EFFICIENT MUSICAL NOISE SUPPRESSION FOR SPEECH ENHANCEMENT SYSTEMS

Thomas Esch and Peter Vary

Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University, Germany
{esch|vary}@ind.rwth-aachen.de

ABSTRACT

Noise reduction techniques that are relying on spectral weighting rules often generate annoying musical noise artifacts in the processed signal. In this paper, we present a postfilter (PF) for the spectral weighting gains that is capable of reducing musical noise in a simple but efficient way. It includes a robust detector for speech pauses and low SNR conditions and adaptively smoothes the weighting gains over frequency based on soft-decisions. Objective and subjective measurements show consistent improvements if the postfilter is applied to conventional noise reduction techniques.

Index Terms— Speech enhancement, musical noise, adaptive smoothing, postfilter

1. INTRODUCTION

In speech communication systems (e.g., mobile communications, hearing aids and hands-free devices), the reduction of background noise in disturbed speech remains a challenging task. Most speech enhancement systems are based on the decomposition of speech and noise in the frequency domain using the Short-Time Fourier Transform (STFT) and the modification of the spectral coefficients with a gain function, e.g., [1], [2], [3]. Although these methods provide an improvement in terms of noise attenuation, they often produce a new randomly fluctuating type of noise, referred to as *musical noise*. This phenomenon can be explained by noise or signal-to-noise ratio (SNR) estimation errors leading to spurious peaks in the processed spectrum. When the enhanced signal is reconstructed in the time domain, these peaks result in short sinusoidals whose frequencies vary from frame to frame. In particular, musical noise is very annoying during speech pauses and in low SNR conditions when it is not masked by the speech signal.

In the literature, a variety of different methods for reducing musical tones has been proposed. A lower limit to the *a priori SNR* was applied in [4] resulting in a flooring of the weighting gains. The well-known *decision directed* approach [3] prevents the musical noise phenomenon by recursive smoothing of the a priori SNR. A time smoothed gain factor was proposed in [5] in order to reduce the dynamics of the weights. In [6], a postprocessing method was presented to suppress the annoying artifacts based on a speech/musical noise classification. Cepstral smoothing was applied to the spectral weighting gains in [7] enabling selective smoothing of speech and musical tones.

In this paper, a postfilter (PF) for the spectral weighting gains is presented that efficiently suppresses musical noise. As it treats the estimation of the initial weighting gains as *black box*, it can be applied to any noise reduction method. The postfilter consists of two steps. In the first step, speech pauses and low SNR regions are robustly detected. Based on these results, adaptive spectral smoothing of the weighting gains is performed in the second step. The remainder of this paper is organized as follows: In Sec. 2, a brief overview of a conventional noise reduction system is given. Section 3 comprises the new postfilter concept in detail. Experimental results are shown in Sec. 4 and conclusions are drawn in Sec. 5.

This work was supported by Nokia, Tampere, Finland.

2. SYSTEM OVERVIEW

A simplified block diagram of a conventional noise reduction system is depicted in Fig. 1(a). The speech signal $s(k)$ is assumed to be degraded by an additive uncorrelated noise signal $n(k)$ producing the noisy speech signal

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

where k is the discrete time index. For the transformation into the frequency domain, the noisy input signal $y(k)$ is first segmented into overlapping frames of length L. After windowing (e.g., applying a Hann window), these frames are transformed via Fast Fourier Transform (FFT) with an FFT length M. The spectrum of the noisy input signal is therefore given by:

$$Y(\lambda, \mu) = S(\lambda, \mu) + N(\lambda, \mu), \tag{2}$$

where $S(\lambda, \mu)$ and $N(\lambda, \mu)$ represent the spectral coefficients of speech and noise at frequency bin μ and frame λ. All statistical estimators that are discussed and evaluated in the following sections require knowledge of the power spectral density (PSD) of the noise signal. As the noise PSD is in general not known a priori, it has to be estimated and updated while executing the noise reduction algorithm. For this purpose, many approaches can be found in the literature, prominent ones are the application of a voice activity detector (VAD) (e.g., [8]) and the minimum statistics approach [9].

Based on the estimate $\hat{\sigma}_N^2$ of the noise PSD, two SNR parameters are estimated, namely the *a posteriori SNR* $\gamma(\lambda, \mu)$ and the *a priori SNR* $\xi(\lambda, \mu)$:

$$\gamma(\lambda, \mu) = \frac{|Y(\lambda, \mu)|^2}{\hat{\sigma}_N^2(\lambda, \mu)} \quad \text{and} \quad \xi(\lambda, \mu) = \frac{\mathcal{E}\{|S(\lambda, \mu)|^2\}}{\hat{\sigma}_N^2(\lambda, \mu)}. \tag{3}$$

The a priori SNR can be estimated using the decision-directed approach [3]. The actual spectral weighting is performed by multiplying the noisy spectrum $Y(\lambda, \mu)$ with a weighting gain $G(\lambda, \mu)$:

$$\hat{S}(\lambda, \mu) = G(\lambda, \mu) \cdot Y(\lambda, \mu). \tag{4}$$

The weighting gains are dependent on the noise reduction algorithm and are usually a function of the noise PSD estimate $\hat{\sigma}_N^2(\lambda, \mu)$ and

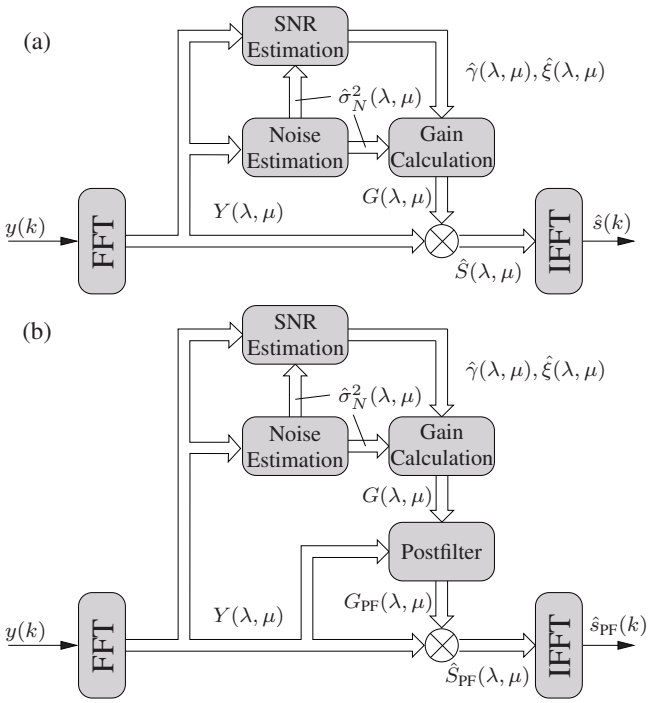

Fig. 1. System block diagram of conventional noise reduction system (a) without and (b) with postfilter.

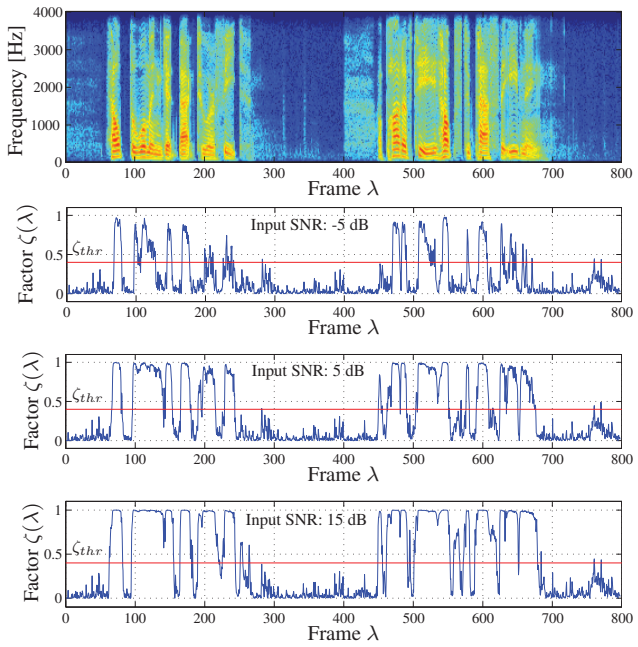

Fig. 2. Example of low SNR detector. *Upper plot*: Spectrogram of the clean speech signal: "Help the woman get back to her feet. A pot of tea helps to pass the evening." (male voice). *Lower plots*: Results for the factor $\zeta(\lambda)$ for different input SNR values (noise type: F16).

the SNR estimates $\hat{\gamma}(\lambda,\mu)$ and $\hat{\xi}(\lambda,\mu)$, as stated before. The spectral weighting results in an estimate $\hat{S}(\lambda,\mu)$ of the clean speech coefficient $S(\lambda,\mu)$. In order to obtain the enhanced signal in the time domain, an Inverse Fast Fourier Transform (IFFT) and overlap-add is applied.

3. POSTFILTER CONCEPT

The main idea of the proposed concept is to reduce the annoying musical tones especially in the mentioned low SNR regions. Therefore, a reliable and robust detector for those regions is required which is presented in the next section. Based on the results of this detector, spectral smoothing of the magnitudes $|G(\lambda,\mu)|$ is performed. Fig. 1(b) illustrates the block diagram of the system that was considered within this work.

3.1. Low SNR Detector

In order to keep the proposed method simple, we directly use the output of the initial noise reduction system, i.e., the weighting gains $G(\lambda,\mu)$ as depicted in Fig. 1(b). It turned out that the power ratio $\zeta(\lambda)$ of the processed signal $\hat{S}(\lambda,\mu)$ and the noisy signal $Y(\lambda,\mu)$ provides a good indicator of speech presence or absence in the current frame λ:

$$\zeta(\lambda) = \frac{\sum_{\mu=0}^{M-1} |G(\lambda,\mu)\cdot Y(\lambda,\mu)|^2}{\sum_{\mu=0}^{M-1} |Y(\lambda,\mu)|^2} = \frac{\sum_{\mu=0}^{M-1} |\hat{S}(\lambda,\mu)|^2}{\sum_{\mu=0}^{M-1} |Y(\lambda,\mu)|^2}. \quad (5)$$

If the frame mainly contains speech (high SNR), the power of the processed frame is equal or only slightly lower to the power of the noisy input frame, i.e., $\zeta(\lambda) \approx 1$. By contrast, the noise reduction system is supposed to strongly attenuate the input signal in low SNR conditions (or during a speech pause), resulting in a power ratio $\zeta(\lambda) \approx 0$.

In order to detect only low SNR regions, a threshold ζ_{thr} is applied to the factor $\zeta(\lambda)$ as follows:

$$\zeta_T(\lambda) = \begin{cases} 1, & \text{if} \quad \zeta(\lambda) \geq \zeta_{thr} \\ \zeta(\lambda), & \text{if} \quad \zeta(\lambda) < \zeta_{thr}. \end{cases} \quad (6)$$

The threshold ζ_{thr} later controls the trade-off between speech distortions and musical noise reduction (cf. Sec. 3.2). An example is depicted in Fig. 2 for a noisy sequence of 8 seconds length. The upper plot shows the spectrogram of the clean speech signal, the lower plots the results for the power ratio $\zeta(\lambda)$ for different input SNR values (-5, 5 and 15 dB) respectively. The speech signal (NTT speech database) was disturbed by F16 noise (NOISEX-92 database) and the weighting gains that were necessary in Eq. 5 were calculated with the Wiener filter rule [2]. As can be seen, good detection results were achieved even at low input SNR values.

The detection of low SNR regions based on the power ratio $\zeta(\lambda)$ performs better than directly utilizing the a priori SNR that has already been estimated in the noise reduction system (cf. Sec. 2). The reason for this lies in an improved estimation due to the additional application of the noise suppression algorithm. Moreover, we can treat the noise reduction system as black box as we only need the noisy input signal and the enhanced output signal or the weighting gains for this detection.

3.2. Adaptive Spectral Smoothing

The aim of the postprocessing method is to retain the naturalness of the background noise and to reduce the occurrence of musical noise in low SNR regions. The power ratio $\zeta_T(\lambda)$ from the previous section provides a reliable method to detect those regions. Based on

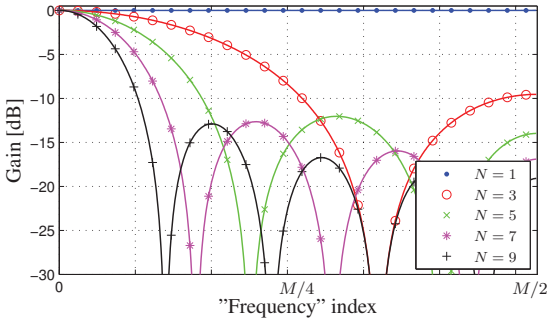

Fig. 3. Fourier transform of $H_\lambda(\mu)$ for different values of N.

$\zeta_T(\lambda)$, the magnitudes of the weighting gains $G(\lambda,\mu)$ of frame λ are adaptively smoothed over frequency using a moving average window. The odd window length $N(\lambda)$ is set to:

$$N(\lambda) = \begin{cases} 1, & \text{if} \quad \zeta_T(\lambda) = 1 \\ 2 \cdot \text{round}\left[\left(1 - \frac{\zeta_T(\lambda)}{\zeta_{thr}}\right) \cdot \Psi\right] + 1, & \text{else.} \end{cases} \tag{7}$$

The term $1 - \frac{\zeta_T(\lambda)}{\zeta_{thr}}$ provides a soft-decision that states the reliability of the low SNR detection. The function round[·] rounds the element to the nearest integer and Ψ is a scaling factor that determines the maximum degree of smoothing. Equation 7 ensures that the more reliable a low SNR frame was detected, the longer the window length resulting in a stronger smoothing of the weighting gains.

Applying a moving average window of length $N(\lambda)$ is equivalent to a linear filtering with the impulse respond $H_\lambda(\mu)$ as follows:

$$H_\lambda(\mu) = \begin{cases} \frac{1}{N(\lambda)}, & \text{if} \quad \mu < N(\lambda) \\ 0, & \text{else} \end{cases}, \quad \text{where} \quad \mu \in [0, M-1]. \tag{8}$$

Fig. 3 depicts the Fourier transform of $H_\lambda(\mu)$ for different values of N. Please note that the term "frequency" in this context is somewhat misleading as $H_\lambda(\mu)$ is already applied in the frequency domain. However, Fig. 3 shows the low-pass characteristic of the filter $H_\lambda(\mu)$ whose cut-off "frequency" is decreasing with an increasing window length N.

Within the postfilter, the weighting gain magnitudes of the initial noise reduction system are convoluted by the low-pass filter $H_\lambda(\mu)$ in every frame λ:

$$G_{\text{PF}}(\lambda,\mu) = |G(\lambda,\mu)| * H_\lambda(\mu). \tag{9}$$

Finally, the new weighting gains $G_{\text{PF}}(\lambda,\mu)$ are applied to the noisy input coefficients $Y(\lambda,\mu)$:

$$\hat{S}_{\text{PF}}(\lambda,\mu) = G_{\text{PF}}(\lambda,\mu) \cdot Y(\lambda,\mu) \tag{10}$$

and the enhanced signal is transformed back into the time domain.

4. EVALUATION

The postfilter that is presented in this paper can be applied to the weighting gains of an arbitrary noise reduction system. In the following, we investigate the postfilter in combination with four statistical noise suppression techniques that were known from literature:

1. Spectral Subtraction [1],
2. Wiener filter [2],
3. MMSE[1] estimator based on a Laplacian model for the speech and a Gaussian model for the noise signal [10],
4. MAP[2] estimator based on a super-Gaussian speech model and a Gaussian noise model [11].

Parameter	Settings
Sampling frequency	8 kHz
Frame length L	160 (20 ms)
FFT length M	256 (including zero-padding)
Frame overlap	50% (Hann window)
Input SNR	-5 dB ... 30 dB (step size: 5 dB)
Noise estimation	Minimum Statistics [9]
SNR estimation	Decision-directed approach [3]
Threshold ζ_{thr}	0.4 (see Fig. 2)
Scaling factor Ψ	10

Table 1. System settings.

In the initial systems, the following commonly used countermeasures were already utilized to avoid musical noise. The a priori SNR was estimated according to the decision-directed approach [3] and a lower limit was applied to the a priori SNR as recommended in [4]. This is equivalent to defining a lower limit G_{min} to the weighting gains $G(\lambda,\mu)$. We set $20\log_{10}(G_{\text{min}}) = -15$ dB. In addition, the dynamics of the weights were reduced by averaging over time according to [5]. On top of that, the proposed postfilter was applied to the resulting weighting gains. Both system setups - with and without the postfilter - are investigated in the following for each noise reduction method.

The evaluation is based on both objective and subjective measurements. The parameters that are used in the simulations are listed in Tab. 1. The values for the threshold ζ_{thr} and the scaling factor Ψ were determined empirically and provide a good compromise between speech distortion and musical noise suppression.

In the simulation, the speech and noise signal can be filtered separately with weighting gains adapted for the noisy signal. Hence, the output signal can additionally be stated as $\hat{s}(k) = \tilde{s}(k) + \tilde{n}(k)$, where $\tilde{s}(k)$ is merely the filtered speech signal and $\tilde{n}(k)$ the filtered noise signal. Based on these quantities, the segmental speech SNR (SegSNR), the cepstral distance (CD) and the segmental noise attenuation (NA) were calculated according to [12]. For the objective evaluation of the noise reduction schemes, five speech signals from the NTT speech database were each degraded by six different noise types (F16, babble, car, factory1, factory2, white), taken from the NOISEX-92 database. Among the five speech signals, there were three sequences from a male and two from a female speaker, each with a length of 8 seconds.

Figs. 4 and 5 depict the averaged results for SegSNR and CD respectively, both plotted over NA with the input SNR as control variable. Thus, a fair comparison with respect to the tradeoff noise attenuation and speech distortion is possible. In Fig. 4, the points of best performance would be placed in the upper right corner, in Fig. 5 in the lower right corner.

The objective measurements show that the presented postprocessing scheme improves the results of all investigated estimators. While keeping the SegSNR or the CD constant for instance, the incorporation of the postfilter increases the NA. The biggest improvement is obtained for the Spectral Subtraction rule. The weighting gains that were additionally smoothed over frequency contribute to the extra NA without affecting the speech quality. This shows that the low SNR detector works very reliable, even if the input SNR is low.

[1] MMSE - Minimum mean square error

[2] MAP - Maximum a posteriori

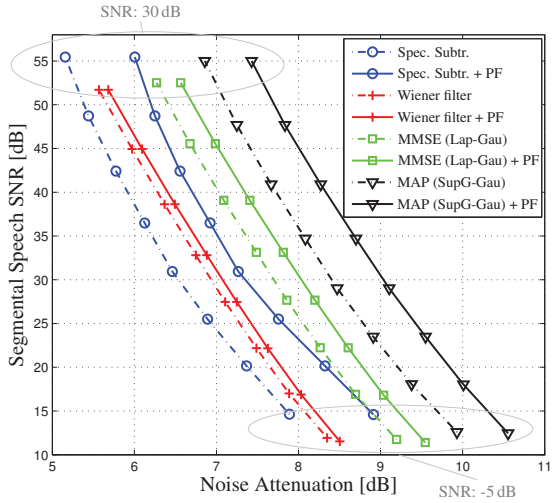

Fig. 4. Segmental speech SNR vs. noise attenuation.

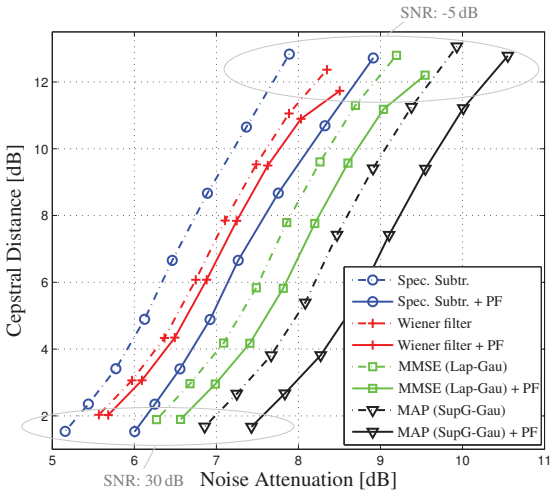

Fig. 5. Cepstral distance vs. noise attenuation.

In addition to the instrumental measurements, an informal listening test was conducted where three different signals were presented to the participants: the noisy signal, the processed signal from method A and the processed signal from method B. The options A and B had been randomly assigned to one of the four statistical estimators with and without the postfilter. The noisy signal consisted of a speech signal randomly taken from the NTT speech database disturbed by a noise signal from the NOISEX-92 database at an input SNR varying from -5 dB to 10 dB. Thirteen experienced listeners were asked to judge the overall speech quality and could choose between 'A sounds better than B', 'B sounds better than A' or 'no preference' if they did not favor one of both methods. Each test person had to judge 16 signals (4 per noise reduction method), i.e., the results are based on $16 \cdot 13 = 208$ votes. The samples could be played ad libitum before the probands had to make their judgments. The results are listed in Tab. 2. In total, approximately 72% of the test listeners preferred the samples that were generated with the new postprocessing technique. As reason, they stated the reduction of musical noise while preserving the speech quality. The results of the Wiener filter slightly differ from the results of the other estimators. This can be explained by the fact that the processed signal of the Wiener filter itself already contains less musical tones than that of the other statistical weighting rules.

Technique	no preference	Conv. NR system	Conv. NR system + postfilter
Spec. Subtr.	7.70 %	15.38 %	76.92 %
Wiener filter	19.23 %	19.23 %	61.54 %
MMSE (Lap-Gau)	9.62 %	11.54 %	78.84 %
MAP (SupG-Gau)	7.69 %	23.08 %	69.23 %
Total	**11.06 %**	**17.31 %**	**71.63 %**

Table 2. Results of the informal listening test.

5. CONCLUSIONS

In this paper, a simple postprocessing method for the spectral weighting gains is presented that efficiently suppresses musical noise. The postfilter adaptively smoothes the weighting gains over frequency based on soft-decisions of a low SNR detector. Instrumental measurements in terms of segmental speech SNR, cepstral distance and noise attenuation show improvements of the new approach when it is applied in addition to commonly used musical noise countermeasures. The objective results were confirmed by an informal listening test.

6. REFERENCES

[1] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Transactions on Speech and Audio Processing*, vol. 27, pp. 113–120, 1979.

[2] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[3] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transaction on Acoustic, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[4] O. Cappe, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–37, 1984.

[5] H. Gustafsson, S. Nordholm, and I. Claesson, "Spectral subtraction with adaptive averaging of the gain function," in *Proc. of EUROSPEECH*, Budapest, Hungary, 1999.

[6] Z. Goh, K.-C. Tan, and B. T. G. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 287–292, 1998.

[7] C. Breithaupt, T. Gerkmann, and R. Martin, "Cepstral Smoothing of Spectral Filter Gains for Speech Enhancement Without Musical Noise," *IEEE Signal Processing Letters*, vol. 14, no. 12, pp. 1036–1039, 2007.

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

[9] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 501–512, 2001.

[10] R. Martin and C. Breithaupt, "Speech Enhancement in the DFT Domain Using Laplacian Speech Priors," in *Proc. of IWAENC*, Kyoto, Japan, 2003.

[11] T. Lotter and P. Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation using a Super-Gaussian Speech Model," *EURASIP Journal on Applied Signal Processing*, pp. 1110–1126, 2005.

[12] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A Psychoacoustic Approach to Combined Acoustic Echo Cancellation and Noise Reduction," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 245–256, 2002.