AN OPTIMAL FILTERING FOR UNMASKED NOISE PREVENTION

Asmaa Amehraye1, Lionel Fillatre2 and Dominique Pastor3
1ESIGETEL, LRIT, 1 rue de port de Valvins, 77210, Avon, France,
2ICD - LM2S - UTT, UMR STMR - CNRS 12 rue Marie Curie - BP 2060 - 10010, Troyes, France
3TELECOM Bretagne, UMR CNRS 3192 Lab-STICC, Technopole Brest-Iroise, 29238 Brest

ABSTRACT
A new estimator, optimal in the frequency domain with respect to the masking properties of the human auditory system, is proposed. This new filtering technique prevents the emergence of post-filtering isolated tonals that increase the musical noise perception. Experimental results by means of objective tests show that this technique improves the enhanced speech quality.

Index Terms — Musical noise, Perceptual filtering, Masking threshold, Speech distortion, masker-to-audible noise

1. INTRODUCTION

Denoising a noisy speech signal is a challenging task because of the trade-off between clean speech distortion and residual noise reduction. In an effort to make residual noise perceptually inaudible, some speech enhancement methods exploit the auditory masking properties [1], [2], [3]. Since human ears cannot perceive noise with level below the noise masking threshold, such perceptual methods basically aim at reducing audible noise only. By so proceeding, these methods reduce speech distortion. However, most of them still return some audible and annoying musical residual noise, even when the noise spectrum and the masking threshold are known.

In this paper, we investigate the unmasked noise phenomenon, which is in part responsible for musical residual noise perception. This phenomenon, initially called the Masker-to-Audible-Noise (MAN) phenomenon in [4], occurs when the attenuation of speech components after perceptual speech enhancement lowers the masking threshold level and may reveal noise components (unmasked noise) initially masked and not processed. The main contribution of this paper is the presentation of a new perceptual filter, namely the Anti-MAN Perceptual Filter (AMPF), derived from a new frequency dependent and non-heuristical criterion. This criterion explicitly takes into account the MAN phenomenon and prevents the emergence of unmasked noise.

Section 2 discusses the MAN phenomenon. Section 3 derives and presents the new perceptual filter AMPF. Section 4 presents objective test results for this new filter. Section 5 concludes this paper.

2. PROBLEM STATEMENT

This section recalls the basics of speech signal filtering and presents the MAN phenomenon.

2.1. Speech signal filtering

The denoising is performed frame by frame such as each frame contains m speech signal samples. In the rest of the paper, we always consider the same given frame. The observed noisy speech signal y is assumed to be some speech signal \( s \) additively corrupted by zero-mean independent noise \( n \) such that \( y = s + n \). The covariance matrices of \( s \) and \( n \) are denoted by \( R_s \) and \( R_n \), respectively. Let \( F = [e^{-2i\pi(p-1)(q-1)/m}/\sqrt{m}]_{1\leq p,q\leq m} \) be the \( m \times m \) discrete Fourier transform matrix where \( i \) is the imaginary unit. The noisy speech signal in the frequency domain is given by

\[
Y = F y = F s + F n = S + N.
\]

(1)

Let \( \lambda_\nu = E[|S_\nu|^2] \) (resp. \( \gamma_\nu = E[|N_\nu|^2] \)) be the power spectrum of the clean speech (resp. the noise) where \( S_\nu \) (resp. \( N_\nu \)) is the spectral component of \( S \) (resp. \( N \)) at frequency bin \( \nu = 0, 1, \ldots, m-1, |z| \) is the modulus of the complex number \( z \) and \( E[\cdot] \) is the expectation operator. This paper considers linear estimates \( \hat{S} \) of \( S \) in the frequency domain such that \( \hat{S} = H Y \) where \( H \) is the known filtering matrix. The corresponding estimation error is given by \( E = \hat{S} - S = \hat{E}_S + \hat{E}_N \) where \( \hat{E}_S = (H - I_m) S \) represents the speech distortion in the frequency domain, \( \hat{E}_N = H N \) represents the residual noise in the frequency domain and \( I_m \) denotes the \( m \times m \) identity matrix.

2.2. Unmasked noise

The masking phenomenon derives from the frequency selectivity of the human auditory system. This paper only deals with the so-called frequency masking. This type of masking occurs when some powerful signal distorts the absolute...
threshold of hearing $T^*_\nu$ inducing the masking threshold $T_\nu$. Signals below $T_\nu$ are inaudible due to the existence of the powerful signal. For the rest of this paper, let us introduce three frequency-dependent masking thresholds which are equally important:

- $T^*$: correspond to the clean speech masking threshold.
- $\bar{T}$: corresponds to the masking threshold after any signal enhancement has been applied but this is now no longer the masking threshold of the clean speech, but the masking threshold of the restored speech. This determines whether noise components before filtering become audible.
- $T^*: This is the absolute threshold of hearing below which any component is inaudible.

Noise components that are not audible because of some maskers in the original noisy signal are still present after perceptual denoising dealing with audible noise only. Besides, due to the additive noise filtering, the speech components are in turn attenuated, which induces a post-filtering masking threshold level $\bar{T}_\nu$ lower than $T_\nu$. Consequently, noise components that are not audible before perceptual filtering can become audible if they are initially above the absolute threshold of hearing $T^*_\nu$ and their maskers are filtered. This is the Masked to Audible Noise (MAN) phenomenon.

3. THE PROPOSED OPTIMAL FILTERING

This section introduces a new perceptual criterion based on the MAN phenomenon and derives the corresponding optimal AMPF filter.

3.1. Anti-MAN criterion for unmasked noise prevention

It is assumed that the clean speech signal $s$ and noise $n$ are asymptotically weakly stationary signals, which means that these signals become weakly stationary when the frame duration diminishes. Hence, the matrices $R_s$ and $R_n$ are asymptotically Toeplitz as the number of samples $m$ becomes large. In practice, note that it is possible to get a large value of $m$ with a small frame duration if the sampling rate $f_s$ is large enough. Under these assumptions, the covariance matrix $\mathbf{F} \mathbf{R}_s \mathbf{F}^\dagger$ (resp. $\mathbf{F} \mathbf{R}_n \mathbf{F}^\dagger$) of the speech signal (resp. the noise) in the frequency domain is asymptotically equivalent [2] to a diagonal matrix whose $(\nu + 1)$-th element is $\lambda_\nu$ (resp. $\gamma_\nu$). Here, $\mathbf{F}^\dagger$ is the conjugate transpose of $\mathbf{F}$. Consequently, in the following, it is assumed that the matrix $\mathbf{H}$ associated to the proposed filter is diagonal, i.e. the gain $h_\nu$, corresponding to the $(\nu + 1)$-th diagonal element of $\mathbf{H}$, is applied to each frequency component individually: $\hat{S}_\nu = h_\nu \gamma_\nu$.

The new perceptual filter $\mathbf{H}$ proposed in this paper intends to process differently the following three areas:

- Audible area : $\mathcal{A} = \{\nu : \gamma_\nu > T_\nu\}$,
- MAN area : $\mathcal{M} = \{\nu : T^*_\nu < \gamma_\nu \leq T_\nu\}$,
- Absolute inaudible area : $\mathcal{I} = \{\nu : \gamma_\nu \leq T^*_\nu\}$.

Noise frequency components in the $\mathcal{M}$ area are typical candidates to the MAN phenomenon. Strictly speaking, the MAN area should actually be a function of $\bar{T}$, i.e. $\mathcal{M}_{\text{ideal}} = \{\nu : \bar{T}_\nu < \gamma_\nu \leq T_\nu\}$. Unfortunately, the problem in this case would be to introduce a dependance between the MAN area $\mathcal{M}_{\text{ideal}}$ and the filter coefficients (via $\bar{T}_\nu$). This may yield to a recursive processing, some convergence problems and numerical instability. Hence, for more clarity, the MAN area $\mathcal{M}$ considered in this paper is larger than $\mathcal{M}_{\text{ideal}}$.

Generally, the mean-square error $\alpha(h_\nu)$ at frequency $\nu$, defined by

$$\alpha(h_\nu)=\mathbb{E}[\hat{S}_\nu-s_\nu]^2 = (h_\nu-1)^2 \lambda_\nu + h_\nu^2 \gamma_\nu,$$

is used to measure the performance of the estimate $\hat{S}_\nu$ with respect to the clean speech signal $s_\nu$. In this paper, in order to take into account the MAN phenomenon, the quality of the estimate $\hat{S}_\nu$ is evaluated differently in each perceptual area $\mathcal{A}$, $\mathcal{M}$ and $\mathcal{I}$ by means of a modified mean-square error.

In the audible area $\mathcal{A}$, the efficiency of the denoising is evaluated by

$$\beta(h_\nu) = (h_\nu-1)^2 \lambda_\nu + h_\nu^2 (\gamma_\nu-T_\nu).$$

Here, $\beta(h_\nu)$ is typically the mean-square error $\alpha(h_\nu)$ where the noise spectrum $\gamma_\nu$ is replaced by the “audible” noise spectrum $\gamma_\nu - T_\nu$.

In the MAN area $\mathcal{M}$, the efficiency of the denoising is measured by

$$\beta(h_\nu) = (h_\nu-1)^2 \lambda_\nu + h_\nu^2 \gamma_\nu = \alpha(h_\nu).$$

In this area, the noise is initially inaudible but it must be reduced in order to avoid the MAN phenomenon after filtering. Here the perceptual auditive properties are ignored and the mean square error $\alpha(h_\nu)$ is kept since there is no reasonable justification to regularize it with a perceptual term.

Finally, in area $\mathcal{I}$, the noise is absolutely inaudible. Consequently, to avoid a too strong and useless attenuation of the clean speech in this area, the efficiency of the denoising is only characterized by the energy of the speech distortion:

$$\beta(h_\nu) = \mathbb{E}[|\xi \hat{S}_\nu|^2] = (h_\nu-1)^2 \lambda_\nu$$

Hence, in a nutshell, it is desirable to minimize the global criterion

$$\beta(\mathbf{H}) = \sum_{\nu=0}^{m-1} \beta(h_\nu)$$

4773
where $\beta(h_\nu)$ is given by:

$$\beta(h_\nu) = \begin{cases} (h_\nu - 1)^2 \lambda_\nu + h_\nu^2 (\gamma_\nu - T_\nu) & \text{if } \nu \in \mathcal{A}, \\ (h_\nu - 1)^2 \lambda_\nu + h_\nu^2 \gamma_\nu & \text{if } \nu \in \mathcal{M}, \\ (h_\nu - 1)^2 \lambda_\nu & \text{if } \nu \in \mathcal{I}. \end{cases} \quad (7)$$

### 3.2. Optimal anti-MAN perceptual filtering

The solution to the minimisation of the global criterion (6) can be found by solving the equation $d\beta(H)/dh_\nu = d\beta(h_\nu)/dh_\nu = 0$ using the method of Lagrangian multipliers. This leads after calculus to the optimum gain $h_\nu$ given by

$$h_\nu = \begin{cases} h_\nu^{(a)} & \text{if } T_\nu < \gamma_\nu, \\ h_\nu^{(b)} & \text{if } T_\nu < \gamma_\nu \leq T_\nu^*, \\ 1 & \text{if } \gamma_\nu \leq T_\nu^*, \end{cases} \quad (8)$$

where

$$h_\nu^{(a)} = \frac{\lambda_\nu}{\lambda_\nu + \max(\gamma_\nu - T_\nu, 0)} \quad \text{and} \quad h_\nu^{(b)} = \frac{\lambda_\nu}{\lambda_\nu + \gamma_\nu}. \quad (9)$$

To summarize, the AMPF filter $H = (h_\nu)_\nu$ performs i) the perceptual denoising $h_\nu^{(a)}$ only for audible noise frequency components (audible area $\mathcal{A}$) and ii) the Wiener denoising $h_\nu^{(b)}$ to attenuate the noisy speech signal within the MAN area $\mathcal{M}$. In this way, the AMPF filter attenuates the non-acceptable noise components that are initially masked by the audible components of the noisy speech signal but may induce the MAN phenomenon after the noise reduction in the audible area. Finally, when the noise spectrum value is below the absolute threshold of hearing (absolute inaudible area $\mathcal{I}$), there is no filtering in order to avoid unnecessary distortion of the clean speech.

### 4. EXPERIMENTAL RESULTS

The AMPF filter is experimentally compared to the following two filters i) the perceptual filter (LPF) proposed in [3] which is given by the diagonal matrix $H$ whose $(\nu + 1)$-th diagonal element is $h_\nu^{(a)}$ if $T_\nu < \gamma_\nu$ and 1 otherwise and ii) the double filtering (DF) proposed in previous work [5] which is the concatenation of two active filters $H = h_\nu^{(a)} h_\nu^{(b)}$.

The MAN phenomenon processing is illustrated in Fig. 1. The car noise signal from the Noisex database is added with SNR=$-5$ dB to one sentence randomly chosen from the Tidigits database downsampld to $f_s = 8$ kHz. Short-time windows (32 ms) of noisy speech, with 50\% overlap, are transformed into the frequency domain ($m = 256$) using the short-time Fast Fourier Transform (FFT). The masking threshold $T_\nu$ is computed by means of the Johnston model [6] applied to the clean speech signal (assumed to be known).

The noise spectrum $\gamma_\nu$ is assumed to be known in order to assess the filtering without taking the risk of introducing any bias due to noise spectrum estimation. The masking threshold $T_\nu$ is directly used to perform both the LPF and AMPF perceptual filterings. To improve the clarity of the figure, the clean and noisy speech signal spectra are not represented and the considered signal spectra are plotted over a reduced interval of frequencies. Initially, certain noise components are masked (they are under the masking threshold $T_\nu$). After LPF filtering, the masking threshold $T_\nu^{\text{LPF}}$ of the denoised speech is plotted and is below the initial masking threshold $T_\nu$. This is due to the attenuation of the clean speech signal after additive noise filtering. Consequently, some residual noise components, which are initially below $T_\nu$, exceed $T_\nu^{\text{LPF}}$ and become audible (this unmasked noise is located in Fig. 1 by ellipses). In the same figure, the residual noise and the
masking threshold $\tilde{T}_{\text{AMPF}}$ after the AMPF filtering are also represented. Thanks to this filtering, the MAN phenomenon is completely avoided: no noise component initially masked by $T_n$ becomes audible after AMPF filtering.

The performance of the proposed approach is evaluated by measuring the Segmental Signal to Noise Ratio (SSNR), the Modified Bark Spectral Distortion (MBSD) and the Perceptual Evaluation of Speech Quality (PESQ). For evaluation purpose, noise signals from the Noisex database (babble and car noise) are added with four SNRs ($-5, 0, 10$ and $15$ dB) to 250 sentences randomly chosen from the TIdigits database downsampled to 8 kHz. The auditory masking threshold is estimated by means of the Johnston model applied to the Wiener estimate of $S$. The enhanced speech signal is obtained using the overlap-and-add approach after transformation back into the time domain via the short-time inverse FFT. Fig. 2 and presents the average MBSDs, PESQs and SSNRs. According to this figure, AMPF demonstrates significant improvement over MMSE-STSA [7], LPF and DF when the speech signal is corrupted by babble noise, whatever the criterion. The same conclusion is true for speech signals corrupted by car noise (see Fig. 3). Those results correlate well with subjective ratings obtained in formal listening tests that All these results confirm the impact of the MAN phenomenon and the relevance of correcting it.

5. CONCLUSION

This paper discusses the idea that additive noise components in a noisy speech signal can be inaudible before perceptual speech enhancement, but become audible after noise reduction due to a change in the masking level as a result of the speech attenuation. To avoid this phenomenon a new optimal perceptual filter dedicated to unmasked noise prevention have been proposed. According to objective measurements, this filter performs better than standard perceptual ones especially the one processing audible noise only.

6. REFERENCES


Fig. 2. Mean MBSD, PESQ and SSNR of MMSE-STSA, LPF, DF and AMPF for speech corrupted by babble noise for the SNR values from $-5$ to $15$ dB.

Fig. 3. Mean MBSD, PESQ and SSNR of MMSE-STSA, LPF, DF and AMPF for speech corrupted by car noise for the SNR values from $-5$ to $15$ dB.