The Correlogram: a visual display of periodicity

Svante Granqvist and Britta Hammarberg***

* *Dept of Speech, Music and Hearing, KTH, Stockholm;*
Electronic mail: svante.granqvist@speech.kth.se

** *Dept of Logopedics and Phoniatrics, Karolinska Institute, Huddinge University Hospital;*
Electronic mail: britta.hammarberg@klinvet.ki.se

Abstract

Fundamental frequency (F0) extraction is often used in voice quality analysis. In pathological voices with a high degree of instability in F0, it is common for F0 extraction algorithms to fail. In such cases, the faulty F0 values might spoil the possibilities for further data analysis. This paper presents the correlogram, a new method of displaying periodicity. The correlogram is based on the waveform matching techniques often used in F0 extraction programs, but with no mechanism to select an actual F0 value. Instead, several candidates for F0 are shown as dark bands. The result is presented as a 3D-plot with time on the x-axis, correlation delay inverted to frequency on the y-axis and correlation on the z-axis. The z-axis is represented in a gray scale as in a spectrogram. Delays corresponding to integer multiples of the period time will receive high correlation, thus resulting in candidates at F0, F0/2, F0/3 etc. While the corrlogram adds little to F0 analysis of normal voices, it is useful for analysis of pathological voices since it illustrates the full complexity of the periodicity in the voice signal. Also, in combination with manual tracing, the correlogram can be used for semi-manual F0 extraction. If so, F0 extraction can be performed on many voices that cause problems for conventional F0 extractors. To demonstrate the properties of the method it is applied to synthetic and natural voices, among them six pathological voices, which are characterized by roughness, vocal fry, gratings/scrape, hypofunctional breathiness and voice breaks, or combinations of these.

Introduction

Fundamental frequency (F_0) is a commonly used parameter being the main acoustic correlate to perceived pitch. In the field of voice quality research, F_0 extraction is particularly relevant, for example for evaluation of pre- and post-treatment of voice disorders and for measuring F_0 perturbation. Fundamental frequency extraction has received a great deal of attention in speech and voice research. Several different algorithms have been invented (e.g. Hess, 1983; Titze & Liang, 1993; Hess, 1995), and the algorithms have been applied both to acoustic waveforms and to electroglottographic (EGG) signals (Rothenberg 1973, Fourcin 1986). Examples of such methods are peak picking and methods based on spectral or cepstral properties of the signal or on waveform matching by means of autocorrelation or autodifference (e.g., Hess, 1983).

The waveform matching technique has many important advantages. For example, it is independent on determination of the instant of excitation and has a low sensitivity to noise (Titze & Liang, 1993). Also, it can offer several estimates of fundamental frequency per period. The basic idea of the waveform matching technique is to compare the signal in two time windows separated by a variable time delay. Certain lengths of this time delay will achieve a high correlation. These delays correspond to multiples of the period time. For example the comparison can be realized in terms of a correlation function, which is a straightforward procedure since few variables are involved. If the waveform matching technique is to be used for F_0 extraction, the F_0 extraction algorithm must select which peak in the correlation function corresponds to the fundamental period time. Normal voices rarely cause selection problems. However, for dysphonic voices with

an unstable period time, different F_0 extraction algorithms will give different results (Rabinov et al., 1995; Karnell et al., 1991). Phonation containing *bicyclic* segments (equivalent to *period doubling*, Titze, 1995) is a typical example; most F_0 extractors will select a fundamental frequency of $F_0/2$ for the bicyclic segments. This is in some sense correct since the period actually is doubled. However, different algorithms require different magnitudes of bicyclicity in order to arrive at this result. Hence different F_0 extraction programs yield different results. This is problematic when the extracted F_0 data are used for deriving perturbation measures, such as jitter. For example, in a study of vocal fry, Blomgren et al. (1998) reverted to a semi-manual extraction method instead of using the automatic methods, since their voice samples were characterized by a high amount of variability.

The problems outlined above do not originate from the waveform-matching algorithm but rather from the selection mechanism. Because fundamental frequency is defined as the inverse of period time, no F_0 exists if the signal is not perfectly repetitive, strictly speaking. Therefore, especially for pathological voices, there will be cases when the task of extracting a single F_0 is ill-defined or unrealistic. In such cases, an improved description of the perturbation itself may be more relevant.

In this paper, we present a display showing the raw correlation functions in a three-dimensional graph. We propose the term *correlogram* for these displays. The result is a picture reflecting periodicity characteristics of a voice rather than an extracted F_0 curve. The correlogram is free from a selection mechanism, leaving to the user to select the F_0 value or to what extent F_0 extraction is at all appropriate. This type of display should be particularly useful for voices where F_0 selections are difficult, that is, all voices with a high amount of F_0 perturbation. The method is tested on synthetic signals and natural voices and compared with other methods.

Descriptions of selected perceptual voice terms

In the case of pathological voices, information about periodicity or lack of periodicity is particularly relevant. Many attempts have been made to correlate F_0 perturbation characteristics with perceptual features. Such correlates are interesting, since a complete understanding of the relationship between perception and acoustics would allow objective measurements of voice qualities. In the following, some perceptual terms frequently used for pathological voices are reviewed together with the typically associated F_0 extraction problems.

Creaky voice, *vocal fry* and *pulsed phonation.* These terms appear to be associated either with low pitch and a prolonged glottal closed phase or by a complex pattern of glottal excitations, giving rise to subharmonics (Titze, 1995; Laver, 1980; Ladefoged, 1988; Hammarberg & Gauffin, 1995).

Roughness. This term also appears to be associated with period time perturbation. However the term appears mostly, but not always, to be linked to a more random perturbation than what is commonly associated with the multi-cyclic type of vocal fry. The term is also sometimes associated with low-frequency noise (Hammarberg & Gauffin, 1995; De Krom, 1995; Titze, 1995; Ishiki et al., 1969; Hillenbrand, 1988; Omori et al., 1997; Imaizumi 1986).

Gratings/scrape is a term mainly used in Sweden (Swedish: *skrap*). The term is often translated to "high-frequency roughness" (Hammarberg & Gauffin, 1995).

Breathiness is caused by soft or incomplete closure of the glottis and is often associated with high-frequency noise. Breathy voice can be produced in both hypo- and hyperfunctional laryngeal settings, which give rise to two different types of breathiness. These modes of phonation correlate strongly to a high or low relative level of the fundamental, respectively (Titze, 1995; Hammarberg & Gauffin, 1995; Hammarberg, 1986).

Voice breaks, vocal breaks or *register breaks* occurs when the vocal folds suddenly switches from one mode of vibration to another, for example between modal and falsetto register (Sundberg, 1987; Hammarberg, 1986; Švec & Pešák, 1994).

Most of the above voice qualities present problems for F_0 extraction. The complex-patterns of glottal excitation that often are associated with vocal fry or gratings/scrape, typically cause octave leaps in the F_0 curve, while pulsed phonation in principle can produce a smooth, continuous curve. Roughness, when associated with a random distribution of period time, mostly generates an unstable F_0 curve, but also, different F_0 algorithms tend to yield

different F_0 values. Most programs generally handle breathy voices successfully since they contain little F_0 perturbation, but for certain algorithms, particularly if event-based, a high noise level can cause errors. F_0 extractors generally succeed in tracking the F_0 in voice breaks. All these problems with F_0 extraction in pathological voices are unfortunate since the F_0 perturbation appears to represent an important characteristic of such voices (Hillenbrand, 1988; Gauffin et al, 1995). Hence, alternative methods for displaying periodicity variation should be useful in the analysis of pathological voices.

Method

The correlogram is based on the correlation between two time windows of the signal (Figure 1). It displays the correlation in a novel manner in terms of a graph showing several such correlation functions, displayed in a gray scale similar to the Fourier transforms in a spectrogram. The method has been implemented by the first author (SG) as a program module of the Soundswell Signal Workstation software (Hitech Development AB, Sweden).

Different waveform matching functions can be used. In this paper we have selected to use the Pearson correlation coefficient:

$$r_{m,n} = \frac{\sum_{k=m}^{m+w-1}(x_k - \overline{x_k}) \cdot (x_{k+n} - \overline{x_{k+n}})}{\sqrt{\sum_{k=m}^{m+w-1}(x_k - \overline{x_k})^2 \cdot \sum_{k=m}^{m+w-1}(x_{k+n} - \overline{x_{k+n}})^2}}$$

where

$$\overline{x_i} = \frac{\sum_{i=m}^{m+w-1} x_i}{w}$$

[1]

or, in computational form

$$r_{m,n} = \frac{w\sum_{k=m}^{m+w-1} x_k x_{k+n} - \sum_{k=m}^{m+w-1} x_k \cdot \sum_{k=m}^{m+w-1} x_{k+n}}{\sqrt{\left(w\sum_{k=m}^{m+w-1} x_k^2 - \left(\sum_{k=m}^{m+w-1} x_k\right)^2\right) \cdot \left(w\sum_{k=m}^{m+w-1} x_{k+n}^2 - \left(\sum_{k=m}^{m+w-1} x_{k+n}\right)^2\right)}} \qquad [2]$$

where x_j is the j:th sample, m is the starting sample number, n is the delay separating the starting points of the windows and w is the window width. This function is normalized, so the result will be restricted to the range $-1 \le r_{m,n} \le 1$. When the delay n corresponds to one or several fundamental periods, a maximum will occur in the correlation coefficient. This is true, regardless of where in the fundamental period the starting point of the window is located, so there is no need to determine the point of excitation.

Note that the use of the Pearson correlation coefficient rather than a simple cross-correlation is advantageous since only the Pearson alternative is insensitive to a DC component of the signal. A DC component adds an increase of the cross-correlation, since the voice signal becomes relatively smaller, as the DC component increases. Such an increase would be irrelevant in periodicity analysis.

For each time value along the x-axis a correlation coefficient is calculated with a starting sample m corresponding to the time coordinate of the x-axis. This correlation coefficient is calculated for different delays n, along the y-axis. The correlation, $r_{m,n}$ in this point, is displayed along a gray scale, with black corresponding to $r_{m,n}=1$ and white for $r_{m,n}\le 0$. If the signal is perfectly periodic with a period time T_0, the correlogram will show a set of horizontal black bands, representing different *candidates* for the fundamental period time C_1, C_2, C_3…, at delays n corresponding to T_0, $2T_0$, $3T_0$…. (Figure 1, middle panel).

Correlograms can also be presented with an inverted y-axis, thus showing frequency rather than time. In this case F_0 candidates, C_1, C_2, C_3… will appear at F_0, $F_0/2$, $F_0/3$ and so on (Figure 1, lower panel). Both these representations have certain advantages. In a time correlogram, the candidates C_n appear as horizontal stripes that are equidistant. It also shows more salient high-order candidates. In a

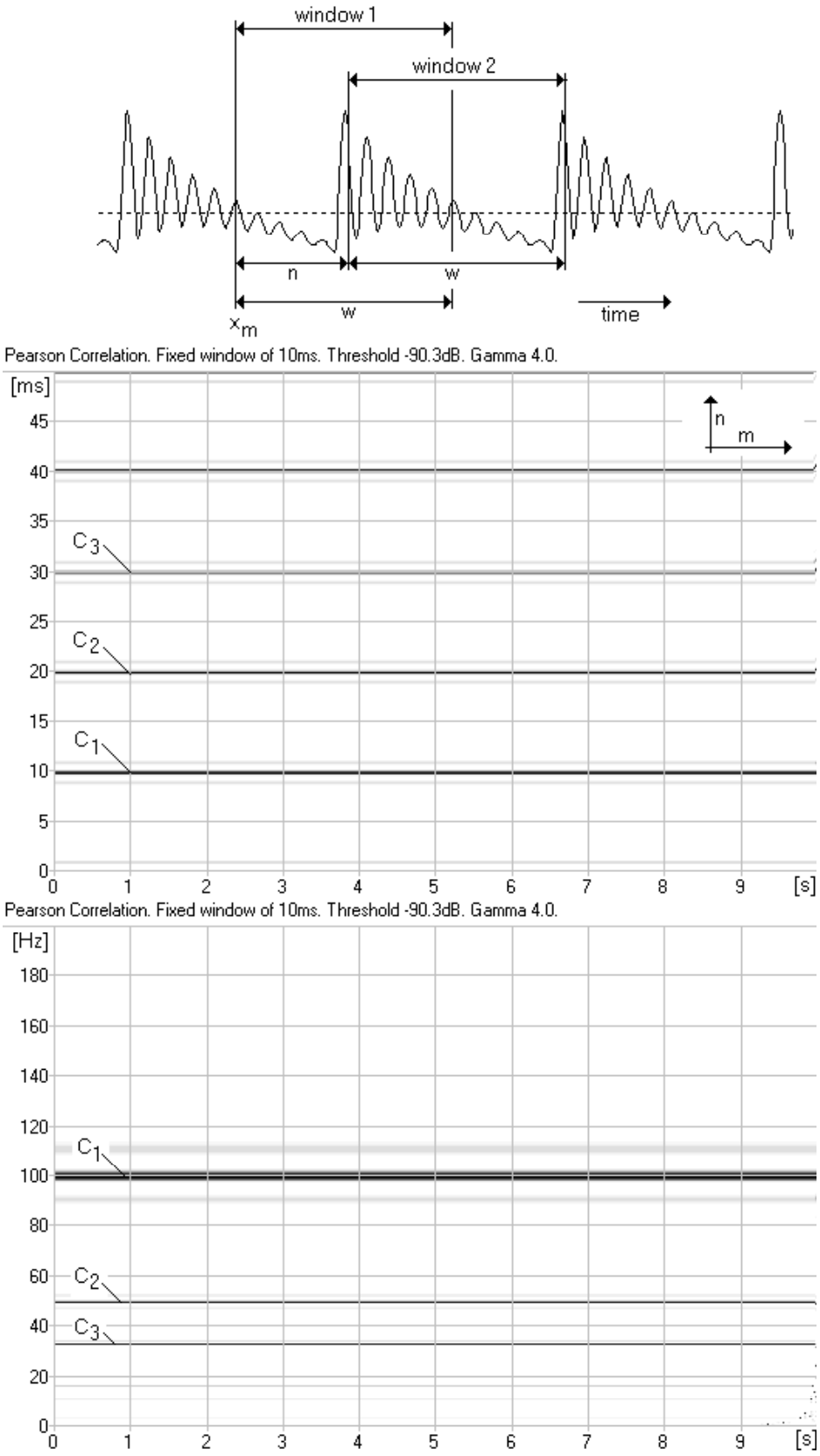

Figure 1. The upper panel shows how the two correlation windows are applied to a perfectly periodic signal (100 Hz sawtooth wave, formant filter at 1000 Hz, bandwidth 100 Hz). Middle and lower panels show the resulting time and frequency correlogram.

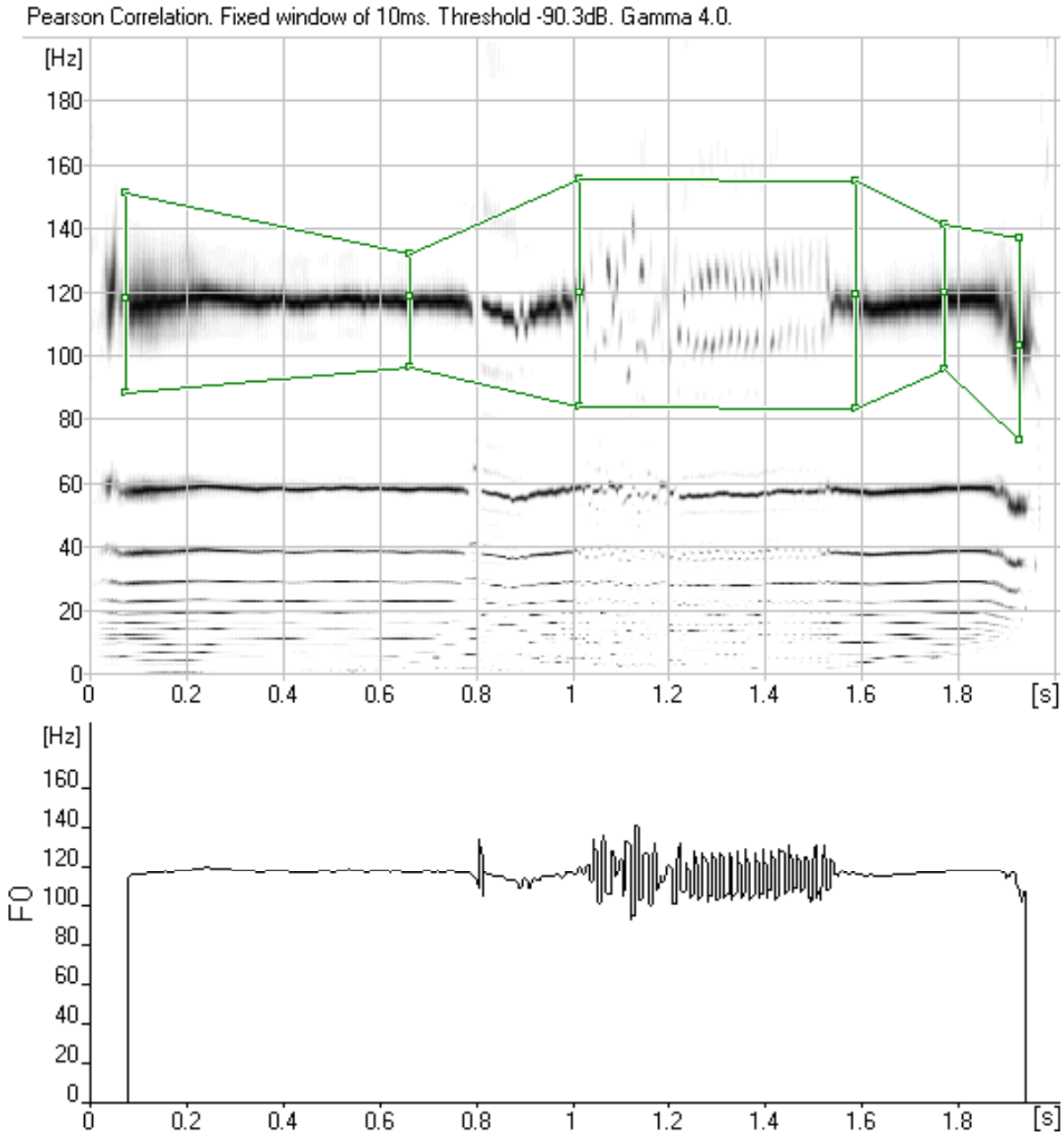

Figure 2. The correlogram used for semi-manual F_0 extraction. The first candidate was traced manually (upper panel) and the computer program then extracted the F_0 value within the traces that corresponded to the highest correlation (lower panel). Note the absence of octave leaps during the bicyclic segment at 1-1.5 s.

frequency correlogram, the strip density is greater at lower than at higher frequencies, but on the other hand, it is more clearly related to pitch. The selection of delay representation is a matter of choice, but since periodicity is mostly expressed as fundamental frequency, rather than fundamental period time, the frequency representation seems intuitively more appropriate.

The length of the correlation window, *w*, will affect the appearance of the correlogram. A shorter window gives better resolution in time, but may also show the first formant in terms of *side bands* surrounding each candidate. Normally, a frame length of about one fundamental period is appropriate. An interesting possibility is to let the window length vary as the correlation delay varies, such that the window length is equal to the delay. With this procedure the window length is automatically adjusted to an appropriate value as fundamental frequency varies. The procedure is computationally less efficient, however, especially if

correlations are calculated for frequencies close to 0 Hz.

For practical reasons, the correlograms show *r* raised to the power of γ for values of r>0. Higher values of γ make the middle levels of gray brighter, and vice versa, see below; γ = 4 has been found to be an appropriate value.

The correlogram allows semi-manual extraction of F_0. In this case the user restricts the range of allowed F_0 values to the range around a candidate (Figure 2). The software then extracts the frequency corresponding to the highest correlation within this range. The manual control allows the user to select the appropriate candidate, thus eliminating the risk that an automatic algorithm selects a faulty candidate and placing the responsibility on the user.

Applications

Synthesized signals

The properties of the correlogram analysis method can be efficiently demonstrated by applying it to synthesized sound, since the properties of such sounds are well-defined while natural voices mostly contain combinations of acoustic properties in unknown magnitudes. All synthesized sound files were generated using a sampling rate of 16 kHz.

Figure 3 illustrates side bands and the effect of the chosen value of γ. The signal was created using a 100 Hz saw-tooth waveform that was fed through a formant filter with a fixed bandwidth of 100Hz and a resonance frequency, F_1, increasing from 0 to 1000 Hz during the 10 s long sound file. In the left panels (γ=1) a

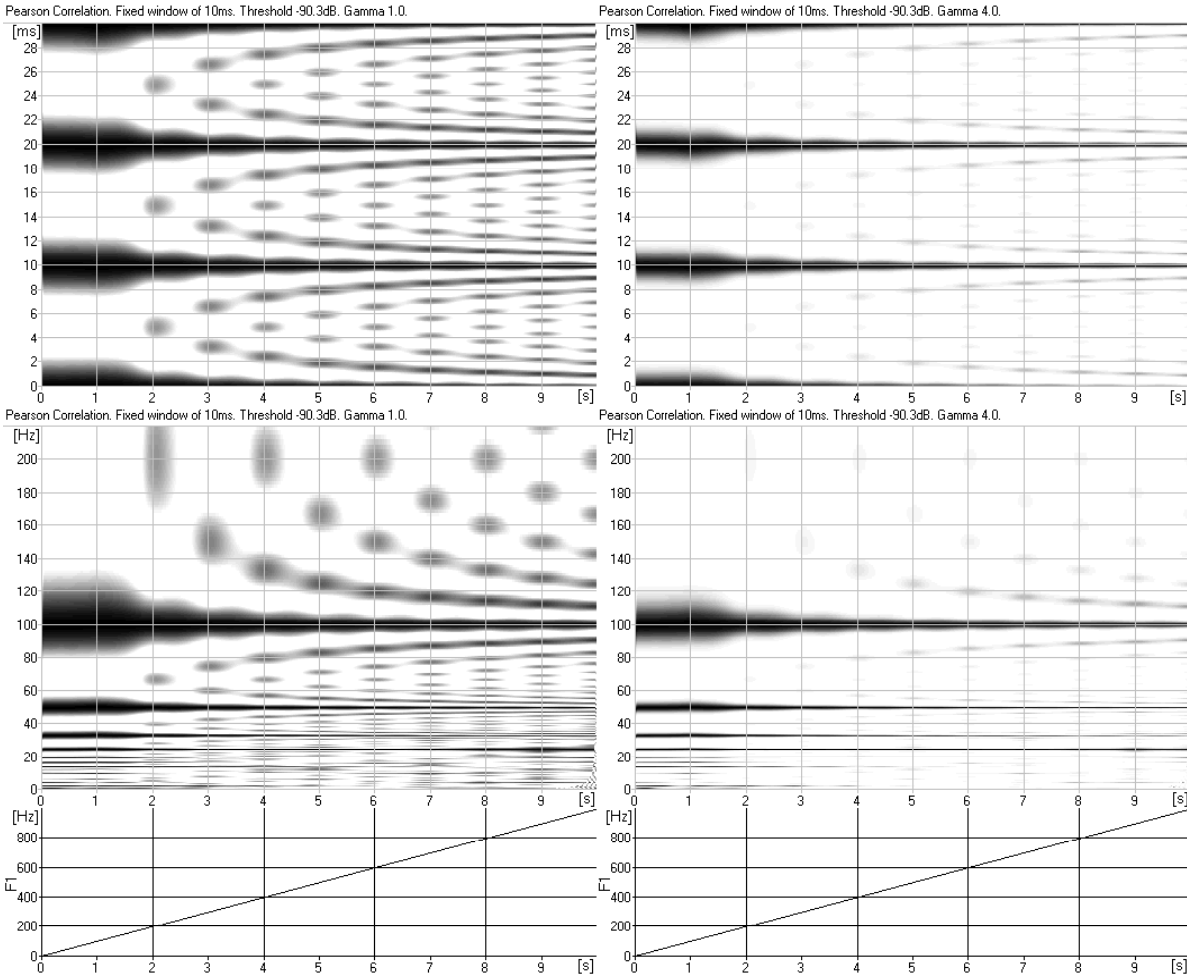

Figure 3. Illustration of side bands. The signal was a 100 Hz saw-tooth wave passed through a formant filter in which the formant frequency was increased from 0 to 1000 Hz (bandwidth = 100 Hz). The side bands are prominent with γ=1 (left panels) but become considerably suppressed by γ=4 (right panels).

segment of a secondary periodicity of 5 ms / 200 Hz appears at 2 s, when F_1 equals 200 Hz. As F_1 is increased, this secondary periodicity shifts and new periodicities appear. The amplitude of the side bands increase when F_1 coincides with a partial. Eventually, the secondary periodicities form continuous side bands around the candidates. If γ is increased to 4 (right panels) the side bands are suppressed such that the candidates can more easily be differentiated from the side bands. The candidates also appear narrower.

Figure 4 illustrates the effects of window length, *w*, on the frequency correlogram. The signal was a saw-tooth wave of 100 Hz, frequency modulated ±10% at 10 Hz and fed through a formant filter of 1000 Hz, bandwidth 100 Hz. For *w*=5 ms (upper left panel) candidates are present at all times, while the side bands are intermittent. Time resolution is good. For *w*=10 ms (upper right panel) the side bands are attenuated and time resolution is still good due to the similarity between T_0 and window length. For *w*=20 ms (lower left panel) the candidates fade at rapid frequency transitions due to poor time resolution, while the side bands are well suppressed. A window length equal to correlation delay *n* (e.g., 10 ms at *n* corresponding to 100 Hz and 20 ms at *n* corresponding to 50 Hz etc.) yields a high time resolution, visible, intermittent side bands at high frequencies, but low time resolution and no side bands at low frequencies (bottom right panel). This is illustrated in terms of the steps in C_2 overlap and extend over longer time than those pertaining to C_1. As seen in the figure, a window length approximating T_0 appears appropriate. In cases of completely unknown T_0, however, a window length equal to correlation delay might be preferable. In this case, C_1 will always be analyzed with a window of length T_0.

Figure 5 illustrates the effect of spectral tilt on side bands. The signal, F_0 = 100 Hz, was created by adding sinusoids according to a spectral tilt, which was continuously varied at a constant rate from 0 to –18 dB/octave over an 18 s long sound file. This signal was fed through a formant filter at 1000 Hz, bandwidth 100 Hz. It can be expected that strong harmonics in the

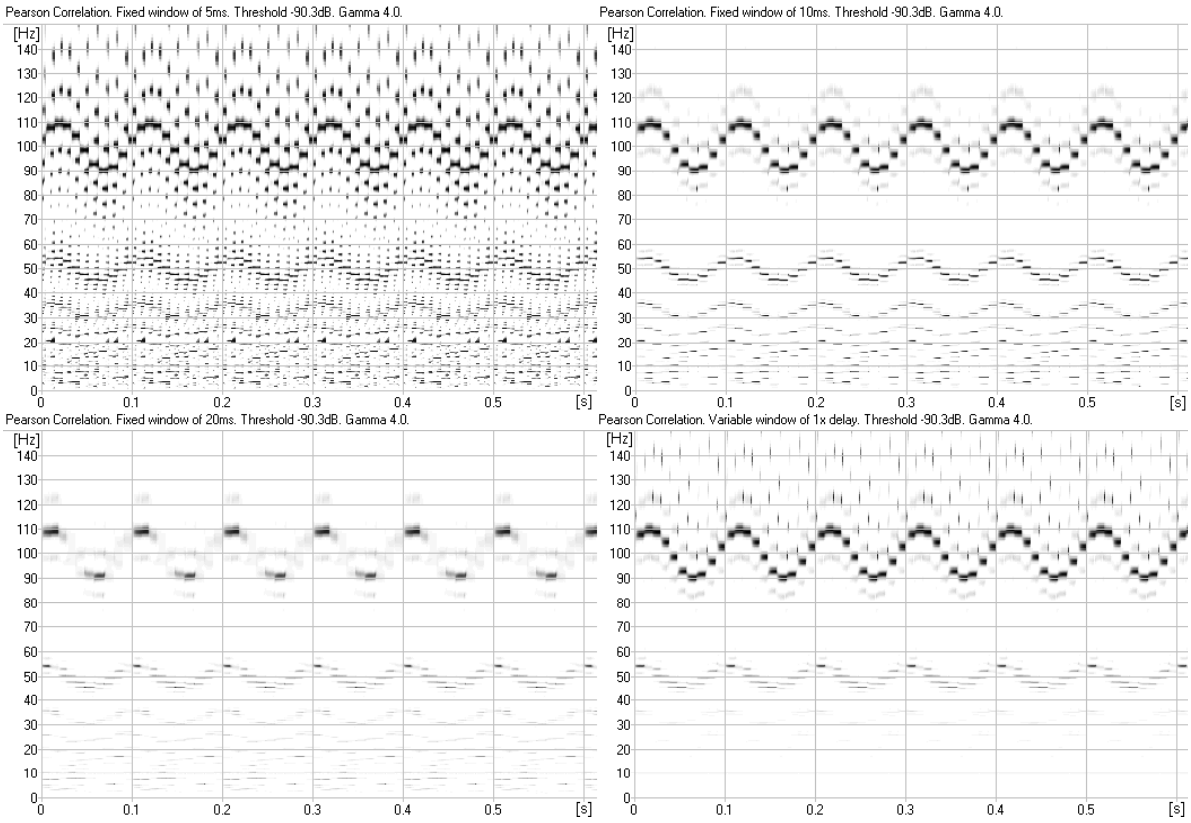

Figure 4. The effects of window length: 5 ms, 10 ms, 20 ms, and variable (upper left, upper right, lower left, and lower right panel, respectively). The signal was a 100 Hz saw-tooth wave, frequency modulated at 10 Hz, ±10 % (formant filter at 1000 Hz, bandwidth 100 Hz).

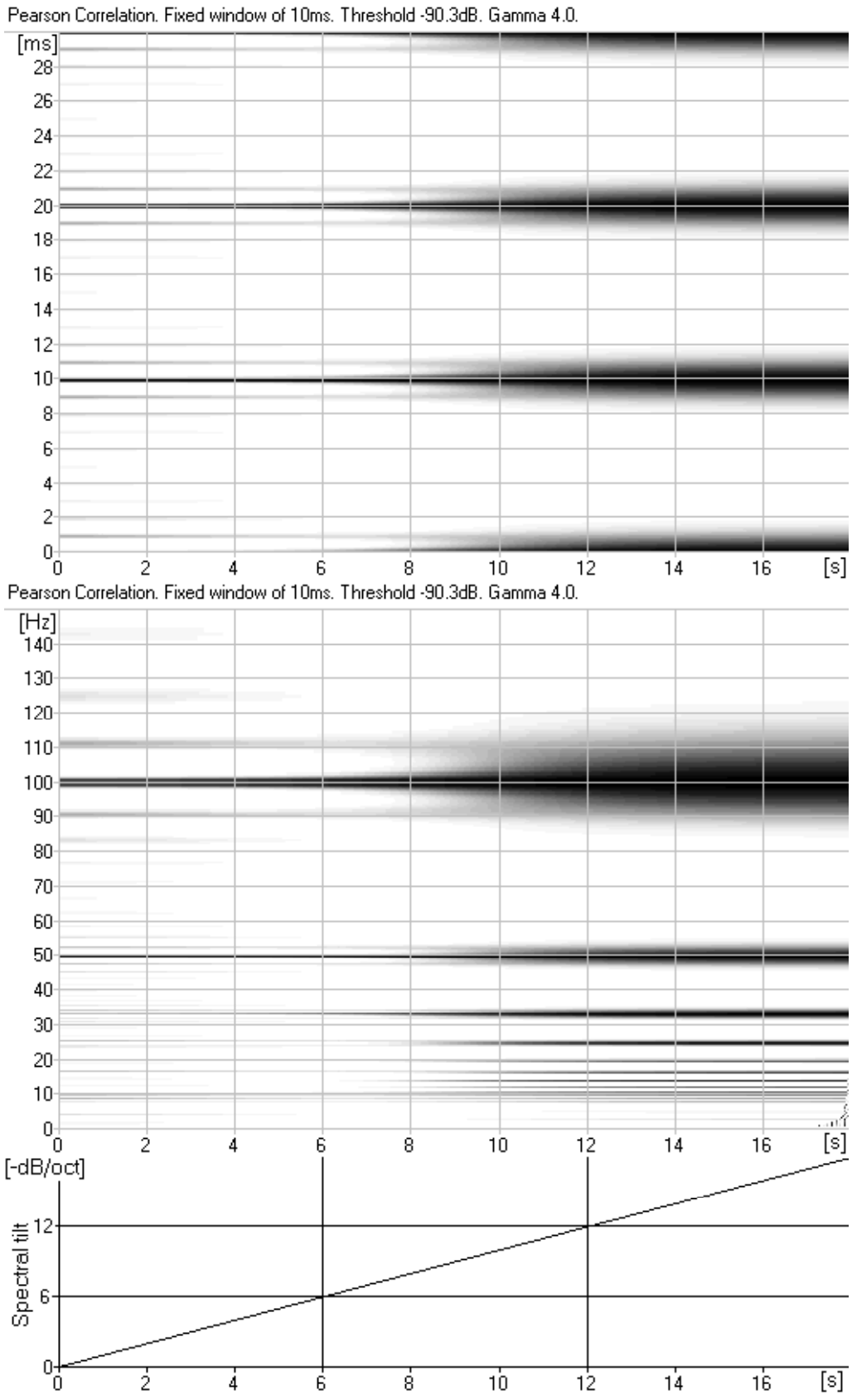

Figure 5. The effect of spectral tilt on side bands. The source signal was created by additive synthesis. The spectral tilt was varied from 0dB/octave at 0 s, to –18dB/octave at 18 s. The source signal was fed through a formant filter at 1000 Hz, bandwidth 100 Hz.

formant region will add extra periodicities that are visible in the correlogram. Correspondingly, if the signal only contains the fundamental, this is the only periodicity that will be visible. These effects can be seen in the figure; a tilt of 0 dB/octave (t=0 s) produces pronounced side bands. The *absolute* width of the candidates is largely determined by the distance between the candidate and the side band. This distance is a direct function of F_1, and thus the candidate width is largely determined by F_1. In other terms, the *relative* width is determined by F_1/F_0. At about -10 dB/octave (t=10 s), candidates grow wide and have no visible side bands. At -18dB/octave (t=18 s) the signal is dominated by the fundamental, the absolute width is mainly determined by F_0, and thus the relative width is nearly independent of both F_0 and F_1. The width's dependance of F_1 can also be seen in figure 3, where the candidates become narrower for higher values of F_1. It should be noted that it is not the increasing spectral tilt *per se* that makes the side bands fade away, but rather the fact that the level of the partial at the formant is reduced. As a rule of thumb, side bands appear if the spectral level at the first formant is near or above the level of the fundamental. However, this is also slightly dependent on the chosen value of γ.

Figure 6 illustrates the effect of adding noise to a periodic signal. At 0 s, the signal, F_0 = 100 Hz, is a saw-tooth waveform, and at 10 s it consists of white noise only. This signal was fed through a 1000 Hz formant filter, bandwidth 100 Hz. The levels of the saw-tooth waveform and the noise were matched, so that the output level of the formant filter was equal at the start and end of the sound file. This means that the harmonics-to-noise ratio (HNR) was infinite at 0 s, 1 (0 dB) at 5 s and 0 at 10 s. The effect of the noise on the correlogram is noticeable at about 2 s, corresponding to a HNR of 4 (12 dB). At about 5 s, or HNR=1 (0 dB), the candidates more or less disappear, the only visible periodicity appearing at 1 ms and created by the F_1 at 1000 Hz.

Figure 7 compares time and frequency correlograms of synthesized saw-tooth waveforms that contain bicyclic F_0 or amplitude variation; this can be seen as a special case of jitter or shimmer. In the jitter case, the period time (mean 10ms) was varied every other period, starting at 0% and ending at ±10%. The jitter (left panels) can be seen as an F_0 fluctuation, the first candidate C_1 splitting into two "stripes" that reach 91 to 111 Hz at the end of the frequency correlogram and 11 to 9 ms at the end of the time correlogram. In the shimmer case (right panels), the amplitude of the periods was varied every other period. The magnitude of the shimmer varies from 0% at the start to ±100% at the end, that is, at the end, every second period has an amplitude of zero, while the intermediate periods have an amplitude twice the original. In this case, the first candidate also shows an oscillating pattern, distinct however, from that characterizing jitter. The odd-order candidates gradually fade as the shimmer quantity increases. The shimmer magnitude is seen less clearly than the jitter magnitude since the former is reflected in terms of the gray scale while the latter is represented by the position along the y-axis.

Figure 8 compares time and frequency correlograms of synthesized saw-tooth waveforms that contain random F_0 or amplitude variation; these represent other types of jitter and shimmer. In the jitter case, the period time (mean 10ms) was randomly distributed within ±10%. The jitter (left panels) can be seen as a random F_0 fluctuation, the first candidate C_1 fluctuating in the range 91 to 111 Hz in the frequency correlogram and 11 to 9 ms in the time correlogram. In the shimmer case (right panels), the ampliude of each period was varied randomly. The magnitude of changes was 100%, in other words, the amplitude was randomly distributed between zero and full scale. In this case, the first candidate also shows a fluctuating pattern, again distinct from that characterizing jitter.

Although the time correlogram is directly linked to the waveform matching function, we shall henceforth focus on frequency correlograms.

Natural voices

Some examples of correlograms and narrow band spectrograms of pathological voices are presented in Figures 9-15. All these figures concern examples of voices that may cause difficulties for F_0 tracking programs. The difficulties are due to ambiguity about whether F_0 is represented by C_1 or by C_2 (Figures 9, 10 and 11), due to a high noise level and unstable C_1 (Figures 12, 13 and 14), or due to the well-excited first formant, which makes the side bands hard to differentiate from C_1 (Figure 15).

For describing the voices, the terminology proposed by Hammarberg and Gauffin (1995)

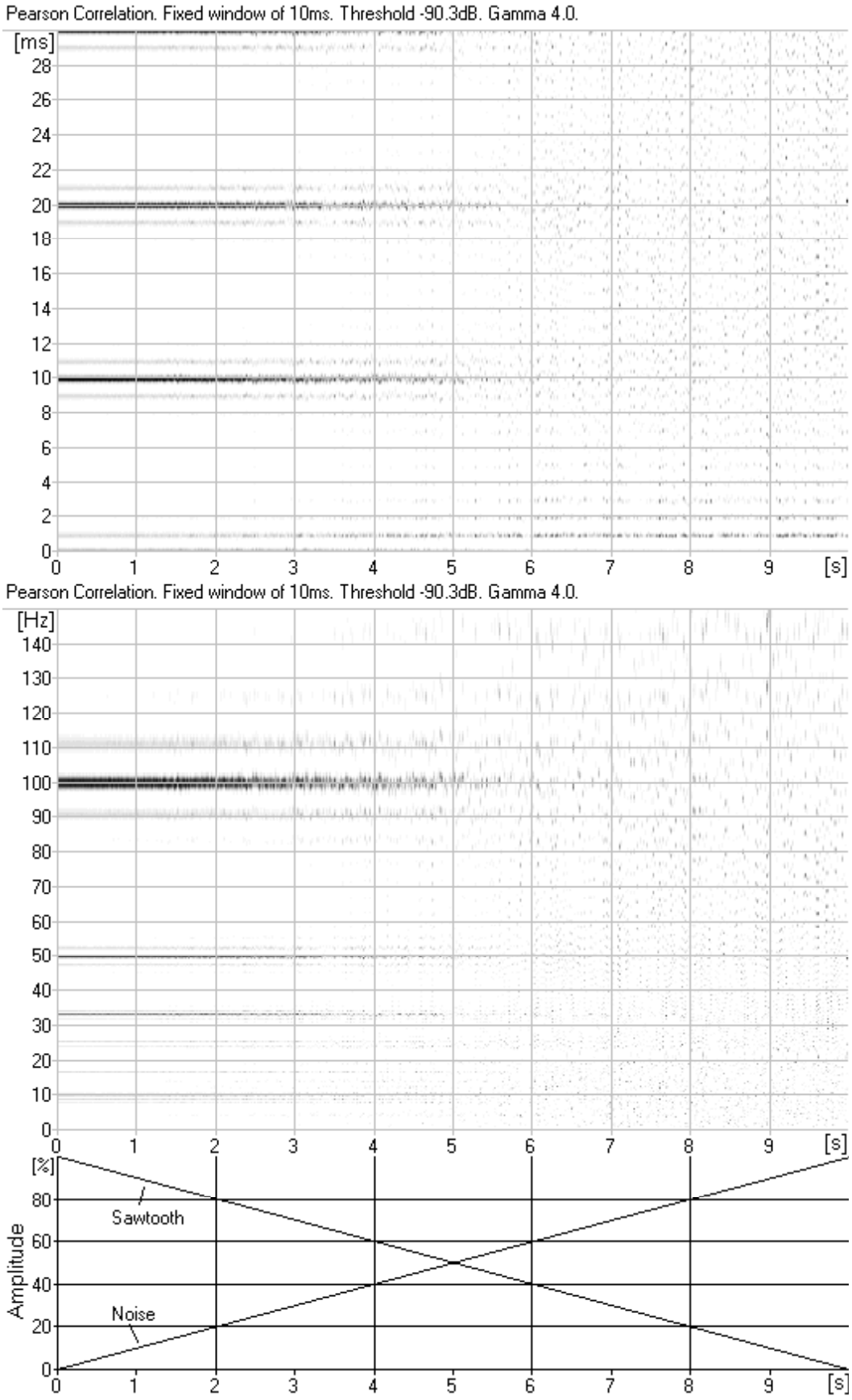

Figure 6. The effect of adding white noise. At 0 s, the source signal consists of a sawtooth wave only, and at 10 s of white noise only. The source signal was passed through a formant filter at 1000 Hz, bandwidth 100 Hz.

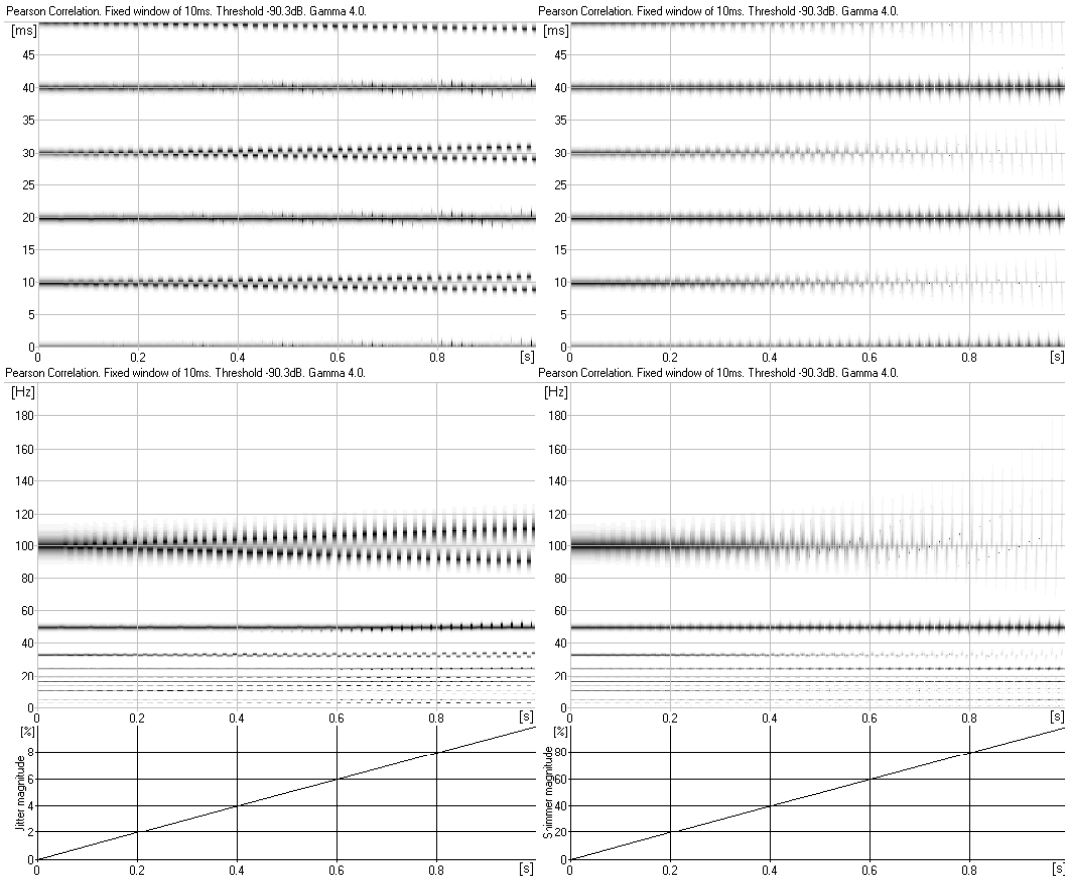

Figure 7. Examples of bicyclic F_0 (left panels) and amplitude variation (right panels) in sawtooth waveforms. F_0 and amplitude variations increased from 0 to 10% and from 0 to 100%, respectively.

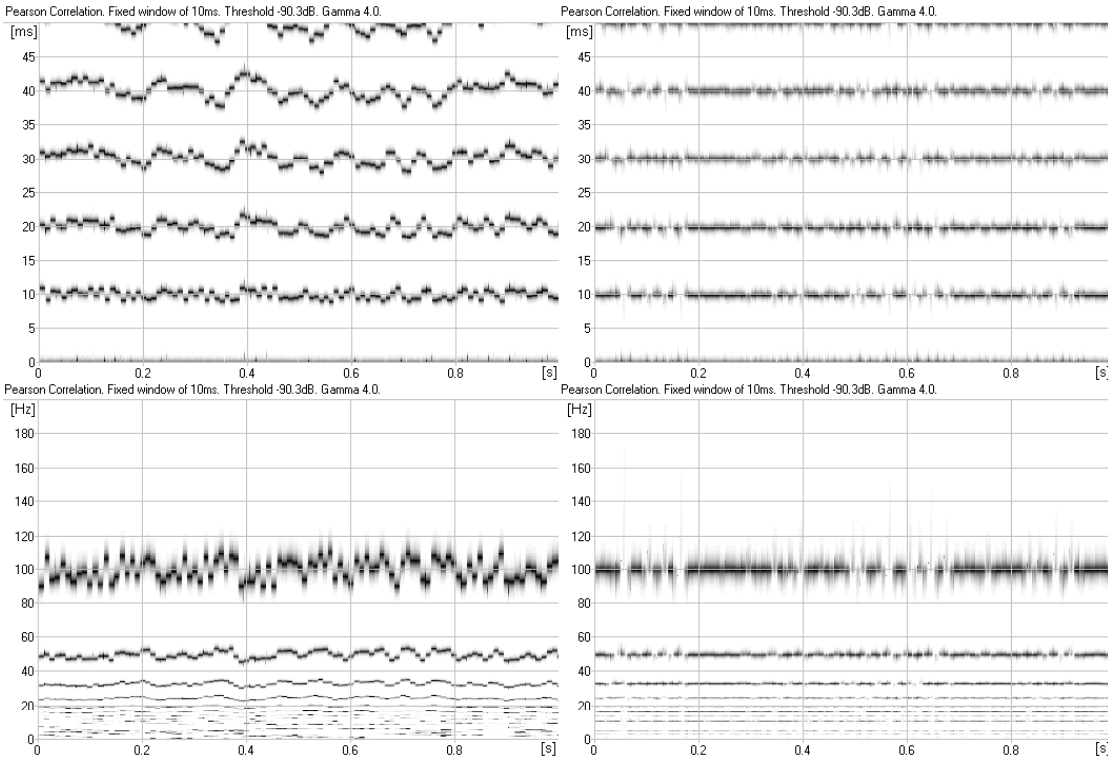

Figure 8. Examples of random F_0 (left panels) and random amplitude variation (right panels) in sawtooth waveforms. F_0 and amplitude variations was 10% and from 0 to 100%, respectively.

was used; all voice examples except one (Figure 15) were taken from Hammarberg's library of pathological voices. The voice samples had been rated by groups of 6 to 14 voice clinicians (speech language pathologists and phoniatricians) using the Stockholm Voice Evaluation Approach (SVEA) assessment protocols (Hammarberg 1986; 2000). These examples have been found useful in teaching as archetypes of various perceived voice properties, as in each example a particular voice quality is dominant over others. However, since all natural voices are perceptually multi-dimensional (Kreiman et al., 1994; 1996), each example still represents more than one single perceptual feature. This well-known fact makes direct mapping of acoustic to perceptual features difficult.

Figure 9 presents the voice of a man, age 41, who was diagnosed with chronic laryngitis and whose voice quality was characterized as rough. This example shows short (about 100 ms) bursts of bicyclicity starting at 0, 0.2 and 0.45 s. Widening of the candidates can be seen at 0.1-0.2 s, 0.6-0.7 s, and 0.85-1 s. The two former cases of widening are probably due to the low amplitude of the overtones in the signal, and the last is probably due to a low first formant. Figure 10 presents the voice of a man, age 32, who was diagnosed with incomplete voice mutation and whose voice quality was characterized as a mixture between vocal fry and gratings/scrape. This example contains bicyclicity throughout most of its duration. Figure 11 presents the voice of a man, age 29, who was diagnosed with a benign tumor, perceived as having gratings/scrape only.

These examples all show short (100 ms) bursts of bicyclicity and the only obvious difference among them is the F_0 at which they occur. In Figure 10 there also is a longer (about 250 ms) bicyclic sequence. Careful inspection of the spectrograms (lower panel) reveals subharmonics coinciding with the bicyclic segments in the correlograms but with poorer time resolution. The subharmonics could be visualized more clearly in the spectrogram if a narrower bandwidth had been chosen. This would, however have further deteriorated the time resolution.

Figure 12 presents the voice of a man, age 50, who was diagnosed with paralytic dysphonia, and whose voice quality was characterized by hypofunctional breathiness with roughness. The voice produces wide and unstable candidates. As sometimes also found in rough voices, an instant of bicyclicity occurs, near t=0.7 s. However, the corresponding subharmonics in the spectrogram (lower panel) are not easily spotted, probably due to the short duration of the bicyclicity.

Figure 13 presents the voice of a man, age 40, who was diagnosed with paralytic dysphonia; the voice was characterized by hypofunctional breathiness. In the narrow-band spectrogram only few harmonics except the fundamental are visible. The correlogram shows wide candidates and no bicyclicity.

Figure 14 presents the voice of a woman, age 75, who was diagnosed with paralytic dysphonia which shows repeated voice breaks between falsetto and modal register with a high degree of instability (Hammarberg, 1986). C_1 suddenly disappears at t=0.5 s and 1.35 s as the voice switches from falsetto to modal.

Finally, Figure 15 presents the voice of an opera singer. Side bands are prominent, indicating a well-excited first formant. As also can be seen in the spectrogram, the singer apparently tuned F_1 to either two or three times F_0, such that either the second or third partial coincides with F_1. This strategy increases the sound pressure level, which is an important ability in operatic singing.

Discussion

Since F_0 fluctuation plays an important role in many different pathological voice qualities, F_0 extraction would be one way to study such voice qualities. Unfortunately, F_0 extraction applied to voices with a high degree of F_0 perturbation presents problems that are not easily solved. The most typical example is bicyclic voice, where F_0 extractors tend to yield $F_0/2$. Since the transition from normal phonation to bicyclicity can be gradual, although without a pitch glide, an F_0 extraction algorithm must determine when to switch from displaying F_0 to displaying $F_0/2$. Such switching results in an octave leap. In the correlogram, this problem is circumvented by eliminating the selection mechanism and displaying raw correlation functions in a three-dimensional graph. Hence, the correlogram can describe highly perturbed voices, even when the value of F_0 is far from obvious.

The appropriateness of extracting F_0 from pathological voices can sometimes be questioned. Pathological voices often show large period-to-period variation, and since the signal is not exactly repetitive, no strict period time

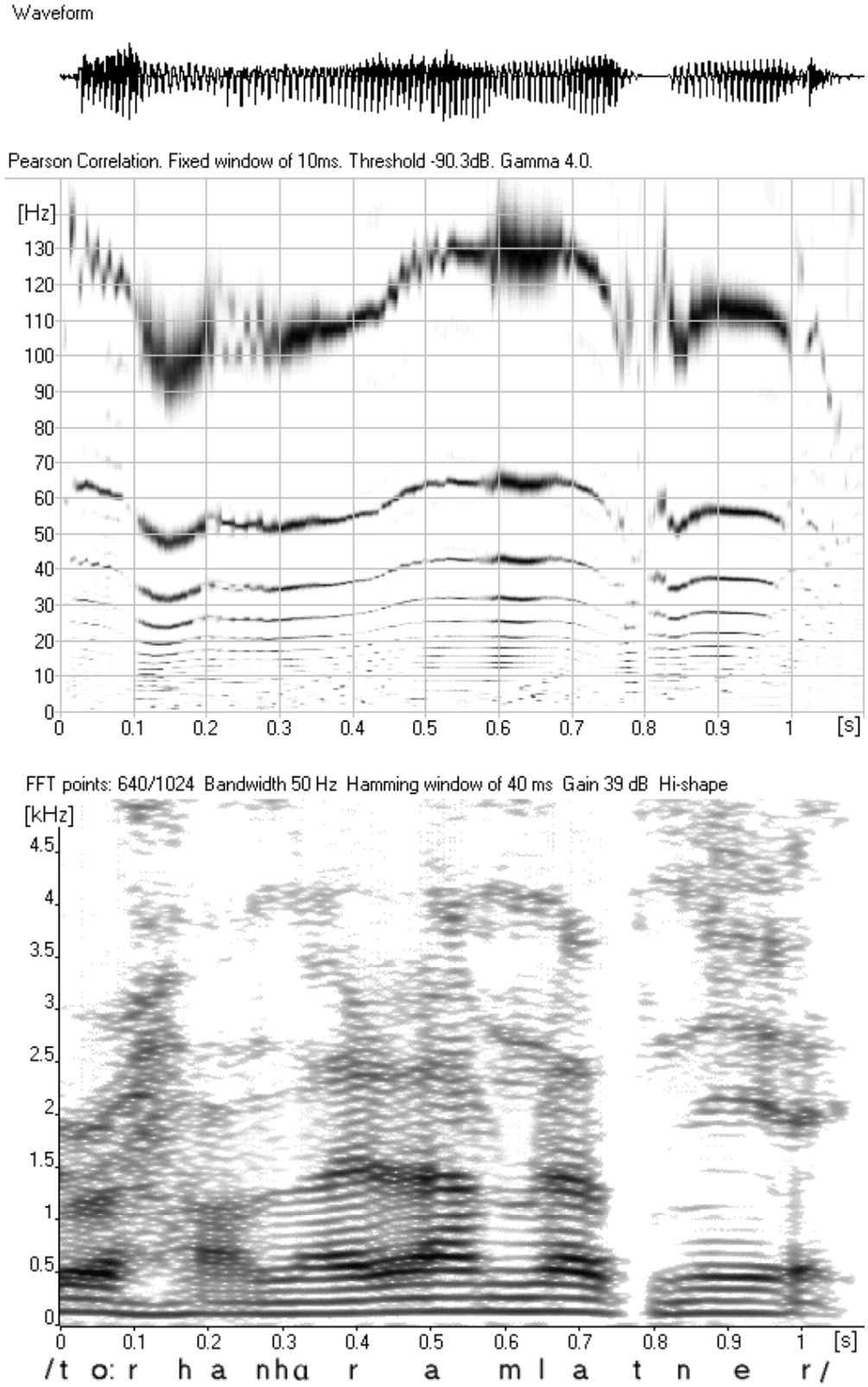

Figure 9. Waveform (top), correlogram (middle) and narrow band spectrogram (bottom) of speech. The voice was characterized by roughness. Bicyclic segments appear around 0.05, 0.25 and 0.5 s. At about 0.8 s, there is a lack of periodicity due to the voiceless consonant /t/.

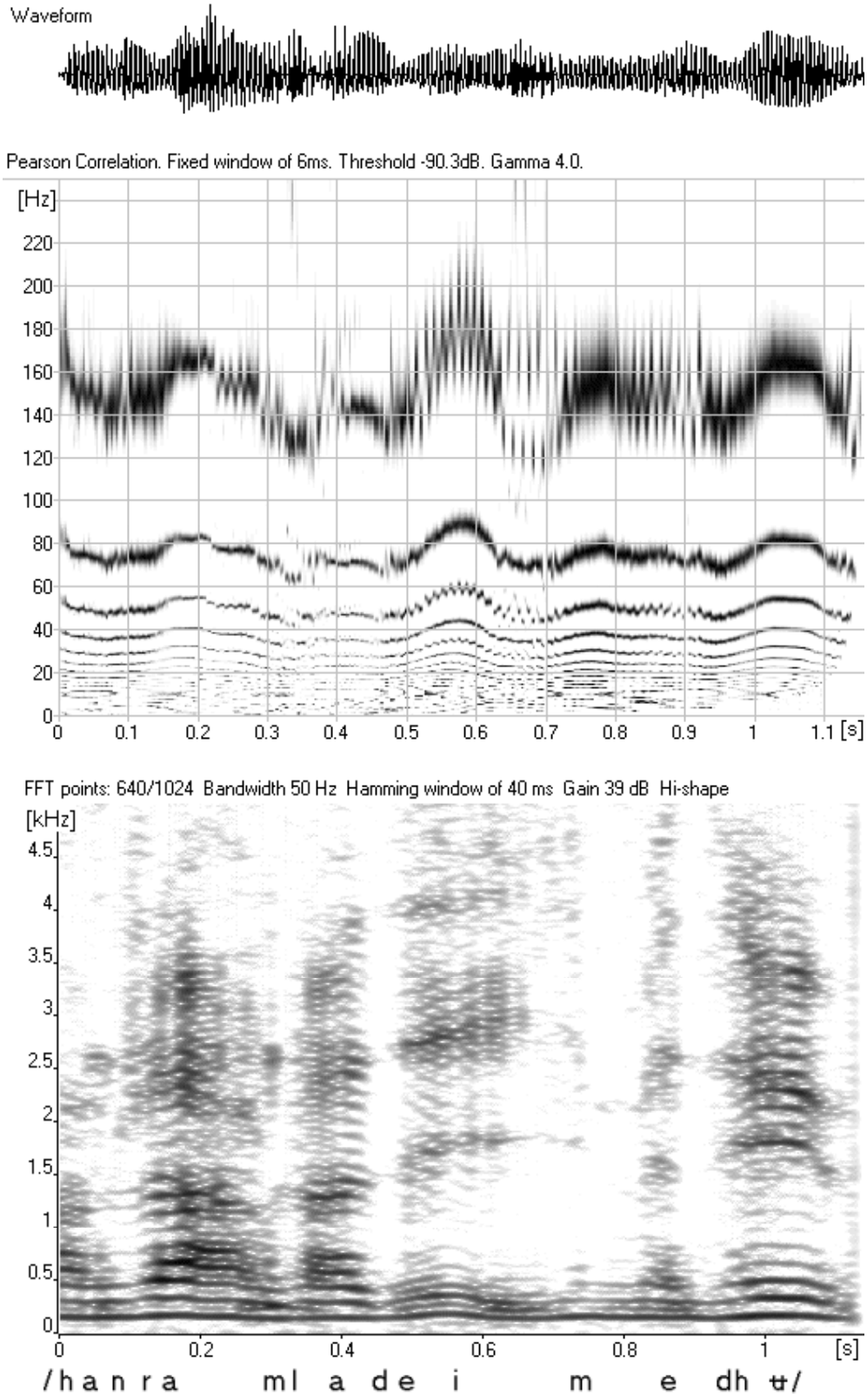

Figure 10. Waveform (top), correlogram (middle) and narrow band spectrogram (bottom) of speech. The voice was characterized by vocal fry and gratings/scrape. Bicyclic segments appear around 0.25 s, at 0.5-0.75 s and at 0.8-0.95 s.

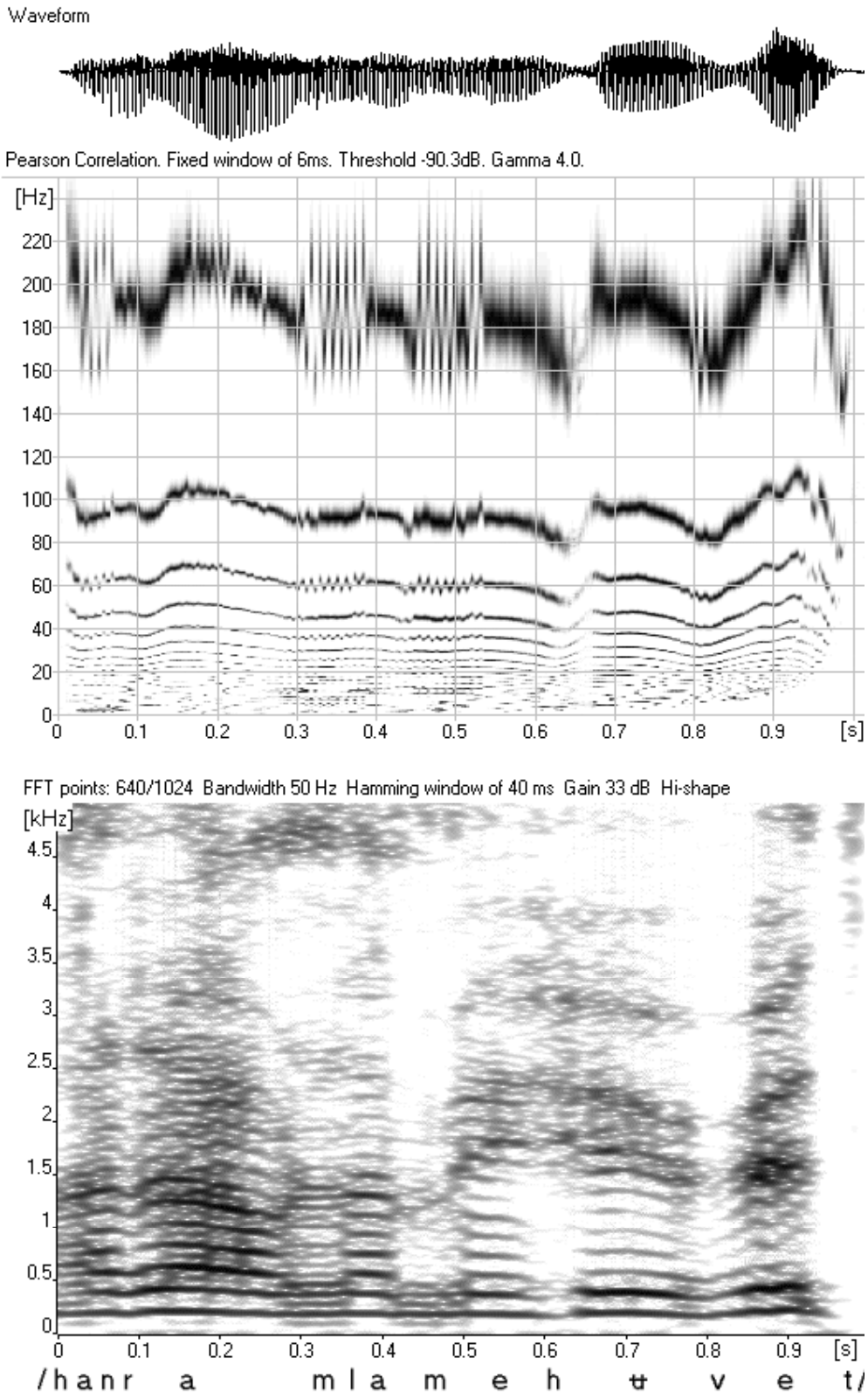

Figure 11. Waveform (top), correlogram (middle) and narrow band spectrogram (bottom) of speech. The voice was characterized by gratings/scrape. Bicyclic segments appear around 0.05s, 0.35 s and 0.5 s and also with less magnitude around 0.2 s and 0.8 s. Note the abnormally high F_0.

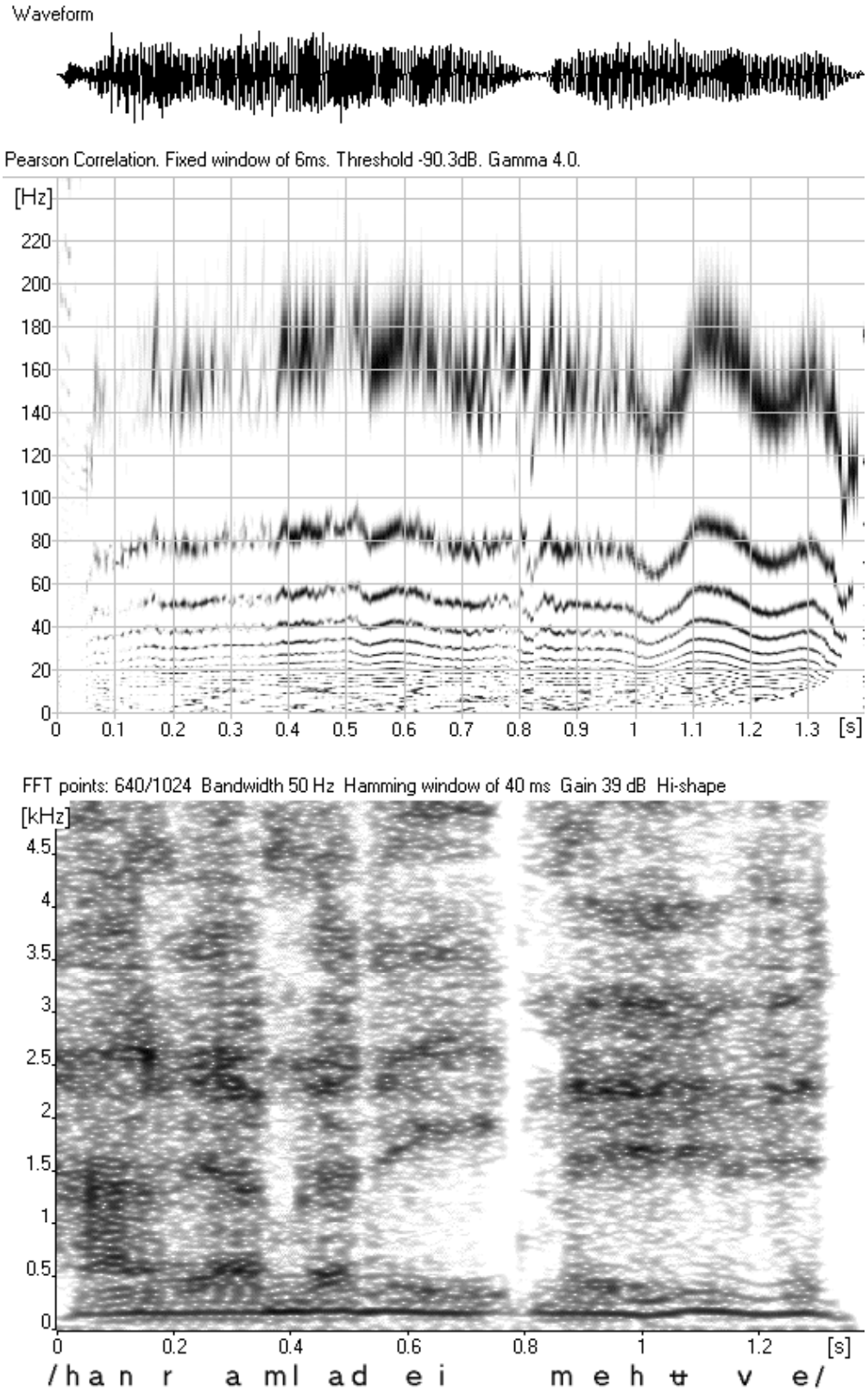

Figure 12. Waveform (top), correlogram (middle) and narrow band spectrogram (bottom) of speech. The voice was characterized by hypofunctional breathiness with roughness. The candidates all are unstable and wide, and a short instance of bicyclicity can be seen at around 0.7 s.

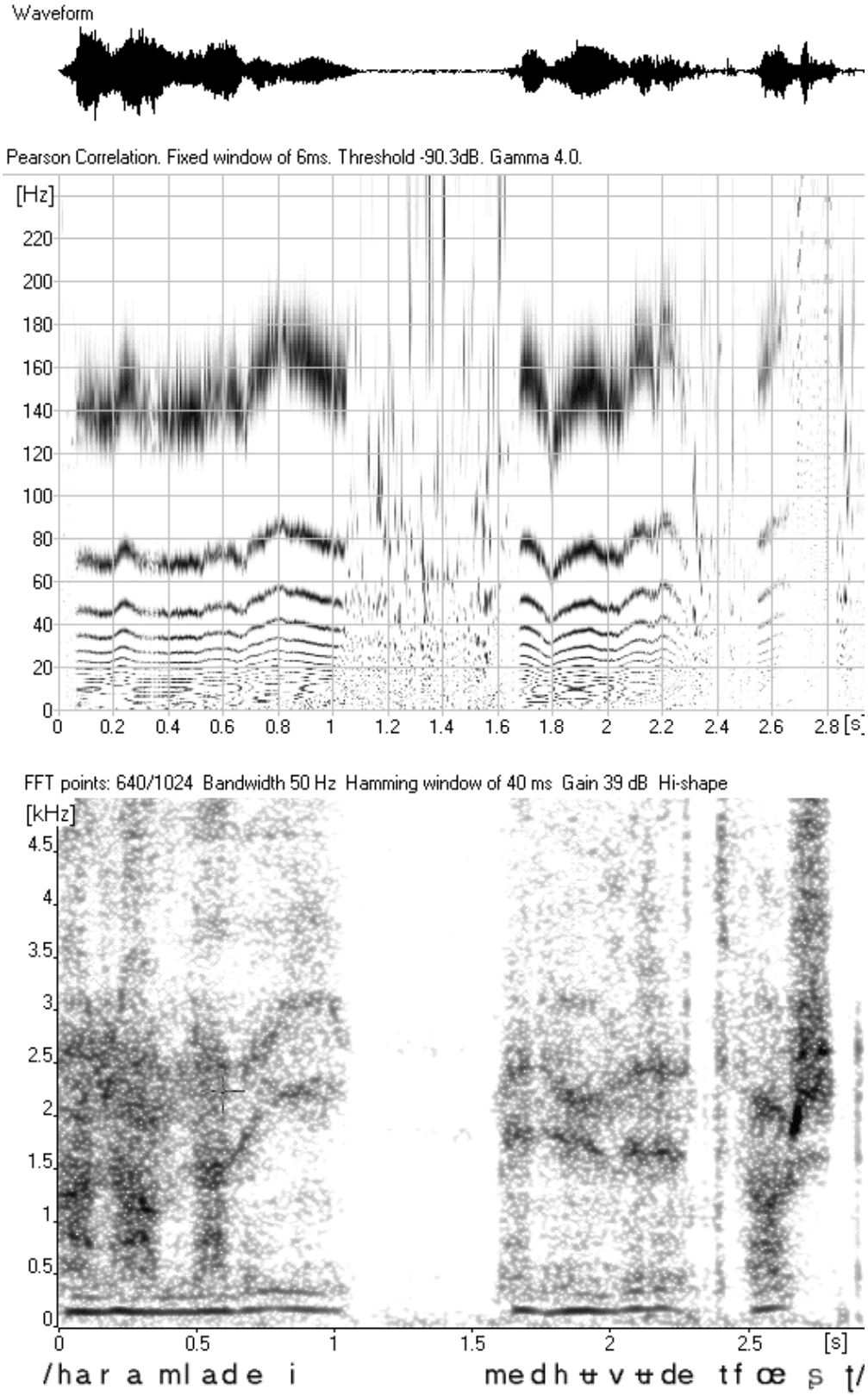

Figure 13. Waveform (top), correlogram (middle) and narrow band spectrogram (bottom) of speech. The voice was characterized by hypofunctional breathiness. All candidates are wide due to the dominant fundamental. The segment between 1.0 and 1.7 s represent silence.

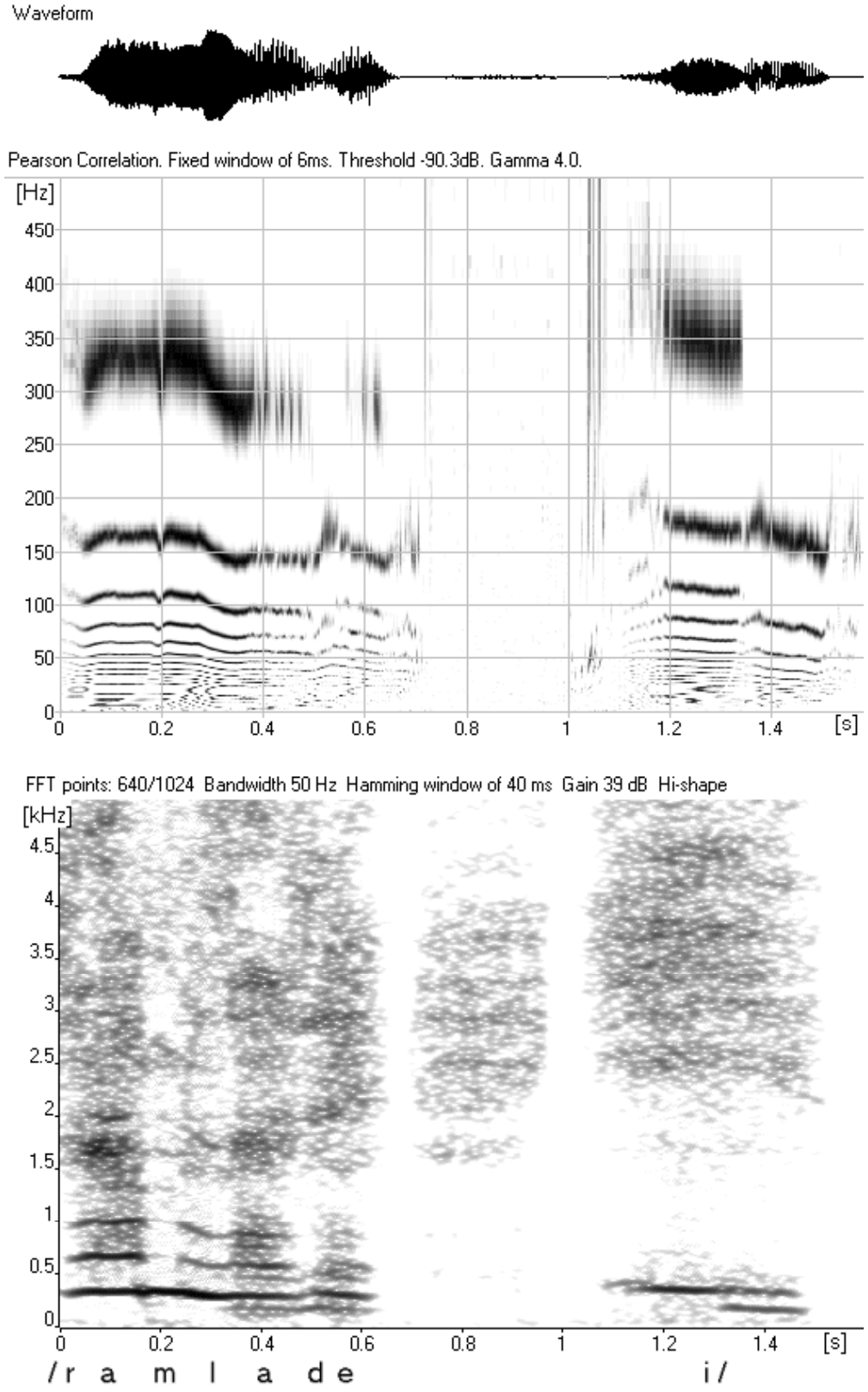

Figure 14. Waveform (top), correlogram (middle) and narrow band spectrogram (bottom) of speech. The voice was characterized by repeated register breaks. At 0.4 s and 1.35 s the voice switched from falsetto to modal. The noise between 0.7 and 0.9 s is inhalatory stridor.

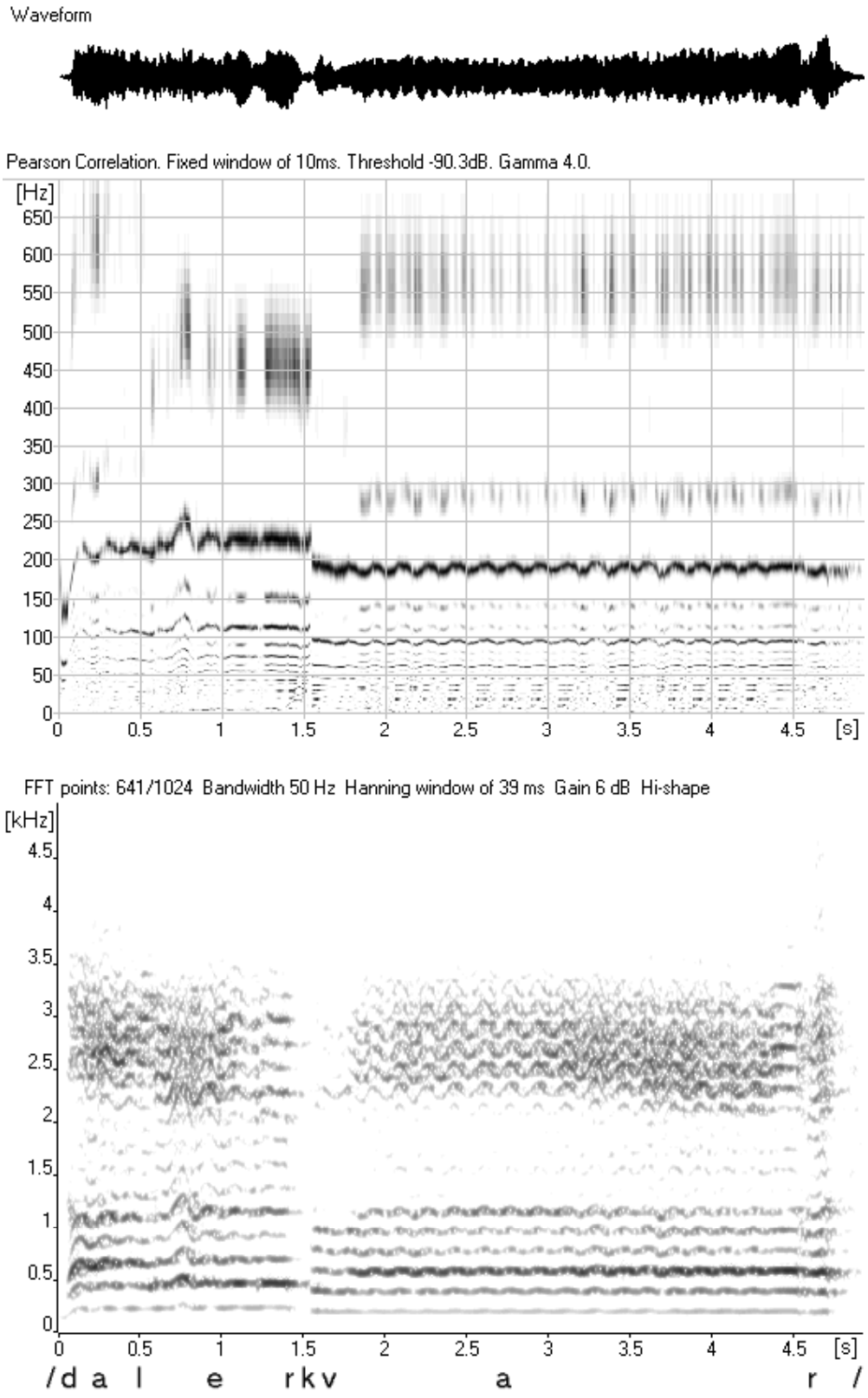

Figure 15. Waveform (top), correlogram (middle) and narrow band spectrogram (bottom) of singing, operatic style. Prominent side bands can be seen either at two or three times F_0 due to a well-excited first formant.

exists; or, an ambiguity exists with regard to F_0. For such voices, it appears appropriate not to enforce an F_0 selection, but rather to display the correlation functions, as in a correlogram. In some cases a perceptual evaluation of pitch may also be a worthwhile alternative.

The interpretation of a correlogram requires some care, however. While the F_0 candidates appear as dark horizontal bands, there can also be side bands, which originate from formant ringings of high amplitudes and are thus not to be considered as candidates. Side bands appear when an overtone coincides with a formant and can mostly be identified from their relatively less dark appearance (i.e. low correlation). In some cases, such as when F_1 is twice F_0, the distinction between side bands and candidates can be less clear.

The presence of side bands can also be used as an indication of a high positive level difference between the first formant (L_1) and the fundamental (L_0). Hence, the presence of side bands may indicate a sonorous or pressed voice with a well-excited first formant. On the other hand, a low or negative L_1-L_0 difference has been shown to be related to hypofunctional breathiness (Hammarberg 1986). In the correlogram this would correspond to a wide candidate. It must be kept in mind, however, that the candidate width also depends on the formant frequencies.

A correlogram is a time-domain analysis tool. This means that it does not directly display spectral properties, such as harmonics, which would require a Fourier transform. It should be noted that the candidates have no direct connection to the harmonics of the signal. It is true that C_1 corresponds to the first harmonic, H_1, but the presence of C_2 does not necessarily indicate the presence of a subharmonic. However, the combined occurrence of a constant C_2 and a varying C_1 would indicate bicyclicity, and a constant C_3 and a varying C_2 and C_1 tricyclicity, etc. These characteristics indicate the presence of subharmonics, although the subharmonics *per se* are not visualized in a correlogram.

Compared to the narrow-band spectrogram, the correlogram shows a better time-resolution, due to the shorter time windows needed. For instance, to display a narrow-band spectrogram with visible subharmonics, the length of the time window must correspond to several fundamental periods, whereas in the correlogram, the time windows typically are as short as one fundamental period. The short time windows have the effect that the correlogram can visualize short bursts of bicyclicity that would not be easily seen in a narrow-band spectrogram.

The correlogram has also been used for extraction of F_0 from violin playing (Gleiser et al 1999) by means of manual tracing. In these experiments, the violin player was accompanied by piano playing, which however was suppressed by placing the microphone on the violin bridge. Violin sound typically presents difficulties in F_0 extraction. However, the correlogram method was surprisingly successful and showed a remarkable insensitivity to the piano sound. In this study, the vibrato rate was also extracted in a second step, by performing correlogram analysis and tracing on the extracted oscillating F_0 curve.

The computation of a correlogram is generally more computationally intensive than the computation of a spectrogram. However, with the increasing power of computers, the computation speed is less of a problem. For example, every correlogram presented above required less than 3 seconds computing time on a 1700 MHz Pentium 4 system running Windows 2000.

These initial applications suggest that the correlogram is useful for future work for refining, revising and standardizing the relations between acoustical voice characteristics and perceived voice quality parameters. Correlograms should be useful also for the training ofan analytic listening to voice qualities. Presenting images of the perturbation of voices together with the sounds seems a valuable opportunity that may pave the way to a better agreement on the meaning of voice terms across the voice community.

The correlogram illustrates the periodicity of the waveform in a robust way, since it lacks the selection mechanism of F_0 extractors. It illustrates differences between periodic and random period-to-period variations. The robustness of the correlogram should make it a particularly valuable tool for periodicity analysis in such cases of pathologic speech where standard F_0 extraction methods fail or where they present ambiguous results.

Conclusions

Correlation functions have previously been used to extract F_0 information from voice signals,

automatically selecting a single value to represent F_0, sometimes even in ambiguous cases. The correlogram method presented here shows the raw correlation functions. In cases of periodic or quasi-periodic phonation, such as in some pathological voices, it displays several F_0 candidates, and leaves the user to select one by tracing, if appropriate. In some cases of quasi-periodic phonation, the correlogram illustrates the type of aperiodicity, differentiating signal characteristics such as multi-cyclic or random perturbations, typically associated with vocal fry or roughness. It should be worthwhile to test the correlogram in cases of quasi-periodic signals where traditional F_0 tracking methods fail.

Acknowledgements

This work was supported by research grants from the Bank of Sweden Tercentenary Foundation and the Swedish Council for Work Life Research. We would also like to thank Jan Gauffin and Stellan Hertegård for valuable advice and Johan Sundberg for discussions and editorial assistance.

References

Blomgren M, Chen Y, Ng M, Gilbert H (1998). Acoustic, aerodynamic, physiologic and perceptual properties of modal and vocal fry registers. *J Acoust Soc Am* 103: 2649-2658.

DeKrom G (1995). Some spectral correlates of pathological breathy and rough voice qualitiy for different types of vowel fragments. *J Speech Hear Res* 38: 794-811.

Fourcin A. (1986). Electrolaryngographic assessment of vocal fold vibration. *J Phonetics* 14: 435-442.

Gauffin J, Granqvist S, Hammarberg B, Hertegård S, Håkansson A (1995). Irregularities in the voice: some perceptual experiments using synthetic voices *Proc ICPhS-95* Vol 2: 242-245

Gleiser J, Friberg A, Granqvist S (1998). A method for extracting vibrato parameters applied to violin performance. *TMH-QPSR, KTH,* 4/1998: 39-44

Hammarberg B. (1986). Perceptual and acoustic analysis of dysphonia. *Doctoral thesis*. Dept of Logopedics and Phoniatrics, Karolinska Institute, Stockholm.

Hammarberg B. (2000) Voice research and clinical needs. *Folia Phoniatr. Logop*. 52:93-102

Hammarberg B, Gauffin J (1995). Perceptual and acoustical characteristics of quality differences in pathological voices as related to physiological aspects. In: Fujimura O, Hirano M (eds). *Vocal Fold Physiology, Voice Quality Control*. San Diego: Singular Publishing Group; 283-303.

Hess W (1983). *Pitch determination of speech signals.* Springer-Verlag. ISBN 0-387-11933-7.

Hess W (1995). Determination of glottal excitation cycles in running speech. *Phonetica* 52: 196-204.

Hillenbrand J (1988). Perception of aperiodicities in synthetically generated vowels. *J Acoust Soc Am* 83: 2361-2371.

Imaizumi S (1986). Acoustic measures of roughness in pathological voice. *J Phonetics* 14: 457-462.

Ishiki N, Okamura H, Tanabe M, Morimoto M (1969). Differential diagnosis of hoarseness. *Folia Phoniatrica* 21: 9-19.

Karnell M, Scherer R, Fischer L (1991). Comparison of acoustic voice perturbation measures among three independent voice laboratories. *J Speech Hear Res* 34: 781-790.

Kreimann J, Gerrat B, Berke G (1994). The multi-dimensional nature of pathologic voice quality. *J Acoust Soc Am* 96: 1291-1302.

Kreimann J, Gerratt, B R (1996). The perceptual structure of pathologic voice quality. *J Acoust Soc Am* 100: 1787-1795.

Laver J. (1980) *The Phonetic Description of Voice Quality*. Cambridge University Press, Cambridge. ISBN 0-521-231760

Ladefoged P (1988). Discussion of phonetics: a note on some terms for phonation types. In: Fujimura O (ed), *Vocal physiology: Voice production, mechanisms and functions* New York: Raven Press; 373-375.

McAllister A (1997). Acoustic, perceptual and physiological studies of ten-year-old chilren's voices. *Doctoral thesis*. Dept of Logopedics and Phoniatrics, Karolinska Institute and Dept of Speech, Music and Hearing, Royal Institute of Technology (KTH), Stockholm.

Omori K, Kojima H, Kakani R, Slavit D, Blaugrund S (1997). Acoustic characteristics of rough voice: Subharmonics. *J Voice* 11: 40-47.

Pabon P (1991). Objective acoustic voice-quality parameters in the computer phonetogram. *J Voice* 5: 203-216.

Rabinov R, Kreiman J, Gerratt B, Bielamowicz S (1995). Comparing reliability of perceptual ratings of roughness and acoustic measures of jitter. *J Speech Hear Res* 38: 26-32.

Rothenberg (1973). A new inverse filtering technique for deriving the glottal airflow waveform during voicing. *J Acousic Soc Am 53:1632-1645*

Sederholm E (1996). Hoarseness in ten-year old children: Perceptual characteristics, prevalence and etiology. *Doctoral thesis*. Dept of Logopedics and Phoniatrics, Karolinska Institute and Dept of

Speech, Music and Hearing, Royal Institute of Technology (KTH),, Stockholm.

Švec J, Pešák J (1994). vocal breaks from the modal to falsetto register. *Folia Phoniatr Logop* 46: 97-103.

Sundberg J (1987). *The Science of the Singing Voice*. Northern Illinois University Press. ISBN 0-87580-120-X.

Titze I (1995). Definitions and nomenclature related to voice quality. In: Fujimura O, Hirano M (eds). *Vocal Fold Physiology, Voice Quality Control*. San Diego: Singular Publishing Group; 335-342.

Titze I, Liang H (1993). Comparison of F_0 extraction methods for high-precision voice perturbation measurements. *J Speech Hear Res* 36: 1120-1133.