The shift-invariant discrete wavelet transform and application to speech waveform analysis

Jörg Enders, Weihua Geng, Peijun Li, and Michael W. Frazier
Department of Mathematics, Michigan State University, East Lansing, Michigan 48824-1027

David J. Scholl
Ford Motor Company, MD3083/SRL Building, Dearborn, Michigan 48121-2053

(Received 17 April 2004; revised 15 October 2004; accepted 24 January 2005)

The discrete wavelet transform may be used as a signal-processing tool for visualization and analysis of nonstationary, time-sampled waveforms. The highly desirable property of shift invariance can be obtained at the cost of a moderate increase in computational complexity, and accepting a least-squares inverse (pseudoinverse) in place of a true inverse. A new algorithm for the pseudoinverse of the shift-invariant transform that is easier to implement in array-oriented scripting languages than existing algorithms is presented together with self-contained proofs. Representing only one of the many and varied potential applications, a recorded speech waveform illustrates the benefits of shift invariance with pseudoinvertibility. Visualization shows the glottal modulation of vowel formants and frication noise, revealing secondary glottal pulses and other waveform irregularities. Additionally, performing sound waveform editing operations (i.e., cutting and pasting sections) on the shift-invariant wavelet representation automatically produces quiet, click-free section boundaries in the resulting sound. The capabilities of this wavelet-domain editing technique are demonstrated by changing the rate of a recorded spoken word. Individual pitch periods are repeated to obtain a half-speed result, and alternate individual pitch periods are removed to obtain a double-speed result. The original pitch and formant frequencies are preserved. In informal listening tests, the results are clear and understandable. © *2005 Acoustical Society of America.* [DOI: 10.1121/1.1869732]

PACS numbers: 43.60.Hj, 43.72.Ar [DOS] Pages: 2122–2133

I. INTRODUCTION

In experimental acoustics, it is common to encounter nonstationary sound waveforms, i.e., those in which the frequency content and amplitude change as a function of time. The conventional approach for analyzing such sounds is to calculate a spectrogram, or short-time Fourier transform (STFT). For a time-sampled waveform $z(t)$, the STFT provides information about the waveform's energy content as a function of both time and frequency, i.e., $F_{\mathrm{STFT}}(z) = E(t,f)$. While the STFT has proven its worth in numerous practical applications, it is ill suited to certain types of sounds, and it lacks some desirable mathematical characteristics. Sounds with frequency content ranging over more than one or two orders of magnitude are often problematic for STFT analysis, because a window long enough to capture low-frequency content (at least one period) will be insensitive to high-frequency sounds of short time duration.

The discrete wavelet transform (DWT) has a severe limitation when used for acoustic waveform analysis: its lack of shift invariance. Let two time-sampled waveforms $z(t)$ and $z'(t)$ be time-shifted copies of one another, such that $z(t) = z'(t+t_0)$ for all t. Calculating the DWT of each, $F_{\mathrm{DWT}}(z) = E(t,n)$, and $F_{\mathrm{DWT}}(z') = E'(t,n)$. Since the DWT is not shift invariant, $E(t,n) \neq E'(t+t_0,n)$. Therefore, the DWT analysis of a sampled sound depends on when the sampling starts, not just when the sound occurs, which is highly undesirable for the study of physical systems. The DWT is critically sampled, i.e., utilizes lower sampling rates (subsampling) for lower-frequency components. The selection of samples to be skipped in the subsampling process is inextricably linked to the time elapsed since the sampling began. Fortunately, by modifying the DWT to retain all possible samples (performing no subsampling), it is possible to obtain explicit shift invariance.[1–3] The resulting shift-invariant discrete wavelet transform (SIDWT) is highly redundant, but since many of the redundant elements are duplicates, the increase in computational complexity is not severe. The full SIDWT may be used as a starting point from which to draw a more efficient representation for lossy compression.[4] The SIDWT may also be used in full (albeit with the duplicate elements grouped and summed), in which form it has been shown to be an isometry, with applications in data visualization.[5] Others have described algorithms which are mathematically equivalent to the SIDWT, but which were developed for applications in exploratory statistics, using different nomenclature, i.e., the stationary wavelet transform[6] and the maximal overlap discrete wavelet transform.[7] The stationary wavelet transform has also been used successfully for waveform denoising.[2]

The use of the SIDWT (and its equivalents) to identify features in a waveform; whether signatures of interesting phenomena, experimental artifacts, or noise, leads naturally to the following question. What would the time-sampled waveform look like (or sound like) if the features were louder, softer, appeared at a different time, or were removed altogether? Performing the desired modifications on the SIDWT output is straightforward; the challenge is reversing the SIDWT to return to a time-sampled waveform. Because of its inherent redundancy, the SIDWT does not have a true inverse in the mathematical sense. However, this fact does not preclude the existence of an algorithm with useful inverse-like behavior. The developers of the stationary wavelet transform also developed such an inverse-like procedure. They showed that averaging together all of the possible shift-induced variations of the IDWT yields intuitively satisfying results.[6] Likewise, a mathematically equivalent procedure was used to invert the maximal overlap discrete wavelet transform, also with intuitively satisfying results.[7] The stationary wavelet transform combined with this inverse-like averaging procedure has also been shown to yield good results in waveform denoising.[2] The wavelet denoising paper states, without proof or discussion, the important mathematical result that the inverse-like procedure is actually the pseudoinverse of the stationary wavelet transform. A recent publication coauthored by one of the present authors describes two examples of sound visualization and modification using the SIDWT and its pseudoinverse (ISIDWT). The discussion and the two examples are narrowly focused on the field of automotive sound quality engineering, and no mathematical material is included.[8]

The goals of this paper are threefold. The first is to describe a newly developed simple and fast convolution algorithm for the ISIDWT, based on the SIDWT algorithm.[5] The SIDWT, the stationary wavelet transform, and the maximal overlap discrete wavelet transform employ significantly different algorithms, so a discussion of computational issues is included. The second goal is to present a simple, self-contained proof that the ISIDWT is the pseudoinverse of the SIDWT. The statement of this result has been published; we believe the details of the proof should be made available as well. The third goal is to illustrate the potential applications of these new analytical methods in the field of acoustics. Section II covers both the theoretical (II A) and the computational (II B) aspects of the SIDWT and its pseudoinverse. In Sec. III, examples of low-level speech waveform processing illustrate the capabilities of the SIDWT/ISIDWT for visualization, feature separation, and analysis/synthesis. One especially promising way to combine these capabilities is to edit (cut and paste sections) of sound recordings in the shift-invariant wavelet domain. While many audible features are easier to recognize in that domain, the primary benefit is the wavelet pseudoinverse transform automatically prevents the occurrence of the audible clicks and pops that are usually produced at section boundaries by time-domain editing. Illustrative examples of waveforms with strong time localization and a wide frequency range can be found in many different technical fields of study. For the development of digital audio effects in music, it is useful to be able to distinguish transient (time-localized) sounds, such as the pluck of a guitar string, from the steady ringing tone (frequency-localized) that follows.[9] Research in wavelet-domain modification of musical sounds began in the early days of wavelet theory, e.g., musical applications of complex wavelets,[10] and continues today, e.g., the use of a ‘‘lapped’’ wavelet transform to stitch together segments of musical waveforms.[11] The examples most familiar to the authors are drawn from the myriad of mechanical sounds produced by motor vehicles, e.g., a momentary rattle excited by (and partially masked by) a car door slam, and the motor whine, blade scrape, and reversal thud of a windshield wiper.[8] The details of speech waveforms, especially the formant resonances modulated by glottal pulses, are also an excellent match to the capabilities of the SIDWT/ISIDWT. The general approach and terminology derives from an *ad hoc* list of recent publications that deal with various details of speech waveforms: pitch period estimation,[12–14] formant modulation,[15] friction noise modulation,[16] voicing onset,[17] glottal characteristics,[18] and waveform irregularities.[19]

II. THE SHIFT-INVARIANT DISCRETE WAVELET TRANSFORM AND ITS PSEUDOINVERSE

A. Theory

Consider a sequence of N physical measurements $z=(z_1,z_2,\ldots,z_N)$, e.g., air pressure measured repeatedly at evenly spaced time intervals. Let $\boldsymbol{S}$ denote the set of all such signals. Since $\Sigma_n z_n^2<+\infty$, the vector z may be regarded as the coordinates of a single point in a finite energy, N-dimensional vector space, $z\in l^2(\mathbb{Z}_N)$. Implicit in $z\in l^2(\mathbb{Z}_N)$ is the assumption that z is a single period of an infinitely long sequence with a periodicity of N. If this assumption is not physically realistic, care must be taken to insure that the conclusions drawn from the analysis are independent of N.

Let $u\in l^2(\mathbb{Z}_N)$ and $v\in l^2(\mathbb{Z}_N)$ represent two digital filters. Denoting the discrete Fourier transform of u by $\hat{u}$, we require the system matrix

$$A(n)=\frac{1}{\sqrt{2}}\begin{pmatrix}\hat{u}(n) & \hat{v}(n)\\ \hat{u}\left(n+\frac{N}{2}\right) & \hat{v}\left(n+\frac{N}{2}\right)\end{pmatrix}, \tag{1}$$

to be unitary for each $n=0,\ldots,N-1$ (Ref. 20, p. 173). Therefore, u is the low-pass filter sequence and v the high-pass filter sequence generating the discrete wavelet transform.

Let $\tilde{u}$ be the complex conjugate reflection of u defined by $\tilde{u}(n)=u^*(N-n)$ for all n. The finite impulse response filtering of z by u is written as a (circular) convolution $z*u$. Most practical applications of these techniques, including the examples presented here, involve only N-element sequences of real numbers, i.e., $x\in\mathbb{R}^N\subset l^2(\mathbb{Z}_N)$. The mathematical results, however, are valid for complex-valued vectors. Assuming that m divides N, a sequence reordering operator R_m, defined by

$$R_m(z)=(z_1,z_{m+1},\ldots,z_{N-(m-1)},z_2,z_{m+2},\ldots,z_{N-(m-2)},\ldots,\\ z_m,z_{2m},\ldots,z_N), \tag{2}$$

in effect, writes the elements of z into an m by N/m matrix by columns and reads the elements out by rows. For example, if $z=(1,2,3,4,5,6,7,8)$, then $R_2(z)=(1,3,5,7,2,4,6,8)$. The inverse of R_m is $R_{N/m}$, i.e., $R_{N/m}(R_m(z))=z$.

From an N-element input vector, given p such that 2^p divides N, the p-stage shift-invariant discrete wavelet transform T produces a $(p+1)$ by N matrix.[5] T may therefore be regarded as a linear map taking each point in $\boldsymbol{S}$ to a point in a $(p+1)N$-dimensional vector space $\boldsymbol{W}$, i.e.,

$$T{:}l^2(\mathbb{Z}_N)\rightarrow l^2(\mathbb{Z}_{(p+1)N}). \tag{3}$$

T of z is given by

$$T(z)=(R_{N/2}(x_1),R_{N/4}(x_2),\ldots,R_{N/2^p}(x_p),R_{N/2^p}(y_p)), \tag{4}$$

where

$$x_1=\frac{1}{\sqrt{2}}R_2(z*\tilde{v}) \quad \text{and} \quad y_1=\frac{1}{\sqrt{2}}R_2(z*\tilde{u}), \tag{5}$$

and for $j=2,...,p$

$$x_j=\frac{1}{\sqrt{2}}R_2(y_{j-1}*\tilde{v}) \quad \text{and} \quad y_j=\frac{1}{\sqrt{2}}R_2(y_{j-1}*\tilde{u}). \tag{6}$$

The set of points mapped by T from $\boldsymbol{S}$ occupies a subspace in $\boldsymbol{W}$ denoted by range (T).

Being a linear map from an N-dimensional space of signals to a space of larger dimension, the SIDWT does not have an inverse. Of all the points in $\boldsymbol{W}$, only those which are in range (T) are directly associated with a point in $\boldsymbol{S}$. The pseudoinverse works around this limitation by providing every point $w\in\boldsymbol{W}$ with an *indirect* association to some point $z\in\boldsymbol{S}$. Every w has a unique nearest (in the standard Euclidean norm) neighbor $w'\in\mathrm{range}(T)$ (possibly itself), and the pseudoinverse associates each w with the z that satisfies $T(z)=w'$. This procedure is mathematically equivalent to finding the least-squares solution to an overdetermined system of linear equations. We now define the ISIDWT

$$S{:}l^2(\mathbb{Z}_{(p+1)N})\rightarrow l^2(\mathbb{Z}_N), \tag{7}$$

a map taking each point in $\boldsymbol{W}$ to a point in $\boldsymbol{S}$. Given $w=(w_1,w_2,\ldots,w_{p+1})\in l^2(\mathbb{Z}_{(p+1)N})$, we compute $S(w)$ by the algorithm

$$\eta_p=\frac{1}{\sqrt{2}}(R_{N/2}(R_{2^p}(w_p))*v+R_{N/2}(R_{2^p}(w_{p+1}))*u) \tag{8}$$

$$\eta_{p-1}=\frac{1}{\sqrt{2}}(R_{N/2}(R_{2^{p-1}}(w_{p-1}))*v+R_{N/2}(\eta_p)*u) \tag{9}$$

$$\vdots$$

$$S(w)=\eta_1=\frac{1}{\sqrt{2}}(R_{N/2}(R_2(w_1))*v+R_{N/2}(\eta_2)*u). \tag{10}$$

The relationship between T and S (as defined above) is established by the following theorem.

Theorem 1. S is the least-squares inverse (pseudoinverse) of T, i.e.,

(i) $ST=\mathrm{id}|_{l^2(\mathbb{Z}_N)}$, and

(ii) TS is the orthogonal projection of $l^2(\mathbb{Z}_{(p+1)N})$ onto range (T).

According to statement (i), for $w\in\mathrm{range}(T)$, S is the inverse of T, and therefore $S(T(z))=z$. Statement (ii) addresses the case of $w\notin\mathrm{range}(T)$, which will apply to virtually all w chosen arbitrarily, i.e., not obtained by $w=T(z)$. In this case, $x=T(S(w))\in\mathrm{range}(T)$ is the unique point in range (T) closest to w, minimizing $\Sigma_n(x_n-w_n)^2$. A proof of Theorem 1 is given in the Appendix.

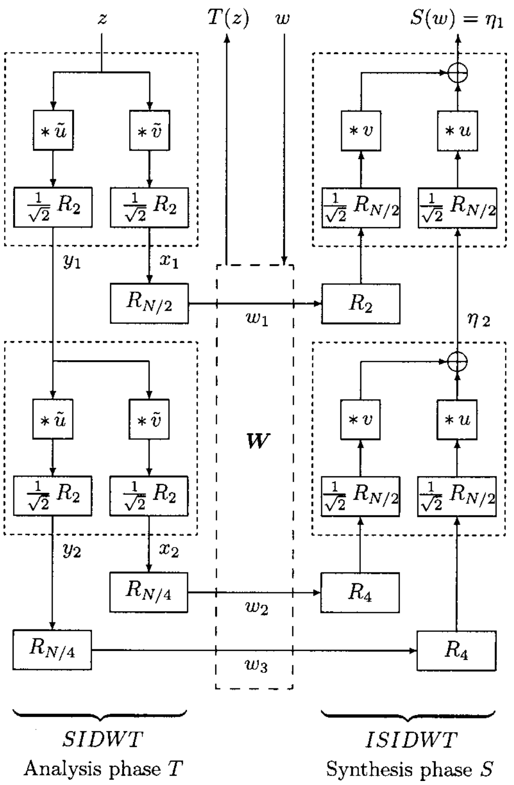

FIG. 1. A block diagram of the SIDWT and ISIDWT for two scale levels, i.e., $p=2$.

B. Computation

The block diagram in Fig. 1 illustrates the processing steps in the SIDWT (analysis phase T) and its pseudoinverse ISIDWT (synthesis phase S) for the two-stage case, i.e., $p=2$. The analysis begins in the upper-left corner with the input waveform z. The four-block clusters inside the dotted-line boxes on the left side of the diagram depict the recursive analysis steps defined in Eqs. (5) and (6). A p-stage transform employs p of these clusters, yielding $p+1$ many series $x_1,x_2,\ldots,x_p,y_p$, each of length N. The final rearrangement step defined in Eq. (4) reaches the vector space $\boldsymbol{W}$, depicted by the central dashed-line box. Completion of the SIDWT is indicated by the vertical arrow leading to $T(z)$ at the top of the diagram. The synthesis begins at the top with w, which in most cases will be a modification of $T(z)$. The first step in the ISIDWT is to undo the final rearrangement step in the SIDWT. After this, the recursive procedure in the dotted-line boxes in the right side of the diagram, as defined in Eqs. (8) and (9), is carried out. The ISIDWT is complete at the top of the diagram where $S(w)=\eta_1$, as in Eq. (10).

Let us look at the computational complexity of the pseudoinverse S. Given $z\in l^2(\mathbb{Z}_N)$, the computational complexity for the transform $T(z)$ is $\mathcal{O}(N\log_2 N)$ according to Ref. 5. We will now show that this result holds for the ISIDWT S as well.

Theorem 2. Let N be a power of 2 and $p \in \mathbb{N}$, with $p \leqslant \log_2 N$ fixed. Then, the total number of complex multiplications required to compute $S(w)$ for $w \in l^2(\mathbb{Z}_{(p+1)N})$ is $\#_N \leqslant 2pN + 3pN \log_2 N$.

Proof: In lieu of a detailed proof, we note that S and the analysis algorithm are essentially the same by symmetry.

From Theorem 2 we obtain the result that the computation of S is an $\mathcal{O}(N \log_2 N)$ operation if p is considered to be a fixed number. If we take the Daubechies D4 u and v and perform the convolutions directly instead of using the FFT, then the computational complexity is just $\mathcal{O}(N)$, since u and v have only 4 nonzero entries. This is the minimal order we can expect when working with signals of length N. Since S is a linear map, the error in the output is bounded by the norm of S (a constant) times the error in the input, which means that the algorithm is numerically stable.

The ISIDWT, the stationary wavelet transform, and the maximal overlap discrete wavelet transform are mathematically equivalent in the sense that they yield the same result. However, they utilize significantly different algorithms, so they are not computationally equivalent in all respects. All three may be calculated with $\mathcal{O}(N \log_2 N)$ computational complexity if p is considered fixed.[5–7] If implemented in a low-level programming language that allows efficient indexing of individual matrix elements, the performance of the three is expected to be essentially equivalent. However, the ISIDWT is significantly easier to implement in an array-oriented scripting language, because it can be constructed by linking together a few of the standard functions that are commonly provided in such languages. In this way, acceptable performance can be obtained without the need to write, compile, and link an external module written in a lower-level language.

III. APPLICATION TO SPEECH WAVEFORM ANALYSIS

To illustrate the application of the SIDWT/ISIDWT to acoustic waveforms, a detailed analysis of a sound recording of a spoken word is presented below. A recording of a male speaker pronouncing the Japanese word ‘‘kaze’’ with a rising intonation from an on-line speech database maintained for phonetic alphabet research[21] is shown in Fig. 2. Voiced speech is produced by periodic glottal closure events, which momentarily interrupt the air flow through the larynx. The frequency at which these events occur, denoted by $F0$, is the fundamental frequency of voiced speech. The rising intonation in this example is reflected in Fig. 2, as $F0 \approx 90$ Hz during the ‘‘a’’ increases to $F0 \approx 120$ Hz during the ‘‘e.’’ The procedure for glottal period estimation is discussed in detail below.

During spoken vowels, the sharp air-pressure transients known as glottal pulses excite pressure oscillations in the volume acoustic resonances of the vocal tract. The frequency content of these resonances, which fall in the range between ~500 Hz to ~8 kHz, depending on the size of the vocal tract and the position of the tongue, jaw, and lips, is the primary factor distinguishing one vowel from another. The significant frequency peaks in the pressure oscillations are known as formants, and are denoted $F1$,$F2$,$F3$,$F4$, in order of increasing frequency. The formant amplitude is highest immediately after the glottal pulse, at which point the glottis is essentially closed. The amplitude decreases rapidly as energy is lost due to air flow between the lips. When the glottis reopens in preparation for the next glottal pulse, the resulting air flow causes the formant amplitude to decrease even more quickly. This strong amplitude modulation of the formants leads to a widely used quasistatic approximation. In this simplified picture, the frequency content of each formant pulse is assumed to be static, and the (relatively slow) motion of the vocal-tract anatomy (tongue, jaw, and lips) is inferred by comparing the frequency content of consecutive formant pulses. The first six formant pulses in kaze may be seen in Fig. 3. The first formant pulse, which signifies the beginning of the vowel sound ‘‘a,’’ occurs at ~0.105 ms. In the sound waveform plot, each formant peak begins at a sharp downward step (a glottal pulse) and oscillates with decreasing amplitude, disappearing before the next glottal pulse. On the spectrogram, labeled ‘‘STFT,’’ each downward step appears as a vertical gray bar; the short time duration of each step maps to broad frequency content. The sampling rate was 44 100 samples/s, and a 352-point Hanning window, shifted in 63-point steps, and zero-padded to 1024-point length, was used in the preparation of spectrograms in this and the next two figures. The gray scale on each spectrogram was adjusted for enhanced contrast.[22] Each formant pulse appears as a pair of dark horizontal bands ~600 Hz apart, beginning at a vertical bar, and ending before the next vertical bar. This formant frequency content of $F1 \approx 500$ Hz and $F2 \approx 1100$ Hz is typical for a male Japanese speaker’s ‘‘a.’’ [23]

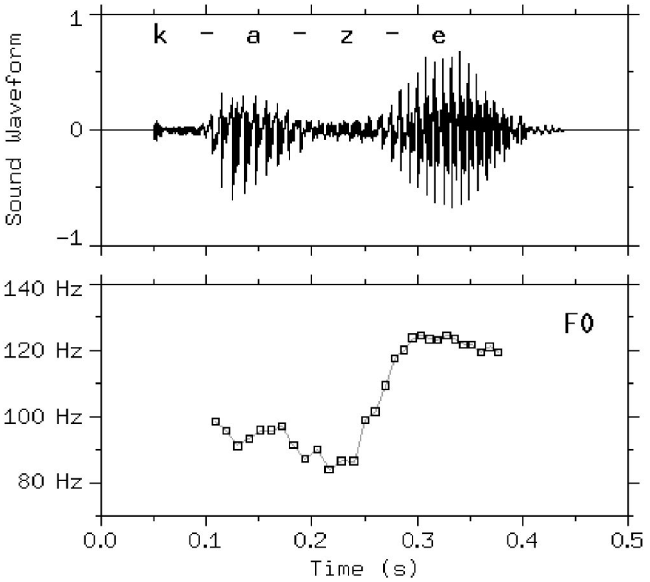

FIG. 2. A sound recording of a male speaker pronouncing the Japanese word ‘‘kaze’’ with rising intonation, and the voiced fundamental frequency $F0$ obtained from glottal period estimates.

The scalogram in Fig. 3 labeled ‘‘DWT’’ is obtained from the conventional, shift-variant, discrete wavelet transform. The 8-tap symlet was used for all examples presented here, but the results do not depend critically on the choice of wavelet. The shape of the wavelet (e.g., symlet vs Daubechies) makes little difference here. Shorter wavelet filters (e.g., 4-tap vs 8-tap) will have increased energy in the

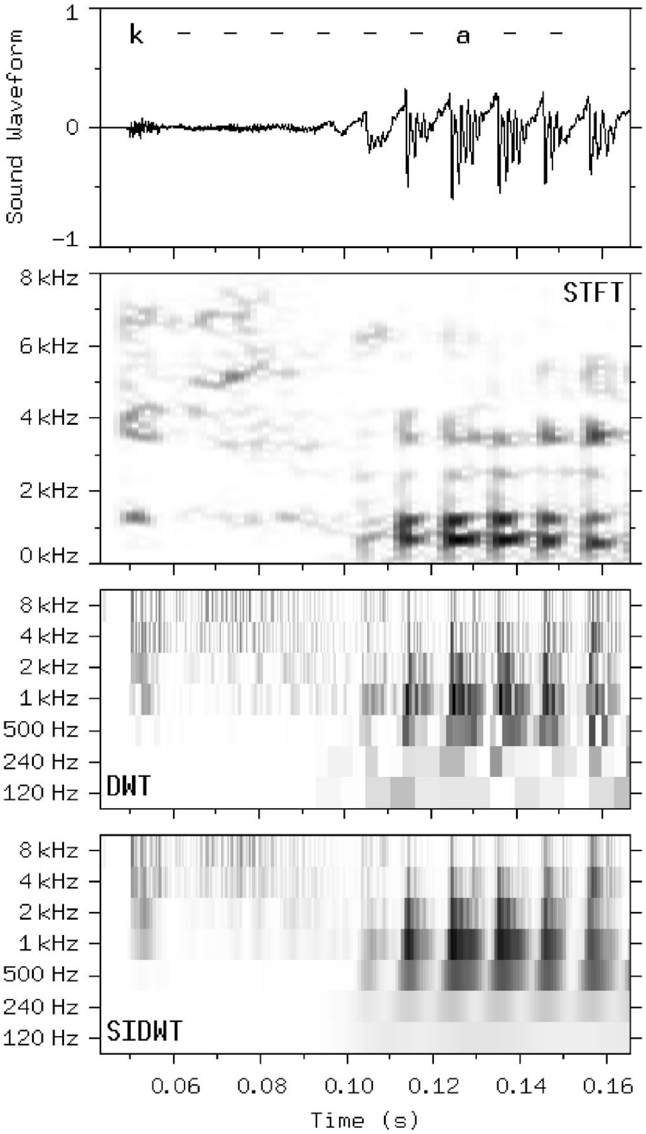

FIG. 3. The “ka” from the Japanese word “kaze,” its spectrogram (STFT), its conventional scalogram (DWT), and its shift-invariant scalogram (SIDWT).

upper sidebands, but provide faster calculations and sharper time resolution. The essential preprocessing step for visualization is to transform the oscillatory coefficients within each scale level of $w \in \text{range}(T)$ to a quadratic envelope.[5] For the coefficients at the mth scale level, w_m, the quadratic envelope $w'_m = w_m^2 + H(w_m)^2$, where H is the Hilbert transform, i.e., a $\pi/2$ phase shift. Following this operation, all of the scalograms presented here were downsampled to fit the available space, and the gray scales were adjusted for enhanced contrast.[22] The formant pulses appear as dark, vertical features extending from the 500-Hz to the 8-kHz bands. Their appearance is more varied than on the spectrogram, due to the shift variance of the DWT. Nevertheless, the DWT has been shown to be a reliable method for identifying glottal pulses for $F0$ estimation.[13,14] The scalogram in Fig. 3 labeled “SIDWT” is obtained from the shift-invariant discrete wavelet transform. In the region corresponding to the “k” sound, the SIDWT and DWT scalograms have a similar appearance. In the region corresponding to the voiced “a,” they are dramatically different. The shift invariance reveals the true reproducibility of the formant pulses, and for each formant pulse shows the amplitude decrease and the gap preceding the next glottal pulse. The voiced region highlights two important differences between the STFT presentation and the SIDWT presentation. First, the STFT has finer frequency resolution than the SIDWT. The two main formants, $F1$ and $F2$, are resolved clearly on the STFT, but are not resolved on the SIDWT due to the single-octave bandwidth of the scalogram levels. The second difference is that the STFT has the same time resolution at all frequencies, so transients appear as vertical features. In contrast, the time resolution of the SIDWT scales inversely with the center frequency of each band. For each step upward to a higher-frequency scalogram band, the time resolution is twice as fine. For this reason, a transient feature tends to have a pyramidal appearance on the SIDWT, with a narrow top on a base that broadens at each next lower level. The practical consequence of these two differences is that the SIDWT is not a substitute or a replacement for the STFT, but rather a complement, and the two techniques can be used effectively together.

Figure 4 shows a similar presentation of the “z” from kaze. This sound is produced by a narrow restriction in the mouth. The frictional (turbulent) loss due to the air flowing through the restriction prevents the build-up of formants. The absence of formants does not imply silence, however, because the turbulence produces audible noise called frication. The loudness of the frication varies with the flow of air through the restriction, which in turn is modulated by periodic glottal closures. The modulated frication appears as bursts of noise (glottis open) separated by momentary silences (glottis closed).[16] It is interesting to contrast this timing to that observed with formants, which are loudest when the glottis is closed, and quiet when the glottis is open.

To complete the presentation of kaze, the final vowel “e” is shown in Fig. 5. The time-domain clarity of the formant peaks in the shift-invariant scalogram (SIDWT), compared to the shift-variant scalogram (DWT), is even more evident here than in Fig. 3. The formant pulses are closer together than in Fig. 3, and they also exhibit a second high-frequency pulse in each glottal period. Secondary glottal pulses such as these are often observed in male speech waveforms, and they can be problematic for glottal period estimation algorithms. Interestingly, the phenomenon is usually vowel-dependent, and only traces of secondary pulses can be seen on the “a” in Fig. 3.

An expanded view of two of the formant pulses from Fig. 3 is shown in Fig. 6, as a time history and as a shift-invariant scalogram. The glottal pulses are indicated by “GP.” A periodic signature with a period of ~1.7 ms is apparent on the 2-kHz scale as alternating bands of light and dark, and is barely visible on the 1-kHz band. The scalogram as prepared for display is a quadratic function of the wavelet coefficients, and this signature is the difference frequency between the two formant peaks, $1.7\ \text{ms} \approx 1/(F2 - F1)$. The difference frequency shows up most clearly on the 2-kHz band because there is significant frequency overlap between adjacent bands, and the strong fundamental $F0$ and its strong

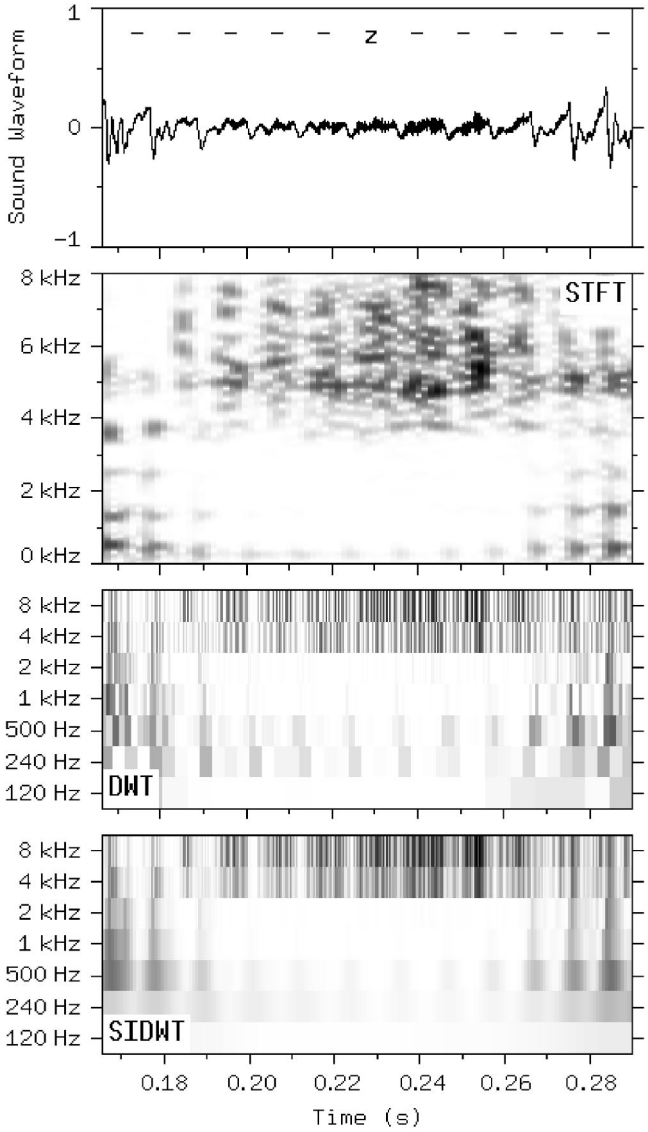

FIG. 4. The "z" from the Japanese word "kaze," its spectrogram (STFT), its conventional scalogram (DWT), and its shift-invariant scalogram (SIDWT).

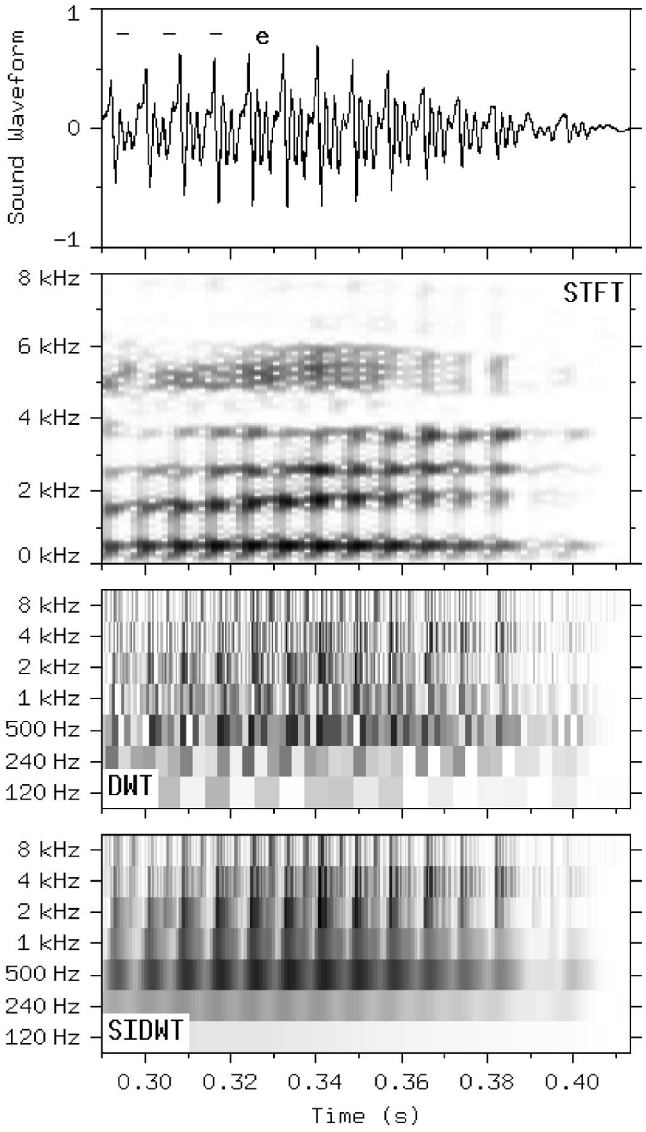

FIG. 5. The "e" from the Japanese word "kaze," its spectrogram (STFT), its conventional scalogram (DWT), and its shift-invariant scalogram (SIDWT).

second harmonic overwhelm the $F2-F1$ difference in the 500-Hz and 1-kHz bands.

A similar view of three of the formant pulses from the vowel sound "e" is shown in Fig. 7, as a time history and as a shift-invariant scalogram. Patterns that appear to be difference frequencies can be seen, but since the formant content of "e" is more complex than the two strong peaks responsible for the signature in Fig. 6, the scalogram signature of the vowel "e" is more complex as well. This expanded view provides a more detailed picture of the secondary glottal pulse labeled "2" in each glottal period, showing that the fundamental periodicity, as well as the gap preceding the glottal closure, are still evident. The $F0$ estimate in Fig. 2 was obtained by finding all occurrences of this formant gap-peak signature. For a list of times $t=(t_1,t_2,\ldots,t_M)$ at which the M occurrences of the signatures were observed, $F0$ at t_n is given by $F0_n=1/(t_{n+1}-t_n)$, where $N=1,2,...,(M-1)$. The times t were obtained by finding local maxima in the sum of the quadratic envelopes of the 1- and 2-kHz bands.

An important category of speech-processing techniques known as PSOLA (pitch-synchronized overLap and add) is based on working with the individual glottal pulses.[24] A typical application of PSOLA might begin with isolating each glottal pulse by multiplying the speech waveform by a rounded window (e.g., Hanning) centered over each pulse in turn. A typical length for the window would be twice the glottal period. The window length represents a compromise: longer windows allow the neighboring pulses to intrude, and shorter windows (or windows with more steeply sloped time-domain cutoffs) increase spectral leakage. After the individual glottal pulses have been processed in the desired manner, they must be recombined to make a single waveform. A variety of approaches has been used to recombine the individual segments,[24] including a technique which utilizes information obtained from wavelet transform analysis.[25] In general, PSOLA produces high-quality results, although sometimes annoying artifacts are present.[26,27] The artifacts are not completely understood, and may be related to the

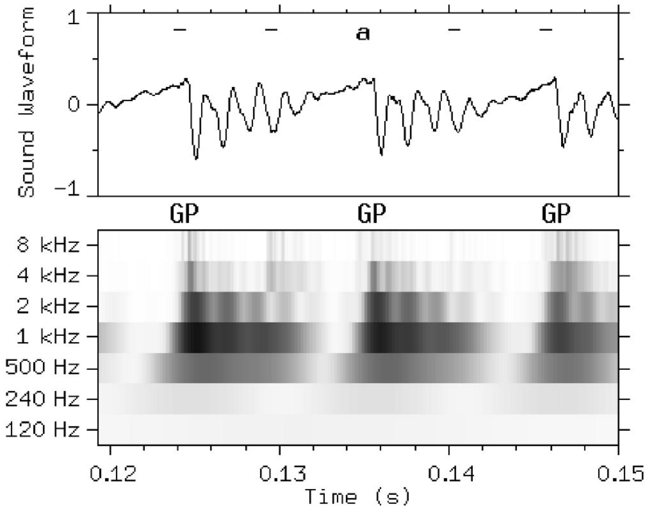

FIG. 6. The Japanese vowel "a," and its shift-invariant scalogram. The glottal pulses are indicated by "GP."

details of how the modified pulses are recombined.

The analysis/synthesis capability of the SIDWT/ISIDWT may be employed to segment and recombine a speech waveform in a manner that is conceptually similar to PSOLA, although the mathematical details are of course quite different. Figures 8 and 9 illustrate the procedure for extracting a single formant pulse from the speech waveform. In Fig. 8 a dashed-line box delineates the region of the scalogram corresponding to the single formant pulse to be extracted, i.e., the time interval $0.133 = t_a < t < t_b = 0.1436$. The edges of the box are aligned with local minima of the sum of the quadratic envelopes of the 2- and 1-kHz bands. The scalogram elements inside this box are preserved, and the remainder of the scalogram is set to zero. Given the speech waveform $z(t)$ over the interval $t_0 \leqslant t \leqslant t_c$, and the scalogram $x(t,n) = T(z)$, a function of both time and scale, this modification produces $w(t,n)$ such that

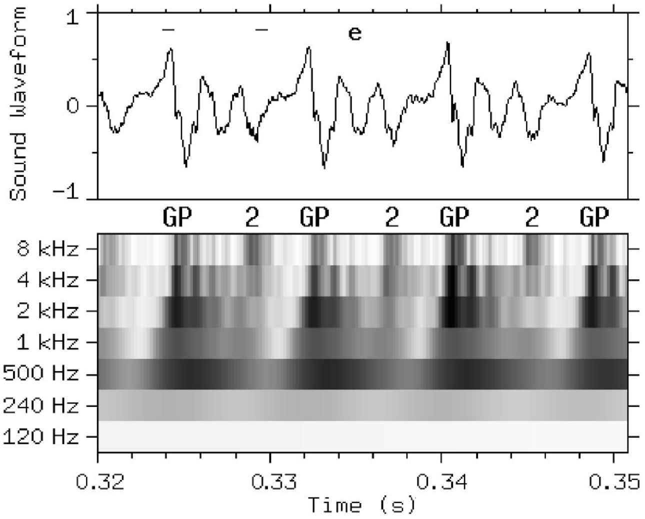

FIG. 7. The Japanese vowel "e," and its shift-invariant scalogram. The glottal pulses are indicated by "GP," with secondary pulses indicated by "2."

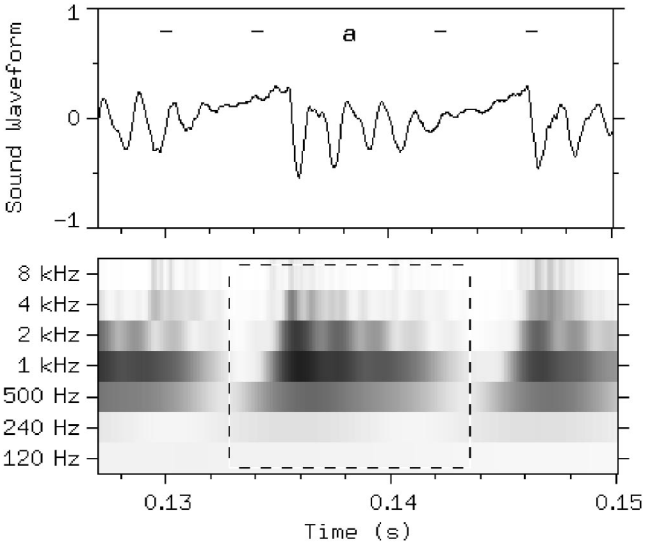

FIG. 8. A single formant pulse signature of the Japanese vowel "a" surrounded by a dashed-line box on the shift-invariant scalogram.

$$w(t,n) = \begin{cases} 0 & : \ t_0 \leqslant t < t_a \\ x(t,n) & : \ t_a \leqslant t \leqslant t_b \\ 0 & : \ t_b < t \leqslant t_c. \end{cases} \tag{11}$$

To complete the procedure, the extracted single formant pulse z' is obtained from $z' = S(x)$. Figure 9 shows z', with the original waveform z in gray for comparison. The scalogram $x' = T(z')$ is also shown in Fig. 9, along with the original dashed-line box. The only significant difference between w and x' is the smoothing of the boundaries at t_a and t_b.

To carry the feature extraction procedure described above to completion, the nonzero elements of a scalogram x are segmented into M pieces $(x'_1, x'_2, \ldots, x'_M)$ such that $\Sigma_n x'_n = x$. Then, by the linearity of S, $\Sigma_n z'_n = z$. The regions of the scalogram where features with strong time localization are absent are segmented at arbitrary time boundaries, with

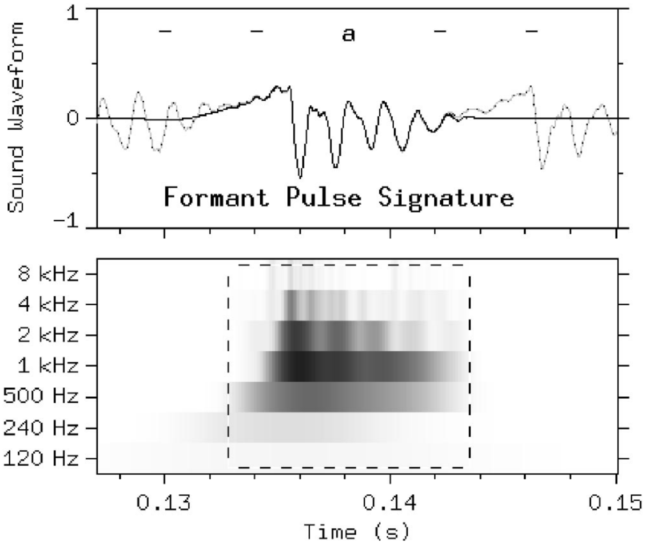

FIG. 9. A single formant pulse of the Japanese vowel "a" extracted by the ISIDWT, with the original waveform in gray for comparison. The shift-invariant scalogram of the reconstructed pulse is shown surrounded by a dashed-line box marking the extracted area.

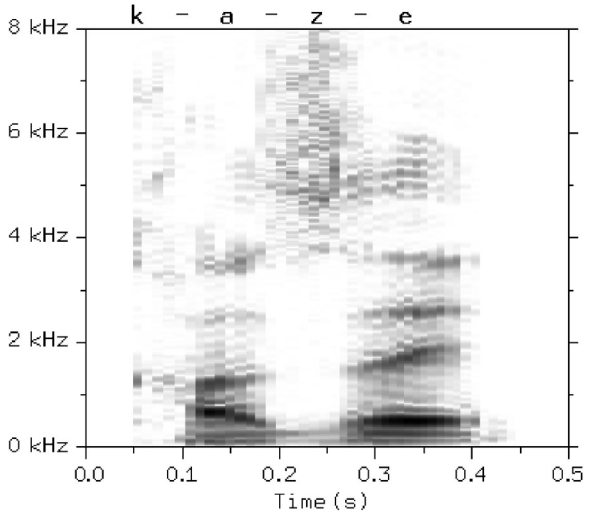

FIG. 10. A pitch-synchronous spectrogram-like display of the Japanese word "kaze."

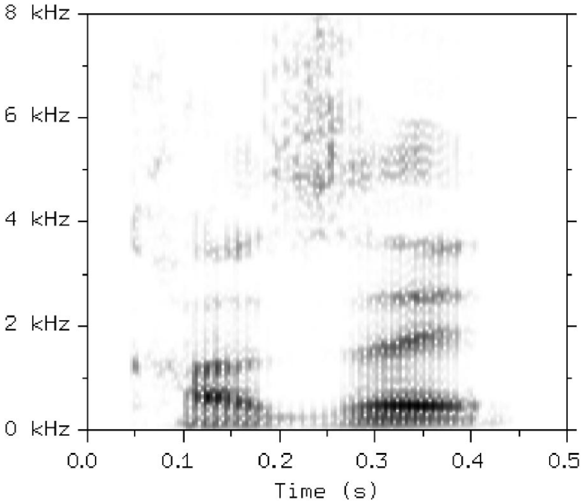

FIG. 11. The spectrogram of the Japanese word "kaze."

the constraint that the arbitrary segmentation lengths $t_b - t_a$ are similar to those used elsewhere in the scalogram. If the segmentation boundaries are aligned with instants of relative quiet, the operation may be considered a type of synchronous windowing, i.e., windowing synchronized with the amplitude modulation inherent in the waveform. In this example, most of the waveform exhibits strong amplitude modulation, and the width of each window $t_b - t_a$ is large relative to the smoothing at the boundaries observed in Fig. 9. Therefore, the time-domain overlap between adjacent segments is negligible, and

$$\sum_{n=1}^{M} F(z_n'(t)) \approx F(z'(t)), \tag{12}$$

even for some nonlinear $F(z)$ that are sufficiently local, e.g., quadratic envelope, or spectral density for frequencies $f > 1/(t_b - t_a)$.

To show how the frequency content of the extracted pulses evolves over time, a spectrogram-like display is presented in Fig. 10. The formant pulses obtained from the ISIDWT were zero padded to 1536-point length, and the energy spectral density of each pulse was calculated via the FFT with no further windowing. For comparison, a conventional spectrogram of the original kaze waveform is shown in Fig. 11. This and subsequent spectrograms were prepared with a 512-point Hanning window shifted in 134-point steps, and each windowed segment was zero padded to 1536-point length before calculating the FFT.

This waveform segmentation procedure is a unique and powerful capability of the SIDWT/ISIDWT. In addition to the analysis methods shown above, it has broad utility for copying, cutting, and pasting sections of sound waveforms. Working with the scalogram $x = T(z)$ rather than the sound waveform z has two advantages. First, for all but the simplest waveforms, it is usually easier to find and delineate features of interest in x. Second, cutting segments from x and joining them to make w doesn't result in audible clicks and pops. In conventional waveform editing, such clicks and pops are caused by steps at boundaries where the final value in one waveform segment differs from the initial value in the next waveform segment. The signature of such a boundary on the scalogram is a peak in the high-frequency scales, reflecting the high-frequency content of the step. It is usually necessary to taper or otherwise reshape the waveforms at each boundary to smooth these steps. However, excessive tapering or reshaping can create other audible artifacts, e.g., a gap in the high-frequency components. In some cases human intervention is required to find the optimal balance. Performing the cutting and joining operations on w creates steps in the scalogram values at boundaries that resemble the steps created on a raw edited sound waveform. However, since they are steps in w, not steps in z, they are not associated with audible clicks and pops. The pseudoinverse smoothes the waveform at the boundary in a way that the scalogram signature of the boundary on x' is a rounded step, as close as possible in the least-squares sense to the original sharp step. The pseudoinverse cannot create a click or a pop at the boundary, because that would require a peak on the scalogram x'. A peak at the step location implies scalogram values with magnitude greater than those on either side of the step, which would never be the solution that minimizes $\Sigma_n (x_n' - x_n)^2$. It should be noted that absence of editing artifacts is no guarantee of realism, since abrupt starts or transitions between sounds of different character may sound false or even unpleasant. Even so, the reliable and automatic prevention of editing artifacts in sound waveforms is a substantial convenience.

A common application of PSOLA is changing the rate of a spoken word without changing the pitch or the frequency content of the formants, i.e., to simulate the same speaker pronouncing the same word, but speaking more rapidly or more slowly. To illustrate the sound waveform editing capabilities of the SIDWT/ISIDWT, the rate of the example waveform kaze has been halved and doubled. The first step in the procedure is to synchronize the analysis with the glottal pulses whenever possible. The spectrogram-like display shown in Fig. 10 was created by identifying M time instants t_i. The t_i in the voiced regions were located at moments of relative quiet in the 1- and 2-kHz bands, and the t_i in the unvoiced regions were merely spaced at regular intervals.

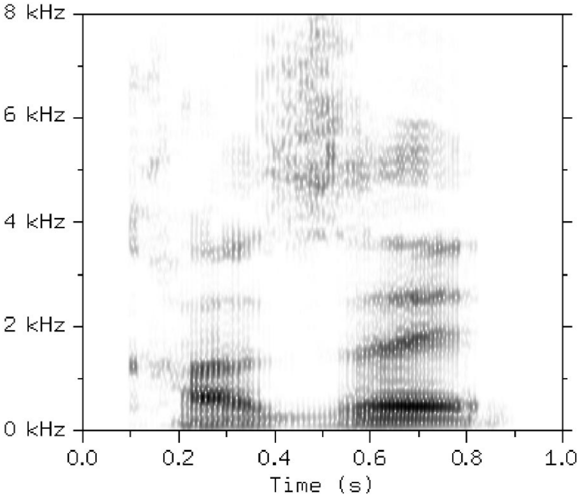

FIG. 12. The spectrogram of the Japanese word ''kaze'' rate-changed to half-speed.

The M time instants t_i were used as delimiters to segment the scalogram x into a set of $M-1$ pieces y_i. The segments y_i correspond to individual glottal pulses, similarly sized sections of unvoiced speech, or silence. The notation y_i omits scale levels for simplicity, but all scale levels are implicitly included. If all segments y_i are concatenated in the proper order, $(y_1, y_2, \ldots, y_{M-1}) = x$, the original scalogram is recovered. The speech rate was halved by simply duplicating each y_i in the proper sequence, $(y_1, y_1, y_2, y_2, \ldots, y_{M-1}, y_{M-1}) = w$. The half-rate waveform z' was then obtained by $S(w) = z'$. A conventional spectrogram of the half-rate waveform is shown in Fig. 12. The speech rate was doubled by concatenating the even-numbered segments $(y_2, y_4, y_6, \ldots) = x$, with $S(w) = z'$ as before. The even-numbered segments were chosen because they included the ''k'' sound; the doubled results from the odd-numbered segments sounded like ''aze.'' A conventional spectrogram of the double-rate waveform is shown in Fig. 13. Informal listening tests found that both examples were clear and understandable. The realism of the half-rate examples was marred slightly by a mild low-frequency artifact, but the realism of the double-rate example was excellent.

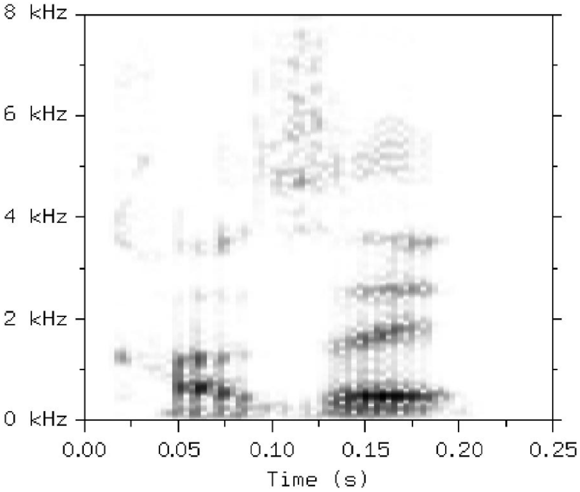

FIG. 13. The spectrogram of the Japanese word ''kaze'' rate-changed to double speed.

IV. CONCLUSIONS

Procedures for calculating and inverting the SIDWT are described above via equations (Sec. II A) and as a block diagram (Sec. II B). A self-contained proof that the ISIDWT is the pseudoinverse of the SIDWT is given in the Appendix. The new SIDWT/ISIDWT algorithm described here is mathematically equivalent to the stationary wavelet transform and the maximal overlap discrete wavelet transform, and is easier to implement efficiently in an array-oriented mathematical scripting language. The recorded speech example discussed in Sec. III demonstrates that the SIDWT is useful for visualization and analysis of complicated, nonstationary, acoustic waveforms. The SIDWT provides a clear picture of the sounds excited and modulated by the opening and closing of the glottis in speech. The SIDWT is complementary to the STFT for visualization, as they provide optimal views of different aspects of the waveform. The SIDWT and ISIDWT together provide the capability to segment and reconstruct the sound from each individual glottal period for further visualization and analysis. Examples of half-rate and double-rate speech modification demonstrate the potential for the SIDWT/ISIDWT to prove useful for applications that currently use one of the PSOLA family of techniques.

The visualization capabilities of the SIDWT are constrained by the properties of the underlying discrete wavelet transform. In particular, the single-octave bandwidth resulting from dyadic wavelet scaling means that the tonal content of sound waveforms cannot be observed in detail. In certain situations, some tonal information can be extracted by analysis of sum and difference frequencies. The relative lack of frequency-domain information provided by the SIDWT is not really a loss, however; it is a trade-off, which enables the SIDWT to provide more detailed time-domain information. For the same reason, segmentation and reconstruction operations provide more detailed control in the time domain than in the frequency domain. The most productive way to use the SIDWT/ISIDWT will likely prove to be in concert with the STFT and other frequency-domain methods, e.g., the pitch-synchronous spectrogram-like display presented in Sec. III. The SIDWT and ISIDWT are based on purely mathematical principles, and have no inherent connection to psychoacoustics, e.g., two sound waveforms that are ''close'' in the sense of human perception of sounds are not guaranteed to be close in the sense of the standard Euclidean norm in redundant wavelet coefficient space. The applicability of the SIDWT/ISIDWT to acoustics can only be judged empirically. The experience to date, while subjective, and limited in scope to the field of automotive sound quality and the two rate-changing examples presented here, has been consistently encouraging. Possible speech-related applications include aspects of automated speech recognition and simplified intonation/formant visualization. Editing sound recordings (cutting and pasting sections) with the SIDWT/ISIDWT is especially convenient since the audible clicks and pops produced at section boundaries by simple time-domain editing procedures are automatically prevented. The rate-modified speech examples presented here show promise, and the SIDWT/ISIDWT may prove useful for other applications that currently employ PSOLA-based methods. Given the broad scope of research in acoustics, there are many potential applications for the SIDWT/ISIDWT in visualization, manipulation, and analysis of nonstationary sound waveforms.

ACKNOWLEDGMENTS

We would like to thank Professor Alan Haddow and Professor William Hartmann of Michigan State University for helpful discussions. We are grateful to Professor Charles MacCluer from the Industrial Mathematics Program at Michigan State University for creating the connections that led to this work.

APPENDIX: PROOF OF THEOREM 1

Our strategy is to prove that T as defined in Sec. II A is an isometry and that S is the adjoint of T. Theorem 1 then follows from a well-known linear algebra fact. According to the definition of T in Sec. II A, we define the mappings

$$T_1 : l^2(\mathbb{Z}_N) \ni z \mapsto (x_1, y_1) \in l^2(\mathbb{Z}_{2N}), \tag{A1}$$

and for each stage $j=2,...,p$

$$T_j : l^2(\mathbb{Z}_{jN}) \ni (x_1,\ldots,x_{j-1},y_{j-1}) \mapsto (x_1,\ldots,x_{j-1},x_j,y_j) \in l^2(\mathbb{Z}_{(j+1)N}). \tag{A2}$$

Additionally, let $R: l^2(\mathbb{Z}_{(p+1)N}) \mapsto l^2(\mathbb{Z}_{(p+1)N})$ be the reordering given by

$$R(x_1,x_2,\ldots,x_p,y_p) = (R_{N/2}(x_1), R_{N/4}(x_2),\ldots, R_{N/2^p}(x_p), R_{N/2^p}(y_p)). \tag{A3}$$

Note that the p-stage SIDWT of $z \in l^2(\mathbb{Z}_N)$ is then given by $T(z) = RT_p \cdots T_2 T_1(z)$. Since convolution and reordering are linear maps, each T_j is linear. Therefore, T is a linear transformation.

We will now show that T is an isometry, i.e., that $\langle Tz_1, Tz_2 \rangle = \langle z_1, z_2 \rangle$ for all $z_1, z_2 \in l^2(\mathbb{Z}_N)$. To avoid considering the case j=1 separately, in the following let $(x_1, x_0, y_0) = y_0 = z$ and $(x_1, x_0) = 0$.

Theorem 3. The transform T is an isometry.

Proof: We show that for each $j=1,...,p$ the mapping T_j is an isometry. Since the reordering operator R is unitary, we have $\langle Rw_1, Rw_2 \rangle = \langle w_1, w_2 \rangle$ for all $w_1,\ w_2 \in l^2(\mathbb{Z}_{(p+1)N})$. As a composition of isometries, T is then an isometry.

Let $j \in \{1,...,p\}$, $z_1 = (x_1,\ldots,x_{j-1},y_{j-1})$ and $z_2 = (\xi_1,\ldots,\xi_{j-1},\eta_{j-1})$. Then

$$\begin{aligned}\langle T_j z_1, T_j z_2 \rangle &= \left\langle \left(x_1,\ldots,x_{j-1}, \frac{1}{\sqrt{2}} R_2(y_{j-1} * \tilde{v}), \frac{1}{\sqrt{2}} R_2(y_{j-1} * \tilde{u}) \right), \right. \\ &\quad \left. \left(\xi_1,\ldots,\xi_{j-1}, \frac{1}{\sqrt{2}} R_2(\eta_{j-1} * \tilde{v}), \frac{1}{\sqrt{2}} R_2(\eta_{j-1} * \tilde{u}) \right) \right\rangle \\ &= \langle (x_1,\ldots,x_{j-1}), (\xi_1,\ldots,\xi_{j-1}) \rangle \\ &\quad + \left\langle \frac{1}{\sqrt{2}} R_2(y_{j-1} * \tilde{v}), \frac{1}{\sqrt{2}} R_2(\eta_{j-1} * \tilde{v}) \right\rangle \\ &\quad + \left\langle \frac{1}{\sqrt{2}} R_2(y_{j-1} * \tilde{u}), \frac{1}{\sqrt{2}} R_2(\eta_{j-1} * \tilde{u}) \right\rangle. \end{aligned} \tag{A4}$$

Using the fact that R_2 is unitary, and applying Parseval's relation, we get

$$\begin{aligned}&\left\langle \frac{1}{\sqrt{2}} R_2(y_{j-1} * \tilde{v}), \frac{1}{\sqrt{2}} R_2(\eta_{j-1} * \tilde{v}) \right\rangle \\ &= \frac{1}{2} \langle y_{j-1} * \tilde{v}, \eta_{j-1} * \tilde{v} \rangle \\ &= \frac{1}{2N} \sum_{n=0}^{N-1} \langle (\widehat{y_{j-1} * \tilde{v}})(n), (\widehat{\eta_{j-1} * \tilde{v}})(n) \rangle \\ &= \frac{1}{2N} \sum_{n=0}^{N-1} \langle \hat{y}_{j-1}(n) \hat{\tilde{v}}(n), \hat{\eta}_{j-1}(n) \hat{\tilde{v}}(n) \rangle \\ &= \frac{1}{2N} \sum_{n=0}^{N-1} |\hat{\tilde{v}}(n)|^2 \langle \hat{y}_{j-1}(n), \hat{\eta}_{j-1}(n) \rangle. \end{aligned} \tag{A5}$$

The same equality holds for $\tilde{u}$ instead of $\tilde{v}$. Making use of the known identity $\hat{\tilde{v}}(n) = \hat{\nu}^*(n)$, we get

$$|\hat{\tilde{v}}(n)|^2 + |\hat{\tilde{u}}(n)|^2 = |\hat{v}(n)|^2 + |\hat{u}(n)|^2 = 2, \tag{A6}$$

since the system matrix $A(n)$ is unitary for all $n=0,...,N-1$. This finishes the proof as follows:

$$\begin{aligned}\langle T_j z_1, T_j z_2 \rangle &= \langle (x_1,\ldots,x_{j-1}), (\xi_1,\ldots,\xi_{j-1}) \rangle \\ &\quad + \frac{1}{N} \sum_{n=0}^{N-1} \langle \hat{y}_{j-1}(n), \hat{\eta}_{j-1}(n) \rangle \\ &= \langle (x_1,\ldots,x_{j-1}), (\xi_1,\ldots,\xi_{j-1}) \rangle \\ &\quad + \frac{1}{N} \langle \hat{y}_{j-1}, \hat{\eta}_{j-1} \rangle \\ &= \langle (x_1,\ldots,x_{j-1}), (\xi_1,\ldots,\xi_{j-1}) \rangle \\ &\quad + \langle y_{j-1}, \eta_{j-1} \rangle \\ &= \langle (x_1,\ldots,x_{j-1},y_{j-1}), (\xi_1,\ldots,\xi_{j-1},\eta_{j-1}) \rangle \\ &= \langle z_1, z_2 \rangle . \Box \end{aligned} \tag{A7}$$

Given a linear mapping $L: l^2(\mathbb{Z}_m) \to l^2(\mathbb{Z}_k)$, $m,k \in \mathbb{N}$, the *adjoint operator*

$$L^\dagger : l^2(\mathbb{Z}_k) \to l^2(\mathbb{Z}_m), \tag{A8}$$

is given by the unique mapping defined by the property $\langle Lz, w \rangle = \langle z, L^\dagger w \rangle$ for all $z \in l^2(\mathbb{Z}_m)$, $w \in l^2(\mathbb{Z}_k)$. The matrix corresponding to $L^\dagger$ is just the conjugate transpose of the matrix corresponding to L.

We will show that the ISIDWT S is the adjoint operator of T. To do so, we first prove a lemma.

Lemma. Let $x,y,v \in l^2(\mathbb{Z}_N)$. Then

$$\langle y * \tilde{v}, x \rangle = \langle y, x * v \rangle. \tag{A9}$$

Proof: Using Parseval's equality and again the identity $\hat{\tilde{v}}(n)=\hat{v}^*(n)$, we get

$$\langle y*\tilde{v},x\rangle=\frac{1}{N}\sum_{n=0}^{N-1}\hat{y}(n)\hat{\tilde{v}}(n)\hat{x}^*(n)$$
$$=\frac{1}{N}\sum_{n=0}^{N-1}\hat{y}(n)\hat{v}^*(n)\hat{x}^*(n)=\langle y,x*v\rangle, \qquad \text{(A10)}$$

which proves that the adjoint of the convolution with v is the convolution with the conjugate reflection $\tilde{v}$. □

Following the definition of S in Sec. II A, we define mappings S_j as follows: Let $w=(w_1,w_2,\ldots,w_{j+1}) \in l^2(\mathbb{Z}_{(j+1)N})$, where $w_i\in l^2(\mathbb{Z}_N)$ for all $i=1,...,j+1$. Then, for all $j=1,...,p$, we define

$$S_j: l^2(\mathbb{Z}_{(j+1)N})\ni w\mapsto(w_1,\ldots,w_{j-1},\eta_j)\in l^2(\mathbb{Z}_{jN}), \qquad \text{(A11)}$$

where

$$\eta_j=\frac{1}{\sqrt{2}}(R_{N/2}(w_{j+1})*u+R_{N/2}(w_j)*v). \qquad \text{(A12)}$$

Noting that $S=S_1S_2\cdots S_pR^{-1}$, where R^{-1} is the inverse of the reordering operator as defined in Eq. (A3), we can now prove the following theorem.

Theorem 4. $S=T^\dagger$.

Proof: Since $T=RT_p\cdots T_2T_1$, it follows from the definition of the adjoint that $T^\dagger=T_1^\dagger T_2^\dagger\cdots T_p^\dagger R^\dagger =T_1^\dagger T_2^\dagger\cdots T_p^\dagger R^{-1}$, where we made use of the fact that the adjoint of the unitary operator R is its inverse R^{-1}. Hence, the theorem follows once we show that $T_j^\dagger=S_j$ for all $j=1,...,p$.

We need to prove that $\langle T_jz,w\rangle=\langle z,S_jw\rangle$ for all $z\in l^2(\mathbb{Z}_{jN})$ and $w\in l^2(\mathbb{Z}_{(j+1)N})$. Let $z =(x_1,x_2,\ldots,x_{j-1},y_{j-1})$ and $w=(w_1,w_2,\ldots,w_{j+1})$. Then

$$\langle T_jz,w\rangle=\left\langle\left(x_1,x_2,\ldots,x_{j-1},\frac{1}{\sqrt{2}}R_2(y_{j-1}*\tilde{v}),\right.\right.$$
$$\left.\left.\frac{1}{\sqrt{2}}R_2(y_{j-1}*\tilde{u})\right),(w_1,w_2,\ldots,w_{j+1})\right\rangle$$
$$=\langle(x_1,x_2,\ldots,x_{j-1}),(w_1,w_2,\ldots,w_{j-1})\rangle$$
$$+\left\langle\frac{1}{\sqrt{2}}R_2(y_{j-1}*\tilde{v}),w_j\right\rangle$$
$$+\left\langle\frac{1}{\sqrt{2}}R_2(y_{j-1}*\tilde{u}),w_{j+1}\right\rangle. \qquad \text{(A13)}$$

Note that by the lemma

$$\left\langle\frac{1}{\sqrt{2}}R_2(y_{j-1}*\tilde{v}),w_j\right\rangle=\left\langle y_{j-1}*\tilde{v},\frac{1}{\sqrt{2}}R_{\frac{N}{2}}(w_j)\right\rangle$$
$$=\left\langle y_{j-1},\frac{1}{\sqrt{2}}R_{\frac{N}{2}}(w_j)*v\right\rangle. \qquad \text{(A14)}$$

The same equation holds for u instead of v. So, we conclude that

$$\langle T_jz,w\rangle=\langle(x_1,x_2,\ldots,x_{j-1}),(w_1,w_2,\ldots,w_{j-1})\rangle$$
$$+\left\langle y_{j-1},\frac{1}{\sqrt{2}}(R_{\frac{N}{2}}(w_j)*v+R_{\frac{N}{2}}(w_{j+1})*u)\right\rangle$$
$$=\left\langle(x_1,x_2,\ldots,x_{j-1},y_{j-1}),\left(w_1,w_2,\ldots,w_{j-1},\right.\right.$$
$$\left.\left.\frac{1}{\sqrt{2}}(R_{\frac{N}{2}}(w_j)*v+R_{\frac{N}{2}}(w_{j+1})*u)\right)\right\rangle$$
$$=\langle z,S_j\,w\rangle.\square \qquad \text{(A15)}$$

We have shown that S is given by $T^\dagger$. It is well known that a 1–1 linear map L has a unique pseudoinverse given by $(L^\dagger L)^{-1}L^\dagger$. Here, T is not only 1–1 but also an isometry, which is equivalent to $T^\dagger T=id\big|_{l^2(\mathbb{Z}_N)}$. Thus, the pseudoinverse of T is given by $T^\dagger=S$, which proves Theorem 1.

Instead of u and v in the definition of S, we can use any a and b in $l^2(\mathbb{Z}_N)$, satisfying the condition

$$\hat{a}(n)\hat{u}^*(n)+\hat{b}(n)\hat{v}^*(n)=2, \qquad \text{(A16)}$$

for all $n=0,...,N-1$ to obtain a mapping $\tilde{S}$ which is still an inverse of T on the image of T. Since $\tilde{S}$ is computed using convolutions in the same way as S, the number of multiplications required to compute $\tilde{S}$ is the same as for S. However, since then $\tilde{S}^\dagger\neq T$, $\tilde{S}$ is no longer the pseudoinverse, i.e., we lose the notion of ''closeness'' in the least-squares sense.

[1] G. Beylkin, ''On the representation of operators in bases of compactly supported wavelets,'' SIAM (Soc. Ind. Appl. Math.) J. Numer. Anal. **29**, 1716–1740 (1992).

[2] R. R. Coifman and D. L. Donoho, ''Translation-invariant de-noising,'' in Wavelets and Statistics, Lecture Notes in Statistics, **103**, edited by A. Antoniadis and G. Oppenheim (Springer, New York, 1995), pp. 125–150.

[3] H. Sari-Sarraf and D. Brzakovic, ''A shift-invariant discrete wavelet transform,'' IEEE Trans. Signal Process. **45**, 2621–2626 (1997).

[4] J. Liang and T. W. Parks, ''A translation-invariant wavelet representation algorithm with applications,'' IEEE Trans. Signal Process. **44**, 225–232 (1996).

[5] D. J. Scholl, ''Translation-invariant data visualization with orthogonal discrete wavelets,'' IEEE Trans. Signal Process. **46**, 2031–2034 (1998).

[6] G. P. Nason and B. W. Silverman, ''The stationary wavelet transform and some statistical applications,'' in *Wavelets and Statistics, Lecture Notes in Statistics*, **103**, edited by A. Antoniadis and G. Oppenheim (Springer, New York, 1995), pp. 281–299.

[7] D. B. Percival and A. T. Walden, *Wavelet Methods for Time Series Analysis* (Cambridge University Press, Cambridge, UK, 2000), pp. 159–205.

[8] D. Scholl and B. Yang, ''Wavelet-based visualization, separation, and synthesis tools for sound quality of impulsive noises,'' SAE 2003-01-1527 (The Society of Automotive Engineers, Warrendale, PA, 2003); ⟨http://www.sae.org⟩

[9] C. Duxbury, M. Davies, and M. Sandler, ''Separation of transient information in musical audio using multiresolution analysis techniques,'' in Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01) Limerick, Ireland (2001).

[10] R. Kronland-Martinet, ''The wavelet transform for analysis, synthesis, and processing of speech and musical sounds,'' Comput. Music J. **12**, 11–20 (1988).

[11] D. Darlington, L. Daudet, and M. Sandler, ''Digital audio effects in the wavelet domain,'' in Proceedings of the 5th Int. Conference on Digital Audio Effects (DAFX-02), Hamburg, Germany (2002).

[12] Y. M. Cheng and D. O'Shaughnessy, ''Automatic and reliable estimation

of glottal closure instant and period,'' IEEE Trans. Acoust., Speech, Signal Process. **37**, 1805–1815 (1989).

[13] S. Kadambe and G. F. Boudreaux-Bartels, ''Application of the wavelet transform for pitch detection of speech signals,'' IEEE Trans. Inf. Theory **38**, 917–924 (1992).

[14] V. N. Tuan and C. d'Alessandro, ''Robust glottal closure detection using the wavelet transform,'' in Proceedings of the European Conference on Speech Technology, EuroSpeech, Budapest, 2805–2808 (1999).

[15] I. Hetrick and H. Ackermann, ''A vowel synthesizer based on formant sinusoids modulated by fundamental frequency,'' J. Acoust. Soc. Am. **106**, 2988–2990 (1999).

[16] P. J. B. Jackson and C. H. Shadle, ''Frication noise modulated by voicing, as revealed by pitch-scaled decomposition,'' J. Acoust. Soc. Am. **108**, 1421–1434 (2000).

[17] A. L. Francis, V. Ciocca, and J. M. C. Yu, ''Accuracy and variability of acoustic measures of voicing onset,'' J. Acoust. Soc. Am. **113**, 1025–1032 (2003).

[18] H. M. Hanson and E. S. Chuang, ''Glottal characteristics of male speakers: Acoustic correlates and comparison with female data,'' J. Acoust. Soc. Am. **106**, 1064–1077 (1999).

[19] J. C. Lucero and L. L. Koenig, ''Time normalization of voice signals using functional data analysis,'' J. Acoust. Soc. Am. **108**, 1408–1420 (2000).

[20] M. W. Frazier, *An Introduction to Wavelets through Linear Algebra* (Springer, New York, 1999).

[21] International Phonetic Association (c/o Department of Linguistics, University of Victoria, Victoria, British Columbia, Canada); ⟨http://web.uvic.ca/ling/resources/ipa/handbook.htm⟩

[22] D. J. Scholl, ''Wavelet-based visualization of impulsive and transient sounds in stationary background noise,'' SAE 2001-01-1475 (The Society of Automotive Engineers, Warrendale, PA, 2001); ⟨http://www.sae.org⟩

[23] P. Mokhtari and K. Tanaka, ''A corpus of Japanese vowel formant patterns,'' Bull. of the Electrotechnical Lab. (Electrotechnical Laboratory, Agency of Industrial Science and Technology, Ministry of International Trade and Industry, Japan), **64**, 57–66 (2000); ⟨www.etl.go.jp/jp/results/bulletin/pdf/64-rinji/05tanaka.pdf⟩

[24] E. Moulines and J. Laroche, ''Nonparametric techniques for pitch-scale and time-scale modification of speech,'' Speech Commun. **16**, 175–205 (1995).

[25] M. Holzapfel, R. Hoffmann, and H. Höge, ''A wavelet-domain PSOLA approach,'' Third ESCA/COCOSDA Workshop on Speech Synthesis, Blue Mountains, NSW, Australia, pp. 283–286 (1998).

[26] R. W. L. Kortekaas and A. Kohlrausch, ''Psychoacoustical evaluation of the pitch-synchronous overlap-and-add speech-waveform manipulation technique using single-formant stimuli,'' J. Acoust. Soc. Am. **101**, 2202–2213 (1997).

[27] R. W. L. Kortekaas and A. Kohlrausch, ''Psychoacoustical evaluation of PSOLA. II. Double-formant stimuli and the role of vocal perturbation,'' J. Acoust. Soc. Am. **105**, 522–535 (1997).