DESIGN OF OPTIMAL WAVELETS FOR DETECTING IMPULSE NOISE IN SPEECH

R. C. Nongpiur, D. J. Shpak, and P. Agathoklis

Department of Electrical and Computer Engineering, University of Victoria, Canada V8W 3P6

ABSTRACT

Removal of impulse noise from speech in the wavelet domain has been found to be very effective due to the multi-resolution property of the wavelet transform and the ease of removing the impulses in that domain. A critical factor that affects the performance of the impulse-removal system is the effectiveness of the impulse detection algorithm. To this end, we propose a new method for designing orthogonal wavelets that are optimized for detecting impulse noise in speech. In the method, the characteristics of the impulse noise and the underlying speech signal are taken into account and a convex optimization problem is formulated for deriving the optimal wavelet for a given support size. Performance comparison with other well-known wavelets show that the wavelets designed using the proposed method have much better impulse detection properties.

***Index Terms*—** impulsive noise detection, wavelet design, speech enhancement

1. INTRODUCTION

The presence of impulse-like noise in speech can significantly reduce the intelligibility of speech and degrade automatic speech recognition (ASR) performance. Impulse noise is characterized by short bursts of acoustic energy having a wide spectral bandwidth and consisting of either isolated impulses or a series of impulses. Typical acoustic impulse noises include sounds of clicks in old phonograph recordings, of rain drops hitting a hard surface like the windshield of a moving car, of popping popcorn, of typing on a keyboard, of indicator clicks in cars, and so on.

Recently, several methods for detection and/or removal of transient and impulse noise have been reported. In [1], impulse noise was removed from audio signals by fusing multiple copies of the same recording, while in [2], the spectral coherence and harmonic property of speech were used to distinguish transient noise from speech. Classical block processing methods such as the STFT algorithm or the linear prediction (LP) algorithm have also been used to detect or remove impulse-like sounds [3, 4, 5]. However, two problems may result if classic block processing techniques are used: the first is determining the exact position of the impulse within the analyzed data-frame – these methods give no straightforward information about the position of the impulse within the analyzed frame. It is possible, however, to reduce the frame size to achieve better resolution in time; but doing this leads to the second problem where we lose the frequency resolution needed to effectively analyze the signal. The wavelet transform overcomes both of these difficulties due to its multi-resolution property [6]. In multi-resolution analysis, the window length or wavelet scale for analyzing the frequency components increases as the frequency decreases. This property enables the wavelet transform to have better time resolution for higher frequency components and better frequency resolution for lower ones. Consequently, by using the wavelet transform we have a relationship between time resolution and frequency resolution that is beneficial for detecting and removing impulse noise.

The use of the Daubechies wavelet has been found to be quite effective in the detection and removal of impulse noise from speech or audio [7, 8]. Though such a wavelet may be very effective in one application, it may not be quite as effective in another where the properties of the impulse noise and the underlying signal are different. Therefore, to enable the designer select the appropriate wavelet for a given application, a connection between certain wavelet features and impulse detection performance was made in our recent work [9]. In that work, we showed how the wavelet impulse-detection features are dependent on the characteristics of the impulse noise and the underlying signal, and provided a procedure for selecting the most appropriate wavelet from a set of pre-designed wavelets. The method, however, has one drawback: the quality of the selected wavelet is dependent on the quality of the wavelets within the set. If none of the wavelets within the set are optimal for the given application, the method will not be effective.

In this paper, we seek to remove the drawback in our previous work [9] by designing wavelets that are most appropriate for a given application. Utilizing the relationships between wavelet features and impulse detection performance [9], we formulated an optimization problem for designing a wavelet of certain support size that is tailored for detecting impulses for a given application. The formulations are framed as a convex optimization problem where the solution obtained corresponds to the FIR filter coefficients of an orthogonal wavelet. The subsequent performance comparison results with other well-known wavelets show that the wavelets designed using the proposed method have much better impulse detection features.

The paper is organized as follows. Section 2 summarizes the wavelet properties that are important for impulse detection and shows their dependence on the nature of the impulse noise and the underlying speech signal. In Section 3, we develop formulations to obtain the filter coefficients of the optimal wavelet for a given support size. Then in Section 4, simulation experiments are presented to compare the impulse detection performance of wavelets derived using the proposed method with other well-known wavelets. Conclusions are drawn in Section 5.

2. DETECTION OF IMPULSE NOISE FROM SPEECH

In this section, we summarize the wavelet properties that influence the detection performance and describe a measure for evaluating the detection performance.

2.1. Wavelet properties and features for impulse detection

A desirable wavelet for impulse detection is one that maximizes the coefficients for the impulse relative to the underlying signal in the finest scale [9]. Such a wavelet will correspondingly have a highpass analysis filter that maximizes the impulse noise relative to the underlying speech and background noise signals. If $P_s(\omega)$ and $P_i(\omega)$ are the power spectrums of the average speech and impulse noise power, respectively, then the ratio between the average impulse noise power and speech power in the finest scale, R_i, is dependent on the wavelet highpass analysis filter and given by

$$R_i = \frac{\sigma_i^2}{\sigma_s^2} \tag{1}$$

where

$$\sigma_i^2 = \int_{-\pi}^{\pi} |G(e^{j\omega})|^2 P_i(\omega) d\omega \approx \sum_i |G(e^{j\omega_i})|^2 P_i(\omega_i) \tag{2}$$

$$\sigma_s^2 = \int_{-\pi}^{\pi} |G(e^{j\omega})|^2 P_s(\omega) d\omega \approx \sum_i |G(e^{j\omega_i})|^2 P_s(\omega_i) \tag{3}$$

and $G(z)$ is the transfer function of the wavelet highpass filter. The design of an optimal wavelet for detecting the impulses should, therefore, seek to maximize R_i.

The other factor that influences the detection performance is the size of the wavelet support, which is dependent on the average width and energy of the impulse noise [9]. One way to determine the correct wavelet support for a given application is to design wavelets that maximize R_i at various wavelet support sizes and then select the one with the best detection performance.

2.2. Metrics to evaluate the detection performance

To determine the most appropriate wavelet for impulse detection, we evaluate the discriminatory capability of the wavelet coefficients in the finest scale, with respect to the impulse noise. This is done by using a stability criterion derived from the scatter matrices [9]. For a one-dimensional, two-class scenario, the separability criterion for feature x is given by

$$J = \frac{n_1(m_1 - m)^2 + n_2(m_2 - m)^2}{\sum_{x\in\omega_1} (x - m_1)^2 + \sum_{x\in\omega_2} (x - m_2)^2} \tag{4}$$

where (m_1, n_1) and (m_2, n_2) are the means and number of feature samples for classes ω_1 and ω_2, respectively. It has been shown [9] that a wavelet with a higher value of J will correspondingly have better detection performance.

3. DERIVING THE OPTIMAL WAVELETS FOR IMPULSE DETECTION

The optimal wavelets are designed to maximize the ratio of impulse noise power to speech power in the finest scale. At the same time, the necessary constraints required for an orthogonal wavelet need to be imposed.

If $H(z)$ corresponds to the transfer function of a lowpass analysis filter of an orthogonal wavelet given by

$$H(z) = h(0) + h(1)z^{-1} + \cdots + h(L-1)z^{-(L-1)} \tag{5}$$

then the highpass counterpart, $G(z)$, can be obtained by taking the alternating flip of $H(z)$ [10]; that is

$$G(z) = -z^{-(L-1)} H(-z^{-1}) \tag{6}$$

where L is assumed to be even. To ensure that the wavelet filterbank is orthogonal, the filter coefficients need to satisfy the *double-shift orthogonality* condition [10], given by

$$\sum_n h(n)h(n-2k) = \delta(k), \text{ for } k = 0, 1, \ldots, (L/2) - 1 \tag{7}$$

where $\delta(k)$ is the delta function. For the existence of the wavelet $\psi(t)$, the following condition must also hold true [11]:

$$H(e^{j\omega})|_{\omega=0} = \sum_n h(n) = \sqrt{2} \tag{8}$$

As in the design of signal-adapted filterbanks by Moulin et al[12], the formulation of the optimization problem becomes more tractable if we use the autocorrelation sequence of the filter coefficients given by

$$r_h(l) = \begin{cases} \sum_{n=0}^{L-l-1} h(n)h(n+l) & l \geq 0 \\ r_h(-l) & l < 0 \end{cases} \tag{9}$$

Therefore, in terms of the aurocorrelation parameters, the double shift orthogonality condition in (7) can be expressed as

$$r_h(2k) = \delta(k), \text{ for } k = 0, 1, \ldots, \left\lfloor \frac{L-1}{2} \right\rfloor \tag{10}$$

and the necessary condition in (8) as

$$\sum_{m=1}^{L-1} r_h(m) = 0.5 \tag{11}$$

by exploiting the orthogonality condition in (7) and the symmetry property in (9). Correspondingly, using (6) and (9) in (2) and (3) the average power of the impulse noise and speech in the finest scale are given by

$$\begin{aligned} \sigma_i^2 &\approx \sum_n \left[r_h(0) + 2\sum_{l=1}^{L-1} (-1)^l r_h(l) \cos(\omega_n l) \right] P_i(\omega_n) \\ &= \mathbf{1}^T \mathbf{C}_i \mathbf{A} \mathbf{r} \end{aligned} \tag{12}$$

$$\begin{aligned} \sigma_s^2 &\approx \sum_n \left[r_h(0) + 2\sum_{l=1}^{L-1} (-1)^l r_h(l) \cos(\omega_n l) \right] P_s(\omega_n) \\ &= \mathbf{1}^T \mathbf{C}_s \mathbf{A} \mathbf{r} \end{aligned} \tag{13}$$

where

$$\mathbf{r} = [r_h(0) \cdots r_h(L-1)]^T \tag{14}$$

$$\mathbf{A} = \begin{bmatrix} a_{00} & \cdots & a_{0(L-1)} \\ \vdots & \vdots & \vdots \\ a_{(N-1)0} & \cdots & a_{(N-1)(L-1)} \end{bmatrix} \tag{15}$$

$$\mathbf{C}_i = \text{diag}(c_0^{(i)}, \ldots, c_{(N-1)}^{(i)}) \tag{16}$$

$$\mathbf{C}_s = \text{diag}(c_0^{(s)}, \ldots, c_{(N-1)}^{(s)}) \tag{17}$$

$$a_{nl} = 2(-1)^l \cos(\omega_n l) \tag{18}$$

$$c_n^{(i)} = P_i(\omega_n), \quad \omega_n \in [-\pi, \pi] \tag{19}$$

$$c_n^{(s)} = P_s(\omega_n), \quad \omega_n \in [-\pi, \pi] \tag{20}$$

and N is the number of samples. The optimization is formulated as the minimization of σ_s^2 while keeping σ_i^2 constant so that R_i in

(1) is maximized. Consequently, after incorporating the double-shift orthogonality constraint in (10) and the necessary condition in (11), the optimization problem is given by

$$\begin{aligned} \text{minimize} \quad & \sigma_s^2 \\ \text{subject to:} \quad & \sigma_i^2 = \text{constant} \\ & r_h(2m) = 0, \text{ for } m = 1, \dots, \left\lfloor \frac{L-1}{2} \right\rfloor \\ & r_h(0) = k \\ & \sum_{m=1}^{L-1} r_h(m) = 0.5k \end{aligned} \tag{21}$$

where k and $r_h(m)$ are optimization variables. Note that the last two equality constraints in (21) ensure that the necessary condition in (11) is satisfied when we set $r_h(0) = 1$. Replacing σ_s^2 and σ_i^2 by their matrix representations, (21) can be expressed as a convex optimization problem given by

$$\begin{aligned} \text{minimize} \quad & \mathbf{1}^T \mathbf{C}_s \mathbf{A} \mathbf{r} \\ \text{subject to:} \quad & \mathbf{1}^T \mathbf{C}_i \mathbf{A} \mathbf{r} = \text{constant} \\ & r_h(2m) = 0, \text{ for } m = 1, \dots, \left\lfloor \frac{L-1}{2} \right\rfloor \\ & r_h(0) = k \\ & \sum_{m=1}^{L-1} r_h(m) = 0.5k \\ & \mathbf{A} \mathbf{r} > \mathbf{0} \end{aligned} \tag{22}$$

where $\mathbf{r}$ and k are optimization variables and $\mathbf{0} \in \mathbf{R}^N$. The inequality constraint in (22) is a positivity constraint to ensure that the magnitude is always positive. Once we obtain the optimal autocorrelation vector $\mathbf{r}_{opt}$, we recover the minimum-phase low-pass wavelet filter coefficients $h_{mp}(n)$ from $\mathbf{r}_{opt}$ using spectral factorization [13]. The filter coefficients obtained are then appropriately scaled so that the necessary condition in (8), or equivalently in (11), is satisfied.

4. EXPERIMENTAL RESULTS

In this section we perform experiments to compare the impulse detection performance of wavelets designed using the proposed method with other well-known wavelets.

To generate the impulse noise signals for carrying out the experiments we use an impulse-noise generation model [14] that has been found to be a good representation for speech signals degraded by clicks. The model, reproduced in Fig. 1, uses two noise generation processes. The first is a binary noise generation process, $i(n)$, that controls a switch. The switch is connected when $i(n) = 1$, thereby enabling a second noise process, $\eta_a(n)$ to be added to the speech signal $x(n)$. As can be seen, the noise produced by such a system occurs in bursts, where its value is precisely zero for at least some of the time. A typical audio signal degraded with impulse noise can have an average impulse width of around 1 ms while the fraction of the signal that is contaminated is usually less than 20 percent [15]. If α is the fraction of signal samples contaminated by impulse noise the average signal to impulse noise ratio is given by [16]

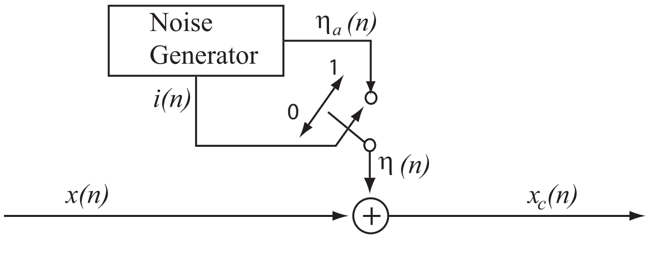

Fig. 1. Impulse noise generation model.

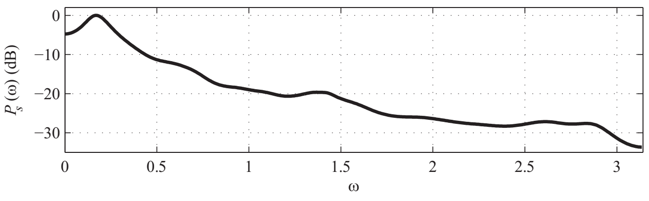

Fig. 2. Normalized average power spectrum of speech. The sampling frequency is 16 kHz.

$$SINR = \frac{P_s}{\alpha P_i} \tag{23}$$

where P_s is the power of the speech signal and P_i is the power of the impulse. For our experiments, we set the contamination level to 5 percent, which is a typical level for audio degraded by impulse noise [16]. The binary noise generation process for $i(n)$ is implemented using a two-state Markov chain where the transition probabilities can be appropriately adjusted to have the desired average impulse width and contamination level. The second noise process, $\eta_a(n)$, is generated using a normal distribution.

To evaluate the detection performance of the wavelets, we compare the discriminatory capability of the impulse-detection features of the wavelet by using the separability criterion J in (4). To compute J, the detection features need to be first classified into either class ω_1 or class ω_2: Class ω_1 if the features correspond to an impulse, and ω_2 otherwise. After the features have been classified, we then use (4) to obtain J.

The signal from the first level, which corresponds to the finest scale, is the one that is used to detect the impulses. To carry out the classification of the detection features in ω_1 and ω_2, the discrete wavelet transform of the clean speech signal and the impulse noise are taken separately. If $x_f^{(s)}(n)$ and $x_f^{(i)}(n)$ are the wavelet coefficients of the clean speech and impulse noise in the finest scale, respectively, the classification of the features in the two classes is given by

$$\mathcal{F}(n) \in \begin{cases} \omega_1 & \text{if } |x_f^{(i)}(n)| > 0 \\ \omega_2 & \text{otherwise} \end{cases} \tag{24}$$

where

$$\mathcal{F}(n) = |x_f^{(s)}(n) + x_f^{(i)}(n)| \tag{25}$$

The speech signal used in the experiments is clean near-microphone speech taken from the ATIS corpus database [17], with a sampling frequency of 16 kHz. The total duration of the signal used for computing J is about 5 minutes long with a total of 3 male and 3 female speakers. In Fig. 2, the average power spectrum of the speech signal, $P_s(\omega)$, is shown. The optimal wavelet filter coefficients are designed as in Section III by solving the optimization problem in (22) to obtain the optimal autocorrelation values and then performing spectral factorization with appropriate scaling to derive the wavelet lowpass filter coefficients. For the optimization, we use the speech power spectrum shown in Fig. 2 to compute $\mathbf{C}_s$ in (17). Since the generated impulse noise has an average spectrum that is flat we normalize

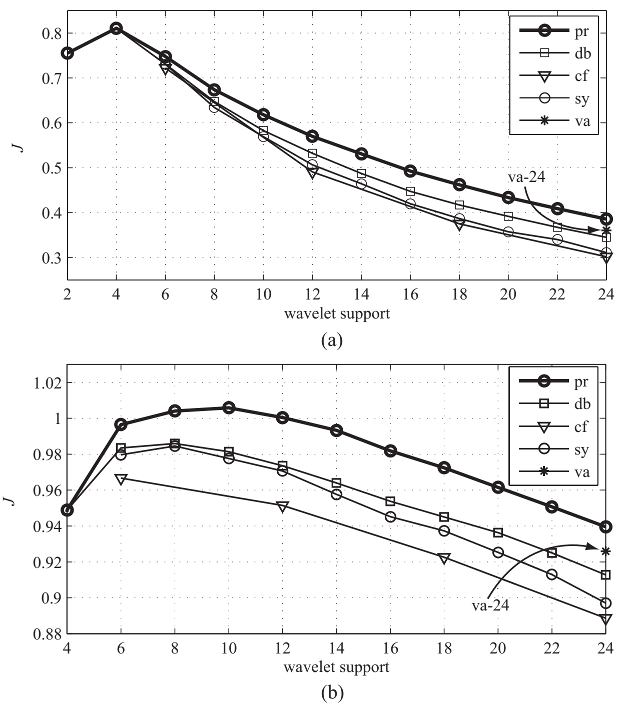

Fig. 3. Comparison plots of J versus support size when the SINR is 10 dB for the cases when (a) the average impulse width = 1 ms (b) the average impulse width = 15 ms. Note that the 'va' wavelet is only a single point with a support size of 24.

$P_i(\omega) = 1$ and, as a result, $\mathbf{C}_i$ in (16) simplifies to an identity matrix. For our experiments, twelve wavelets ranging from orders 2 to 24 were designed and their corresponding low-pass filter coefficients have been made available online [18]. In the figures, the wavelets designed using the proposed approach are denoted as 'pr'.

For the comparison, we consider various wavelets taken from either the WAVELAB toolbox [19, 20] or the MATLAB Wavelet Toolbox: Daubechies ('db') orders 2-24, Coiflet ('cf') orders 6-24, Symmlet ('sy') orders 6-24, and Vaidyanathan ('va') order 24.

Two experiments are carried out to compare the wavelet impulse-detection performance. In the first experiment, we compare the detection performance using impulse noise with two different average widths while keeping the SINR constant. In the second experiment, we compare the detection performance for impulse noises with different SINR levels but having the same average widths.

4.1. Experiment 1

In this experiment, we consider two impulse noises that have the same SINR but different average widths and use them to compare the detection performance of the wavelets for different support sizes. The first impulse noise has an average impulse-width of 1 ms while the second has a width of 15 ms. The SINR is set to 10 dB in both cases. In Figs. 3(a) and (b), the separability parameter, J, is compared for different wavelet support sizes. As can be seen from the figures, the performance of wavelets designed using the proposed method is equal to or better than all of the competing wavelets. We also observe that this performance improvement tends to get better relative to the other wavelets as the support size increases; this is because the increase in wavelet support corresponds to an increase in the number of wavelet filter coefficients, thereby allowing more degrees of freedom in the optimization. Furthermore, comparing the plots between Figs. 3(a) and (b) we observe that the optimal wavelet support size is larger for the impulse noise that has larger average impulse width. This is in accordance with the conclusions drawn in our previous work [9].

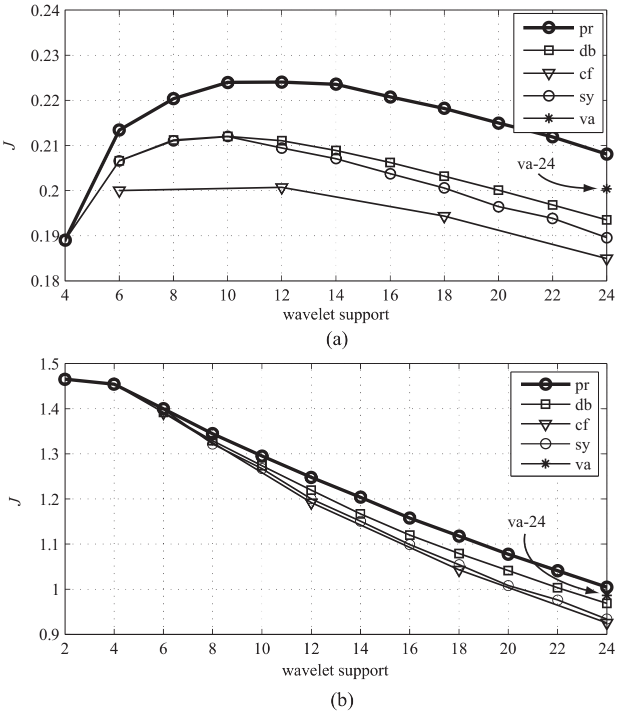

Fig. 4. Comparison plots of J versus support size when the average impulse width is 5 ms for the cases when (a) the SINR is 20 dB (b) the SINR is 0 dB. Note that the 'va' wavelet is only a single point with a support size of 24.

4.2. Experiment 2

In this experiment, we consider two impulse noises that have the same impulse width but different SINRs and use them to compare the detection performance of the wavelets with different support sizes. The first impulse noise has an SINR of 0 dB while the second has an SINR of 20 dB. The average impulse width is set to 5 ms in both cases. In Figs. 4(a) and (b), curves of the separability parameter, J, versus the wavelet support size are plotted for the various wavelets. As can be seen, the curve corresponding to the wavelets designed using the proposed method show the highest separability at all of the wavelet support sizes. And as in Experiment 1, the improvement over the competing wavelets tends to get better as the support size increases. Comparing the plots between Figs. 4(a) and (b) we observe that the optimal wavelet support size is larger for the impulse noise with larger SINR, in accordance with the results in our previous work [9].

5. CONCLUSION

A new method for designing orthogonal wavelets that are optimized for detecting impulse noise in speech has been described. In the method, the characteristics of the impulse noise and the underlying speech signal are taken into account and a convex optimization problem was formulated for deriving the optimal wavelet for a given support size. Performance comparison with other well-known wavelets showed that the wavelets designed using the proposed method have superior impulse detection properties.

6. REFERENCES

[1] P. Sprechmann, A. Bronstein, J.-M. Morel, and G. Sapiro, "Audio restoration from multiple copies," *Proceedings of ICASSP 2013*, pp. 878-882.

[2] C. Zheng, X. Chen, S. Wang, R. Peng, and X. Li, "Delayless method to suppress transient noise using speech properties and spectral coherence," *Proceedings of the 135th AES Convention*, New York, USA (2013).

[3] Z. Liu, A. Subramanya, Z. Zhang, J. Droppo, and A. Acero, " Leakage model and teeth clack removal for air- and bone-conductive integrated microphones," in *Proceedings of ICASSP 2005*, vol. 1, pp. 1093-1096.

[4] S. V. Vaseghi and R. Frayling-Cork, "Restoration of old gramophone recordings," *J. Audio Eng. Soc.*, **40**, 791-801 (1992).

[5] J. A. Moorer, "Dsp restoration techniques for audio," *Proceedings of ICIP 2007*, vol. 4, pp. 5-8.

[6] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed. (Academic, San Diego, 1998), pp. 221-228.

[7] S. Montresor, J. C. Valiere, J. F. Allard, and M. Baudry, "The restoration of old recordings by means of digital techniques," in *Proceedings of the 88th AES Convention*, Montreux, Switzerland (1990).

[8] R. C. Nongpiur, "Impulse noise removal in speech using wavelets," *Proceedings of ICASSP 2008*, pp. 1593-1597.

[9] R. C. Nongpiur and D. J. Shpak, "Impulse-noise suppression in speech using the stationary wavelet transform,"*J. Acoust. Soc. Am.*, **133**(2), 866-879 (2013).

[10] G. Strang and T. Nguyen, *Wavelets and filter banks*, Wellesley-Cambridge Press (1997).

[11] C. S. Burrus, R. A. Gopinath, and H. Guo, *Introduction to Wavelets and Wavelet Transforms*, (Prentice Hall, Upper Saddle River, NJ, 1998), pp. 53.

[12] P. Moulin and M. K. Mihcak, "Theory and design of signal-adapted FIR paraunitary filter banks", *IEEE Trans. Signal Processing*, **46**(4), 920-929 (1998).

[13] A. H. Sayed and T. Kailath, "A survey of spectral factorization methods", *Numer. Linear Algebra Appl.*, **8**, 467-496 (2001).

[14] S. J. GodSill and P. J. W. Rayner, "A Bayesian approach to the restoration of degraded audio signals", *IEEE Trans. Speech Audio Process.*, **3**, 267-278 (1995).

[15] S. V. Vaseghi and P. J. W. Rayner, "Detection and suppression of impulse noise in speech communication systems,"*IEE Proc.*, **137**, 38-46 (1990).

[16] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, 4th ed. (Wiley, Chicheser, UK, 2008), pp. 349-355.

[17] C. Hemphill, J. Godfrey, and G. Doddington, "The ATIS spoken language systems pilot corpus," *Proceedings of the DARPA Speech and Natural Language Workshop*, Hidden Valley, PA (1990), pp. 96-101.

[18] [Online]. Available: www.ece.uvic.ca/~rnongpiu/icassp2014/wavelet_lowpass_filter_coefficients.pdf

[19] J. B. Buckheit and D. L. Donoho, in "WaveLab and Reproducible Research," *Wavelets and Statistics,* edited by A. Antoniadis and G. Oppenheim, Lecture Notes Statistics, Vol. 103 (Springer, New York, 1995), pp. 55-81.

[20] [Online]. Available: http://www-stat.stanford.edu/~wavelab/