A NEW APPROACH TO NOISE REDUCTION BASED ON AUDITORY MASKING EFFECTS

Stefan Gustafsson and Peter Jax

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 an algorithm for the reduction of stationary noise in speech signals is proposed.

In contrast to the conventional trade-off between distortions of the speech and noise components the primary goal of the presented algorithm is to attenuate the noise by a pre-defined factor and to conceal artifacts and distortions of the residual noise by taking advantage of the auditory masking threshold. In this way it can be achieved that the spectral colour of the attenuated noise is the same as that of the background noise in the input signal.

Informal listening tests and instrumental evaluations with several quality measures have been performed to compare the new weighting rule with conventional approaches. Audio demonstrations are available from http://www.ind.rwth-aachen.de.

1. INTRODUCTION

Noise reduction is still a challenging field of signal processing. At present the need for robust and high-quality noise reduction algorithms even increases due to the widespread use of mobile speech communication systems such as mobile phones or hands-free telephone sets in noisy environments.

Common solutions to the problem of noise reduction mostly rely on spectrum-oriented approaches, e.g. by minimizing the mean squared error between the spectra of the output signal and the clean speech component [1, 2]. The ultimate goal of these methods is to identically reconstruct the spectrum of the original, undisturbed speech. Sophisticated algorithms incorporate models of the signal sources (i.e. speech and/or noise generation) to enhance estimation results and thereby improve signal quality (e.g. [3, 4, 5]). The main drawback of most of these spectrum-oriented methods is that annoying residual noise (so-called *musical tones*) often remains in the processed signal.

If the output signal is intended for a human listener, the properties of the human ear can be exploited. Although psychoacoustic effects can not always be described in exact mathematical terms, some quite reliable models have been developed in the past. The usage of such models makes it possible to derive noise reduction algorithms which achieve better results in a perceptual sense.

A number of methods relying on the masking properties of the human ear have been proposed in the last years (e.g. [6, 7, 8, 9]). The common idea behind these methods is that those parts of the background noise which are already masked by the speech do not have to be attenuated. Most of the algorithms use the additional masking information to control a conventional weighting rule which balances speech distortions versus residual noise.

The algorithm proposed in this paper concentrates on the properties of the residual background noise. The primary goal is that the noise in the output signal shall sound exactly like the input background noise — just attenuated by a constant factor. Speech distortions are considered only implicitely by choosing from the set of possible solutions fulfilling the primary goal that one which minimizes the speech attenuation. By concentrating on noise characteristics, annoying artifacts of the residual noise can be avoided completely.

2. PSYCHOACOUSTICS

The perception of audio signals is a quite complicated process involving a number of acoustical, physiological and psychological effects. Some of these effects are not investigated or fully understood yet. A general problem with auditory effects is that they differ between individuals and often depend on absolute sound pressures.

One property of the auditory system which has been widely exploited in several audio and speech processing algorithms is the *masking effect*. A model of the auditory system can be used to estimate a so-called *masking threshold*. The masking threshold at a particular frequency depends on the power spectral density (PSD) of the masker at the same and nearby frequencies as well as on the maskers noise- or tone-like character. An additional signal can be assumed to be inaudible to a human listener if its PSD is lower than the estimated masking threshold at *all* frequencies. In Fig. 1 the PSD of a speech frame and the calculated masking threshold are illustrated.

The masking effect is used e.g. in wideband audio coding applications [10] to reduce the perceivable effect of quantization noise for a given bitrate. In the noise reduction method presented in this paper, the masking effect is utilized to conceal artifacts and distortions of the residual background noise.

3. DERIVATION OF A WEIGHTING RULE

We will now derive and analyse a weighting rule for a frequency domain noise reduction system. It is motivated by the psychoacoustic masking effect described in the previous section. The algorithm has previously been introduced in [11].

In most situations a complete removal of existing background noise is not just unnecessary, but even undesired. For example, a low level of remaining natural sounding noise will give the user of a telephone set an impression of the atmosphere at the other end of the line, and will also avoid the impression of an interrupted connection. With this motivation we define the factor ζ as the desired noise attenuation. Relevant values of ζ lie in the range of $20 \lg \zeta = -20 \ldots -10$.

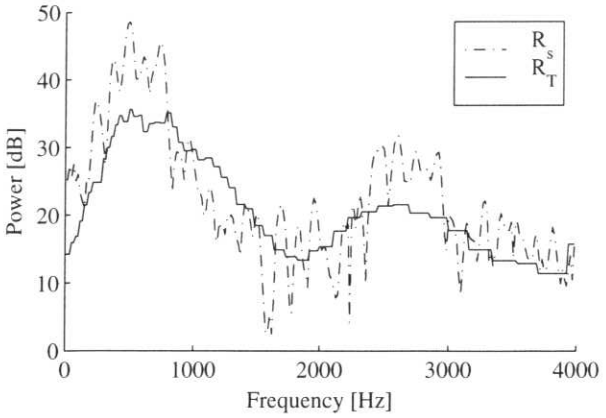

Figure 1: Example of an estimated masking threshold R_T for a given power spectral density R_s of the masker.

3.1. Weighting Rule Based on Auditory Masking

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

The desired output signal $\tilde{S}(\Omega_i)$ of the noise reduction system consists of the original speech, to which the noise multiplied with the factor ζ is added,

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

The actual output of the system $\hat{S}(\Omega_i)$ is, however, the sum of speech and noise multiplied with the real-valued weighting coefficients $H(\Omega_i)$,

$$\hat{S}(\Omega_i) = H(\Omega_i)[S(\Omega_i) + N(\Omega_i)]. \tag{2}$$

Let us define the error $E(\Omega_i)$ between the desired and the actual outputs as

$$\begin{aligned} E(\Omega_i) &= \tilde{S}(\Omega_i) - \hat{S}(\Omega_i) \\ &= [1 - H(\Omega_i)]S(\Omega_i) + [\zeta - H(\Omega_i)]N(\Omega_i). \end{aligned} \tag{3}$$

Because the speech and the noise are assumed to be uncorrelated, the power spectral density $R_e(\Omega_i)$ of the error $E(\Omega_i)$ can be split into two components,

$$\begin{aligned} R_e(\Omega_i) &= R_{e_s}(\Omega_i) + R_{e_n}(\Omega_i) \\ &= [1 - H(\Omega_i)]^2 R_s(\Omega_i) + [\zeta - H(\Omega_i)]^2 R_n(\Omega_i), \end{aligned} \tag{4}$$

where $R_s(\Omega_i)$ and $R_n(\Omega_i)$ are the PSDs of the speech and the noise, respectively.

The first term, $R_{e_s}(\Omega_i) = [1 - H(\Omega_i)]^2 R_s(\Omega_i)$, is the distortion of the speech and is minimized by choosing $H(\Omega_i) = 1$. The second term, $R_{e_n}(\Omega_i) = [\zeta - H(\Omega_i)]^2 R_n(\Omega_i)$, can be considered as the "distortion" of the residual noise, which is about the same as the difference between the actual noise power and the desired noise power, and is minimized by $H(\Omega_i) = \zeta$. Both $R_{e_s}(\Omega_i)$ and $R_{e_n}(\Omega_i)$ are quadratic functions of $H(\Omega_i)$, so the total error $R_e(\Omega_i)$ will be minimized for some $H^{opt}(\Omega_i)$ in $\zeta \leq H^{opt}(\Omega_i) \leq 1$.

Our goal is to achieve a perceived noise attenuation equal to ζ. With this in mind we define our new weighting rule to put the distortion of the residual noise, $R_{e_n}(\Omega_i)$, exactly on the masking threshold $R_T(\Omega_i)$ of the speech component,

$$[\zeta - H(\Omega_i)]^2 R_n(\Omega_i) \stackrel{!}{=} R_T(\Omega_i). \tag{5}$$

Solving for $H(\Omega_i)$ with the constraint $\zeta \leq H(\Omega_i) \leq 1$ and calling the solution $H^{JND}(\Omega_i)$, where JND stands for Just Noticeable Distortion, we achieve

$$H^{JND}(\Omega_i) = \min\left(\sqrt{\frac{R_T(\Omega_i)}{R_n(\Omega_i)}} + \zeta,\ 1\right). \tag{6}$$

Note that $H^{JND}(\Omega_i)$ is a function of only two variables: the masking threshold $R_T(\Omega_i)$ and the noise PSD $R_n(\Omega_i)$.

Although the speech distortion is not explicitly considered in the development of the weighting rule, the solution minimizes it for the specified degree of noise reduction; if $H(\Omega_i)$ is given a greater value, the distortion of the speech will certainly be smaller, but more noise will be audible. On the other hand, if $H(\Omega_i)$ is smaller, the speech distortion will rise without any perceivable improvement of the noise reduction.

One of the most important properties of H^{JND} is how artifacts are avoided. With most spectral weighting techniques, artifacts arise when an estimation $\hat{R}_n(\Omega_i)$ of the noise PSD is subtracted from the input signal PSD $R_y(\Omega_i)$. Because of the statistical nature of the noise, after the subtraction there will remain short bursts of sinusoidal noise distributed randomly over time and frequency [12]. These bursts will be perceived as tones of short duration, thus the common term "musical tones". Especially during speech pauses such artifacts may be heard clearly when, for example, the standard spectral subtraction [1] is applied as a weighting function.

In contrast, for H^{JND} we see directly from Eq. (6) that if no speech is present, the masking threshold $R_T(\Omega_i)$ is zero and the weighting coefficient is equal to the constant ζ. Hence the output signal is identical to the input signal up to a scalar factor – the characteristics of the background noise are perfectly preserved. Furthermore, in Eq. (6) there is no subtraction at all. Moderate estimation errors of the masking threshold or the noise PSD will only have a minor impact on the output signal quality.

3.2. Comparison with the Wiener Rule

A common optimization criterion in the field of digital signal processing is the minimization of the mean squared error $\mathcal{E}\{|E(\Omega_i)|^2\}$, where $\mathcal{E}\{\cdot\}$ denotes the expectation operator. In our case, the optimal solution can be expressed as the ratio between the cross-correlation spectrum of the speech and the microphone signal, and the auto-correlation spectrum of the microphone signal, see for example [13]. This leads to the well-known Wiener weighting rule $H(\Omega_i) = R_s(\Omega_i)/(R_s(\Omega_i) + R_n(\Omega_i))$. If the desired output signal is not the speech, but the speech and a proportion ζ of the noise, as defined in Eq. (1), we will instead obtain a modified Wiener rule

$$H^W(\Omega_i) = \frac{R_s(\Omega_i) + \zeta R_n(\Omega_i)}{R_s(\Omega_i) + R_n(\Omega_i)}. \tag{7}$$

The magnitude of the frequency response of $H^W(\Omega_i)$ for a fixed but arbitrary frequency Ω_i is plotted in Fig. 2 (the solid

line) as a function of the instantaneous signal-to-noise ratio $R_s(\Omega_i)/R_n(\Omega_i)$. We see that

$$\lim_{R_s/R_n \to \infty} H^W(\Omega_i) = 1, \tag{8}$$

which is in accordance with the "standard" Wiener rule, and that

$$\lim_{R_s/R_n \to 0} H^W(\Omega_i) = \zeta, \tag{9}$$

which is the consequence of the additional term $\zeta R_n(\Omega_i)$ in the numerator of Eq. (7).

In contrast, the magnitude of $H^{JND}(\Omega_i)$ is only indirectly a function of $R_s(\Omega_i)/R_n(\Omega_i)$, since the masking threshold $R_T(\Omega_i)$ is not only a function of $R_s(\Omega_i)$ at a single frequency, but also nearby ones, and furthermore of the properties of the masker.

To illustrate this, we define θ as the relative distance between the masking threshold and the speech PSD,

$$\theta = \frac{R_T(\Omega_i)}{R_s(\Omega_i)}. \tag{10}$$

Common values for θ lie in the range of 0 dB to -20 dB, compare Fig. 1. In other words, any additive, uncorrelated signal component at the frequency Ω_i won't be perceived if its relative level to $R_s(\Omega_i)$ is lower than θ.

In Fig. 2 some curves for $H^{JND}(\Omega_i)$ are plotted, for which $R_T(\Omega_i)$ in Eq. (6) has been replaced with $\theta \cdot R_s(\Omega_i)$. Apart from the distinctive break at $H^{JND}(\Omega_i) = 0$ dB, they look similar to the attenuation curve of the modified Wiener rule. The main difference is that the attenuation of H^{JND} depends on several other factors, as mentioned above. For example, because θ is greater for noise-like speech components (fricative consonants) such as /f/ and /s/, these are in general less attenuated than tonal speech (vocals). Fricatives are considered very important for the speech intelligibility. The effect of the explicit usage of the masking threshold is that the weighting coefficients depend on the signal properties. In Fig. 2 this could be illustrated with a curve moving from left to right when θ decreases.

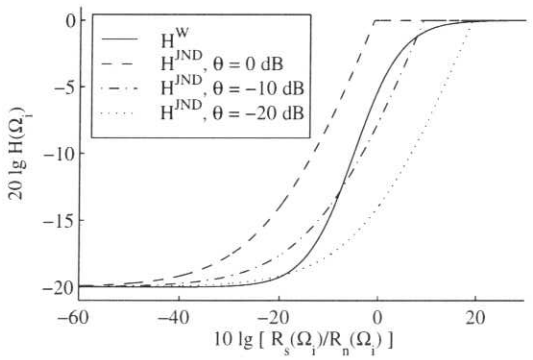

Figure 2: Magnitude of the filter response at a fixed but arbitrary frequency Ω_i for the Wiener filter $H^W(\Omega_i)$, and $H^{JND}(\Omega_i)$ with $\theta = 0$, -10, and -20 dB.

4. RESULTS

4.1. The Simulation System

The proposed weighting rule has been used in a noise reduction system as illustrated in Fig. 3, which is based on a standard analysis/synthesis system with input data frames multiplied by a Hamming window, zero-padded and then transformed into the frequency domain by means of the FFT. The synthesis consists of the inverse FFT and overlap-and-add. The chosen parameters for the sampling rate 8 kHz are: decimation ratio $M = 128$, frame length $L = 256$, and FFT length $N = 512$.

The noise PSD is estimated by a modified version of the algorithm described in [5]. The estimate is used to obtain a preliminary estimation $\check{S}(\Omega_i)$ of the speech spectrum. This is performed by a conventional noise reduction algorithm, using for example the Wiener rule [2, 13] or the MMSE-STSA/LSA [3, 4]. Experiments have shown that a relatively rough estimate of the speech is sufficient under most circumstances, but the more accurate the speech estimate is, the better will the overall system performance in terms of noise reduction and speech distortion be. The estimation of the masking threshold is based on the ISO-model [10].

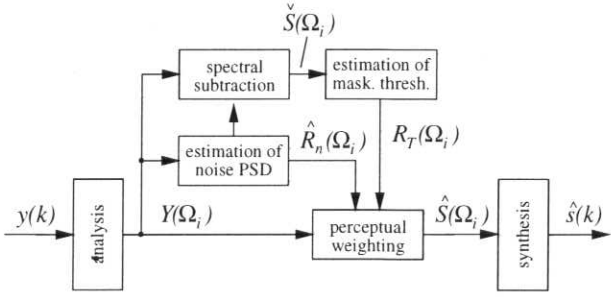

Figure 3: Overview of the noise reduction system.

4.2. Instrumental Evaluations

The instrumental evaluations were performed by processing a set of test sequences spoke by both male and female speakers. Different noise signals were added to achieve signal-to-noise ratios in the range of -10 to 20 dB. The noise attenuation was measured as the logarithmic ratio between the mean noise power P_n in the input signal and the mean noise power $P_{n'}$ in the output signal,

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

Similarly, the attenuation of the speech is defined as the ratio between the mean speech power in the input signal (P_s) and in the output signal ($P_{s'}$),

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

Though the speech attenuation doesn't describe the distortion of the speech satisfactorily, it gives an indication of the effective noise reduction $NA - SA$.

The H^{JND} weighting rule has been compared with several conventional methods. In this paper we present the result of H^{JND} together with those of MMSE-LSA [4] modified to attenuate the noise no more than the factor ζ. The MMSE-LSA is chosen as a reference because it has been extensively studied (see for example [14, 15]) and is considered to be a state-of-the-art conventional noise reduction method.

Fig. 4 shows the mean noise and speech attenuations. We see that the noise attenuation is almost exactly the same for both methods. Especially for $H^{JND}(\Omega_i)$, a high attenuation of the speech components can be observed for very low input signal SNR, which reduces the effective noise reduction $NA - SA$. The

reason for this is that the masking threshold of a given masker is constant and independent of the noise power. Thus, the amount of noise which can be masked is also invariable and $H^{JND}(\Omega_i)$ will be closer to ζ the stronger the noise is. The mean speech attenuation can be reduced by choosing a higher ζ. For $20 \lg \zeta = -20$ the H^{JND} weighting rule is most effective when the input SNR is above 0 dB.

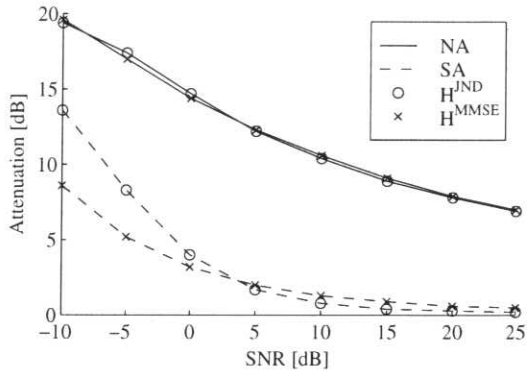

Figure 4: Noise attenuation (NA) and speech attenuation (SA) with ζ = -20 dB for the new $H^{JND}(\Omega_i)$ weighting rule and the MMSE-LSA algorithm as a function of the input signal SNR.

Some different instrumental measures were used to assess the distortion of the speech, for example SEGSNR, basilar distance and the cepstral distance. We found that the speech distortion using the new proposed weighting rule is about the same as when applying the MMSE-LSA weighting rule. Only at very low SNR the global attenuation of the speech caused by H^{JND} is reflected in values indicating larger distortions.

4.3. Listening Tests

All instrumental assessments should be interpreted cautiously. For example, no standard instrumental measure describes the subjective quality of the background noise. However, informal listening tests confirmed that the achieved noise reduction using the new weighting rule indeed is in accordance with the noise attenuation factor ζ. Furthermore, the spectral characteristics of the original noise is retained and no artifacts are audible.

The listening test also verified that the distortions of the speech are similar to those of MMSE-LSA, and that the high attenuation at low SNR principally is the effect of a global attenuation which only has a relatively low impact on the spectral shape of the speech. For the H^{JND} weighting rule, the tradeoff between noise and speech attenuation can be easily controlled by the factor ζ.

5. CONCLUSION

In this paper we have presented a novel noise reduction algorithm based on auditory masking effects. In contrast to other algorithms, our design object was to achieve a perceived noise reduction equal to a pre-defined factor. The distinguishing feature of the algorithm is that it succeeds in preserving the naturalness of the background noise, whereas other conventional or psychoacoustically motivated techniques often introduce artifacts such as musical tones. Although the speech distortion is not considered explicitly, it remains comparable to those of conventional noise reduction methods, apart from a global attenuation of the speech component at very low input signal-to-noise ratios.

6. REFERENCES

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 27, no. 2, pp. 113 – 120, April 1979.

[2] R.J. McAulay and M.L. Malpass, "Speech enhancement using a soft-decision noise suppression filter", *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 28, no. 2, pp. 137 – 145, April 1980.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator", *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 32, no. 6, pp. 1109–1121, December 1984.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator", *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 33, no. 2, pp. 443 – 445, April 1985.

[5] R. Martin, "Spectral subtraction based on minimum statistics", in *Proceedings EUSIPCO-94*, September 1994, pp. 1182 – 1185, Edinburgh, UK.

[6] D.E. Tsoukalas, J. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression", *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 6, pp. 497 – 514, November 1997.

[7] N. Virag, "Speech enhancement based on masking properties of the auditory system", in *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, May 1995, pp. 796–799, Detroit, USA.

[8] H. Reininger and C. Kuhn, "Signalverbesserung durch gehörgerechte Spektrale Subtraktion", in *Proceedings 9. Aachener Kolloquium*, March 1997, pp. 259 – 262, Aachen, Germany.

[9] C. Beaugeant, V. Turbin, P. Scalart, and A. Gilloire, "New optimal filtering approaches for hands-free telecommunication terminals", *Signal Processing*, vol. 64, no. 1, January 1998.

[10] ISO/IEC 11172-3:1993, "Information technology – coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part 3, Audio", 1993.

[11] S. Gustafsson, P. Jax, and P. Vary, "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics", in *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, May 1998, Seattle, USA.

[12] P. Vary, "Noise suppression by spectral magnitude estimation – mechanism and theoretical limits –", *Signal Processing*, vol. 8, pp. 387 – 400, 1985.

[13] S.V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction*, John Wiley and Teubner, 1996.

[14] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor", *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 345 – 349, 1994.

[15] P. Scalart and J. Vieira Filho, "Speech enhancement based on a priori signal-to-noise estimation", in *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, May 1996, pp. 629 – 632, Atlanta, USA.