Impulse-noise suppression in speech using the stationary wavelet transform

R. C. Nongpiur[a] and D. J. Shpak
Department of Electrical and Computer Engineering, University of Victoria, Victoria, British Columbia V8W 3P6, Canada

(Received 30 June 2012; revised 22 November 2012; accepted 4 December 2012)

An approach for detecting and removing impulse noise from speech using the wavelet transform is proposed. The approach utilizes the multi-resolution property of the wavelet transform, which provides finer time resolution at higher frequencies than the short-time Fourier transform to effectively identify and remove impulse noise. The paper then describes how the impulse-detection performance is dependent on certain wavelet features and their relationships with the impulse noise and the underlying speech signal. Performance comparisons carried out with an existing method show that the wavelet approach yields much better features for detecting the impulses. To remove the impulses, an algorithm that uses the stationary wavelet transform has been developed. The algorithm uses a two-step approach where the wavelet coefficients corresponding to the impulses are suppressed in the first step and then substituted by suitable coefficients located within the vicinity of the impulse in the second step. Performance evaluations with an existing method show that the proposed algorithm gives superior results. © *2013 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4773264]

PACS number(s): 43.60.Bf, 43.60.Hj, 43.50.Pn [SAF] Pages: 866–879

I. INTRODUCTION

The presence of impulse-like noise in speech can significantly reduce the intelligibility of speech and degrade automatic speech recognition 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.

One difficulty with discerning impulse noise from speech is the wide temporal and spectral variation between different parts of speech, such as the periodic and low-frequency nature of vowels and the random and high-frequency nature of consonants. An effective algorithm should, therefore, consistently detect and remove the impulse noise whether it falls in vowels, consonants, or silent portions of speech. For audio signals, several time-domain algorithms have been developed to detect and remove impulse noise.[1–3] However, these algorithms do not exploit the differences in spectral and temporal characteristics of speech and impulse noise to maximize the detection performance.

Classical block processing methods such as the short-time Fourier transform (STFT) algorithm or the linear prediction (LP) algorithm have also been used to detect or remove impulse-like sounds.[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 the 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.

A wavelet approach for the detection and removal of impulse noise in degraded old analog recordings has been reported,[7] whereby the wavelet coefficient corresponding to the scale where the audio signal is weak in comparison to the impulse noise is rectified, smoothed, and then a peak detector is applied to detect the impulses. However, since the peak detector uses a fixed threshold to detect the impulses, false detection may occur on occasions where the speech signal has high-frequency energy such as during consonants and fricatives; the other possibility is that it may fail to detect the smaller impulses that can be quite audible in regions where there is little or no speech signal. Further, the removal of the impulses in the method is done by substituting with uncorrupted wavelet coefficients from a nearby signal using autocorrelation properties. Although the approach works well if the impulses are sparsely located, substitution of the coefficients can be troublesome if a number of impulses are located in the same vicinity, an issue that is not considered in the method. Furthermore, the method uses the

[a]Author to whom correspondence should be addressed. Electronic mail: rajeevcn@gmail.com

dyadic wavelet transform, which is not translation invariant, and prone to artifacts when the coefficients are modified.[8] In another wavelet approach,[9] a variable threshold has been used to detect the impulses by taking advantage of the slow time-varying nature of speech relative to the duration of an impulse; in the approach, the detected impulses are suppressed by decreasing the amplitude of the wavelet coefficients corresponding to the impulses.

In this paper, we describe the wavelet properties that are important for detecting the impulses in speech and show how the detection of impulses is dependent on the nature of impulse noise and the underlying speech signal. Comparisons with an existing method then show that the wavelet approach yields much better features for detecting the impulses. To remove the impulses, we develop a new algorithm that uses the stationary wavelet transform (SWT). The algorithm uses a two-step approach where the wavelet coefficients corresponding to the impulses are suppressed in the first step and then they are substituted by suitable coefficients located within the vicinity of the impulse in the second step. Performance comparisons with an existing method show that the new algorithm gives far superior results.

The paper is organized as follows. Section II discusses the use of wavelets for impulse detection in speech. In Sec. III, we establish the wavelet properties that are important for impulse detection and show their dependence on the nature of the impulse noise and the underlying speech signal. We then describe two metrics that are used to evaluate the suitability of the detection features, followed by an example of a simple detector that is based on the median filter. In Sec. IV, we describe the new impulse-noise removal algorithm. Then in Sec. V, simulation experiments are presented to compare the impulse-detection and removal performances with existing methods. This is followed by experiments that illustrate how the detection performance is dependent on certain wavelet features and their relationship with the impulse noise and the underlying speech signal.

II. USING WAVELETS TO DETECT IMPULSE NOISE IN SPEECH

A speech signal can be considered to be broadly made up of vowels, consonants, and silence portions. The vowel portion is generated by periodic pulses from the vocal chords, which are then low-pass filtered by the vocal tract. As such, vowels are usually harmonically rich with an upper cutoff frequency that does not exceed 5 kHz. The consonants, on the other hand, are generated by constriction in the mouth; they are usually anharmonic with a spectrum that can extend up to 20 kHz. The silence portion of speech is essentially background noise that is random in nature. An important feature that distinguishes impulse noise from speech is the slow time-varying nature of the temporal and spectral envelope of speech in comparison to that of an impulse; this slow-time varying nature is because variations are generated by the movements of muscles in the mouth and vocal tract, which is a relatively slow process.

An impulse is characterized by a sudden change in the signal amplitude or a sudden shift in the signal mean value. If the continuous wavelet transform (CWT) of a signal with impulse noise is taken, large magnitude coefficients, termed modulus maxima, will be present at time points where the impulses have occurred.[6] Impulses are distinguishable from noise by the presence of modulus maxima at all of the scale levels; noise, on the other hand, produces modulus maxima only at finer scales. Mallat and Hwang[10] developed a method for detecting singularities by analyzing the evolution of the wavelet modulus maxima across scales for a CWT. However, in practical applications the SWT is preferred over a CWT due to its lower computational effort. Additionally, the dyadic discrete wavelet transform could be used if only impulse detection is required. But if both impulse detection and removal are required the SWT is preferred over the discrete wavelet transform due to absence of aliasing artifacts after synthesis.[8]

Having large wavelet coefficients for impulses in the finest scale is beneficial since it leads to better detection of the impulses. Apart from the impulses, even some components of speech, such as high-frequency fricatives, are characterized by relatively larger coefficients at the finer scales. However, compared to a fricative or other high-frequency noises, an impulse has significantly higher energy compacted within a short time interval, e.g., the typical time interval for impulse noise in speech is usually less than 20 ms. Therefore, with an appropriate wavelet it is possible to transform this energy into coefficients that are correspondingly compacted and much larger in comparison to those of the fricatives or high-frequency noises, in the finest scale.

III. DETECTION OF IMPULSE NOISE FROM SPEECH

There are two aspects in the detection of impulses in speech. The first is the selection of the appropriate wavelet for impulse detection and the second is the design of the impulse-detection algorithm. In this section, we describe the wavelet properties that influence the detection performance and present two measures for evaluating the performance. A simple impulse-detection algorithm is then described, providing a framework for comparing the detection performance between the wavelets and making the evaluation process more comprehensive. It should be pointed out that in this section we will focus more on the selection of the most appropriate wavelet and on the feasibility of the wavelet coefficients as a feature for impulse detection, with little emphasis on the implementation aspect of the impulse-detection algorithm. The selection of a particular detection algorithm for a specific application is highly dependent on the context of the application and is therefore beyond the scope of this paper.

A. Wavelet properties and features for impulse detection

For impulse detection, it is important to select a wavelet that maximizes the finest scale coefficients for impulses relative to those of the underlying speech signal and background noise. As will be seen, this depends not only on the nature of the impulse noise, but also on the spectral characteristics of the underlying signal.

The size of the wavelet support has two important effects on an impulse: (a) A smaller wavelet support corresponds to a shorter analysis filter and, therefore, lesser temporal smearing of the wavelet coefficients corresponding to the impulse. (b) A larger wavelet support, on the other hand, corresponds to a longer analysis filter and, therefore, better frequency selectivity for separating the impulse noise from speech, but more temporal smearing.

1. Frequency selectivity

For a certain wavelet support size, a desirable wavelet for impulse detection is one that maximizes the impulse coefficients relative to those of the underlying signal in the finest scale. Such a wavelet will correspondingly have an analysis filter that maximizes the impulse noise relative to the underlying speech and background-noise signal. Consequently, for a given support size, the selection of such a wavelet would be dependent on the spectral properties of the impulse noise and the underlying speech and background noise.

2. Wavelet support versus impulse energy

When the energy of the impulse noise is weak in comparison to the speech energy, having good frequency selectivity to enhance separation of the impulse noise from speech is more important than minimizing the temporal smearing of the coefficients, and, therefore, a wavelet with a larger support size is desirable. On the other hand, if the impulse energy is strong in comparison to the speech signal, we get larger magnitudes for impulse wavelet coefficients by reducing the temporal smearing at the expense of frequency selectivity, and therefore a smaller support size is more appropriate. Consequently, the most appropriate wavelet support size is dependent on the average energy of the impulse noise relative to the underlying speech signal.

3. Wavelet support versus impulse width

For good detection performance, the size of the wavelet support, or alternatively wavelet filter length, is also dependent on the width of an impulse burst. Once the length of the analysis filter gets longer than the impulse width, the impulse wavelet coefficients get relatively smaller. This is because the convolution of the filter with the impulse noise also includes portions outside of the impulse, thereby reducing the contribution of the impulse. On the other hand, a longer analysis filter improves the frequency selectivity for impulse noise versus speech. Consequently, as the impulse width increases, the optimal filter length that maximizes the impulse coefficients correspondingly increases.

4. Sampling frequency

A wavelet that is optimal for detecting the impulses at one sampling frequency will not usually remain optimal if the signal is processed at another sampling frequency. This is because the change in sampling frequency alters the spectral characteristics of the impulse noise and the underlying signal, which in turn changes the frequency response of the optimal high-pass analysis filter. In addition, the change in sampling frequency also scales the wavelet support size relative to the average width of the impulse noise, thereby making the support size sub-optimal at the new sampling frequency.

B. Metrics to evaluate the detection performance

To determine the most appropriate wavelet for impulse detection, the discriminatory capability of the wavelet coefficients in the finest scale with respect to the impulse noise is evaluated. A well-known measure in statistics that quantifies the discriminative power of a feature is a separability criterion derived from the scatter matrices[11] and described in greater detail in Appendix A; for a one-dimensional, two-class scenario, the separability criterion for feature x is given by

$$J = \frac{n_1(m_1 - m)^2 + n_2(m_2 - m)^2}{\sum_{x\in\omega_1}(x - m_1)^2 + \sum_{x\in\omega_2}(x - m_2)^2}, \tag{1}$$

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, and m is the overall mean.

If the performance of the wavelet coefficients is to be evaluated against a competing method that has a different feature for discriminating the impulses, the separability measure may not give the complete picture as it only measures the discriminative power between the features. Another measure that can be included in addition to the separability measure is the mutual information (MI) measure[12] between the feature and the quantity to be detected,[13–15] which in this case is the impulse noise. The MI includes all the linear and non-linear dependencies and gives a lower bound of the best achievable performance of a feature for detecting the impulses;[13] therefore, it is an appropriate condition to determine the quality of the features. It should be noted, however, that the bound gives a necessary but not a sufficient condition, that is, a large MI alone will not guarantee good detection performance. If X and Y are two random variables having possible outcomes, or alphabets, in the sets χ and γ, respectively, the MI between them is given by

$$\mathrm{MI}(X;Y) = \sum_{x\in\chi}\sum_{y\in\gamma} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}, \tag{2}$$

where $p(x)$ and $p(y)$ are the probability density functions of X and Y, respectively. In the context of this paper, X can correspond to the feature for detecting the impulse noise and Y to the outcome of the impulse-noise process. In Appendix B, a procedure for computing the MI between the feature and the quantity to be detected is described in more detail.

C. A simple impulse detector

As mentioned in Sec. II, the temporal and spectral envelope of speech is slowly time-varying in comparison to an impulse. This property is used to detect the wavelet coefficients that correspond to an impulse. Therefore, what is

needed is a dynamic threshold that varies in proportion to the smooth envelope of the absolute wavelet coefficients values, but, at the same time, is not affected by impulse noise. That is, for the finest scale s_f and sample n, such a dynamic threshold, $\Gamma(n, s_f)$, can be defined as

$$\Gamma(n, s_f) = k_f \cdot \mathrm{Env}[|Wf(n, s_f)|], \tag{3}$$

where $Wf(n, s)$ are the wavelet coefficients of $f(n)$ at scale s, Env[·] is the envelope of the signal that is unaffected by impulse noise, and k_f is a factor that is determined empirically on the basis of the type of wavelet used and the nature of the impulse noise. A wavelet coefficient would be considered to be that of an impulse if its absolute value is greater than $\Gamma(n, s_f)$. That is,

$$\mathrm{detector}(n) = \begin{cases} \mathrm{TRUE} & \text{if } |Wf(n, s_f)| > \Gamma(n, s_f) \\ \mathrm{FALSE} & \text{otherwise.} \end{cases} \tag{4}$$

The operator Env[·] is implemented by a median filter[9] as it possesses the property where step-function type signals are preserved while at the same time being robust to impulse noise;[16] that is,

$$\Gamma(n, s_f) = k_f \,\mathrm{MED}[|Wf(n-K, s_f)|, \ldots, |Wf(n, s_f)|, \ldots, |Wf(n+K, s_f)|]. \tag{5}$$

The length $L_{\mathrm{med}} = 2K+1$ of the median filter is adjusted so that it is sufficiently long in comparison to an impulse but short in comparison to a vowel or consonant. For a median filter of length $2K+1$, impulses shorter than $K+1$ will be removed.[16] So if the maximum width of the impulses that can occur in a signal is $K_{\max}$ samples, the median filter length should satisfy

$$L_{\mathrm{med}} > 2K_{\max} + 1. \tag{6}$$

Of course, L_{med} should also be small enough in comparison to the length of the vowels and consonants, which are usually above 20 ms. It should also be noted that there are components in speech that may be falsely detected as impulses due to similar time duration and spectral properties; fortunately, components having sufficiently short duration are not common in normal speech. For example, if the maximum width of an impulse is 2 ms, which corresponds to $K_{\max} = 32$ at 16 kHz sampling frequency, setting $L_{\mathrm{med}} = 100$ would be quite adequate for removing the impulses.

In Fig. 1, typical waveforms of the various parameters dealing with the detection of the impulses are shown. As can be seen, a positive detection by the detector corresponds to the location of the impulses along $\eta(n)$.

It should be noted that although the threshold estimate in Eq. (5) is quite simple, it can nevertheless be used to determine which wavelet has better features for impulse detection for the same support size. However, if a robust detector is to be designed, more sophistication can be incorporated into the threshold estimator, if the computational resources permit. For example, a vowel/consonant/background-noise detector

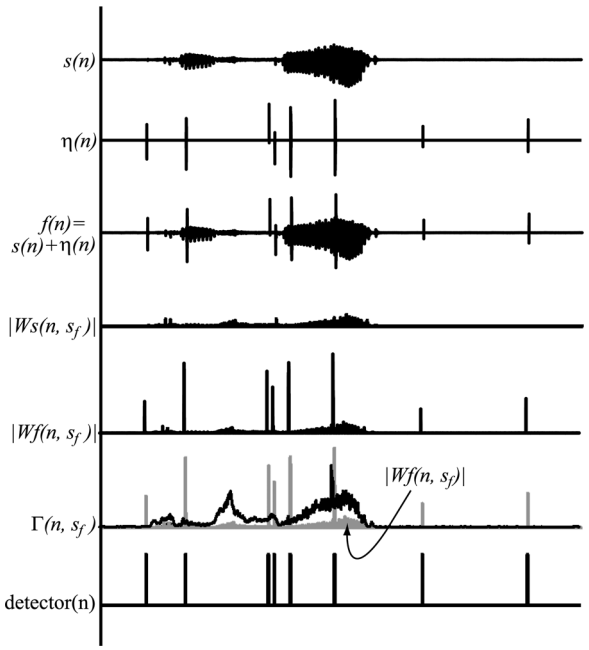

FIG. 1. Typical waveforms of the parameters dealing with the detection of the impulses. A non-zero value at the detector output corresponds to the presence of an impulse.

can be included to appropriately adjust k_f and/or the window length L_{med} in Eq. (5) according to the statistics and signal level of the speech components and the impulse noise. Alternatively, a lookup table or code-book[17] for k_f and L_{med} that is optimized for a particular wavelet and impulse-noise type may be pre-computed for the various speech components and types of background noises that can be encountered in the intended application. Additionally, the median filter in Eq. (5) may be replaced by a more sophisticated filter, such as a trimmed-mean filter,[18] to provide more optimal estimates. Consequently, it is apparent that the design of a robust detector is very much application specific and closely dependent on the nature of the impulse noise, the background noise, the speech signal, and the wavelet used.

IV. REMOVAL OF IMPULSE NOISE FROM SPEECH

In the approach by Nongpiur,[9] the wavelet coefficients at the coarser scales are suppressed by thresholding the wavelet coefficients. One drawback of this approach is that the coefficients corresponding to the impulses are smeared over a wider time span as the scales get coarser. Consequently, using the thresholding method to suppress the impulse at coarser scales will not be particularly effective and will result in some phase distortion of the speech signal. Since the human perception of speech is quite sensitive to phase distortion below 1 kHz, we can conclude that the thresholding method will not be as effective for coarser scales that correspond to 1 kHz and below.

In our proposed method, we use the finest scale to detect the location of the impulse in the signal. Once the start and end positions of an impulse have been identified, the wavelet coefficients between those positions are then replaced by the most similar section located in the vicinity of the impulse. Since an increase in the number of scales results in wider

temporal smearing of the wavelet impulse coefficients due to increase filtering operations, only two levels of wavelet scales are used as shown in Fig. 2. This ensures that the impulse energy is localized within a small time interval, thereby making the impulse-removal algorithm more effective and efficient. Consequently, the two-level SWT of $f(n)$ denoted as $Sf(n, l)$ is defined as

$$Sf(n,l) = \begin{cases} Wf(n, 2^{-1}) & \text{if } l = 1 \\ Vf(n, 2^{-1}) & \text{if } l = 2, \end{cases} \tag{7}$$

where $Vf(n, s)$ are the scaling coefficients of $f(n)$ at scale s.

The removal of the impulses is done using a two-step process. In the first step, the impulse coefficients are supressed by soft-thresholding the coefficients $Sf(n, l)$, and in the second step the impulse coefficients that have been suppressed are replaced by suitable coefficients obtained from the vicinity of the impulse. Though the first step may seem to be unnecessary, it helps to minimize the artifacts when the coefficient replacement algorithm is not completely accurate, which may happen when the coefficients adjacent to the impulse are also corrupted by impulse noise.

For the first step, we proceed to soft-threshold the coefficients of $Sf(n, l)$ corresponding to the impulse as follows. After the start and end locations of the impulses have been obtained from the impulse detector, the portion of the signal where the impulse is not present is set to zero to give a new signal, $g(n)$, where

$$g(n) = \begin{cases} f(n) & \text{if detector}(n) = \text{TRUE}, \\ 0 & \text{otherwise}. \end{cases} \tag{8}$$

To obtain the value of the threshold, a two-scale level SWT of $g(n)$, denoted as $Sg(n, l)$, is taken and the envelope of the absolute values of the coefficients for a particular wavelet scale is computed using prior and aft sliding windows given by

$$\xi_g(n, l) = \min\{\phi_p(n, l), \phi_a(n, l)\}, \tag{9}$$

where

$$\phi_p(n, l) = \max[|Sg(n - K, l)|, \ldots, |Sg(n, l)|], \tag{10}$$

$$\phi_a(n, l) = \max[|Sg(n, l)|, \ldots, |Sg(n + K, l)|], \tag{11}$$

and $2K$ is the length of the sliding windows. In a similar manner, the envelope of $|Sf(n, l)|$ is computed and denoted as $\xi_f(n, l)$. Using $\xi_g(n, l)$ as the threshold, the coefficients where the impulses are located are attenuated by soft-thresholding the coefficients, given by

$$\widehat{Sf}(n, l) = \begin{cases} Sf(n, l) & \text{if } |Sf(n, l)| > \xi_g(n, l), \\ \dfrac{Sf(n, l)}{|Sf(n, l)|}\xi_g(n, l) & \text{otherwise}. \end{cases} \tag{12}$$

In the second step, the section along the wavelet level where the impulse is located is replaced using the most similar section in the vicinity of the impulse. To avoid audible artifacts, the substitution is done by smoothly blending the coefficients at the boundaries.

To carry out the substitution, we first construct a comparison template that will be used to find a section in the vicinity of the impulse with the best match. The comparison template is constructed by taking the section of $\widehat{Sf}(n, l)$ where the impulse is located, and extending the section at the front and back by L_f and L_b samples, respectively. That is, if $n_s^{(i)}$ and $n_e^{(i)}$ are the start and end locations of the ith impulse, the comparison template for that impulse is the section from $(n_s^{(i)} + L_b)$ to $(n_e^{(i)} + L_f)$.

Within the comparison template, we need to disregard the coefficients where the impulse energy is greater than the speech energy since they do not accurately represent the underlying speech signal. To do this, we construct a template mask, $\Xi(n, l)$, given by

$$\Xi(n, l) = \begin{cases} \dfrac{\xi_f(n, l) - \xi_g(n, l)}{\xi_f(n, l)} & \text{if } \xi_f(n, l) > \xi_g(n, l), \\ 0 & \text{otherwise}, \end{cases} \tag{13}$$

where the product, $\Xi(n, l)\widehat{Sf}(n, l)$, disregards the coefficients that have been corrupted by the impulse. In Fig. 3, typical waveforms for the various parameters are shown for an impulse that is detected in the middle of a speech vowel. The coefficients shown correspond to the second level of the two-level SWT.

To find the most similar pattern, the search window is extended by L_{wb} to the left and L_{wf} to the right. As can be seen in Fig. 4, the search region results in two search windows, w_1 and w_2, that may overlap; the template then slides along the two search windows to find the most correlated pattern. The degree of correlation between the template for the ith impulse at wavelet level l and a section along the search window of the same length and starting at n is given by

$$\varrho(n, i, l) = \frac{1}{\varepsilon_d^{(i)}\varepsilon_t^{(i)}} \sum_{k=1}^{K_i} \Xi(n'', l)\widehat{Sf}(n', l) \times \Xi(n'', l)\widehat{Sf}(n'', l), \tag{14}$$

where

$$\varepsilon_d^{(i)} = \sqrt{\sum_{k=1}^{K_i} \left|\Xi(n'', l)\widehat{Sf}(n', l)\right|^2}, \tag{15}$$

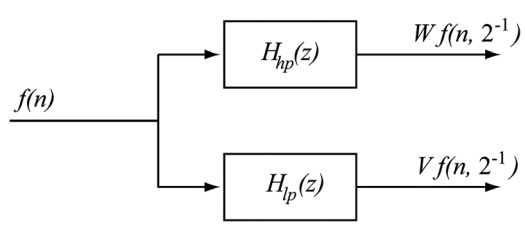

FIG. 2. A SWT of signal $f(n)$ with two levels of wavelet scales that is implemented using a two-band analysis filterbank; the outputs of the high-pass and low-pass filters are the wavelet and scaling coefficients, respectively.

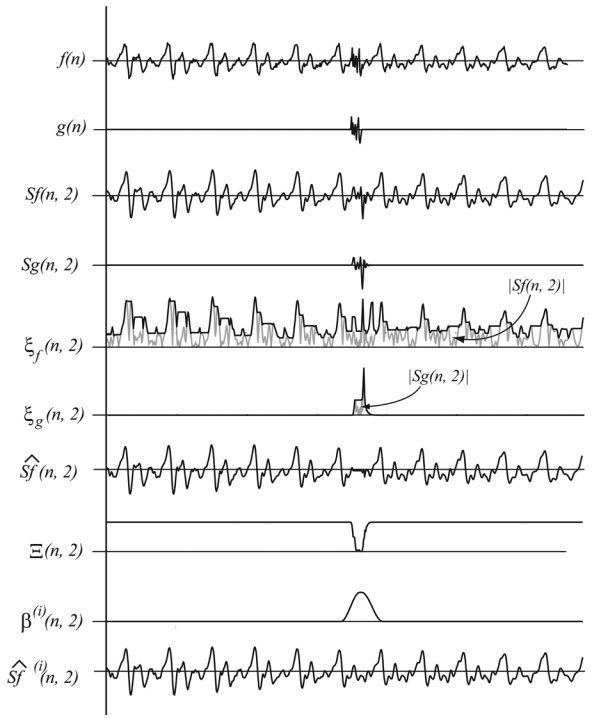

FIG. 3. Typical waveform plots of the various parameters used in the impulse-noise removal algorithm for an impulse that is located in the middle of a vowel. Apart from $f(n)$ and $g(n)$, the other parameters correspond to the second level of the two-level SWT; the parameters for the first level are also computed in a similar manner.

$$\varepsilon_t^{(i)} = \sqrt{\sum_{k=1}^{K_i} \left|\Xi(n'', l)\widehat{Sf}(n'', l)\right|^2}, \tag{16}$$

$$n' = n + k, \tag{17}$$

$$n'' = n_s^{(i)} - L_b + k, \tag{18}$$

$$K_i = n_e^{(i)} - n_s^{(i)} + L_b + L_f. \tag{19}$$

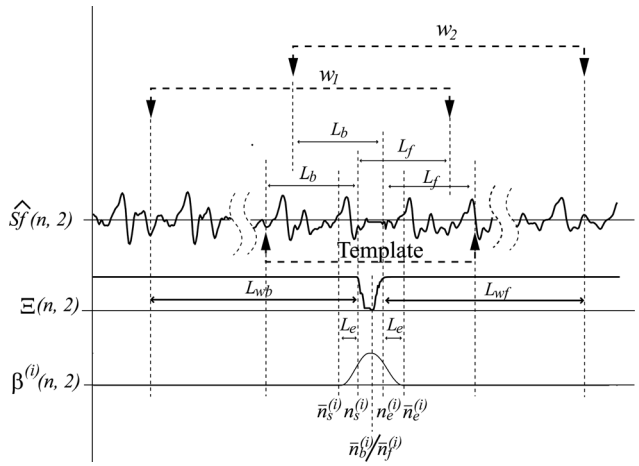

FIG. 4. A more detailed illustration of the relationship between the template mask, $\Xi(n, l)$, and blending mask, $\beta^{(i)}(n, l)$, with the corresponding coefficients, $\widehat{Sf}(n, l)$, for $l = 2$.

Since the template energy, $\varepsilon_t^{(i)}$, does not vary with n, it can be considered a constant and, therefore, need not be computed during implementation.

If $n_{\max}^{(i)}$ corresponds to the starting point of the section with the highest value of $\varrho(n, i, l)$ within the search window, the start and end locations of the section that is used for substitution is given by

$$\begin{aligned} \hat{n}_s^{(i)} &= n_{\max}^{(i)} + L_b, \\ \hat{n}_e^{(i)} &= n_{\max}^{(i)} + n_e^{(i)} - n_s^{(i)} + L_b. \end{aligned} \tag{20}$$

However, to minimize the artifacts, the substituted section needs to blend smoothly at the boundaries. To do this, the section to be substituted is extended on both sides to cause an overlap for smooth blending; if the section is extended by L_e samples on both sides, the new start and end limits become

$$\begin{aligned} \bar{n}_s^{(i)} &= n_s^{(i)} - L_e, \\ \bar{n}_e^{(i)} &= n_e^{(i)} + L_e. \end{aligned} \tag{21}$$

A blending mask, $\beta^{(i)}(n, l)$, that extends from $\bar{n}_s^{(i)}$ to $\bar{n}_e^{(i)}$ is constructed to smoothly blend the substituting section at the impulse location. To do this, the minimum value of $\Xi(n, l)$ between $\bar{n}_s^{(i)}$ and $\bar{n}_e^{(i)}$ is located. Assuming that the minimum value may not be unique, the locations of the first and last minimum values are denoted as $\bar{n}_b^{(i)}$ and $\bar{n}_f^{(i)}$, respectively. Note that in Fig. 4 the minimum value is unique and $\bar{n}_b^{(i)}$ and $\bar{n}_f^{(i)}$ correspond to the same location. Using the minimum values as inner limits, raised-cosine smoothing is incorporated at the edges of the blending mask, given by

$$\beta^{(i)}(n, l) = \begin{cases} \cos\dfrac{\pi\left(n - \bar{n}_s^{(i)}\right)}{N_b} & \text{if } n \in [\bar{n}_s^{(i)}, \bar{n}_b^{(i)}] \\ 1 & \text{if } n \in [\bar{n}_b^{(i)}, \bar{n}_f^{(i)}] \\ 1 - \cos\dfrac{\pi\left(\bar{n}_e^{(i)} - n\right)}{N_f} & \text{if } n \in [\bar{n}_f^{(i)}, \bar{n}_e^{(i)}] \\ 0 & \text{otherwise,} \end{cases} \tag{22}$$

where $N_b = 2\left(\bar{n}_b^{(i)} - \bar{n}_s^{(i)}\right)$ and $N_f = 2\left(\bar{n}_e^{(i)} - \bar{n}_f^{(i)}\right)$. Consequently, the substitution of the ith impulse by the correlated section is given by

$$\begin{aligned} \widehat{Sf}^{(i)}(n, l) = &\left(1 - \beta^{(i)}(n, l)\right)\widehat{Sf}(n, l) \\ &+ \beta^{(i)}(n, l)\widehat{Sf}\left(n + \hat{n}_s^{(i)} - n_s^{(i)}, l\right). \end{aligned} \tag{23}$$

During experimentation, it has been observed that using the finest scale with only the soft-thresholding method in Eq. (12), without the need for the coefficient substitution method in Eq. (23), still results in comparable performance. Therefore, we refer to the impulse-removal method where Eq. (23) is applied to both wavelet levels as "proposed-variant1"; and the method where Eq. (23) is applied to only the second coarser level as "proposed-variant2."

The length of the search window and template length is dependent on the width of the impulse and the minimum pitch frequency assumed. For example, if the maximum impulse width, $\tau_{d\max}$, is 8 ms and the minimum pitch frequency, $f_{\min}$, is 80 Hz then the following conditions should apply on the search window and template length parameters:

$$(L_f \geq \tau_{f\min}) \vee (L_b \geq \tau_{f\min}) = \text{True},$$
$$(L_{wf} \geq \tau_{f\min} + \tau_{d\max}) \vee (L_{wb} \geq \tau_{f\min} + \tau_{d\max}) = \text{True}, \tag{24}$$

where $\tau_{f\min} = 1000/f_{\min} = 12.5$ ms and $\tau_{d\max} = 8$ ms. In certain implementations, which may require low delay or low computational effort, it may not be possible to use the signal that occurs after the impulse for substitution. In such a case, only window w_1 is used and L_f and L_{wf} are set to 0.

The synthesis of the modified SWT coefficients is done using the inverse-SWT algorithm.[19] If the algorithm is to be implemented for real-time applications, overlap and add methods similar to the STFT may be adopted.[9,20]

V. EXPERIMENTAL RESULTS

The experiments are divided into three sections. In Sec. V A, we carry out experiments to compare the performance of the impulse-detection features with an existing method. Then, in Sec. V B we compare the performance of the impulse-removal algorithm with an existing method. In Sec. V C, we perform experiments to validate the important wavelet features for detecting impulses.

A. Comparison of the impulse-detection features

Here we perform two experiments to evaluate the performances of the impulse-detection features. In the first experiment, experiment A-1, we compare the discriminative performances of the impulse-detection features between the proposed method and an existing method. Then, in the second experiment, experiment A-2, we use the MI measure to compare the feasibility of the impulse-detection features between the two methods.

To generate the impulse-noise signals for carrying out the experiments we use an impulse-noise generation model[21] that has been found to be a good representation of speech signal degraded by clicks. The model, reproduced in Fig. 5,

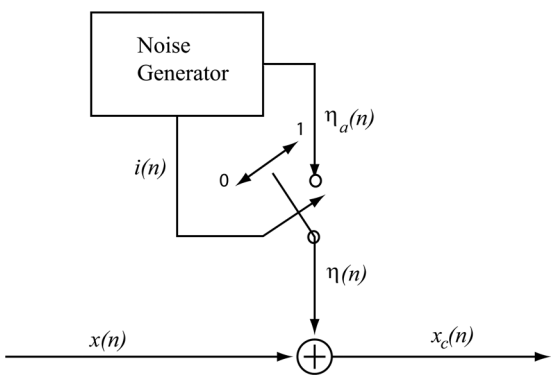

FIG. 5. Impulse-noise generation model.

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 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%.[1] If α is the fraction of signal samples contaminated by impulse noise the average signal to impulse noise ratio (SINR) is given by[22]

$$\text{SINR} = \frac{P_s}{\alpha P_i}, \tag{25}$$

where P_s is the power of the speech signal and P_i is the power of the impulse. For our experiments, we consider speech degraded by impulse noise that has a SINR of 10 dB with 5% contamination.[22] The binary noise generation process for $i(n)$ is implemented using a two-state Markov chain, with the transition probabilities adjusted so that the average impulse width is 1 ms with 5% contamination. The second noise process, $\eta_a(n)$, is generated using a normal distribution.

The speech signal used in the experiments is clean near-microphone speech taken from the ATIS corpus database.[23]

1. Experiment A-1

In this experiment, we compare the discriminatory capability of the impulse-detection features of the wavelet approach with an existing method, using the separability criterion J in Eq. (1). 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 if not an impulse. After the features have been classified, we then use Eq. (1) to obtain J.

To obtain the detection features for the wavelet approach, we use the Daubechies wavelet of order 4. As will be seen in Sec. V B, the order of 4 has been found to be most appropriate among the Daubechies wavelets when the impulse noise is white and the SINR is 10 dB. Using the SWT, the signal is analyzed into two levels. 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 SWT 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

$$\mathcal{F}(n) \in \begin{cases} \omega_1 & \text{if } |x_f^{(i)}(n)| > 0 \\ \omega_2 & \text{otherwise,} \end{cases} \tag{26}$$

where

$$\mathcal{F}(n) = |x_f^{(s)}(n) + x_f^{(i)}(n)|. \tag{27}$$

For the comparison, we use the detection features of an existing impulse-detection method developed by Vaseghi

and Rayner.[1,22] In this method, the signal is divided into blocks and the linear prediction coefficients (LPCs) for each of the blocks is computed. Using the LPCs, an inverse filter is applied on the block, followed by matched filtering. The output from the matched filter is then used for detecting the impulses by the algorithm. To carry out the classification for the competing method, the inverse and matched filters for each block are computed from the corrupted signal, which is the sum of the clean signal and the impulse noise. The clean signal is then processed separately through the filters obtained to give $x_m^{(s)}(n)$; likewise, the impulse noise is processed through the same filters to give $x_m^{(i)}(n)$. The classification of the detection features is then initiated using

$$\mathcal{G}(n) \in \begin{cases} \omega_1 & \text{if } |x_m^{(i)}(n) > 0| \\ \omega_2 & \text{otherwise,} \end{cases} \tag{28}$$

where

$$\mathcal{G}(n) = |x_m^{(s)}(n) + x_m^{(i)}(n)|. \tag{29}$$

In Table I, the values of J for sampling frequencies of 8 and 16 kHz are tabulated. As can be seen, the proposed method has significantly higher values of J for both sampling frequencies. This significantly higher separability shows that better detection can be achieved if the wavelet method is used.

In Fig. 6, a typical speech sample that is contaminated with impulse noise is processed by the two detection algorithms and the respective absolute values of the processed speech signal and impulse noise just before detection by the threshold detector are plotted. Comparing Figs. 4(c) and 4(d), it is apparent that the processing done by taking the SWT results in greater amplification of the impulse wavelet coefficients, relative to that of the speech coefficients, than the LPC method in Vaseghi and Rayner.[1]

2. Experiment A-2

In this experiment, we evaluate the suitability of the impulse-detection features of the wavelet approach with the method developed by Vaseghi and Rayner,[1] by comparing the MI between their impulse-detection features and the impulse-noise signal. To compute the MI numerically, the impulse-detection feature X and the corresponding impulse-noise signal Y are used as training data for deriving the Gaussian-mixture-model probability-density-function as in Eq. (B2). For the experiment, we set the number of Gaussian mixtures L to 10 because we observed that increasing L above 10 makes little or no difference.

TABLE I. Comparison of separability J between the impulse-detection features of the proposed method and a competing method, at sampling frequencies of 8 and 16 kHz.

Detection method	J $f_s = 8$ kHz	J $f_s = 16$ kHz
Proposed method	0.47	0.81
Competing method[a]	0.08	0.22

[a]Reference 1.

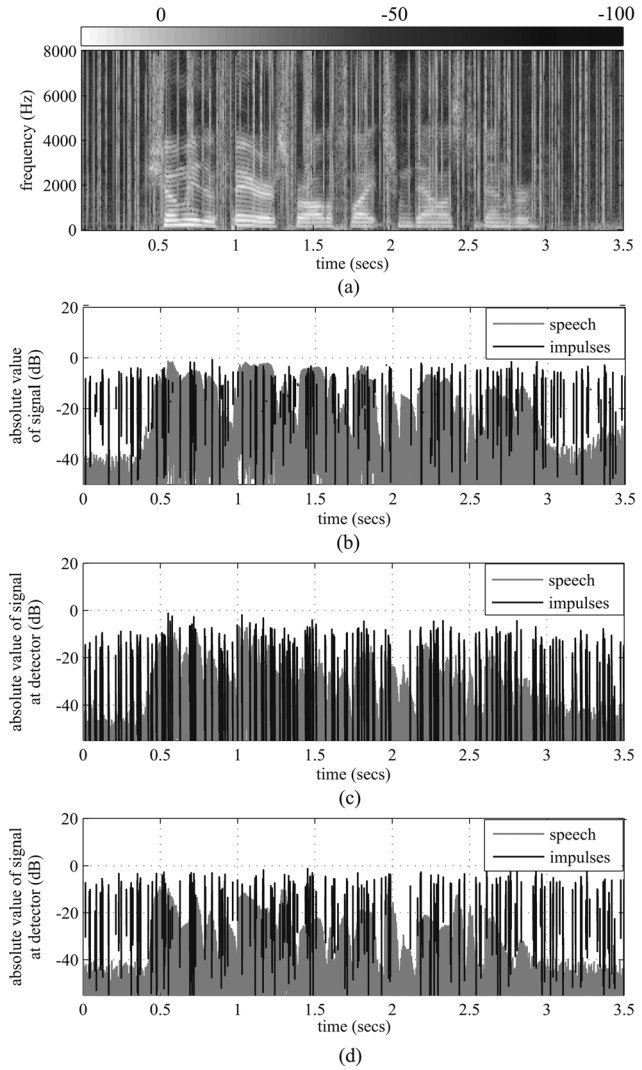

FIG. 6. (a) Spectrogram of a typical speech signal contaminated with impulse noise at $f_s = 16$ kHz. (b) Absolute value of the speech signal and the impulse noise. (c) Speech and impulse noise processed using the LPC method just before detection by the threshold detector. (d) Speech and impulse noise processed using the Daubechies SWT of order 4 just before detection by the threshold detector.

To obtain the impulse-detection feature for the wavelet approach we use the same procedure as in experiment A-1 to get $x_f^{(s)}(n)$ and $x_f^{(i)}(n)$, denoting the detection features as random variable $X_f = x_f^{(s)}(n) + x_f^{(i)}(n)$ and the corresponding impulse-noise signal as random variable $Y_f = \eta(n)$, where $\eta(n)$ is generated using the impulse-noise model in Fig. 5. Likewise, for the competing method we use the same procedure in experiment A-1 to get $x_m^{(s)}(n)$ and $x_m^{(i)}(n)$, giving random variables $X_m = x_m^{(s)}(n) + x_m^{(i)}(n)$ and $Y_m = \eta(n)$. The random variable pairs (X_f, Y_f) and (X_m, Y_m) are then used for computing the MI measure for the respective methods, using the procedure outlined in Appendix B.

In Table II, the MI measure for sampling frequencies of 8 and 16 kHz is tabulated. As can be seen, the wavelet approach has significantly higher values for both sampling frequencies. Consequently, from these results we can infer

TABLE II. Comparison of the MI measures between the two methods.

Detection method	MI (bit) $f_s = 8$ kHz	MI (bit) $f_s = 16$ kHz
Proposed method	1.92	1.66
Competing method[a]	1.32	1.05

[a]Reference 1.

that the wavelet feature has a stronger relationship with the impulse noise and, hence, better suited as an impulse-detection feature.

B. Comparison of the impulse-removal algorithm

In this experiment, we perform an objective comparison between the impulse-removal algorithm described in Sec. IV and the existing method described in the work by Vaseghi and Rayner.[1,22] Since the aim of this experiment is to compare only the performance of the impulse-removal algorithm, we assume that the impulse detection is working perfectly, which implies that the impulse-removal algorithm has knowledge of the exact location of the impulse.

For the proposed method, the signal with impulse noise is analyzed into two levels using a SWT that utilizes the Daubechies wavelet of order 4. In an actual implementation, the first level, which corresponds to the finest scale, would be used for detecting the location of the impulses and then we would compute $g(n)$ as in Eq. (8). However, in this experiment the exact location of the impulses in the signal is assumed to be known, so $g(n)$ is exact. As described in Sec. IV, the SWT of $g(n)$ is then taken and is used along with $Sf(n, l)$ for removing the impulses. The search window length parameters, L_{wb} and L_{wf}, are both set to 20.5 ms and the template length parameters, L_b and L_f, are set to 0 and 12.5 ms, respectively. Note that the selected values of L_{wb}, L_{wf}, L_b, and L_f satisfy the conditions in Eq. (24) with the assumption that the maximum width of an impulse, $\tau_{d\max}$, is 8 ms and the minimum pitch frequency, $f_{\min}$, is 80 Hz. The amount of overlap for smooth blending, L_e, is 15 samples, which corresponds to approximately 0.94 ms (or 11.8% of $\tau_{d\max}$) at 16 kHz.

For the method by Vaseghi and Rayner,[1] the signal where the impulse is located is reconstructed by taking portions of the signals before and after the impulse and performing a least-square error linear-prediction interpolation. To ensure that the voiced portions of speech are also accurately interpolated, the pitch period just before the start of the impulse is determined and a long-term predictor that is adjusted to the length of the pitch period is also included. As in their paper,[1] the order of the LPC model used for the short and long term predictors is 20 and 7, respectively.

To evaluate the closeness of the reconstructed speech to the original speech signal, we use the rms log-spectral distortion (LSD) measure[24] given by

$$d^2_{\mathrm{LSD}} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(20\log_{10}\frac{\rho}{|A(\omega)|} - 20\log_{10}\frac{\bar{\rho}}{|\bar{A}(\omega)|}\right)^2 d\omega, \quad (30)$$

where $\rho/A(\omega)$ and $\bar{\rho}/\bar{A}(\omega)$ are the spectral models of the original and reconstructed signals. From Parseval's theorem, the LSD can be expressed in the cepstral domain as

$$d^2_{\mathrm{LSD}} = \left(\frac{10}{\log_e 10}\right)^2\left((c_0 - \bar{c}_0)^2 + 2\sum_{i=1}^{\infty}(c_i - \bar{c}_i)^2\right), \quad (31)$$

where the cepstral coefficients c_0, c_1,… are calculated from the LP coefficients using the recursive equation given in Gray and Markel.[24] As shown in their work,[24] sufficient accuracy is still maintained if the number of cepstral coefficients is truncated to the order of the LP coefficients. The order of the LP coefficients and number of cepstral coefficients are both 20, and each block is 45 ms long with a 35 ms overlap.

For this experiment, the impulse noise is generated as in experiments A-1 and A-2 with an SINR value of 10 dB, 5% contamination, and average impulse-noise width of 1 ms. The algorithm is also tested at different levels of background noise by adding white Gaussian noise to the impulse-noise corrupted signal. As in the previous experiments, the speech signal is clean near-microphone speech taken from the ATIS corpus database. The duration of the signal is about 5 min long, with five male and five female speakers.

The two variants of the proposed algorithm, proposed-variant1 and proposed-variant2, are compared with the existing method. Audio examples for the impulse-noise removal algorithm can be accessed online.[25] In Table III, the LSD measure for different background-noise levels is tabulated for the various algorithms. As can be seen, the proposed-variant1 method has the best performance with the smallest LSD measure followed by proposed-variant2. In all three methods, the LSD measure increases with an increase in signal to noise ratio (SNR), thereby implying that artifacts would be more perceptible at lower background-noise levels, which is not surprising since background noise is an effective masker.[26] In Fig. 7, we show comparisons of the spectrograms of a speech sample corrupted with impulse noise after it has been processed by the proposed and conventional impulse-noise removal algorithms. The spectrogram of the clean speech signal is also included as a reference. From the plots, we observe that the impulse noise is significantly reduced by all three methods. Upon more careful comparison with the spectrogram of the clean speech signal, it can be observed that the spectrograms corresponding to proposed-variant1 and proposed-variant2 contain lesser artifacts than the spectrogram processed by the method by Vaseghi and Rayner.[1] And among the two proposed methods, the one processed by proposed-variant1 is slightly better.

TABLE III. LSD measure for different background-noise levels.

SNR (dB)	LSD measure of proposed-variant1	LSD measure of proposed-variant2	LSD measure of the competing method[a]
>30	0.93	1.22	3.89
20	0.74	0.92	2.84
10	0.65	0.79	2.34

[a]Reference 1.

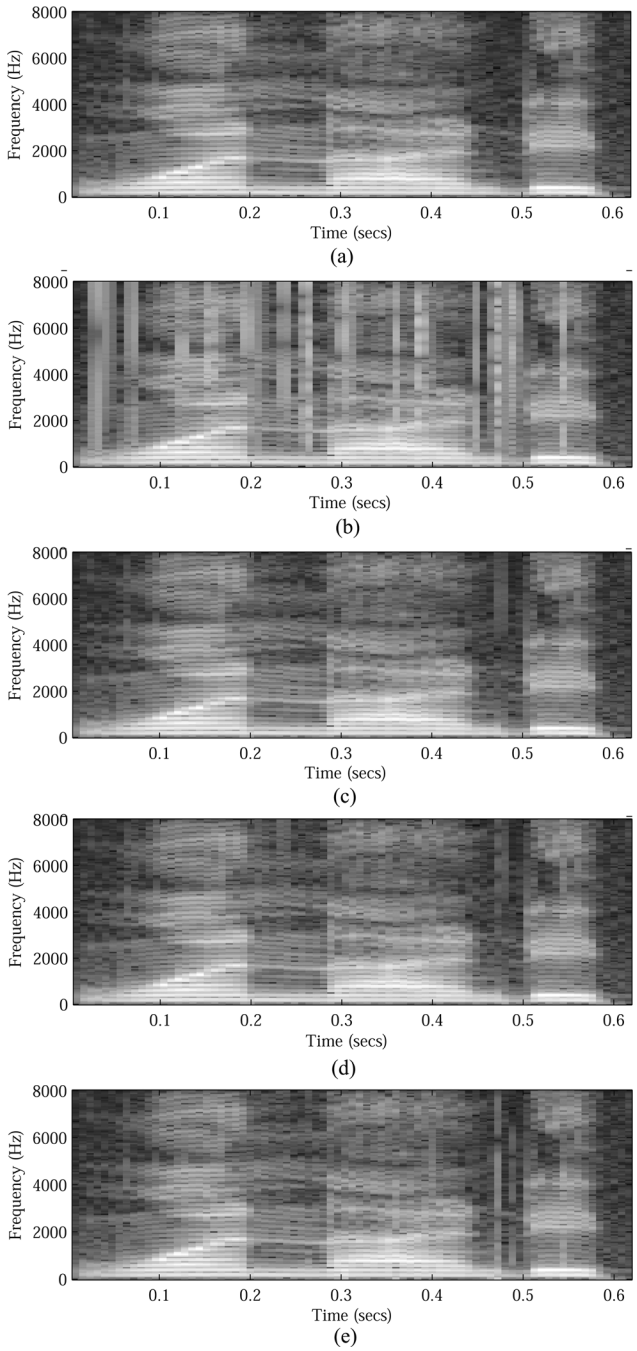

FIG. 7. Spectrograms of (a) the clean speech signal, (b) the clean speech signal + impulse noise, (c) after processing by "proposed-variant1," (d) after processing by "proposed-variant2," and (e) after processing by the competing method (Ref. 1). The Fourier-transform window length for the spectrograms is 256 with 50% overlap.

C. Wavelet performance comparison for impulse detection

Here we perform experiments to show how certain aspects of the wavelet influence the detection performance. We carry out three experiments. In the first, experiment C-1, we show how the detection performance is dependent on the frequency responses of the wavelet, the impulse noise, and the speech signal. Then in the second experiment, experiment C-2, we show how good detection is dependent on the support size relative to the impulse width of the wavelet. And in the third experiment, experiment C-3, we show how the support size is dependent on the strength of the impulse noise for good detection.

1. Experiment C-1

In this experiment, we study how the detection performance is dependent on the frequency response of the wavelet, the impulse noise, and the speech signal. To ensure that differences in wavelet support size do not influence the result, we use wavelets of the same order for comparison. We consider wavelets with orders of 24 and select the following wavelets: Daubechies order 24 (db24), Coiflet order 24 (cf24), Symmlet order 24 (sy24), Vaidyanathan order 24 (va24), and Battle-Lemarie order 23 (bl23). Note, however, that the Battle-Lemarie wavelet of order 24 is not defined in MATLAB or WAVELAB;[27,28] the closest one available is order 23.

To obtain the test signal we combine the speech signal with artificially generated impulse noise. The impulse noise is generated in the same manner as in experiments A-1 and A-2 of Sec. V A. The SINR is set to 10 dB with 5% contamination and an average impulse width of 1 ms.

To test the detection performance, we use the condition in Eq. (4) to decide if an impulse is present. The detection performance is then compared with the ideal result template, which is obtained by running the same detector in Eq. (4) on the impulse noise signal only. An impulse will be assumed to have been correctly detected if the detector output for that impulse corresponds to the location of the impulse in the template output.

Using Monte Carlo simulation, we determined that the width of the generated impulse noise is less than 7 ms for 99.9% of the occurrences. Therefore, for the experiment we set the maximum impulse width $K_{\max}$ to 112 samples, which corresponds to 7 ms at 16 kHz. And to ensure that the condition for the median filter length in Eq. (6) is satisfied, we set the median filter length, L_{med}, to the shortest possible length, which is $2K_{\max}+1=225$ samples.

To get the optimal detection result for a given wavelet, the detection error is computed for different values of k_f in Eq. (5). An optimal value of k_f is defined as the value of k_f where the total detection error is at a minimum. The total detection error, $\epsilon_t(k_f)$, is the normalized sum of the number of false detection of impulses and the number of non-detection of impulses. That is,

$$\epsilon_t(k_f)=\epsilon_f(k_f)+\epsilon_n(k_f), \tag{32}$$

where $\epsilon_f(k_f)$ and $\epsilon_n(k_f)$ are, respectively, the ratio of the number of false detections and non-detections to the total number of impulses present. The detection error is computed for the five wavelet types at sampling frequencies of 16 kHz, using MATLAB in combination with the WAVELAB toolbox. The speech signal has a mixture of male and female speech and is about 5 min long. Figures 8(a)–8(c) show plots of $\epsilon_t(k_f)$, $\epsilon_f(k_f)$, and $\epsilon_n(k_f)$, respectively. As can be seen, the Vaidyanathan wavelet gives the best total detection performance. We also observe that as k_f increases, $\epsilon_f(k_f)$ decreases and $\epsilon_n(k_f)$ increases for all the wavelets. Note that in applications

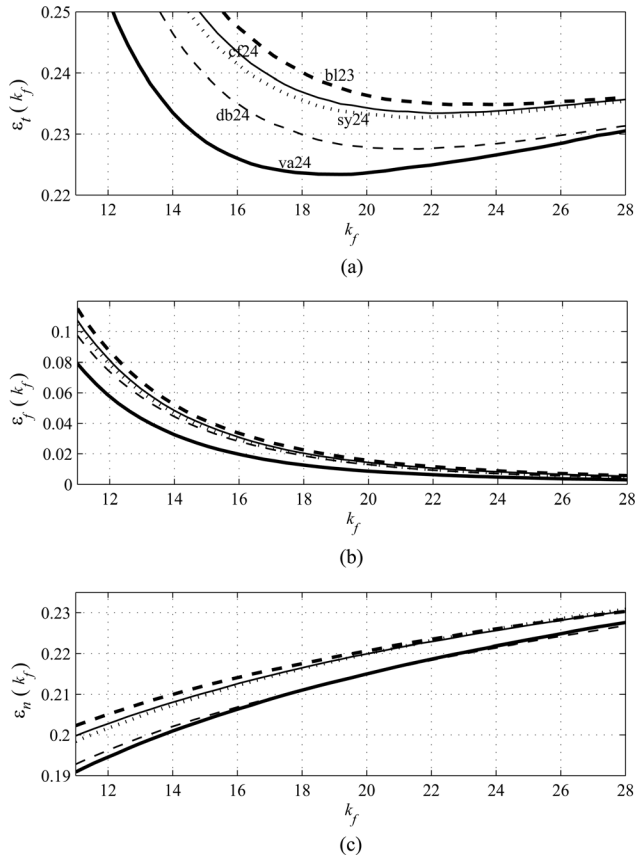

FIG. 8. Plots of $\epsilon_t(k_f)$, $\epsilon_f(k_f)$, and $\epsilon_n(k_f)$ at sampling frequency of 16 kHz; the average width of the impulse noise is 1 ms with a SINR of 10 dB and 5% contamination, and the median filter in the detector is 225 samples long.

where the removal of the detected impulses results in an unacceptable level of distortion, having a smaller value of $\epsilon_f(k_f)$ at the expense of larger $\epsilon_n(k_f)$ may be preferable to obtaining the minimum value of $\epsilon_t(k_f)$. Conversely, if the impulse-removal algorithm causes little or no distortion it may be preferable to reduce $\epsilon_n(k_f)$ at the expense of an increase in $\epsilon_f(k_f)$, although this will require more overall computational effort since more impulses will be removed.

Next, the separability measure J defined in Eq. (1) is computed for the various wavelets in the same manner as in experiment A-1, and the values are plotted in Fig. 9 for the different wavelets. From the plot, we observe that the Vaidyanathan wavelet gives the highest value of J. In Fig. 10 plots of the amplitude responses for the first stage of the wavelet high-pass filter for a sampling frequency of 16 kHz are shown. Also included in the plot is the average spectral

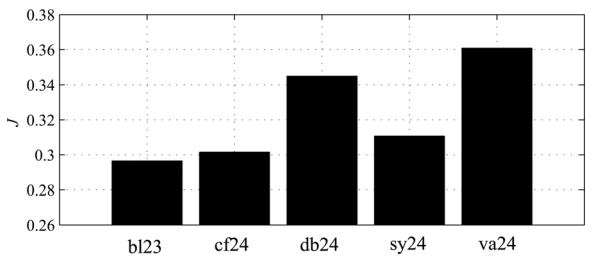

FIG. 9. Plots of J for wavelets with order 23 or 24; the average width of the impulse noise is 1 ms with a SINR of 10 dB and 5% contamination.

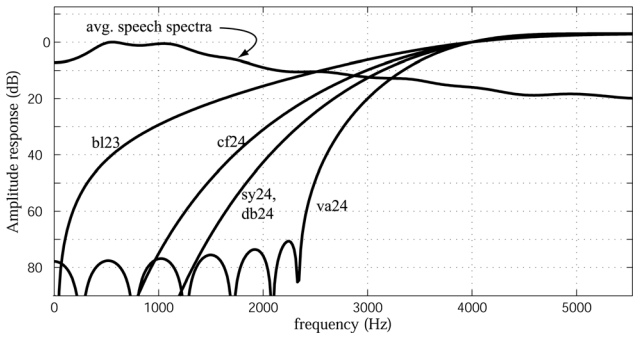

FIG. 10. Plots showing amplitude response curves of the first stage of the various wavelet high-pass filters, for 16 kHz sampling frequency; also included is the average spectral energy plot of a typical speech signal.

energy of a typical speech signal computed using a 20th-order LP model. The ratio of the average energy of the impulse noise to that of the speech signal in the finest scale, $\mathcal{R}_f$, is given by

$$\mathcal{R}_f = \frac{\int_{-\pi}^{\pi} E_i(\omega)|H_h(\omega)|d\omega}{\int_{-\pi}^{\pi} E_s(\omega)|H_h(\omega)|d\omega}, \tag{33}$$

where $E_s(\omega)$, $E_i(\omega)$, and $|H_h(\omega)|$ are the average energy spectrum of the speech signal, the average energy spectrum of the impulse noise, and the amplitude response of the wavelet high-pass filter, respectively. Since the impulse noise used in the experiment is generated from a Gaussian white noise process, the average frequency response of the impulse noise is a flat spectrum and, therefore, $E_i(\omega) = \text{const}$. In Table IV, values of $\mathcal{R}_f$, J, and the detection error for the different wavelets are listed. From Table IV, we notice a strong correlation between all three parameters, that is, the higher the value of $\mathcal{R}_f$, the higher J is and the smaller the detection error. We can also infer that for the same wavelet support size, the wavelet high-pass filter that maximizes the impulse signal in relation to the speech signal will give the largest value of J and, in turn, the best detection performance.

From Fig. 10 and Table IV we observe that wavelets db24 and sy24 have a high-pass filter with almost identical amplitude response and similar values of $\mathcal{R}_f$, yet their values of J and detection error are different, with db24 having a slightly better performance. We attribute this difference to the fact that the db24 filter is minimum phase while sy24 is not,[6] and therefore the energy of the db24 filter is optimally concentrated at the start of the impulse response; this higher

TABLE IV. Comparison of separability J, the minimum detection error, and R_f for various wavelets with equal (or nearly equal) wavelet support.

Wavelet type	Wavelet support	J	Minimum detection error	R_f (dB)
bl23	23	0.297	0.235	4.51
cf24	24	0.302	0.234	4.96
db24	24	0.345	0.228	5.08
sy24	24	0.311	0.233	5.08
va24	24	0.361	0.223	5.21

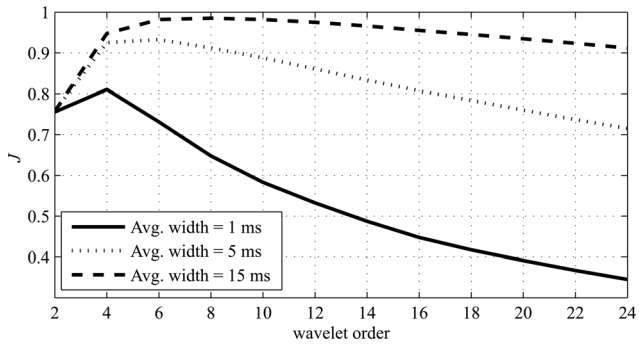

FIG. 11. Plots of J versus the wavelet support for the Daubechies wavelet at 16 kHz sampling frequency, for impulse noise with average widths of 1, 5, and 15 ms; the SINR is 10 dB with 5% contamination.

concentration of temporal energy results in lesser smearing of the impulse wavelet coefficients and, consequently, better detection.

2. Experiment C-2

In this experiment, we show how the support size of the wavelet is dependent on the impulse width for optimal impulse-detection performance. To show the dependency between wavelet support size and impulse width, we compare the separability measure, J, of the Daubechies wavelet at different support sizes by varying the order of the Daubechies wavelet. The comparison is carried out for three impulse-noise widths: The first has an average impulse width of 1 ms, the second has 5 ms, and the third has 15 ms. As in the other experiments, the SINR is set to 10 dB with 5% contamination. The values of J for the three impulse-noise widths are plotted versus the wavelet support size, or alternatively wavelet order, in Fig. 11. As can be seen from the plots, the optimal value of the wavelet support size increases as the average width of the impulse noise increases.

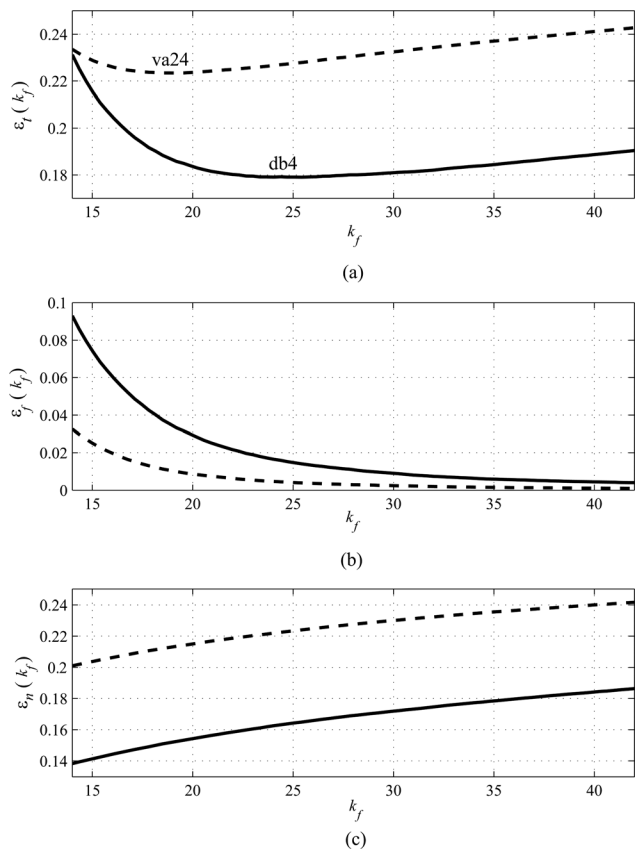

FIG. 12. Plots of the detection errors $\epsilon_t(k_f)$, $\epsilon_f(k_f)$, and $\epsilon_n(k_f)$ for the va24 and db4 wavelets at 16 kHz sampling frequency; the impulse noise has an average width of 1 ms with a SINR of 10 dB and 5% contamination.

TABLE V. Comparison of separability J, the minimum detection error, and R_f for two wavelets with different support. The impulse noise has an average width of 1 ms with a SINR of 10 dB and 5% contamination.

Wavelet type	Wavelet support	J	Minimum detection error	R_f (dB)
db4	4	0.811	0.180	3.53
va24	24	0.361	0.223	5.21

Next, we keep the average impulse-noise width at 1 ms and compare the separability measure, J, and the detection error between the Daubechies wavelet of order 4 (db4) and the Vaidyanathan wavelet of order 24 (va24). In experiment C-1, we determined the va24 wavelet to be most appropriate among wavelets with order 24, for detection of impulse noise. However, as can be seen from Fig. 12 and Table V, the db4 wavelet has a much higher separability measure and smaller total detection error when compared to the va24 wavelet. Therefore, we can conclude that for impulse noise that has a flat spectrum with an average width of 1 ms and SINR of 10 dB, the lower temporal smearing of the db4 wavelet is more critical than the frequency selectivity of the va24 wavelet for impulse detection.

3. Experiment C-3

In this experiment, we show how the support size of the wavelet is dependent on the strength of the impulse noise for optimal impulse-detection performance. To show the dependency between wavelet support size and impulse-noise strength, we compare the separability measure, J, of the Daubechies wavelet at different support sizes by varying the order of the Daubechies wavelet. The comparison is carried out for three impulse-noise strengths: The first has a SINR of

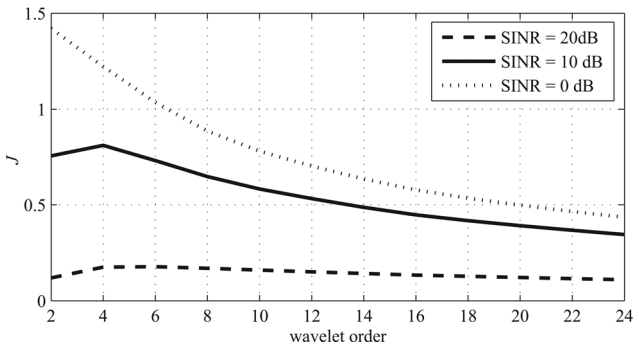

FIG. 13. Plots of J versus the wavelet order (or support size) for the Daubechies wavelet at 16 kHz sampling frequency, for impulse noise with SINR of 20, 10, and 0 dB; the average impulse width is 1 ms with 5% contamination.

20 dB, the second has 10 dB, and the third has 0 dB. The average width of the impulse noise is set to 1 ms with 5% contamination. The values of J for the three impulse-noise strengths are plotted versus the wavelet support size in Fig. 13. As can be seen from the plots, the optimal value of the wavelet support size decreases as the strength of the impulse noise increases.

VI. CONCLUSION

A new method for detecting and removing impulse noise from speech in the wavelet transform domain has been described. The method utilizes the multi-resolution property of the wavelet transform, which provides finer time resolution at high frequencies, to effectively identify and remove the impulse noise. We then established how the impulse-detection performance is dependent on certain wavelet features and their relationship with the impulse noise and the underlying speech signal. Performance evaluations carried out with an existing method showed that the wavelet approach gives much better features for detecting the impulses. To remove the impulses, a new algorithm that uses the stationary wavelet transform has been developed. The algorithm uses a two-step approach where the wavelet coefficients corresponding to the impulses are suppressed in the first step and then replaced by suitable coefficients located within the vicinity of the impulse in the second step. Performance evaluation with an existing method showed that the new algorithm gives superior results.

ACKNOWLEDGMENTS

The authors are grateful to the Natural Sciences and Engineering Research Council of Canada for supporting this work.

APPENDIX A: SEPARABILITY MEASURE

The separability measure is built upon information related to the way feature vectors are scattered in space. If n_i is the number of vector features $\mathbf{x}$ in class ω_i, the mean, $\mathbf{m}_i$, and scatter matrix, $\mathbf{S}_i$, of the class are defined as

$$\mathbf{m}_i = \frac{1}{n_i}\sum_{x\in\omega_i}\mathbf{x}, \tag{A1}$$

$$\mathbf{S}_i = \frac{1}{n_i}\sum_{\mathbf{x}\in\omega_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^T. \tag{A2}$$

If $n_T = \sum_{i=1}^{c} n_i$ is the total number of samples and $p_i = n_i/n_T$ is the *a priori* probability of class ω_i, the within-class scatter matrix $\mathbf{S}_W$ and between-class scatter matrix $\mathbf{S}_B$ are, respectively, given by

$$\mathbf{S}_W \sum_{i=1}^{c} p_i\mathbf{S}_i, \tag{A3}$$

$$\mathbf{S}_B = \sum_{i=1}^{c} p_i(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^T, \tag{A4}$$

where

$$\mathbf{m} = \sum_{i=1}^{c} p_i\mathbf{m}_i. \tag{A5}$$

Consequently, a popular separability measure for the feature vector $\mathbf{x}$ is given by[29]

$$J = \mathrm{trace}\{\mathbf{S}_W^{-1}\mathbf{S}_B\}. \tag{A6}$$

For a one-dimensional two-class problem, J simplifies to

$$J = \frac{n_1(m_1-m)^2 + n_2(m_2-m)^2}{\sum_{x\in\omega_1}(x-m_1)^2 + \sum_{x\in\omega_2}(x-m_2)^2}. \tag{A7}$$

An important advantage of the measure in Eq. (A6) is that it is invariant under linear transformations.[29]

APPENDIX B: THE MUTUAL INFORMATION MEASURE

The MI expression in Eq. (2) can be alternatively expressed as

$$\mathrm{MI}(X;Y) = E_p\left\{\log\frac{p(x,y)}{p(x)p(y)}\right\}, \tag{B1}$$

where E_p is the expectation operator with probability density function (PDF) p. The joint PDF $p(x,y)$ is approximated by a Gaussian mixture model (GMM), which is a sum of L weighted Gaussian densities $\mathcal{N}(\cdot)$ with mean vectors μ_l and covariance matrices Σ_l, given by

$$\bar{p}(x,y) = \sum_{l=1}^{L}\rho_l\mathcal{N}(x,y;\mu_l,\Sigma_l) \approx p(x,y), \tag{B2}$$

where ρ_l are the scalar weights. The parameters ρ_l, μ_l, and Σ_l are trained by the expectation maximization algorithm. From the estimated GMM $\bar{p}(x,y)$, the MI is computed numerically using

$$\mathrm{MI}(X;Y) \approx \frac{1}{N}\sum_{k=1}^{N}\log\frac{\bar{p}(\bar{x}(k),\bar{y}(k))}{\bar{p}(\bar{x}(k))\bar{p}(\bar{y}(k))}, \tag{B3}$$

where the pairs $\{\bar{x}(k),\bar{y}(k)\}$ are generated from the GMM $\bar{p}(x,y)$. The computation is performed with $N = 10^6$ generated pairs.

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

[2] P. Esquef, M. Karjalainen, and V. Valimaki, "Detection of clicks in audio signals using warped linear prediction," in *14th International Conference on Digital Signal Processing* (2002), Vol. 2, pp. 1085–1088.

[3] C. Chandra, M. S. Moore, and S. K. Mitra, "An efficient method for the removal of impulse noise from speech and audio signals," in *Proceedings of the International Symposium on Circuits and Systems (ISCAS 1998)*, Vol. 4, pp. 206–208.

[4] 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 the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, Vol. 1, pp. 1093–1096.

[5]S. V. Vaseghi and R. Frayling-Cork, "Restoration of old gramophone recordings," J. Audio Eng. Soc. **40**, 791–801 (1992).
[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. R. Coifman and D. L. Donoho, "Translation invariant de-noising," in *Wavelets and Statistics*, edited by A. Antoniadis and G. Oppenheim (Springer, New York, 1995), pp. 125–150.
[9]R. C. Nongpiur, "Impulse noise removal in speech using wavelets," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008)*, pp. 1593–1597.
[10]S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," IEEE Trans. Inf. Theory **38**, 617–643 (1992).
[11]S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 3rd ed. (Academic, San Diego, 2006), pp. 224–231.
[12]T. M. Clover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley Interscience, Hoboken, NJ, 2006), pp. 13–37.
[13]R. Battiti, "Using mutual information for selecting features in supervised neural net learning," IEEE Trans. Neural Networks **5**, 537–550 (1994).
[14]N. Kwak and C.-H. Choi, "Input feature selection for classification problems," IEEE Trans. Neural Networks **13**, 143–159 (2002).
[15]H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," IEEE Trans. Pattern Anal. Mach. Intell. **27**, 1226–1238 (2005).
[16]N. C. Gallagher, Jr. and G. L. Wise, "A theoretical analysis of the properties of median filters," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-29**, 1136–1141 (1981).
[17]I. Cohen, J. Benesty, and S. Gannot, *Speech Processing in Modern Communication*, 2nd ed. (Springer, Berlin, 2010), Chap. 7, pp. 183–198.
[18]J. B. Bednar and T. L. Watt, "Alpha-trimmed means and their relationship to median filters," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-32**, 145–153 (1984).
[19]G. P. Nason and B. W. Silverman, "The stationary wavelet transform and some statistical applications," in *Wavelets and Statistics*, edited by A. Antoniadis and G. Oppenheim, Lecture Notes in Statistics Vol. 103 (Springer, New York, 1995), pp. 281–299.
[20]P. Rajmic and J. Vlach, "Real-time audio processing via segmented wavelet transform," in *Proceedings of the 10th International Conference on Digital Audio Effects*, Bordeaux, France (2007).
[21]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).
[22]S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, 4th ed. (Wiley, Chicheser, UK, 2008), pp. 349–355.
[23]C. Hemphill, J. Godfrey, and G. Doddington, "The ATIS spoken language systems pilot corpus," in *Proceedings of the DARPA Speech and Natural Language Workshop*, Hidden Valley, PA (1990), pp. 96–101.
[24]A. H. Gray, Jr. and J. D. Markel, "Distance measures for speech processing," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-24**, 380–391 (1976).
[25]www.ece.uvic.ca/~rnongpiu/jasa.html (Last viewed July 12, 2012).
[26]H. Fastl and E. Zwicker, *Psychoacoustics—Facts and Models*, 3rd ed. (Springer, Berlin, 2007), Chap. 7, pp. 174–202.
[27]J. B. Buckheit and D. L. Donoho, "WaveLab and Reproducible Research," in *Wavelets and Statistics*, edited by A. Antoniadis and G. Oppenheim, Lecture Notes in Statistics Vol. 103 (Springer, New York, 1995), pp. 55–81.
[28]http://www-stat.stanford.edu/~wavelab/ (Last viewed July 12, 2012).
[29]A. R. Webb, *Statistical Pattern Recognition*, 2nd ed. (Wiley, Chichester, UK, 2002).