Speech Signal Enhancement Using Masking Effects of Hearing Physiology

Kristian Kroschel
Institut fuer Nachrichtentechnik
Institut fuer Automation und Robotik
Universitaet Karlsruhe
Kaiserstrasse 12, D-76128 Karlsruhe

Germany

Abstract

Classical methods for speech signal enhancement such as spectral subtraction techniques do not use effects of hearing physiology. Thus the result for signal enhancement approved by a human listener might not be satisfying even if the signal-to-noise ratio enhancement which is used as an objective figures of merit yields high scores. In this paper a new method for speech signal enhancement based on the masking effect is introduced which results in both, a higher signal-to-noise enhancement and a better quality of the processed speech.

1 Introduction

Modern communication systems prefer handset free operation because of comfort and the possibility to use the hands for other operations such as driving a car, keying in messages in front of a terminal etc. Often background noise will disturb the transmitted speech so that the far-end listener will have problems in understanding. To enhance intelligibility, noise reduction techniques are provided for modern handsfree communication systems. Mostly they are based on signal estimation by Wiener filters [2] which is known in literature as spectral subtraction [8].

The figure of merit to optimize such a system is the signal-to-noise ratio enhancement. Unfortunately a high figure of the signal-to-noise ratio enhancement does not mean that good speech quality is guaranteed because the subjective evaluation has proven to be not highly correlated with the signal-to-noise ratio enhancement figure. Another drawback of spectral subtraction techniques is that residual noise known as musical tones is corrupting the processed speech signal which makes it sound very unnatural. Furthermore a speech pause detector is necessary to initiate the estimation of the noise power density in speech pauses.

To overcome the latter drawbacks, array-based systems [3] have been proposed in literature. They are consisting of a minimum of two microphones and deliver in parallel an estimate of the power density spectrum of the corrupted speech and the noise. Thus there is no need for a speech pause detector and in addition the residual noise components are reduced in comparison to spectral subtraction. Nevertheless, also in this case the figures of signal- to-noise ratio enhancement are not very high and in some cases the speech quality does not prove to be satisfactory due to a specific background noise situation.

Thus there is need for new techniques which take into account the peculiarities of the human ear. One of the effects of hearing physiology that can be used for speech signal enhancement are masking effects. The idea is that the speech signal has to be processed in such a way that a listener evaluates the resulting speech signal to be of high quality. How this can be done will be explained in more detail in the sequel. But first the classical Wiener filter approach for speech signal enhancement has to be presented because it acts as a reference for the new system.

2 Wiener Filters for Speech Signal Enhancement

This is a quick review of the classical signal estimation technique based on Wiener filtering [2] applied to speech signal enhancement. Assuming that the power density function of the sampled input speech signal is given by $S_{SS}(e^{j\Omega})$ and the power density function of the sampled corrupting noise is given by $S_{NN}(e^{j\Omega})$ the transfer function of the Wiener Filter of the infinite lag

type is given by:

$$H(e^{j\Omega}) = \frac{S_{SS}(e^{j\Omega})}{S_{RR}(e^{j\Omega})} = \frac{S_{SS}(e^{j\Omega})}{S_{SS}(e^{j\Omega}) + S_{NN}(e^{j\Omega})} \quad (1)$$

where it is assumed that the noise process $N(k)$ is added to the speech signal process $S(k)$ resulting in the corrupted speech signal process

$$R(k) = S(k) + N(k). \quad (2)$$

Unfortunately in real world situations the pure speech signal process is not available so that the power density $S_{SS}(e^{j\Omega})$ in the numerator of the expression for the transfer function $H(e^{j\Omega})$ cannot be calculated. Furthermore, only estimates of the power density spectra can be derived due to the fact that both processes, the speech and the noise process, are instationary. Thus an estimate can be calculated only within a finite time interval. During speech activity an estimate of the power density of the corrupted speech signal is given by the squared magnitude of the spectrum calculated by the fast Fourier transform (FFT):

$$\hat{S}_{RR}(n) = |R(n)|^2, \quad 0 \leq n \leq N-1 \quad (3)$$

where $R(n)$ is the sample of the spectrum $R(e^{j\Omega})$ at the frequency index $n = \Omega \cdot N/2\pi$. $R(n)$ is calculated by the FFT from the time sequence $r(k), 0 \leq k \leq N-1$ of the corrupted speech signal within the actual observed time interval. During speech pauses an estimate of the noise power density can be determined and is updated during the following speech pause indicated by the index i:

$$\hat{S}_{NN}(n,i) = \beta \cdot \hat{S}_{NN}(n,i-1) + (1-\beta) \cdot |N(n,i)|^2 \quad (4)$$

with β being an update constant depending on the stationarity of the noise process $N(k)$. If the noise is highly stationary β is close to 1, otherwise it is close to 0. From eqn. (3) and (4), an estimate of the power density of the uncorrupted speech signal can be calculated using eqn. (2):

$$\hat{S}_{SS}(n) = \hat{S}_{RR}(n) - \hat{S}_{NN}(n,i). \quad (5)$$

With the estimates $\hat{S}_{SS}(n)$ and $\hat{S}_{NN}(n,i)$ the transfer function of the Wiener filter with the two parameters a and b can be given by:

$$H(n) = \begin{cases} \frac{\hat{S}_{RR}(n) - a \cdot \hat{S}_{NN}(n,i)}{\hat{S}_{RR}(n)} & \frac{\hat{S}_{RR}(n) - a \cdot \hat{S}_{NN}(n,i)}{\hat{S}_{RR}(n)} \geq b \\ b & \text{elsewhere.} \end{cases} \quad (6)$$

In this equation a is the so-called overestimate factor which is used to control the uncertainty about the fact whether the estimate $\hat{S}_{NN}(n,i)$ corresponds with the noise power density imposed on the actual speech signal. This uncertainty is due to the fact that the noise power has been calculated in the past speech pause interval and the noise might have been changed in its character till the actual time epoch. Typical values for the parameter a are in the interval $2 \leq a \leq 4$.

The upper expression in eqn. (6)

$$\frac{\hat{S}_{RR}(n) - a \cdot \hat{S}_{NN}(n,i)}{\hat{S}_{RR}(n)} \quad (7)$$

is valid only if it is larger or equal to b where b is a positive value known as the spectral floor and ranging between 0.1 and 0.3. Because of a mismatch between the estimates of the power densities of the corrupted speech signal and the noise, the expression given in eqn. (7) might result in a value which is less than b. This is in contradiction with the real world because this expression is a power density which is always positive by definition. The constant b represents the minimum power of the transmitted noise. It is made sure by this expression that a minimum power unequal zero is transmitted so that the far-end listener never has the impression that the transmission path is interrupted.

Because the calculation in eqn. (6) is executed for each frequency index n individually, it might happen that isolated spectral lines determine the transmitted signal. This results at the output of the system in sinusoidal signals which are switched on and off and produce a noise signal known as musical tones. To avoid this effect, many modifications of the transfer function given in eqn. (6) are published in literature. Recently some non-linear modifications of this expression have been published [6].

To evaluate the result of the noise reduction by a Wiener filter the signal-to-noise ratio enhancement is used. This expression is defined by:

$$SNRE = \frac{SNR_O}{SNR_I} \quad (8)$$

with SNR_O the signal-to-noise ratio at the output and SNR_I the signal-to-noise ratio at the input of the system. If Wiener filters are applied to mobile communication systems used in cars, $SNRE$ ranges around 5 dB.

Unfortunately this figure does not correlate significantly with the subjective evaluation of a listener, i.e. high scores do not mean that a listener qualifies the processed speech signal to be intelligible, to sound naturally etc. Since no subjective criteria are known which can be calculated easily from the parameters of

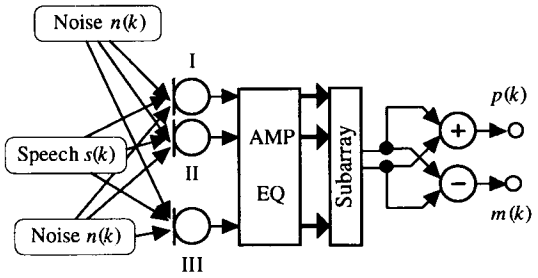

Figure 1: Microphone array consisting of three microphones

the estimated speech signal and thus can be used to optimize a noise reduction system, $SNRE$ is normally used for performance evaluation.

There are two drawbacks of spectral subtraction systems: first the power spectra of the speech and the noise process are not estimated for the same time epochs, second the design does not take subjective criteria based on hearing physiology into account.

First, a front end for noise reduction systems will be presented which delivers simultaneously estimates of the spectral densities of the enhanced speech signal and the corrupting noise. For this purpose a minimum of two microphones and a time delay operation to synchronize the speech at the output of the two microphones is necessary [3]. Alternatively, three microphones which are positioned appropriately and which do not need a costly time delay compensation can be used [4].

3 Estimation of the Speech and Noise Power Densities

Since both, speech and noise, are instationary processes, a simultaneous estimation of the power spectra of both processes should be implemented in a noise reduction system based on spectral subtraction. Approaches which fulfil this requirement are known from literature [3] and are based on microphone arrays. The disadvantage of most of these systems is that they require a time delay compensation of the speech signals carried in the microphone paths which needs expensive real-time hardware for implementation. A robust and cheap approach is based on three microphones as given in Fig. 1.

The distance of the microphones is 5 cm between microphone I and II and 10 cm between microphone II and III, respectively. First the output signals are amplified (AMP) and equalized (EQ) to adapt the signal power, to compensate the transfer paths from the speaker's mouth to the microphones, and to install a static delay compensation of the speech signals conveyed in the paths. This is when the system is initiated by a set-up procedure. In the subarray the output of the microphones I and III is filtered by low pass filters, the output of the microphones II and III by band pass filters, and the output of the microphones I and II by high pass filters. By this the difference between the travel time of the speech signals from the speakers mouth to a pair of microphones multiplied by the center frequency of the associated filters remains more or less the same if the speaker does not move his mouth too far away from the position which he had during the set-up procedure. The output of one low pass filter, one band pass filter, and one high pass filter are added together to form a new full-band signal. The same synthesis is executed with the other subband signals so that finally two full-band signals are available at the output of the subarray. With this subarray approach there is no need for dynamic delay compensation.

Pairs of the preprocessed signals are added and subtracted. The summation channel carries the speech signal $p(k)$ which is enhanced maximally by 3 dB, whereas in the difference channel the equivalent noise signal $m(k)$ is available. From these outputs estimates of the power spectrum $\hat{S}_{RR}(n)$ of the enhanced speech signal and $\hat{S}_{NN}(n)$ of the noise can be calculated as in eqn. (3):

$$\hat{S}_{RR}(n) = |P(n)|^2, \quad 0 \leq n \leq N-1 \qquad (9)$$

$$\hat{S}_{NN}(n) = |M(n)|^2, \quad 0 \leq n \leq N-1. \qquad (10)$$

Using these entities a Wiener filter according to eqn. (6) can be calculated. In this case the estimate for the power density of the noise process is not calculated and updated in the time epoch with index i as in eqn. (4) which is not identical to the time interval in which the estimate of the power spectrum of the speech signal is determined. Instead, both signals, the speech signal and the corresponding noise signal, are available at the same time instant.

4 Phenomena of Hearing Physiology

For noise reduction the masking effect of the human hearing system can be exploited. Masking means that a tone of higher intensity masks a tone of lower intensity so that the latter one is not audible. Whether or not this effect is active depends on the frequency of the tone, the frequency distance of the two tones, their relative intensity, and their spectral character. A

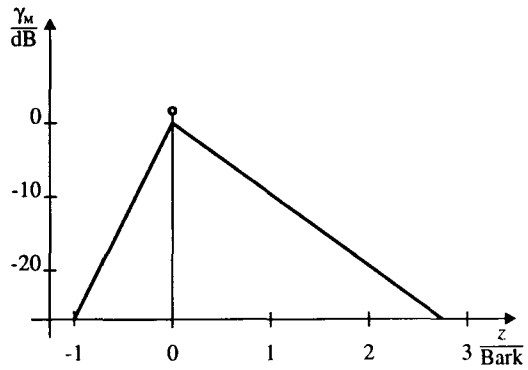

Figure 2: Model of the monitor threshold γ_M measured in dB as a function of the tonality z measured in Bark

parameter describing this effect is the monitor threshold γ_M which also depends on all the items mentioned above.

The total frequency band analysed by the human ear stretches from roughly 20 Hz up to 16 kHz and is cut into 24 subbands logarithmically spaced along the frequency axis. By this a new frequency measure called tonality z [9] is introduced with the unit 1 Bark for one of these subbands. For speech input systems used for telephone applications the frequency band from 300 Hz up to 3400 Hz is used into which fall 14 of these subbands or 14 Bark.

The masking effect depends on the fact into which and into how many of these subbands the masking signal falls. The simplest case is given if the masking signal is a sine wave or a narrow-band noise process covering just one of the subbands. Even in these cases the masking effect depends on the frequency measured in Bark and the intensities of the masking and the masked signal, respectively. For simplicity all these influences are omitted in the sequel, and a simple model will be used to describe the masking effect in the telephone channel. This model is given by Fig. 2 in which the level of the monitor threshold γ_M measured in dB is plotted as a function of tonality z measured in Bark.

The value 0 Bark corresponds with the frequency of the masking signal and the slopes right and left of 0 Bark determine the level of the masked signal. The level of the masked noise has to be 2 dB below the speech signal at 300 Hz if the noise is masked by a speech signal of the same frequency. At 3 kHz this level is increased to 4 dB. Independent of the absolute level of the noise and the speech, the type of noise and speech, and the absolute frequency the slope of the monitor threshold is 27 dB/Bark for noise signals with lower frequency than the speech signal and the decay is -10 dB/Bark for higher frequencies.

Besides the static masking effect described here there is also a dynamic masking effect. In this case the levels of consecutive components of speech and noise determine whether a masking effect appears or not. If a speech signal of high intensity is applied to the ear and immediately afterwards a noise signal of lower intensity enters the ear, the noise will not be audible. This situation can be described to be a short-time deafness. This effect will not be used in the sequel.

5 The Masking Filter for Speech Signal Enhancement

Exploiting the masking effect as described in the preceding section the transfer function of the noise reduction system called masking filter can be calculated in a first approach similar to that of the Wiener filter given in eqn. (6):

$$H(n) = \begin{cases} 1 & \text{if } \hat{S}_{RR}(n) \geq \gamma_M \cdot \hat{S}_{NN}(n) \\ b & \text{if } \hat{S}_{RR}(n) \geq \gamma_M \cdot \hat{S}_{NN}(n). \end{cases} \quad (11)$$

The estimates of the power spectra are calculated as in eqn. (9) and (10) and the monitor threshold γ_M is determined according to the rules derived in the preceding section. To improve the estimation of the power spectra given in eqn. (9) and (10), for the Wiener filter approach often an exponential average over a couple of data blocks of length N is executed. This is not helpful for the calculation of the transfer function in eqn. (11) since the masking effect is based on sudden changes in the level of the noise and the speech, respectively.

In contrast to the Wiener filter approach in eqn. (6) where the input signal is attenuated by a factor depending on the specific signal-to-noise ratio of the observed spectral line with the index n, the input signal is passing the transfer function in eqn. (11) without any attenuation if the monitor threshold γ_M is passed. This is done because the noise component corrupting the spectral line with index n is masked by the speech component. Thus the spectral floor b can be set to a value which is smaller than in eqn. (6). This design philosophy is similar to the one applied to data rate compression techniques for high quality music signals [7].

5.1 Determination of the Monitor Threshold

The estimates for the power spectra used in eqn. (11) are calculated by the FFT with N samples, i.e. a finite

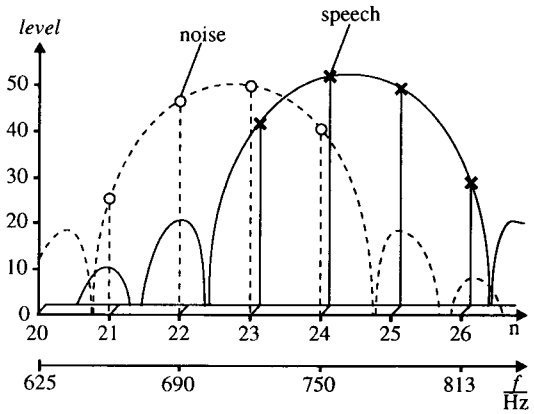

Figure 3: Speech and noise represented by spectral lines and their mapping into the FFT grid

observation interval. This results in the well-known leakage effect [1] which broadens up the spectrum of an analysed sine wave, e.g. To reduce this effect, windows such as the Hanning window are applied. In the case of a Hanning window four spectral lines fall into the main lobe so that the power of the original spectral line is distributed over these four spectral lines. Therefore it is not clear whether the monitor threshold given in Fig. 2 is still valid.

How this question can be answered is discussed by the example given in Fig. 3. It is assumed that the speech signal is modelled by the sinusoid at 758 Hz and the noise by the component at 711 Hz, both having the same level of 50 dB. The figure shows the spectral lines within the grid n of the FFT together with the spectral representation of the Hanning window. The frequency difference is 47 Hz and since both frequencies fall into the subband stretching from 630 Hz to 770 Hz being equal to 1 Bark these 47 Hz correspond to

$$\frac{758\text{ Hz} - 711\text{ Hz}}{770\text{ Hz} - 630\text{ Hz}} = \frac{47\text{ Hz}}{140\text{ Hz}} \hat{=} 0.336\text{ Bark}. \quad (12)$$

Observing that the speech component is of higher frequency than the noise component one can read from Fig. 2 that the monitor threshold is roughly 7 dB. With this value the noisy spectral lines with index $n = 21$ till $n = 23$ are set to b according to eqn. (11) since they are below the monitor threshold whereas the spectral lines with index $n = 24$ till $n = 26$ pass the noise reduction system without any attenuation. The level of the speech spectral lines at these frequencies is 49 dB, 46 dB and 31 dB, whereas the level at frequency index $n = 23$ was 42 dB. This results in an attenuation of the speech signal of

$$\frac{10^{49/20} + 10^{46/20} + 10^{31/20}}{10^{42/20} + 10^{49/20} + 10^{46/20} + 10^{31/20}} = 0.804 \hat{=} -1.89\text{ dB} \quad (13)$$

so that the level of the speech signal after noise reduction is given by

$$50\text{ dB} - 1.89\text{ dB} = 48.11\text{ dB}. \quad (14)$$

Before noise reduction the noisy spectral lines were of the same level as the speech spectral lines. Therefore the attenuation calculates to be

$$\frac{10^{42/20}}{10^{42/20} + 10^{49/20} + 10^{46/20} + 10^{31/20}} = 0.196 \hat{=} -14.16\text{ dB} \quad (15)$$

which results in the noise level

$$50\text{ dB} - 14.16\text{ dB} = 35.84\text{ dB}. \quad (16)$$

Taking into account that the speech signal at 758 Hz has to exceed the noise level at the same frequency by 3 dB and that the corrupting noise signal has a distance of 0.336 Bark and that its frequency is below that of the speech signal the maximum level of the masked noise is given by

$$48.11\text{ dB} - 3\text{ dB} - 0.336\text{ Bark} \cdot 27\text{ dB/Bark} = 36.04\text{ dB}. \quad (17)$$

Since the actual noise level is 35.84 dB by eqn. (16), the noise will be masked in this case and the signal enhancement is equal to 48.11 dB - 35.84 dB = 12.27 dB.

This example is far from reality because the noise normally will be broadband and the speech signal is not just one sine wave but at least a combination of sine waves in case of a vowel. In these more realistic cases the calculation of the masking effect is quite tedious and cannot be done without a computer program. Therefore a program was written to simulate up to 18 spectral lines for the speech signal and 22 spectral lines for the noise signal. Since the sampling frequency is $f_s = 8$kHz for simulations of the telephony channel and the block length of the FFT was chosen to be $N = 256$ the spacing between spectral samples is 31.25 Hz. A frequency band covering 16 spectral samples has been chosen for analysis which is equivalent to a bandwidth of $16 \cdot 31.25$Hz = 500Hz. This restriction of the bandwidth is due to the fact that spectral representations of vowels are at least 150 Hz

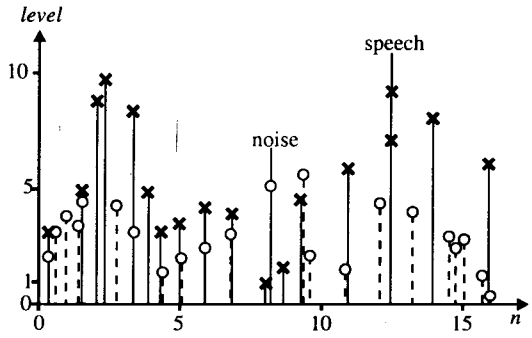

Figure 4: Simulation of a spectrum for a speech signal corrupted by noise

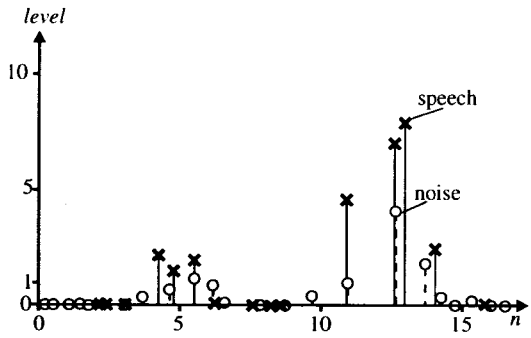

Figure 5: Filtered version of the spectrum given in Fig. 4 with a monitor threshold of 8.9 dB

apart from each other and cover less or equal 500 Hz. Furthermore the masking effect given in Fig. 2 covers a frequency band of roughly 3.5 Bark which is equivalent to a maximum of 300 Hz within the telephone channel.

By this simulation it has been shown that the monitor threshold γ_M has to be set to 10.4 dB at low frequencies around 300 Hz and to 8.9 dB for high frequencies around 3 kHz. For a block of 10 spectral samples each a value of the monitor threshold has been determined and between these blocks the threshold values have been interpolated linearly. The signals have been restricted to the telephone channel with 300Hz $\leq f \leq$ 3400Hz which yields with the sampling frequency f_s = 8kHz n = 100 spectral lines. The resulting values of the monitor threshold γ_M are summed up in Tab. 1.

Table 1: Monitor threshold γ_M as a function of frequency

frequency Hz	frequ. index n	γ_M dB	γ_M
312.5	10	10.4	11.96
625	20	10.25	11.59
937.5	30	10.1	11.23
1250	40	9.95	10.89
1562.5	50	9.8	10.55
1875	60	9.65	10.23
2187.5	70	9.5	9.91
2500	80	9.35	9.6
2812.5	90	9.1	9.13
3125	100	8.9	8.76

Fig. 4 shows a simulation example with randomly chosen speech and noise components with a level scaled from 0 to 10 and frequency components with a minimum distance of $0.1 \cdot 31.25\text{Hz} = 3.125\text{Hz}$.

Applying the masking filter defined in eqn. (11) with the spectral floor $b = 0$ and using the experimentally derived values for the monitor threshold given in Tab. 1 results in the spectrum plotted in Fig. 5. As can be seen the noise is masked totally.

So far the monitor threshold was derived using synthetic data. For test purposes the design rules of the masking filter given in this section have been applied to real-world data which is described in the following section.

5.2 Test of the Masking Filter

It is assumed that the noise reduction system is used for mobile communication in a car. Therefore speech data of one female and three male speakers as well as noise signals generated by traffic, the engine of the car etc. have been picked up by the microphone array described in section 3. The signal-to-noise ratio SNR_I in a car typically ranges from -5 dB up to 10 dB. As a typical case $SNR_I = 0$ has been taken for test purposes, and the spectral floor has been set to $b = 0.2$. For spectral analysis a block length of $N = 256$ and a four-fold overlap has been chosen so that every 8 ms a data block is analysed. For these parameters Fig. 6 shows the result of the masking filter for $l = 40$ blocks or a time interval of 0.32 s of the input signal. The blocks with index l are plotted as a function of the frequency index n. The marks represent the time instants and frequency values at which the input signal passes the filter without any attenuation.

The pattern is characterized by short chains of marks being often interrupted on one hand, and isolated marks on the other. Because of the fourfold overlap a frequency component at the input results in maximally four marks in the plot. Isolated marks and short chains thus are either caused by noise or by artefacts of the filtering process. Taking into account that the duration of a vowel is typically 200 ms and that of plosives is 10 to 20 ms then this result is not

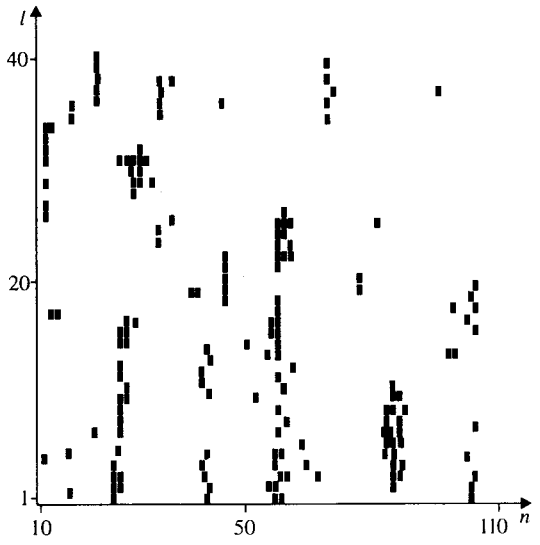

Figure 6: Marks representing time-frequency passes of the masking filter

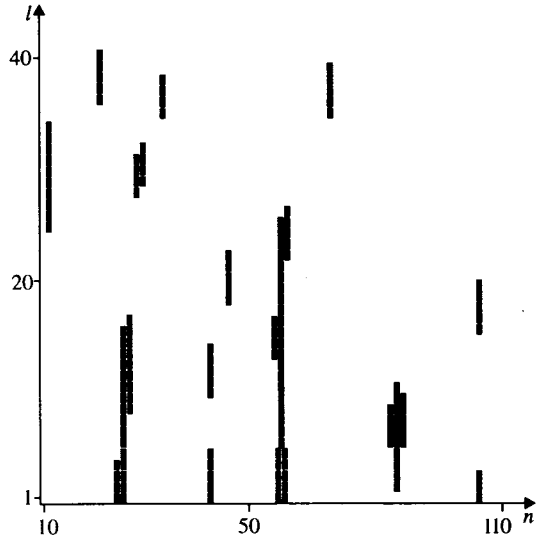

Figure 7: Output of the post-processor with the input given in Fig. 6

satisfying. This impression is underlined by informal listening tests which showed many musical tones as residual noise.

For comparison the enhancement defined in eqn. (8) for the Wiener filter given in eqn. (6) and the masking filter in eqn. (11) has been calculated. The Wiener filter yields $SNRE = 4.78$dB whereas the improvement by the masking filter is $SNRE = 7.05$dB and the speech signal component at the output of the masking filter is attenuated by 0.69 dB compared to the Wiener filter.

To improve this result a post-processor has been designed which extracts short interruptions of chains of up to two missing marks. Furthermore if there are at least three marks, a fourth one is added. Thus the incoming chain $\ldots b111b1bb\ldots$ with one missing mark is replaced by $\ldots b1111111b\ldots$ and the chain $\ldots b111bb11bb\ldots$ with two missing marks followed by at least two marks is replaced by $\ldots b111111111b\ldots$. If the input signal as specified in Fig. 6 is applied to the post-processor the output is given by Fig. 7.

Obviously the pattern is much less random and longer chains representing vowels can be detected. Almost all of the original noise is rejected and only few musical tones are disturbing the processed speech. The listening test yields a much better subjective impression. The signal enhancement has improved to $SNRE = 9.48$dB but the speech level is reduced by 0.17 dB in comparison to the result given in Fig. 6. The noise level was reduced by 11.1 dB which is not too far away from the maximum reduction of $20 \cdot \log 0.2 \hat{=} 14$dB which is caused by the spectral floor $b = 0.2$.

6 Conclusion

The application of the masking filter to real-world scenarios has shown that phenomena of the human hearing system can be exploited for noise reduction. The noise can be reduced significantly without distorting the speech signal too much. This is in contrast to the Wiener filter which attenuates all of the spectral components more or less.

The subjective evaluation of the system has shown that it is recommended to use the spectral floor b even if with $b \neq 0$ the figure of merit $SNRE$ is less than for $b = 0$ because then the musical tones are reduced.

For low signal-to-noise ratios SNR_I a combination of the Wiener filter and the masking filter might be preferable because then the pure masking filter would cut away too many spectral lines and the processed speech would sound quite unnatural. For the selection of the parameters subjective listening tests would be most appropriate in this case.

There is one important disadvantage of the masking filter: the time delay. In the implementation in section 5.2 the delay was two data blocks or 16 ms. A longer observation of the past might have improved the speech quality but was not investigated because of the longer delay. If such a system is used in two-way communication the delay is annoying for the customers and hurts regulations of communication companies.

Another improvement of the noise reduction sys-

tem could be achieved by exploitation of the dynamic masking effect mentioned in section 4. But this would increase further the processing time so that ITU recommendations might prevent this approach.

Acknowledgement

The author thanks the Deutsche Telekom AG for the support of a part of this work by a research grant. Furthermore he wants to thank Gunther Muth, Keld Lange and Marc Ihle who significantly have contributed to the results presented in this paper.

References

[1] Kammeyer, K.D.; Kroschel, K.: *Digitale Signalverarbeitung. Filterung und Spektralanalyse. Teubner, Stuttgart 1992*

[2] Kroschel, K.: *Statistische Nachrichtentheorie, Signalschaetzung. 2nd ed. Springer, Heidelberg 1988*

[3] Kroschel, K.: *Arraytechniken zur Geraeuschreduktion bei Sprachuebertragung. Proc. Elektronische Sprachsignalverarbeitung. Ed.: Fellbaum, K.-R., Berlin 1990, pp. 102-110*

[4] Kroschel, K.; Lange, K.: *Subband array processing for speech enhancement. Proc. Eurospeech'93, Berlin 1993, pp. 621-624*

[5] Kroschel, K.: *Speech input systems for multimedia applications. Annual Report of the French German Institute for Automation and Robotics, Karlsruhe 1995, to be published*

[6] Lockwood, P.; Boudy, J.: *Experiments with a non-linear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars. Speech Communication 11 (1992), pp. 215-218*

[7] Theile, G.; Stoll, G.; Link, M.: *Low bit rate coding of high-quality audio signals (MUSICAM). EBU Technical Review, no. 230, August 1988, pp. 71 -94*

[8] Vary, P.: *Verfahren zur digitalen Verbesserung gestoerter Sprache. TEKADE, Technische Mitteilungen 1983, pp. 70-76*

[9] Zwicker, E.: *Psychoakustik. Springer, Berlin 1982*