A NOVEL PSYCHOACOUSTICALLY MOTIVATED AUDIO ENHANCEMENT ALGORITHM PRESERVING BACKGROUND NOISE CHARACTERISTICS

Stefan Gustafsson, Peter Jax and Peter Vary

Institute of Communication Systems and Data Processing,
RWTH Aachen, Templergraben 55, D-52056 Aachen, Germany
E-mail: {gus, jax}@ind.rwth-aachen.de

ABSTRACT

In this paper we propose an algorithm for reduction of noise in audio signals. In contrast to several previous approaches we do not try to achieve a complete removal of the noise, but instead our goal is to preserve a pre-defined amount of the original noise in the processed signal. This is accomplished by exploiting the masking properties of the human auditory system.

The speech and noise distortions are considered separately. The spectral weighting rule, adapted by utilizing only estimates of the masking threshold and the noise power spectral density, has been designed to guarantee complete masking of distortions of the residual noise.

Simulation results confirm that no audible artifacts are left in the processed signal, while speech distortions are comparable to those caused by conventional noise reduction techniques. Audio demonstrations are available from http://www.ind.rwth-aachen.de.

1. INTRODUCTION

The enhancement of noisy speech has gained an increasing interest in recent years. This is mainly due to the driving forces in the area of mobile communications, where speech enhancement algorithms could be integrated in e.g. hands-free telephony devices. The availability of powerful digital signal processors is also supporting these activities.

Until some years ago, noise reduction algorithms were in general based upon some form of spectral subtraction [1, 8]. The drawback of these methods is that a very unpleasant residual noise in form of musical tones remains in the processed signal, and that the speech is distorted. Some algorithms have partly met this problem utilizing modified weighting rules and a more advanced estimation of the momentary signal-to-noise ratio (SNR) [3].

Common to all algorithms is, however, that their performance strongly depends on how well the power spectral density (PSD) of the noise can be estimated. The better the estimation is, the more natural the residual noise sounds (with fewer musical tones) and the lower the distortion of the speech is. As discussed in e.g. [2], the estimation of the momentary SNR as presented in [3] is a key factor for the improved results. The SNR estimation procedure tends to reduce the susceptibility of the weighting rule to estimation errors.

Still in an embryonal stage are speech enhancement methods relying on psychoacoustical considerations. Most contributions in this area exploit the masking properties of the auditory system. In principle, they make use of various linear or nonlinear weighting rules, which are adjusted according to the noise masking threshold to reduce solely the audible part of the noise power spectrum [9] or to find the best tradeoff between noise reduction and speech quality [10]. The results were reported to be better under consideration of the perceptual properties, though some musical tones as well as distortions of the speech were still audible.

In this paper we discuss a new approach to speech enhancement using masking properties. In contrast to previous methods, the proposed one does not use the masking threshold to modify a standard spectral weighting rule, but uses it in a direct manner to calculate the weighting coefficients, such that the perceived noise suppression will always be equal to a predefined level. Actually, in this process the distortion of the speech is not explicitly considered.

2. PSYCHOACOUSTICAL SPECTRAL WEIGHTING

The perception of an audio signal is the result of various physiological and psychological effects, which are not fully understood yet. Nevertheless some models to describe these effects have been developed in the past [11]. Especially the known phenomenon of auditory masking has been exploited successfully in signal processing systems, e.g. in the field of wideband audio coding [5, 6]. For this purpose a model of the auditory system is used to calculate a spectral masking threshold. A human listener will not perceive any additive signal components as long as their power spectral density lies completely below the masking threshold. It must, however, be emphasized that conclusions about the subjective perception of partially masked signals can not be easily drawn from the knowledge of the masking threshold alone.

In most situations a complete removal of the noise is neither necessary nor desirable. In a telephone application, for example, a retained low-level natural sounding background noise will give the far end user a feeling of the atmosphere at the near end and also avoids the impression of an interrupted transmission. Consequently, it is only desired to reduce the noise level by a pre-defined amount. However, in this step the spectral characteristics (i.e. the colour) of the noise shall be maintained. With this motivation we define a residual noise level $\zeta = 10^{-NR/20}$, where *NR* is the desired noise reduction in dB.

2.1. Definition of Speech and Noise Distortions

Now, in terms of short-time spectral analysis, let $S(\Omega_i)$ and $N(\Omega_i)$ denote the discrete and complex Fourier transformations of the speech $s(k)$ and the additive, statistically independent noise $n(k)$, respectively, with $\Omega_i = 2\pi \frac{i}{M}, i \in \{0, 1, \ldots, M-1\}$. Then the

0-7803-4428-6/98 $10.00 © 1998 IEEE

desired output signal $\tilde{S}(\Omega_i)$ and the noise $\tilde{N}(\Omega_i)$ to be suppressed are

$$\tilde{S}(\Omega_i) = S(\Omega_i) + \zeta N(\Omega_i) \quad (1)$$

$$\tilde{N}(\Omega_i) = (1-\zeta) N(\Omega_i), \quad (2)$$

with PSDs

$$R_{\tilde{s}}(\Omega_i) = R_s(\Omega_i) + \zeta^2 R_n(\Omega_i) \quad (3)$$

$$R_{\tilde{n}}(\Omega_i) = (1-\zeta)^2 R_n(\Omega_i), \quad (4)$$

where $R_s(\Omega_i)$ and $R_n(\Omega_i)$ are the PSDs of the speech and noise, respectively. The estimation $\hat{S}(\Omega_i)$ of the speech is the result of multiplying the filter input $S(\Omega_i) + N(\Omega_i) = \tilde{S}(\Omega_i) + \tilde{N}(\Omega_i)$ with a real-valued weighting function $H(\Omega_i)$ limited to $0 \leq \zeta \leq H(\Omega_i) \leq 1$. The PSD of the actual output signal is then

$$R_{\hat{s}}(\Omega_i) = H^2(\Omega_i)(R_s(\Omega_i) + R_n(\Omega_i)). \quad (5)$$

The estimation error $E(\Omega_i)$ is the difference between the estimated speech $\hat{S}(\Omega_i)$ and the desired output signal $\tilde{S}(\Omega_i)$. The PSD of the error can be expressed as the sum of two components, $R_E(\Omega_i) = R_{E_s}(\Omega_i) + R_{E_n}(\Omega_i)$, where

$$R_{E_s}(\Omega_i) = R_s(\Omega_i)(H(\Omega_i) - 1)^2 \quad (6)$$

$$R_{E_n}(\Omega_i) = R_n(\Omega_i)(H(\Omega_i) - \zeta)^2. \quad (7)$$

The component $R_{E_s}(\Omega_i)$ describes the distortion of the speech, which will be minimized by choosing the weighting factor $H(\Omega_i)$ to 1. The other component $R_{E_n}(\Omega_i)$ can be interpreted as the difference between the desired noise power and the actual noise power and is minimized by a weighting factor $H(\Omega_i) = \zeta$. We call $R_{E_n}(\Omega_i)$ the distortion of the residual noise – recall that the PSD of the desired residual noise is exactly $\zeta^2 R_n(\Omega_i)$. As both distortion components are quadratic functions of $H(\Omega_i)$, the minimal total error $R_E(\Omega_i)$ will be found for some $H^{opt}(\Omega_i)$ in $\zeta \leq H^{opt}(\Omega_i) \leq 1$, compare Fig. 1. In fact, with $\zeta = 0$ this $H^{opt}(\Omega_i)$ is equal to the Wiener-filter solution [4].

2.2. Development of a Psychoacoustically Motivated Spectral Weighting Rule

For our problem of noise reduction the speech signal $S(\Omega_i)$ is taken as the masker. The distortions corresponding to $R_{E_s}(\Omega_i)$ and $R_{E_n}(\Omega_i)$ that are produced by the process of spectral weighting are then interpreted as additive signal components. In the ideal case both distortions should be masked and thus be inaudible to the listener. However, in most real cases a complete masking of all distortions cannot be guaranteed because it is not always possible to adjust the weighting vector $H(\Omega_i)$ in such a way that $R_E(\Omega_i)$ falls below the masking threshold $R_T(\Omega_i)$. This is due to the fact that the minimum of $R_E(\Omega_i)$ is greater than zero for non-trivial signals, as can be seen in Fig. 1. Hence we must say goodbye to the idea of a *perfect* noise reduction by means of spectral weighting.

Another possibility would be to ensure the best possible speech quality by trying to mask the speech distortions $R_{E_s}(\Omega_i)$. However, this method leads to the problem that the achieved noise reduction varies widely over time. During periods of speech inactivity other actions would be necessary, as the speech distortion is undefined in this case.

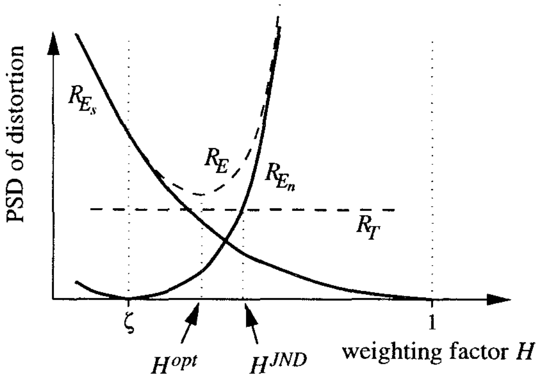

Figure 1: Components of the distortion that is produced by the process of spectral weighting. The scaling of the parabolas depends on the power of the speech and noise signal respectively.

A third option consists of trying to mask the distortions of the residual noise, but allowing a variable speech distortion. It is sufficient to keep $R_{E_n}(\Omega_i)$ exactly at the masking threshold, thereby minimizing the speech distortion. This solution is realized by choosing $H(\Omega_i)$ such that the PSD, $R_{E_n}(\Omega_i)$, of the difference between the desired and the actual noise level lies exactly at the masking threshold

$$R_{E_n}(\Omega_i) = R_n(\Omega_i)(H(\Omega_i) - \zeta)^2 \overset{!}{=} R_T(\Omega_i). \quad (8)$$

Solving this equation for $H(\Omega_i)$ with the constraint $\zeta \leq H(\Omega_i) \leq 1$ leads to the weighting rule

$$H^{JND}(\Omega_i) = \min\left(\sqrt{\frac{R_T(\Omega_i)}{R_n(\Omega_i)}} + \zeta, 1\right). \quad (9)$$

We call this weighting factor $H^{JND}(\Omega_i)$ (JND standing for Just Notable Distortion) and it is a function only of the calculated masking threshold $R_T(\Omega_i)$ and the noise PSD $R_n(\Omega_i)$, which both have to be estimated.

Fig. 1 helps to explain how the weighting factor is chosen. The two error components and the sum $R_E(\Omega_i)$ are plotted as a function of $H(\Omega_i)$ for a constant ζ and for some fixed frequency Ω_i. The scaling of the parabolas depends on the power of the speech and noise signal, respectively, and the level of the masking threshold depends on the speech PSD exactly at and in a neighbourhood of the frequency Ω_i. $H^{JND}(\Omega_i)$ is chosen at the crossing point of $R_T(\Omega_i)$ and $R_{E_n}(\Omega_i)$.

First consider a relatively strong speech and thus a high masking threshold. The weaker the noise is, the closer $H^{JND}(\Omega_i)$ will be to 1. In such situations $H^{JND}(\Omega_i)$ is often greater than $H^{opt}(\Omega_i)$, thus reducing the speech distortion compared to the Wiener rule. Actually the crossing point can be above $H(\Omega_i) = 1$. This means that the noise is already fully masked and no noise suppression has to be performed. Hence $H^{JND}(\Omega_i)$ will be set to one.

If, on the other hand, the speech is weak and therefore $R_T(\Omega_i)$ is low, then $H^{JND}(\Omega_i)$ might be well below $H^{opt}(\Omega_i)$ when a strong noise is present. This will lead to a larger distortion of the speech, but also to a stronger suppression of the noise. The unavoidable tradeoff is easily controlled by the factor ζ.

By applying this weighting rule, the resulting signal exhibits – in a perceptual sense – a constant noise reduction. Although the speech distortions are not explicitly considered by the weighting rule, the solution nonetheless reduces them to the smallest possible value for the specified amount of noise reduction: if a greater weighting factor $H(\Omega_i)$ is chosen, the distortions of the speech will be further reduced but at the same time the residual noise will become audible. If the weighting factor is set smaller, the speech distortions will rise without any improvement of the perceived noise reduction.

2.3. Avoidance of Musical Tones

Most algorithms performing some kind of spectral subtraction utilize the PSD, $R_{s+n}(\Omega_i)$, of the input signal explicitly in the weighting rule (e.g. [1, 2, 3, 8, 9, 10]). Even small errors in the estimation of the noise PSD will then have a major impact on the processed signal. The results will suffer from a residual noise in form of short sinusoids distributed over time and frequency (musical tones).

The proposed algorithm avoids such artifacts efficiently as the weighting factor calculation only implicitly uses the PSD, $R_{s+n}(\Omega_i)$, of the input signal in form of the masking threshold. Therefore, small estimation errors of the noise PSD will only have a minor impact on the output signal quality.

If no speech at all is present the masking threshold will be zero, from which $H^{JND}(\Omega_i) = \zeta$ follows, i.e. the processed signal is identical to the input signal up to the attenuation factor, thus preserving all characteristics and avoiding any artifacts.

3. IMPLEMENTATION

For evaluating the new weighting rule, it has been embedded into a noise reduction system as shown in Fig. 2, designed for a telephone application with sampling frequency $f_s = 8\text{kHz}$.

The basis of the system is the commonly known *filter bank overlap-and-add method*, here with a decimation ratio of $M = 128$, a frame length of $L = 256$, an FFT length of $N = 512$, and a Hamming window function $w(k)$ for input signal weighting. The estimation of the power spectral density of the noise is performed by a modified version of the algorithm proposed by Martin [7].

For the estimation of the masking threshold various methods have been proposed in the past (e.g. [5, 6]). Most of these algorithms share a similar model of the human auditory system. The estimation method implemented in the noise reduction algorithm described in this paper is a mixture of the Johnston and the ISO models.

For our noise reduction application, the chosen masking model raises the problem that the masking signal is not explicitly available in the system and thus has to be estimated. To do this we use a conventional noise reduction filter. Simulations have shown that the demands on the performance of this filter are quite low and hence a simple spectral subtraction can be used. The position of this spectral subtraction filter is shown in Fig. 2.

4. SIMULATIONS AND RESULTS

The new weighting rule has been compared with a wide range of conventional noise reduction techniques based on spectral subtraction, including some psychoacoustically oriented types. Input signals were generated by mixing a number of different test sequences spoken by female and male speakers with several noise signals of different spectral characteristics and at various signal to noise ratios (-10 ... 20 dB).

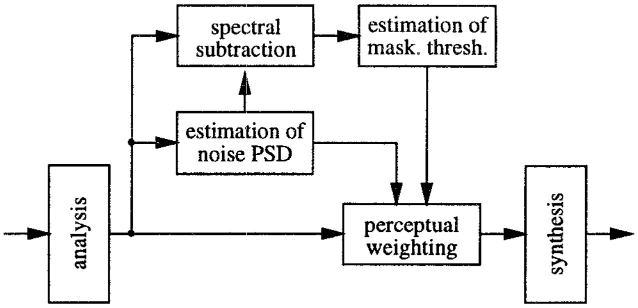

Figure 2: Overview of the proposed noise reduction system.

The noise attenuation NA is measured with

$$NA = 10\lg(P_n/P_{n'}), \quad (10)$$

where P_n is the mean noise power in the input signal and $P_{n'}$ is the mean noise power in the processed signal. In the same way the speech attenuation is defined as

$$SA = 10\lg(P_s/P_{s'}), \quad (11)$$

with P_s and $P_{s'}$ denoting the mean powers of the speech in the input and the processed signals, respectively. In Fig. 3 the noise and speech attenuation for the proposed algorithm and for a conventional one (MMSE STSA [3] modified according to [2] to preserve the pre-defined amount ζ of residual noise) are illustrated as a function of the input SNR. As predicted, the noise attenuation is somewhat lower using $H^{JND}(\Omega_i)$. For very low input signal SNR, the speech components suffer from a global attenuation, though no significant spectral distortion of the speech arises, and thus the effective noise attenuation $NA - SA$ is reduced. With $H^{JND}(\Omega_i)$ the effective noise attenuation actually decreases for SNR < -5 dB, which is no crucial disadvantage, since the performance of the noise PSD estimation procedure decays with lower SNR and then the noise attenuation should be performed more cautiously.

The reason for the attenuation of the speech is, that independent of the noise level the masking threshold for a given masker is fixed, and thus the amount of noise power which can be masked is constant. Therefore, if a strong noise is present, only a minor part of the noise will be masked and thus the weighting factor $H^{JND}(\Omega_i)$ will be close to ζ, leading to the attenuation of the speech. However, the attenuation can be reduced by choosing a greater ζ for input signals with very low SNR.

In addition to extensive informal listening tests, the degradation of the speech quality was assessed by different instrumental measures, e.g. SEGSNR, cepstral distance and basilar distance. The basilar distance is calculated as the mean difference between the basilar excitation of the original and the filtered speech signals. The speech distortions induced by the new weighting rule are comparable to those caused by MMSE STSA or similar methods based on spectral subtraction, see Fig. 4. The effect of the global attenuation as discussed above explains the fast increase of the basilar distance towards lower SNR.

The results of these instrumental evaluations should be interpreted with care, because they take psychoacoustical aspects only partially into account.

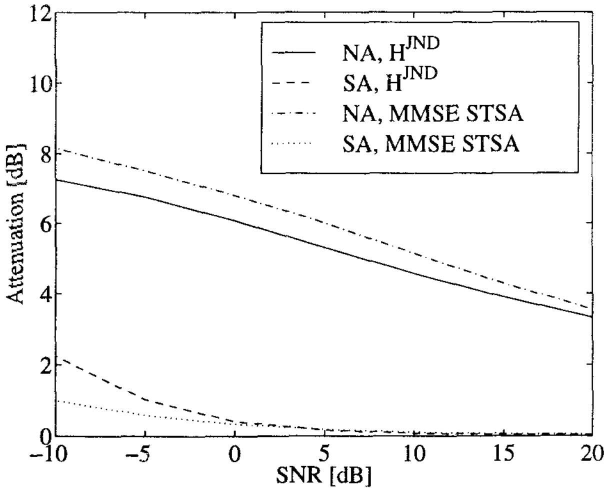

Figure 3: Noise attenuation (*NA*) and speech attenuation (*SA*), with $\zeta = 0.1$, for the new perceptual weighting rule $H^{JND}(\Omega_i)$ and a state-of-the-art conventional algorithm (MMSE STSA [3]) as a function of the input signal SNR.

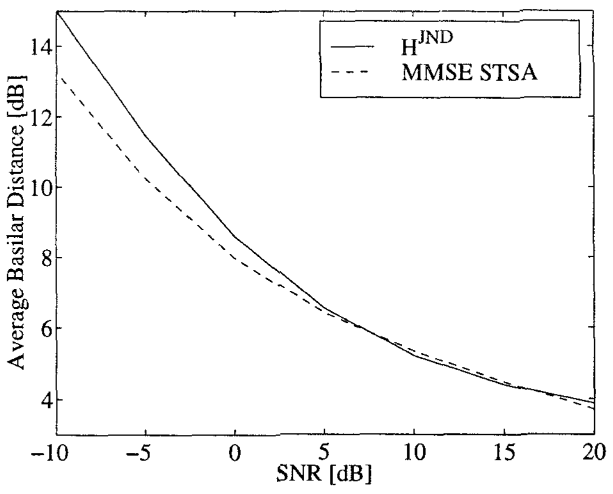

Figure 4: Average basilar distance between the original and the processed speech signal in dependence of the SNR of the input signal. Shown are the curves for the new perceptual weighting rule $H^{JND}(\Omega_i)$ and for the MMSE STSA algorithm.

Standard instrumental speech quality measures do not reflect the naturalness and subjective qualities of the residual background noise. Therefore we have relied on listening tests to assess such qualities. The comparison of the input signal mix with the processed signal reveals a perceived noise reduction in accordance to the predefined level ζ. The spectral character of the noise is hereby preserved, i.e. the processed noise sounds like a version of the original background noise which has been multiplied by the factor ζ. It should be stressed that no artifacts are audible.

These positive properties of the residual noise component are a consequence of the uncompromising design of the new weighting rule solely for masking the distortions of the residual noise. As described in section 2.3, this leads to a high robustness of $H^{JND}(\Omega_i)$ against estimation errors of the PSD of the noise.

5. CONCLUSIONS

The proposed method allows a noise reduction which is free of noise distortions and artificial sounds such as musical tones. Regarding speech distortions, the new algorithm behaves similar to conventional ones based on spectral subtraction. The global attenuation of speech ocurring at low SNR can be avoided by choosing a smaller noise reduction *NR*.

The algorithm has been found to be very robust against estimation errors of the noise PSD and the masking threshold. It also promises good results for suppression of speech-like disturbances. The application of residual echo reduction is currently examined.

6. REFERENCES

[1] S. Boll. "Suppression of Acoustic Noise in Speech Using Spectral Subtraction". *IEEE Transactions on Speech and Audio Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[2] O. Cappé. "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor". *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, April 1994.

[3] Y. Ephraim and D. Malah. "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 6, December 1984.

[4] S. Haykin. *Adaptive Filter Theory, 3rd Edition*, Prentice Hall, 1996.

[5] Draft Standard ISO 11172-3 MPEG Audio, London, November 1992.

[6] J. Johnston. "Transform Coding of Audio Signals Using Perceptual Noise Criteria". *IEEE Journal on Selected Areas of Communication*, vol. 6, pp. 314–323, February 1988.

[7] R. Martin. "Spectral Subtraction Based on Minimum Statistics". *Proc. EUSIPCO*, pp. 1182–1185, September 1994.

[8] R. McAulay and M.Malpass. "Speech Enhancement Using a Soft-Decision Noise Suppression Filter". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 2, pp. 137–145, April 1980.

[9] D. Tsoukalas, J. Mourjopoulos and G. Kokkinakis. "Improving the Intelligibility of Noisy Speech Using an Audible Noise Suppression Technique". *Proc. Eurospeech*, vol. III, pp. 1415–1418, 1997.

[10] N. Virag. "Speech Enhancement Based on Masking Properties of the Auditory System". *Proc. ICASSP*, pp. 796–799, 1995.

[11] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and Models*, Springer, New-York, 1990.