ADAPTIVE PSYCHOACOUSTIC FILTER FOR BROADBAND NOISE REDUCTION IN AUDIO SIGNALS

Ismo Kauppinen and Kari Roth

University of Turku, Department of Applied Physics, FIN-20014 Turku, Finland
iska@utu.fi, kari.roth@utu.fi

Abstract: A novel adaptive noise reduction filter for audio signals is presented in this paper. The main aim of the method is to achieve a balanced tradeoff between noise suppression level and audible quality loss. The noise reduction is based on frequency domain processing of the low energy frequency bands. Perceptual model is applied to achieve the best quality judged by the listener. Unlike in the conventional spectral subtraction methods, in this method, the estimate of the noise amplitude spectrum is not required.

1. INTRODUCTION

A common method for broadband noise reduction is the spectral subtraction method [1], which offers a simple and computationally efficient frame-by-frame tool for additive noise reduction in audio signals. The noisy signal model is given by

$$y(n) = s(n) + d(n), \tag{1}$$

where $y(n)$ is the observed signal, $s(n)$ is the original signal and $d(n)$ is random noise. Signal and noise are assumed to be uncorrelated. The key idea for the spectral subtraction techniques is to estimate the magnitude spectrum of the background noise $|\hat{D}(f)|$ and then to subtract it from the magnitude spectrum of the noisy signal $|Y(f)|$. The frequency domain representation of $y(n)$ is in practice achieved by short-time fast Fourier transform (STFFT) and the processing is done in the frequency domain, where the estimate of the noiseless magnitude spectrum is obtained by

$$|\hat{S}(f)| = |Y(f)| - |\hat{D}(f)|. \tag{2}$$

The phase spectrum of the noisy signal $\Theta(f)$ is not modified and it is used in the synthesis of the noiseless signal estimate which is obtained by using inverse fast Fourier transform (IFFT)

$$\hat{s}(n) = \text{IFFT}\{|\hat{S}(f)|e^{i\Theta(f)}\}. \tag{3}$$

The noise suppression can be interpreted as time-varying linear filtering, where the observed signal magnitude spectrum $|Y(f)|$ is multiplied by a real valued gain function $H(f)$, known as the suppression rule:

$$|\hat{S}(f)| = H(f)|Y(f)|, \qquad 0 \leq H(f) \leq 1. \tag{4}$$

In conventional short-time spectral attenuation (STSA) techniques, the function $H(f)$ is heavily dependent on the noise magnitude spectrum estimate $|\hat{D}(f)|$. This noise magnitude spectrum is estimated from observations in the absence of the underlying signal (*e.g.* in the beginning of a tape recording or in between words in speech signal).

Our method is based on selecting the frequency bins that clearly contain components of the original signal by using a highly adaptive threshold function. These selected frequency bins are left untouched and attenuation is performed to the noisy spectral bins by using certain attenuation rules. Perceptual models are taken into account in the development of these rules.

The aim of our method is not to get mathematically best approximation of the original signal $s(n)$, but to get an approximation $\hat{s}(n)$ that sounds closest to $s(n)$ when perceptually analyzed. We apply the principle of "least treatment" which means in practice that only the audible components of the noise are processed and the inaudible noise components in the signal are left unprocessed.

The remaining text is organized as follows. In Section 2, the adaptive threshold function is introduced. In Section 3, the perceptual model is presented. In Section 4, subjective experiments of the method are discussed. Conclusions are drawn in Section 5.

2. ADAPTIVE THRESHOLD CURVE

The problem can be formulated as finding a decision tool to classify each frequency bin f of the noisy signal amplitude spectrum $|Y(f)|$ to contain either desired signal components or noise. The frequency bins that clearly contain desired signal components are left unprocessed.

The classification of each frequency bin f is done by forming an adaptive threshold curve given by

$$T(f) = \alpha[\text{median}\{A_p(f)\}] - \frac{\beta}{2j+1}\sum_{k=f-j}^{f+j} Y_p(f) + \lambda, \tag{5}$$

where $Y_p(f) = |Y(f)|^2$ is the short-time power spectrum (STPS) and $A_p(f)$ is a subset of the STPS components given by

$$A_p(f) = \{Y_p(f-i), ..., Y_p(f), ..., Y_p(f+i)\}, \tag{6}$$

α and β are a long term median and a short term average scaling factors respectively, $2i+1$ and $2j+1$ $(i > j)$ are the lengths of the median filter and the average respectively and λ is a threshold offset. The median of a set is

defined as the middlemost value of an ordered table of the set values. The decision that a frequency bin f contains desired signal components is done when a value of the STPS $Y_p(f)$ exceeds the value of the adaptive threshold curve $T(f)$. The processing scene can be formulated as:

$$|\hat{S}(f)| = \begin{cases} |Y(f)|, & |Y(f)|^2 \geq T(f) \\ H(f)|Y(f)|, & |Y(f)|^2 < T(f), \end{cases} \quad (7)$$

where the suppression function $H(f)$ is chosen to meet psychoacoustic conditions.

The development of the threshold curve is visualized in Fig. 1, where an FFT analysis is performed to 8192 samples long frame of a music signal, which is corrupted by tape hiss (that is related to $1/f$ noise), to obtain STPS $Y_p(f)$. The long term median filter component of the threshold curve $T(f)$ is visualized in the upmost graph by setting $\alpha = 16$, $\beta = 0$, and $\lambda = 0$ in Eq. 5. The length of the median filter is 201. This component gives the background shape of the noisy signal power spectrum. The short term average component in Eq. 5, is demonstrated in the middle graph by setting $\alpha = 0$, $\beta = -10$ and $\lambda = 0$ in Eq. 5. The length of the average is 21. The purpose of the short term average is to lower the threshold curve significantly at the locations of the frequency bins that contain mostly desired signal components. This is achieved by subtracting the short time average component from the median filter component. In the bottom graph these components are subtracted resulting as the threshold curve $T(f)$ presented in Eq. 5 with $\alpha = 16$, $\beta = 10$, and $\lambda = 0.2$. The short term average makes the threshold lower at the locations of the signal components and even some components (indicated by arrows) below the noise floor survive in the processed spectrum.

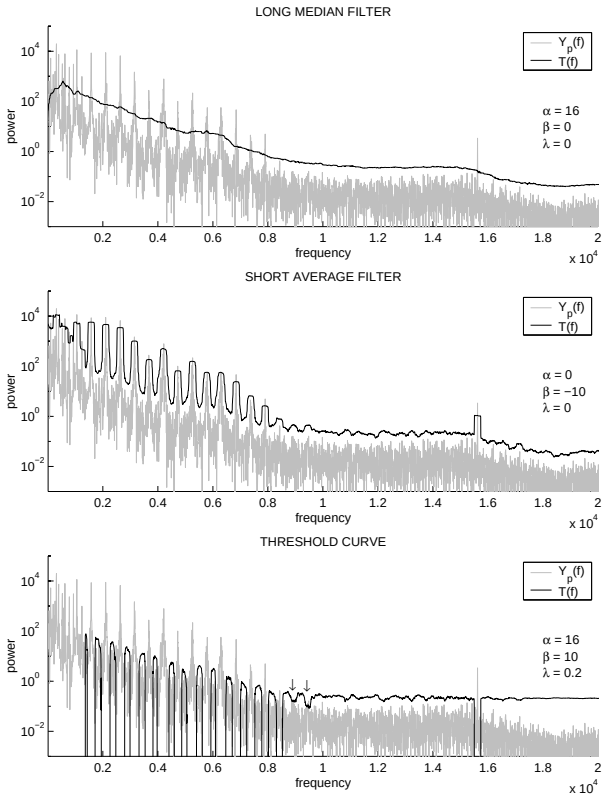

Fig. 1. Visualization about the components of the threshold curve.

3. THE PERCEPTUAL MODEL

Successful results of applying perceptual models to audio signal have been obtained in audio coding [2]. A psychoacoustic phenomenon known as masking can be defined as a process by which the threshold of audibility is raised by the presence of another (masking) sound [3]. Masking is a fundamental aspect of the human auditory system and it is typically expressed in dB and its effect is known as the masking threshold. Masking occur in both the frequency and time domain. The effect of the masking signal depends on the frequency, level, and the structure of both the masker and the masked signal.

3.1. Calculation of the masking threshold

The masking threshold is calculated in the Bark scale known as the critical band rate (CBR) that is based on the assumption that the basilar membrane in the hearing mechanism analyzes the incoming sound through a spatial spectral analysis. An approximation for the transformation from frequency to Bark scale is given by [4]

$$z = 13\tan^{-1}\left(\frac{0.76f}{1000}\right) + 3.5\tan^{-1}\left\{\left(\frac{f}{7500}\right)^2\right\} \quad (8)$$

where z is the mapped frequency in Barks. The base of the masking threshold analysis in the frequency domain is the energy per critical band defined as

$$E(z) = \sum_{f=lz}^{hz} Y_p(f) \quad (9)$$

where lz and hz are the lower and higher boundaries in each critical band z. The spread masking across critical bands can be obtained by a convolution given by

$$C(z) = E(z) * B(z), \quad (10)$$

where $B(z)$ is an approximation of the basilar membrane spreading function. This function, expressed in dB, is obtained by [5]

$$B_{\text{dB}}(z) = 15.91 + 7.5(z+0.474) - 17.5\sqrt{1+(z+0.474)^2}. \quad (11)$$

To form an estimate of the masking threshold level, two different cases are to be considered: noise masking tone (NMT) and tone masking noise (TMN). The threshold levels are defined to be 5.5 dB and $(14.5 + z)$ dB below $C(z)$ respectively. A spectral flatness measure (SFM) is used to determine which is the case for each critical band. The SFM is defined as the ratio of the geometric mean (Gm) of the power spectrum to the arithmetic mean (Am) of the power spectrum. In this use, the SFM is converted to decibels, *i.e.*,

$$SFM = 10\log_{10}\left(\frac{Gm}{Am}\right) \quad (12)$$

This is further used to generate an index for tonality given by

$$\delta = \min\left(\frac{SFM}{SFM_{\max}}, 1\right) \tag{13}$$

where $SFM_{max} = -60$ dB. The offset for the masking threshold is defined as [2] [5]

$$O_{\mathrm{dB}}(z) = \delta(14.5 + z) + (1 - \delta)5.5 \tag{14}$$

and in order to form the masking threshold the offset is subtracted from the spread masking threshold $C(z)$ and a normalization technique is used to keep the threshold in the desired level. The masking threshold is given by

$$M(z) = \frac{10^{\left(\log_{10} C(z) - \left(\frac{O_{\mathrm{dB}}(z)}{10}\right)\right)}}{N} \tag{15}$$

where N is the number of points in each critical band z.

The masking function calculated for the same signal frame whose power spectrum is used in Fig. 1, is presented in Fig. 2, with the difference that the frequency is mapped in logarithmic scale. It can be seen that above 500 Hz the critical band width is approximately constant for each band in logarithmic frequency scale.

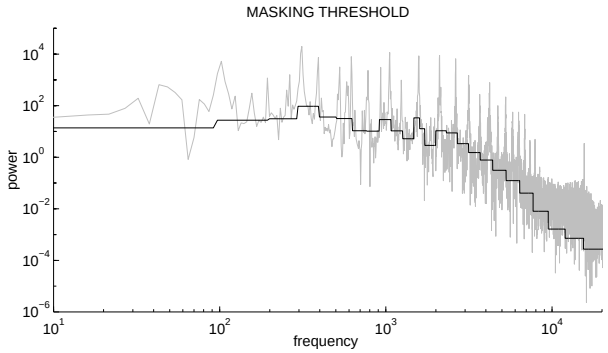

Fig. 2. The masking threshold calculated for the same signal frame as in Fig. 1.

3.2. The suppression function

The aim is to reduce the audible noise components to the level where they become inaudible. However, the masking threshold function $M(f)$ is proportional to the STPS $Y_p(f)$ and if we change the power spectrum this changes also the masking threshold function. This leads us to iteration. The processing scene for each iteration step is defined as

$$|\hat{S}^k(f)| = \begin{cases} |Y^k(f)|, & Y_p^k(f) \geq T(f) \text{ or} \\ & Y_p^k(f) < M^k(f), \\ H^k(f)|Y^k(f)|, & Y_p^k(f) < T(f) \text{ and} \\ & Y_p^k(f) \geq M^k(f), \end{cases} \tag{16}$$

where the suppression function is defined as

$$H^k(f) = \frac{\sqrt{M^k(f)}}{|Y^k(f)|}. \tag{17}$$

The superscript k in Eqs. 16 and 17 denotes the number of current iteration step. In each iteration step frequencies f that are selected as undesired noise components are attenuated to the approximated masking threshold level $M(f)$ and in the next step the masking threshold is updated to represent the masking threshold of the spectrum where the noise components are attenuated. The same noisy frequencies are further attenuated to the level of the new masking threshold.

4. EXPERIMENTS

The noise reduction method was subjectively evaluated by applying the method to various different types of high-fidelity audio signals which were additively corrupted by Gaussian random noise. Best results were obtained when the tonality factor δ was fixed to unity so that the case where tone is masking signal (TMS) was assumed for each signal frame. In this case the experiments show that the number of iterations required in the reduction is two and even with one iteration the method performs well. If the threshold offset λ is given a too low value, it results as "musical noise" that is commonly encountered in conventional spectral subtraction techniques. This can be overcome by increasing the offset value λ until the "musical noise" disappears completely. Over 50% overlapping between adjacent processing frames should be used to avoid sudden changes of the filter.

The masking threshold is a constat average value over each critical band. This introduces a small side effect in the low frequency region where low level noise can be heard pumping (modulated by the signal). This phenomenon can be overcome by adding over attenuation to the few first critical bands.

In the conventional spectral subtraction method the processed signal usually sounds as low pass filtered since the high frequencies suffer from heavy attenuation. In our method the frequencies that are found to be signal components are left unprocessed and this makes the high frequencies also survive without attenuation resulting in high-fidelity in the processed signal.

Since the noise power spectrum estimate is not required in our method, also voice activity detectors (VAD) are not required when the method is applied to speech signals. A speech signal additively corrupted by Gaussian random
noise with a signal-to-noise ratio (SNR) of 0 dB is enhanced and demonstrated in Fig. 3.

5. CONCLUSIONS

In this paper we have presented an adaptive broadband noise reduction technique based on selecting the noise components of the signal power spectrum and attenuating them by using perceptual criteria. The main advance of the method compared to conventional STSA techniques is that the estimate of the noise power spectrum is not required. The optimal application for the method is to apply it to high quality music recordings that are corrupted by additive noise such as tape hiss. The proposed method gives superior results also when applied to speech signals with low SNR's.

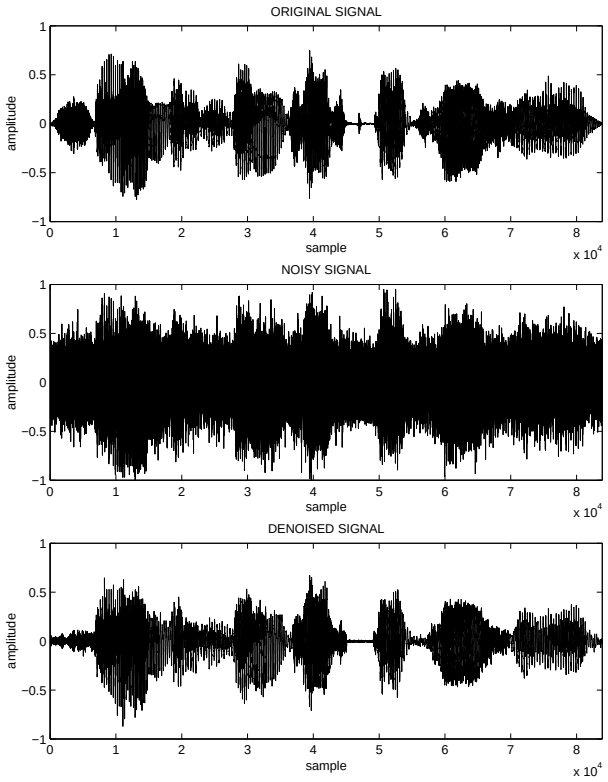

Fig. 3. Original speech signal, same signal corrupted by additive noise at a SNR = 0 dB, and enhanced signal.

Future plans for the authors are to apply a more detailed masking threshold and use variable frame length, determined from the stationarity of the signal, instead of constant fixed frame length.

6. REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 27, No. 2, April 1979.

[2] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria", *IEEE Journal on Selected Areas in Communications*, Vol. 6, February 1988, pp. 314-323.

[3] B. C. J. Moore, *An introduction to the psychology of hearing*, Academic Press, 4th ed., 1997.

[4] E. Zwicker and H. Fastl, *Psychoacoustics*, Springer Verlag, 2nd ed., 1999.

[5] J. Mourjopoulos and D. Tsoukalas, "Neural network mapping to subjective spectra of music sound", *J. Audio Eng. Soc.*, Vol. 40, April 1992, pp. 253-259.