technical Report TR-93-012

International Computer Science Institute,
Berkeley, California, USA