<table>
<thead>
<tr>
<th>Title</th>
<th>Improved speech presence probability estimation based on wavelet denoising</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Lun, DPK; Shen, TW; Hsung, TC; Ho, DKC</td>
</tr>
<tr>
<td>Issued Date</td>
<td>2012</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10722/202308">http://hdl.handle.net/10722/202308</a></td>
</tr>
<tr>
<td>Rights</td>
<td>IEEE International Symposium on Circuits and Systems Proceedings. Copyright © IEEE.; ©2012 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.; This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.</td>
</tr>
</tbody>
</table>
Abstract—A reliable estimator for speech presence probability (SPP) can significantly improve the performance of many speech enhancement algorithms. Previous work showed that a good SPP estimator can be obtained by using a smooth a-posteriori signal to noise ratio (SNR) function, which can be achieved by reducing the noise variance when estimating the speech power spectrum. In this paper, a wavelet based denoising algorithm is proposed for such purpose. We first apply the wavelet transform to the periodogram of a noisy speech signal to generate an oracle for indicating the locations of the noise floor in the periodogram. We then make use of that oracle to selectively remove the wavelet coefficients of the noise floor in the log multitaper spectrum (MTS) of the noisy speech. The remaining wavelet coefficients are then used to reconstruct a denoised MTS and in turn generate a smooth a-posteriori SNR function. Simulation results show that the new SPP estimator outperforms the traditional approaches and enables a significantly improvement in the quality and intelligibility of the enhanced speeches.

I. INTRODUCTION

For many frequency domain speech enhancement algorithms [1]-[2], clean speech estimators are often derived under the assumption that speech is always present. It is indeed not true in speech pauses or between spectral bins of the harmonics of a voiced speech. Speech presence probability (SPP) estimator is thus used [3] to help in detecting the non-speech frequency components for further suppression. However, traditional SPP estimators are not always accurate. Speech components can be wrongly suppressed and leads to large distortion in the enhanced speech. In [4], it was suggested that a good SPP estimator can be achieved by smoothing the a-posteriori SNR function, both temporally and spectrally, before applying to the estimation of the SPP. Temporal smoothing is achieved by using a time averaging method performed across speech frames. Spectral smoothing is achieved by using a pair of local and global filters applied to the noisy a-posteriori SNR function in each frame. The resulting SPP estimator achieves probabilities close to zero for speech absence and probabilities close to one for speech presence. Although such feature is extremely useful for suppressing frequency components of noise, it has the side effect that any error in the estimation of speech absence can lead to a sudden jump in the SPP function and give rise to the musical noise [1] in the enhanced speech. In fact, the local and global filters used in [4] resemble a multiresolution filter bank, but limited to two resolutions. It is obviously that a much higher degree of freedom will be acquired if the true wavelet filter banks are used. Besides, we shall benefit from the studies in the wavelet community when determining the various thresholds used in the smoothing process. In [5], we proposed a wavelet based denoising algorithm for improving the a-priori SNR estimation given a noisy speech. In that approach, the wavelet transform and the multitaper spectrum (MTS) estimation techniques were used together to improve the estimation of the true power spectrum of speech. In this paper, we modify that approach to improve the estimation of the a-posteriori SNR and apply to the estimation of SPP. The new SPP estimator is adopted in different speech enhancement algorithms, such as the MMSE-LSA [2]. It leads to better performance comparing with the traditional SPP estimators evaluated using different standard measures such as the segmental SNR (segSNR) [1] and the Perceptual Evaluation of Speech Quality (PESQ) [6].

II. SPP ESTIMATION FOR SPEECH ENHANCEMENT

Given that a speech signal \( x \) is contaminated by additive noise \( n \) such that \( y = x + n \). In frequency domain, the process can be modeled as: \( Y = X + \Psi \) where \( Y, X \) and \( \Psi \) are the Fourier transform of \( y, x \), and \( n \), respectively. To enhance the speech, traditional approaches first frame the noisy speech using different windowing techniques. A gain function is then derived with an aim to suppress the frequency components with low SNR in each frame. For instance, for the traditional MMSE-LSA approach, a gain function \( G_{\text{LSA}} \) is derived to multiply to the observed noisy speech spectrum as follows:

\[
\hat{X}(k,i) = G_{\text{LSA}}(k,i)Y(k,i)
\]

(1)

where \( k \) and \( i \) is the frequency and frame index, respectively. The resulting \( \hat{X} \) will be used to reconstruct the enhanced speech. As mentioned above, clean speech estimator such as \( X \) in (1) is derived under the assumption that speech is actually present, which is not true in practice. Recently, SPP estimator is used in many speech enhancement algorithms to allow speech and noise frequency components to be treated
differently and appropriately. For instance, in Cohen’s algorithm given in [3], the gain function is modified as:

$$G_{LSA+SP}(k) = \left(G_{LSA}(k)\right)^{p(k)} \cdot \left(G_{noise}(k)\right)^{1-p(k)}$$

(2)

where $p(k)$ is the SPP and $G_{noise}$ is chosen to be a small constant less than 1. The frame index $i$ has been dropped to simplify the presentation. Many approaches have been suggested for estimating $p(k)$. Recently, Gerkmann et al. proposed an improved SPP estimator [4] and suggested that a good SPP estimator can be obtained by smoothing the a-posteriori SNR function, both temporally and spectrally, before applying to the estimation of the generalized likelihood ratio (GLR):

$$\Lambda = \frac{q}{1 - q} \frac{p(\gamma | H_0)}{p(\gamma | H_1)}$$

(3)

where $\Lambda$ is the GLR and $q$ is the a-priori speech presence probability. $\gamma$ is the a-posteriori SNR defined as follows:

$$\gamma_k = \left|p(k)\right|^2 / \hat{S}_n(k)$$

(4)

where $\hat{S}_n$ is the estimated noise power spectrum. $p(\gamma | H_1)$ is the probability density function (pdf) of $\gamma$ under the hypothesis $H_1$, i.e. speech is present. Similarly, $p(\gamma | H_0)$ is the pdf of $\gamma$ under the hypothesis $H_0$, i.e. speech is absent. The SPP can be computed based on the GLR as follows:

$$spp = P(H_1 | \gamma) = \frac{\Lambda}{1 + \Lambda}$$

(5)

In [4], the temporal smoothing is achieved by using a time averaging method performed across speech frames. Spectral smoothing is achieved by using a pair of local and global filters applied to $\gamma$ in each frame. They have different orders for smoothing the spurious impulses and strong spectral variations in $\gamma$. However, due to various reasons, the filter pair can leave behind many spurious impulses in $\gamma$ that often lead to the musical noise problem in the enhanced speeches.

III. SMOOTHING THE A-POSTERIORI SNR

As a smooth $\hat{S}_n(k)$ can be relatively easy to obtain [5], $\gamma$ can be smoothed if we can reduce the noise variance in $\left|p(k)\right|^2$. One of the traditional approaches is to use the MTS estimation technique plus wavelet denoising [7]. First, for a noisy frame $y$, we obtain its MTS as follows:

$$\hat{S}_n^{mt}(k) = \frac{1}{L} \sum_{l=1}^{L} \hat{S}_n^l(k)$$

(6)

where $\hat{S}_n^l(k) = \left|\sum_{n=0}^{M-1} a_l(n) y(n) e^{-j2\pi kn / M}\right|^2$. Since the tapers $a_l$, for $l = 1 \ldots L$, are designed to be orthonormal, the noise variance of $\hat{S}_n^{mt}$ is reduced by $L$ times [7].

To further reduce the noise variance, the wavelet shrinkage with either the universal threshold [8] or the SURE shrink threshold [9] is applied to the log MTS of the noisy speech. Nevertheless, this approach was found to be ineffective [5]. For a noisy speech power spectrum, we can often find spectral peaks contributed by the speech and/or the input additive noise (for certain kinds of colored noise). On the other hand, we can also find regions where no spectral peaks can be found. Let us call these regions as the noise floor. It is important to have a smooth noise floor since any large variance noise exists on the noise floor will likely contribute to the annoying musical noise in the enhanced speech. Although the noise floor contains no spectral peak, we have shown in [5] that its log MTS can have large wavelet coefficients with magnitude similar to those of speech. Directly thresholding the wavelet coefficients in the log MTS domain cannot smooth the noise floor.

IV. PROPOSED SMOOTHING ALGORITHM

Similar to the idea in [5], we solve the abovementioned problem by first deriving from the periodogram of the noisy speech an oracle that can indicate the locations of the wavelet coefficients of the noise floor in the log MTS. More specifically, let $\hat{S}_n^{y}(k)$ be the periodogram of a noisy speech frame $y$ generated using a taper $a_t$ and $w_{j}^l = W[l] \hat{S}_n^{y}(k)$ be its level $j$ wavelet coefficients, where $W[l]$ is the wavelet transform. While $\hat{S}_n^{y}$ behaves also like a noisy signal, $w_{j}^l$ will be scatter for all $j$ with magnitude depends on the local variance of $\hat{S}_n^{y}$.

IV. PROPOSED SMOOTHING ALGORITHM

Similar to the idea in [5], we solve the abovementioned problem by first deriving from the periodogram of the noisy speech an oracle that can indicate the locations of the wavelet coefficients of the noise floor in the log MTS. More specifically, let $\hat{S}_n^{y}$ be the periodogram of a noisy speech frame $y$ generated using a taper $a_t$ and $w_{j}^l = W[l] \hat{S}_n^{y}$ be its level $j$ wavelet coefficients, where $W[l]$ is the wavelet transform. While $\hat{S}_n^{y}$ behaves also like a noisy signal, $w_{j}^l$ will be scatter for all $j$ with magnitude depends on the local variance of $\hat{S}_n^{y}$. Let us denote $(\sigma_{nfloor}^{y})^2$ to be the variance of the wavelet coefficients of the noise floor at level $j$. Hence if $\hat{S}_n^{y}$ contains a spectral peak at frequency bin $k$, then

$$|w_{j}^l(k)| > \sigma_{nfloor}^{y} \quad \forall j$$

(7)

This allows us to use a simple thresholding scheme to identify the wavelet coefficients of the noise floor. Firstly we need to have a good estimation of $\sigma_{nfloor}^{y}$. Following the same argument as in [10], it is reasonable to regard $w_{j}^l(k)$ as approximately Gaussian distributed since $w_{j}^l(k)$ are just linear combinations of $\hat{S}_n^{y}$, which are independent random variables. Now let us assume we have a Voice Activity Detector (VAD) (such as [11]) which can help us identify the noise floor. Let $w_{j}^{y}(k)$ be the wavelet coefficients of those noise frames. Hence based on robust statistics,

$$\sigma_{nfloor}^{y} = \text{median}\left\{w_{j}^{y}(k)\right\} / 0.6745 \quad \forall j$$

(8)

Here we assume the noise power spectrum contains only a limited number of spectral peaks. Such assumption applies to white noise and many kinds of colored noise such as pink noise, high frequency noise, etc. Since spectral peaks are rare, most of the wavelet coefficients are on the noise floor. Hence the median of their magnitude is close to their standard derivation. Based on $\sigma_{nfloor}^{y}$, we can develop a threshold $\text{thr}_j$ such that if $|w_{j}^l(k)| < \text{thr}_j$, we consider $w_{j}^l(k)$ to belong to the noise floor. Since $w_{j}^l(k)$ is approximately Gaussian distributed, we propose to use the following level dependent universal threshold to carry out the above classification:
where $M_j$ is the number of wavelet coefficients at level $j$. The universal threshold is chosen because, given a set of normal distributed random variables, the universal threshold is their maximum limit asymptotically [8]. Note that the threshold $thr_j$ is level dependent since $S_y^u$ is in general not white. To further remove the outliers in the classification, the results of using two orthonormal tapers can be combined as follows:

\[
V_j(k_j) = \begin{cases} 
1 & \text{if } \left| w_{y,j}^1(k_j) \right| > thr_j \text{ and } \left| w_{y,j}^2(k_j) \right| > thr_j \\
0 & \text{otherwise}
\end{cases}
\]  
(10)

Note that the actual denoising operation is performed in the log MTS domain. With the locality property of the wavelet transform, the oracle (10) can also be used for the classification of the wavelet coefficients in the log MTS domain. Let $w_{y,j}^{lm}(k_j) = W_j \log(S_y^{lm})$. A hard thresholding procedure is applied to $w_{y,j}^{lm}(k_j)$ based on the oracle $V_j$:

\[
\hat{w}_{y,j}^{lm}(k_j) = \begin{cases} 
w_{y,j}^{lm}(k_j) & \text{if } V_j(k_j) = 1 \\
w_{y,j}^{lm}(k_j) & \text{if } \left| w_{y,j}^{lm}(k_j) \right| > thr_2j \text{ and } V_j(k_j) \geq \varepsilon = 1 \\
0 & \text{otherwise}
\end{cases}
\]  
(11)

for all $j$. In general if the oracle $V_j(k_j) = 1$, it indicates that the wavelet coefficient $w_{y,j}^{lm}(k_j)$ at $k_j$ likely belongs to a true spectral peak. Hence it should be kept. Besides, the wavelet coefficients $w_{y,j}^{lm}(k_j)$ in the vicinity of $k_j$ also has a good chance to belong to a true spectral peak, particularly if it has a large magnitude. For the proposed hard thresholding procedure in (11), $w_{y,j}^{lm}(k_j)$ is kept if the oracle $V_j(k_j) = 1$.

Besides, if $\left| w_{y,j}^{lm}(k_j) \right| > thr_2j$, $w_{y,j}^{lm}(k_j)$ should also be kept if in the vicinity of $k_j$, i.e. $k_j \pm \varepsilon$, there is a $V_j(k_j \pm \varepsilon)$ equals to 1. The threshold $thr_2j$ is obtained using the standardSUREshrink approach [9] due to the fact that the log MTS of a speech signal in general is not white. The SUREshrink usually can give a more accurate threshold than the universal threshold particularly for colored noises. The limit of $\varepsilon$ is selected to be $\varepsilon \leq \lfloor l_a/2 \rfloor$  
(12)

where $l_a$ is the length of the wavelet filter and $\lfloor x \rfloor$ stands for the nearest integer smaller than $x$. $\varepsilon$ is selected as in (12) because it can be shown that if there is a signal change in $\hat{S}_y^u$ and $\log(S_y^{lm})$ at frequency index $k$ that will lead to strong wavelet coefficient $w_{y,j}^1(k_j)$ and $w_{y,j}^{lm}(k_j)$, respectively, at level $j$, then $\left| k_a - k_j \right|$ is bounded approximately by $\lfloor l_a/2 \rfloor$. The resulting wavelet coefficients $\hat{w}_{y,j}^{lm}(k_j)$ are then used to reconstruct $\hat{S}_y$. Note that the noise variance of $\hat{S}_y$ is significantly reduced since, firstly the variance of the noise floor is suppressed by (11); secondly the noise variance of the spectral peaks is also reduced due to the use of the MTS estimation technique. A smooth a-posteriori SNR function can then be generated as follows:

\[
\hat{\gamma}_k = \frac{\hat{S}_y(k)}{S_n(k)}
\]  
(13)

Following the same approach as in [4], $\hat{\gamma}_k$ is further smoothed temporally by averaging with the $\hat{\gamma}_k$ obtained in the last 4 frames. Then it can be used for the estimation of GLR. Since the smoothing procedure has been changed, the procedure for estimating the GLR also needs to be slightly revised. To be specific, $p(\gamma | H_k)$ needs to be estimated by directly computing the histogram of $\gamma$ in noise frames detected using a VAD. $p(\gamma | H_k)$ can be estimated using the same approach as in [4]. However, the degree of freedom parameter needs to be adjusted to 5 to adapt to the change in the smoothing procedure. Due to page limit, detailed analysis about such modification cannot be included here. Fig.1 shows a comparison of the SPP estimated by using different approaches for a typical noisy speech frame. As can be seen in Fig.1b, the SPP given by [4] can contain large spurious impulses (such as near $\pi/2$). They often lead to the musical noise problem in the final enhanced speech. On the other hand, the SPP estimator using the smoothing method in [7], as shown in Fig.1c, can merge all spectral peaks of the speech. It is due to the universal threshold adopted in [7] over-kills many speech wavelet coefficients. Since the proposed approach can accurately estimate the spectral locations of the noise floor, the resulting SPP as shown in Fig.1a can largely preserve the spectral peaks while achieve a good control of the spurious impulses on the noise floor.

### V. SIMULATION RESULTS

A series of simulations have been performed for comparing the performance between different SPP estimators when applying to the traditional MMSE-LSA [2]. They include the LSA+SPP4 approach which uses the traditional SPP estimator given by Cohen [3]; the LSA+uthFPSPP approach that uses the SPP estimator given in [4]. Both the LSA+uthFPSPP and the proposed LSA+2sFPSPP approaches use a similar framework for SPP estimation. They are different only in the way to denoise the speech power spectrum, i.e. using either the proposed wavelet denoising algorithm or the wavelet based MTS denoising with universal thresholding [7]. The basic simulation settings are similar to those described in [1]. For the testing data set, we arbitrarily selected 40 male and 40 female test speeches from the TIMIT database [12]. White noise and different kind of colored noises adopted from the NOISEX-92 database [13] were added to the speeches with different input segSNRs. The resulting enhanced speeches generated by all algorithms were evaluated using standard measures including segSNR [1] and PESQ [6]. The results are shown in Fig.2. It can be seen that the performance of the proposed algorithm is always one of the best. More specifically, when comparing with LSA+FPSPP [4], the proposed LSA+2sFPSPP always gives a similar PESQ score but an about 0.4dB improvement in segSNR. When comparing with LSA+uthFPSPP, the proposed LSA+2sFPSPP gives a similar segSNR but an about 0.2 improvement in PESQ score. Finally, the proposed LSA+2sFPSPP always improves over LSA+SPP4 both in segSNR and PESQ score.
In Fig.3, the spectrograms generated by different enhancement algorithms are compared. The original speech is added with color (pink) noise at input segSNR 0dB. It can be seen that the proposed algorithm in general preserves better the speech contents while effectively removing the background noise. Improvement can easily be seen in particularly the circled part in the spectrograms.

VI. SUMMARY

In this paper, we proposed a new algorithm for the estimation of SPP of a noisy speech signal. Rather than using a pair of arbitrarily defined local and global filters as in [4], the proposed approach makes use of the wavelet transform and the MTS estimation technique to smoothen the a-posteriori SNR. It significantly reduces the spurious impulses in the function. As compared with the previous wavelet based approach, the proposed approach better preserves the spectral peaks which are important to the intelligibility of the enhanced speech. When applying the new SPP estimator to the MMSE-LSA speech enhancement algorithm, significant improvement in segSNR and PESQ was noted as compared with other SPP estimators for different noises at different noise levels.

REFERENCES