Behavior Research Methods
2010, 42 (4), 1030-1041
doi:10.3758/BRM.42.4.1030

Integrated software for analysis and synthesis of voice quality

Jody Kreiman, Norma Antoñanzas-Barroso, and Bruce R. Gerratt
University of California, Los Angeles, California

Voice quality is an important perceptual cue in many disciplines, but knowledge of its nature is limited by a poor understanding of the relevant psychoacoustics. This article (aimed at researchers studying voice, speech, and vocal behavior) describes the UCLA voice synthesizer, software for voice analysis and synthesis designed to test hypotheses about the relationship between acoustic parameters and voice quality perception. The synthesizer provides experimenters with a useful tool for creating and modeling voice signals. In particular, it offers an integrated approach to voice analysis and synthesis and allows easy, precise, spectral-domain manipulations of the harmonic voice source. The synthesizer operates in near real time, using a parsimonious set of acoustic parameters for the voice source and vocal tract that a user can modify to accurately copy the quality of most normal and pathological voices. The software, user's manual, and audio files may be downloaded from http://brm.psychonomic-journals.org/content/supplemental. Future updates may be downloaded from www.surgery.medsch.ucla.edu/glottalaffairs/.

Although much is known about the psychoacoustics of pitch and loudness, the auditory perception of sound quality is relatively poorly understood. Our insufficient information about how listeners assess the quality of a given acoustic signal is a particular problem in the study of voice. Voice quality has been implicated as a perceptual cue in studies of lexical access and spoken word recognition (Goldinger, Pisoni, & Logan, 1991; Nygaard & Pisoni, 1998), many aspects of prosody (e.g., Cutler, Dahan, & van Donselaar, 1997), and animal communication (e.g., Fitch, Neubauer, & Herzel, 2002; Volodina, Volodin, Isaeva, & Unck, 2006). A clinician's judgments about voice quality are also essential in caring for patients with voice disorders (e.g., Colton, Casper, & Leonard, 2005; Gerratt, Till, Rosenbek, Wertz, & Boysen, 1991). However, our ability to elucidate the role played by the voice in these functions is limited by a lack of understanding of the psychoacoustic nature of voice quality. Most studies rely on correlations between acoustic measures and listeners' ratings of voices on scales for attributes such as breathiness and roughness (e.g., Kempster, Gerratt, Verdolini Abbott, Barkmeier-Kraemer, & Hillman, 2009; Kreiman, Gerratt, Kempster, Erman, & Berke, 1993; Shrivastav, 2003), but this approach is undermined by concerns about listener reliability, by lack of evidence concerning the psychological reality of the rating scales, by lack of theoretical models motivating the choice of acoustic measures, and by the inability of such experiments to demonstrate that changes in the acoustic parameter actually cause the correlated changes in voice quality.

Speech synthesis offers a means of directly testing hypotheses about the relationship between acoustic parameters and voice quality perception; thus, it provides one possible solution to these difficulties. Synthesis software typically provides control over some aspects of the voice source. For example, the well-known Klatt synthesizer includes parameters that control fundamental frequency ($f0$), the amplitudes of voicing pulses and of aspiration noise, the percentage of time during which the vocal folds are open during a phonatory cycle, and overall spectral tilt (Klatt, 1980; Klatt & Klatt, 1990), with additional parameters (FL, DI) controlling cycle-to-cycle variations in $f0$ and/or amplitude. However, the limited number of fixed parameters controlling the shape of the individual source pulses in such systems (five in the Liljencrants–Fant [LF] source model, Fant, Liljencrants, & Lin, 1985; six in the KGLOTT88 source model, Klatt & Klatt, 1990) in turn limits their ability to capture many of the details of naturally occurring vocal sources (Bangayan, Long, Alwan, Kreiman, & Gerratt, 1997; Henrich, d'Alessandro, & Doval, 2001). Alternative synthesis approaches using vocoders (e.g., the Straight analysis–synthesis system; Kawahara, 1997) have the flexibility to capture fine acoustic details, but at the cost of a very large number of parameters that are hard to interpret acoustically, physiologically, or perceptually. The present article describes the UCLA voice synthesizer, a tool for analysis and synthesis of voice quality that addresses these theoretical and methodological concerns. This description presupposes a basic knowledge of the variables involved in speech analysis and synthesis and is aimed primarily at researchers

J. Kreiman, jkreiman@ucla.edu

studying voice, speech, and other vocal behavior. Extensive tutorial descriptions of these variables and functions are included in the user's manual that accompanies the software.

The synthesizer provides two approaches to parametric source modeling: a low-dimensional time-domain approach and a spectral-domain approach, whereby users can define as many or as few parameters as are needed to achieve the desired result. The latter functionality is, to our knowledge, unique among currently available voice synthesizers. Further, because the synthesizer operates in near real time, it can easily be applied in studies that assess the importance of specific acoustic parameters to perceived voice quality, either in method-of-adjustment tasks (Gerratt & Kreiman, 2001) or as a tool for creating series of stimuli for testing just-noticeable differences (JNDs). The software is provided as open-source freeware. C++ code, executable files, documentation (including tutorial material on voice acoustics, analysis, and synthesis), and two sample voices (one male and one female) are available for free download at the Psychonomic Society supplemental archive or at www.surgery.medsch.ucla.edu/glottalaffairs/.

OVERVIEW OF THE ANALYSIS–SYNTHESIS PROCESS

The UCLA voice synthesizer is a formant synthesizer, specialized for accurately modeling a wide range of voice qualities, including matching of personal quality. This synthesizer differs from other formant synthesizers in the precision with which the source can be modeled and in the integration of analysis and synthesis functions. It also differs from other synthesizers in that it is limited at present to modeling vowels with steady-state resonances, although details of the vocal source function and its instabilities can be modeled in great detail.

The synthesizer is based on the source/filter theory of speech production (Fant, 1960). In source/filter theory, vibrations of the vocal folds (the glottal source) excite the resonances of the vocal tract above the vocal folds (often modeled as an all-pole filter, shaping the input glottal source). The resulting voice signal is radiated outward from the lips (modeled by a differentiator, which increases the output sound energy level by 6 dB/octave). Following this model, synthesizing a voice requires an estimate of the glottal source and the vocal tract resonance frequencies (with associated bandwidths) as input. To recover the glottal source from a recorded audio signal, the vocal tract transfer function is canceled by applying an all-zero filter to the speech signal (the inverse of the vocal tract model; hence the name *inverse filtering*). This process removes the effects of the transfer function from the signal, leaving behind an estimate of the glottal-flow derivative. If the acoustic effects of sound radiation from the lips are also canceled by integration, an estimate of the actual glottal pulse shape is generated. These parameters are estimated in the inverse-filter program and then exported to the synthesizer program, as described below.

Technical Details

UCLA voice analysis and synthesis software comprises three separate but integrated programs: an interactive inverse filter, an interactive formant synthesizer with extensive source-manipulation capabilities, and an acoustic analysis program featuring a wider-than-usual selection of measures of voice. When a new case is inverse filtered, the software automatically estimates starting synthesis values for the source and vocal tract models and creates all of the needed input files and directories, along with the appropriate subdirectories for the voice in question. This integrated approach to analysis and synthesis distinguishes this software from other programs that parameterize the voice source but do not synthesize (e.g., TKK Aparat; Airas, 2008), or that synthesize but require the user to estimate and input parameter values (e.g., KLSYN; Klatt & Klatt, 1990).

The software operates under Windows (XP or later, although the C++ code could also be compiled to run under Mac OS or Linux) and requires a sound card. Only the inverse filter and voice synthesizer are described here. The acoustic analysis program is documented (along with additional details of the inverse filter and synthesizer) in the manual available on the software Web site.

During software development, a number of standard validation studies were undertaken to ensure accuracy. These included reverse-analyzing synthetic voice signals to recover the input source pulse shapes and spectra, formants, and bandwidths, $f0$ and amplitude contours, and noise spectra. When possible, output was compared with results from comparable algorithms in MATLAB. In several other studies, natural voice samples were copy synthesized to form the best possible matches to the original tokens. These synthetic samples were presented to listeners, who were unable to distinguish the natural and synthetic tokens in a same/different (A/X) task (Gerratt & Kreiman, 2001; Kreiman & Gerratt, 2005), confirming the validity of the analysis–synthesis process. Additional studies demonstrating the synthesizer are described in the Application Examples section at the end of this article.

THE ANALYSIS AND SYNTHESIS PROCESSES

Recording Voice Samples for Inverse Filtering

Voice recording for inverse filtering must preserve phase relationships among the different spectral components in order to recover the shape of the glottal pulse (or its derivative, referred to as the *flow derivative*) accurately in the time domain. Voices for our own studies are transduced with a 0.5-in. Brüel and Kjær condenser microphone (model 4193; Norcross, GA) and directly digitized. Signals are sampled at 20 kHz, with 16-bit resolution. They are subsequently downsampled to 10 kHz for analysis (all algorithms currently assume a 10-kHz sample rate). However, signals may also be recorded using a pneumotachographic mask and a differential pressure transducer, as described by Rothenberg (1973, 1977). This method has a poor high-frequency response, but provides

accurate information about low-frequency components of the airflow that arise when the glottis fails to close completely. In our experience, loss of high-frequency information from flow-mask data and resulting difficulties estimating vocal tract resonance frequencies have proven to be far more problematic than is contamination by low-frequency noise in microphone signals. For this reason, we will not discuss flow-mask data any further here, although procedures for analyzing such data are described in the software manual.

Inverse Filtering the Voice Signal

The inverse-filtering protocol uses the method described by Javkin, Antoñanzas-Barroso, and Maddieson (1987). Formant frequencies and bandwidths are estimated by using linear predictive coding analysis. Because a successful result depends mostly on correctly specifying values for formant frequencies and bandwidths, the inverse filter features an interactive process whereby users can manipulate resonances by clicking and dragging the peaks on the spectral display, by dragging sliding cursors to change resonance frequencies or bandwidths, or by typing new values into a table (Figure 1). Formants can also be added or deleted either by pointing to the desired location in the spectral display and clicking or by typing the desired values into an empty cell in the table. In all cases, the impact of changes on the recovered glottal pulse is apparent in real time as manipulations are made, so that it is easy to determine when the best possible outcome (usually defined as an output pulse with minimal residual formant ripple and a smoothly decreasing source spectrum conforming to theoretical expectations; Fant, 1979) has been achieved.

Figure 1 illustrates this process. Figure 1A shows preliminary analysis results based on automatic measures of formant frequencies and bandwidths. The trace labeled "A" represents the glottal waveform, "B" is the glottal-flow derivative, and "C" shows the spectrum of the flow derivative. This result is not very satisfactory: A large amount of ripple remains in the flow derivative, and the flow derivative spectrum does not decrease as smoothly as one would expect. An examination of the formant values in the top left panel of the figure ("D") shows the reason for this: A resonance has incorrectly been placed at 441 Hz, below the true first formant (*F*1) at 837 Hz. Excess ripple like this is nearly always caused by the presence of a spurious low-frequency resonance in the vocal tract model. Such large errors can also be identified based on phonetic knowledge (a typical *F*1 value for the vowel /a/ is about 700–800 Hz). In addition, a vowel synthesized with these formant frequencies will have an obviously incorrect vowel quality. When this resonance is deleted (by double right-clicking), the result improves dramatically (Figure 1B). Other examples of such manipulations are included in the software manual.

When the inverse filtering result is satisfactory, users can issue a "save" command, which creates the directories needed by the synthesizer and saves files containing the original voice sample and preliminary estimates of the source and vocal tract parameters. There is no way to know with certainty that the inverse-filtering process has recovered the "true" or "correct" shape of the glottal pulses, even when conventional criteria for success are met. However, the main purpose of this process is to generate preliminary parameter estimates for the synthesizer. In that context, accuracy (however defined) is not a primary concern. Pulse shapes can be manipulated in the synthesizer, so the synthetic voice sample matches the target voice in quality, but the relationship of these perceptually validated pulses to underlying physiological vocal function remains unknown.

Synthesizing a Voice

Overview. Figure 2A shows the synthesizer interface, and Figure 2B shows details of the toolbar. The process of generating the initial synthetic version of a voice is nearly automatic and takes less than 1 min. The first step is to smooth and parameterize the glottal waveform that has been imported from the inverse filter. This is accomplished by fitting an LF model to the output of the inverse filter ("A" in Figure 2A). Next, *f*0 is tracked on the natural voice sample, and the track is imported to model slow changes in *f*0 and amplitude (the pitch contour; vibrato or tremor). The spectral shape of the inharmonic part of the source (the noise excitation) is then estimated and modeled. Sliding cursors controlling these values are labeled "D" in Figure 2A. Finally, parameters are combined with an initial vocal tract model (imported from the inverse filter) to synthesize a preliminary version of the voice. The vocal tract model is labeled "B" in the figure, and the voice spectrum is labeled "C." All the synthesizer parameters can then be adjusted as needed to improve the acoustic and perceptual match of the synthetic token to the original voice. Each of these steps is described in more detail in the following sections.

Modeling the harmonic vocal source. Because accurate modeling of the voice source is an essential part of accounting for variations in voice quality (e.g., Ananthapadmanabha, 1984; Karlsson, 1991), the voice synthesis process begins by generating an estimate of the harmonic vocal source. If the goal is to model a specific voice, this can be accomplished through inverse filtering the target voice sample, as described in the previous section. A source can also be imported from outside the program (for example, as a list of values in a text editor). Alternatively, one of the supplied sample voices can be opened in the synthesizer, and its source can be edited (as described below) until it has the desired characteristics.

Despite an experimenter's best efforts at inverse filtering, the recovered glottal flow waveform often includes ripples, bumps, and other theoretically undesirable—but, in practice, unavoidable—features. One common approach to coping with this situation is to fit the output of the inverse filter with a theoretical model of the glottal flow pulse. In practice, substituting the modeled flow for the experimentally derived flow eliminates errors, wiggles, bumps, and excess high-frequency formant ripple and attendant high-frequency distortion, while preserv-

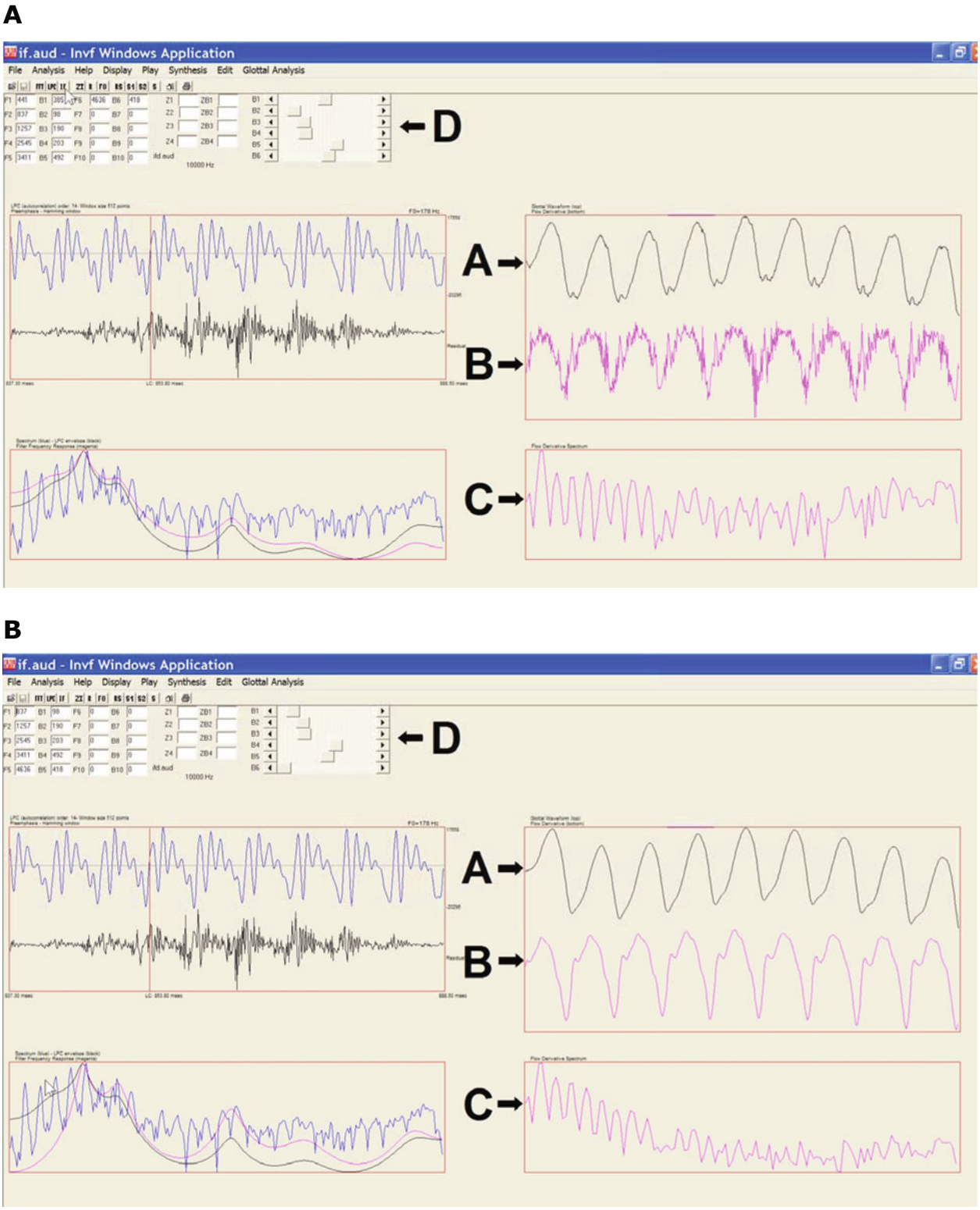
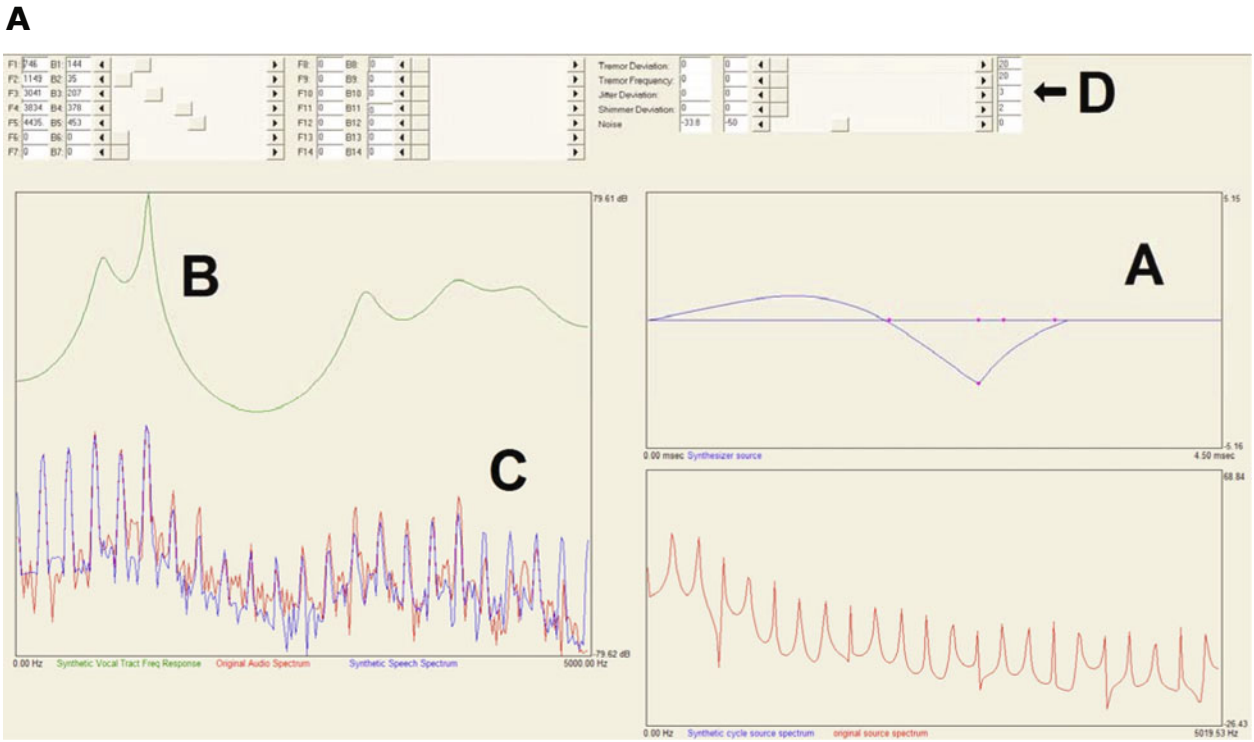


Figure 1. The inverse-filtering process. (A) Initial output of the analysis. (B) When the spurious low-frequency resonance is removed, the inverse-filtering result improves significantly. In each panel, the curve labeled "A" is the glottal waveform; "B" is the glottal-flow derivative; "C" shows the spectrum of the flow derivative; and "D" shows the resonance and bandwidth frequencies used to construct the filter.



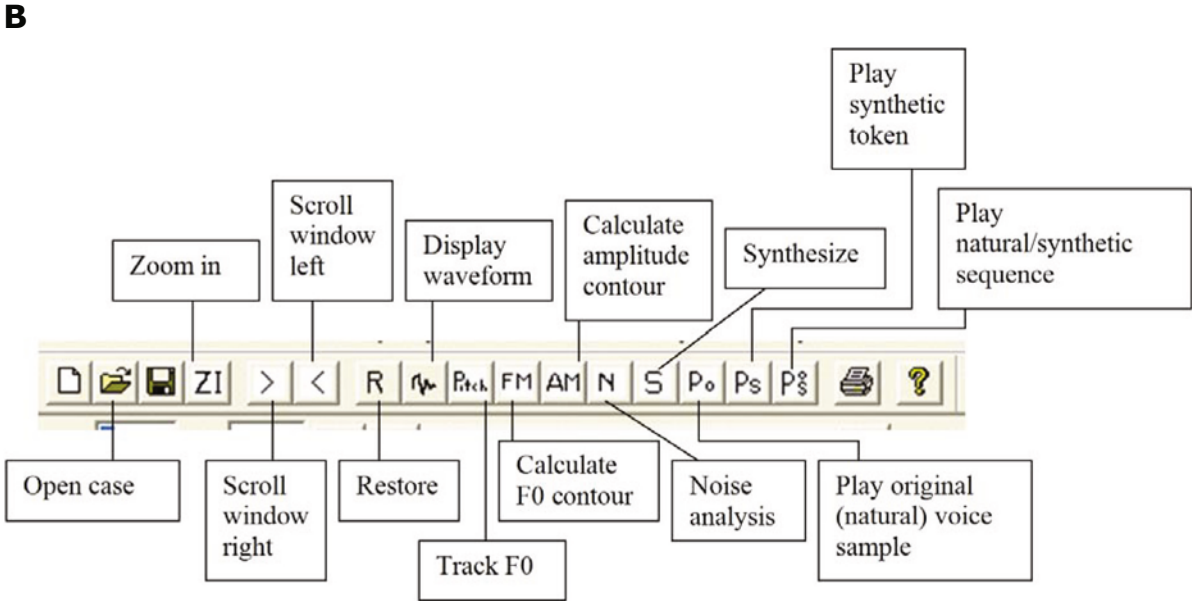


Figure 2. The synthesizer interface. (A) The work window. (B) Detail of the toolbar. Time-domain source manipulations are accomplished by clicking and dragging the different Liljencrants–Fant parameters, which are plotted with dots on the flow derivative pulses in the middle right frame in the synthesizer window (labeled "A" in Figure 2A). "B" shows the vocal tract model, and "C" shows the superimposed spectra of the natural and synthetic voices. Adjustments can be made to vocal tract resonances and bandwidths, the noise-to-signal ratio, jitter level, and/or shimmer level, either by sliding the appropriate cursor or by typing the desired value into the appropriate cell in the table, labeled "D." See the text for more discussion.

ing most of the important features of the pulse shapes. Experiments with synthetic voices have further shown that smoothing with a theoretical model increases the accuracy with which various parameters of the glottal source can be estimated (Strik, 1998).

Many time-domain source models have been proposed to model patterns of glottal opening and closing or to characterize changes in airflow through the glottis over time (for reviews, see Fujisaki & Ljungqvist, 1986; Ní Chasaide & Gobl, 1997). The present synthesizer uses a

slightly modified version of the popular LF model (Fant et al., 1985), which parameterizes the derivative of the glottal-flow waveform.[1] Modeling glottal pulses in the derivative rather than in the time domain has the advantage that acoustically important rapid changes in pulse shape around the time of glottal closure are emphasized. In particular, the rate and moment of glottal closure and the timing of peak airflow are easy to specify in this model. The LF model is fitted to the output of the inverse-filter by iterative least-squares minimization performed on major features of the time-domain LF curve (Figure 3). The spectrum of the LF-fitted pulse is calculated and displayed (shown as "A" in Figure 3), along with the raw output of the inverse filter and the LF-fitted pulse (shown as "B"). These data serve as starting values for source modeling.

Modeling the inharmonic voice source. Because the human voice is not perfectly periodic, accurately modeling the voicing source requires modeling of both harmonic (periodic) and inharmonic (noise) components of the glottal excitation. Noise components contribute substantially to acoustic excitation of the vocal tract and are an important part of a complete source model.

Traditionally, two sources of spectral noise have been distinguished. The first is noise related to irregularities in vocal-fold vibration (jitter and shimmer, representing random variability in the period and amplitude of glottal pulses, respectively; see, e.g., Baken, 1987, for a review). Noise also emerges due to turbulence generated during the open phase of the glottal cycle and/or flow through a persistent glottal gap (especially for female and pathological voices). Noise is often measured separately from jitter and shimmer with a harmonics-to-noise ratio or with a bundle of measures representing the relative noise components in different frequency bands (e.g., Michaelis, Gramss, & Strube, 1997; Yumoto, Gould, & Baer, 1982).

In the synthesizer, the inharmonic part of the source (the noise excitation) is estimated through application of a cepstral-domain comb filter (i.e., a *lifter*, like that described by de Krom, 1993; see also Qi & Hillman, 1997). Cepstral analysis is performed on a 204.8-msec segment of the original voice sample. The *f*0 is estimated using an algorithm based on Pearson correlations between successive cycles and is used to construct a lifter to remove the "rahmonics" (the cepstral-domain equivalent of harmonics). This process filters out the periodic energy in the voice, leaving only the noise. The residual signal is transformed back into the frequency domain, producing the spectrum of the noise component of the voice. Finally, the estimated noise spectrum is smoothed and fitted with a piecewise linear approximation, the number of pieces being specified by the user. A 100-tap finite impulse response filter is synthesized for the noise spectrum, through which white noise is passed to create a spectrally shaped noise time series. It is also possible to import a specific desired noise spectrum, or to copy the noise spectrum from one voice to another. Other currently available synthesizers do not allow users similar control over the noise spectrum. For example, KLSYN models the inharmonic source via the parameter AH, which controls the level of spectrally flat noise added to the harmonic source.

Jitter (in percent) is modeled using a shape-preserving interpolation algorithm that expands or contracts the glot-

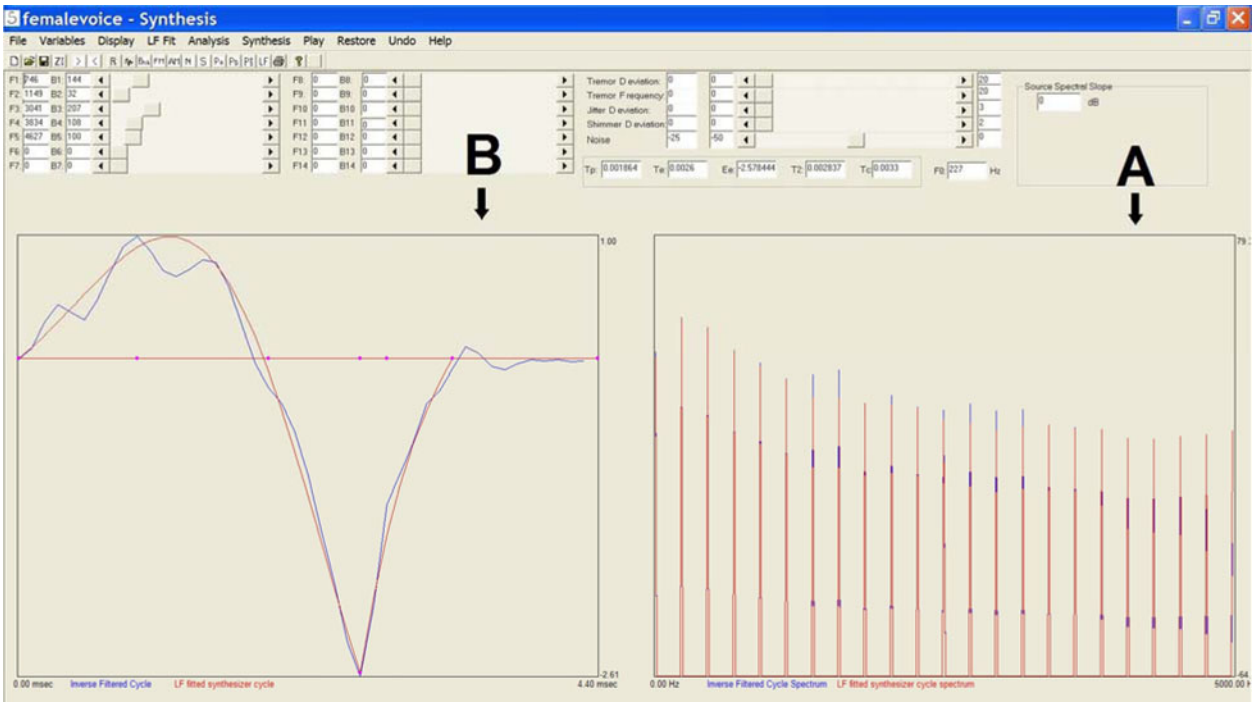


Figure 3. The Liljencrants–Fant (LF) model fitting process. The spectrum of the LF-fitted pulse is calculated and displayed (shown as "A") along with the raw output of the inverse filter and the smooth, LF-fitted pulse ("B").

tal pulses to obtain the desired amount of random frequency perturbation. The shimmer parameter varies the amplitude of each glottal pulse by applying a random gain to obtain the specified level of random amplitude perturbation (in dB).

Modeling intonation and loudness contours. The third step in voice modeling is assessment of long-term patterns of *f*0 and amplitude modulation (i.e., pitch and loudness variability). Several approaches are available. *f*0 and amplitude can be tracked within the synthesizer, and the contours (labeled "A" and "B" in Figure 4, respectively) can be smoothed to a degree specified by the user. In most cases, this provides a very good match to the overall prosodic shape of the natural voice sample. Alternatively, pitch and/or amplitude tracks can be imported from outside the synthesizer to create a specific desired contour, or users can model the contours using two synthetic tremor models (one that models sinusoidal modulations and one that provides more irregular pitch contours).

Modeling the vocal tract filter. Finally, users model the vocal tract response by specifying formant frequencies and bandwidths. Initial values are usually imported from the inverse filter or from a file created in a text editor.

Synthesizing the voice. Because the synthesizer sampling rate is currently fixed at 10 kHz, the following procedure is applied to overcome quantization limits on modeling *f*0. First, a plot of *f*0 versus time is generated for the duration of the 1-sec token to be synthesized, taking into account the desired pitch contour, tremor, jitter, and shimmer. For each successive glottal pulse, *f*0 is determined by interpolating this curve at the starting instant for that pulse. A glottal pulse is then generated and upsampled by a factor chosen by the user; the default is 4, resulting in a sampling rate of 40 kHz. This pulse is stretched or compressed according to the *f*0 desired for that cycle. The process repeats until the complete *f*0 time series has been calculated, at which point it is downsampled to 10 kHz. (Pulses can also be downsampled, one by one, when it is important to preserve the exact location of LF features; see the manual for details.) The overall effect is equivalent to digitizing an analog pulse train with pulses of the exact desired frequencies at the fixed 10k sample rate.

The LF pulse train is added to the noise time series to create a complete glottal source time series. The ratio of noise to LF energy (the noise-to-signal ratio) is set to the default value of −25 dB. Finally, the complete synthesized source is filtered through the current vocal tract model to generate a preliminary version of the synthetic voice. At any point in this process, the synthetic and/or natural stimuli can be played individually or as a pair by clicking the appropriate "play" button (Figure 2B).

MODIFYING THE SYNTHETIC VOICES

After a synthetic voice has been created (or a case opened), the operator is able to adjust all parameters to modify the voice as desired. These modifications can serve a number of purposes, including creation of stimuli to test listeners' sensitivity to different acoustic parameters and their assessment of the accuracy and validity of acoustic measurement techniques. Some of these applications are described later in the present article.

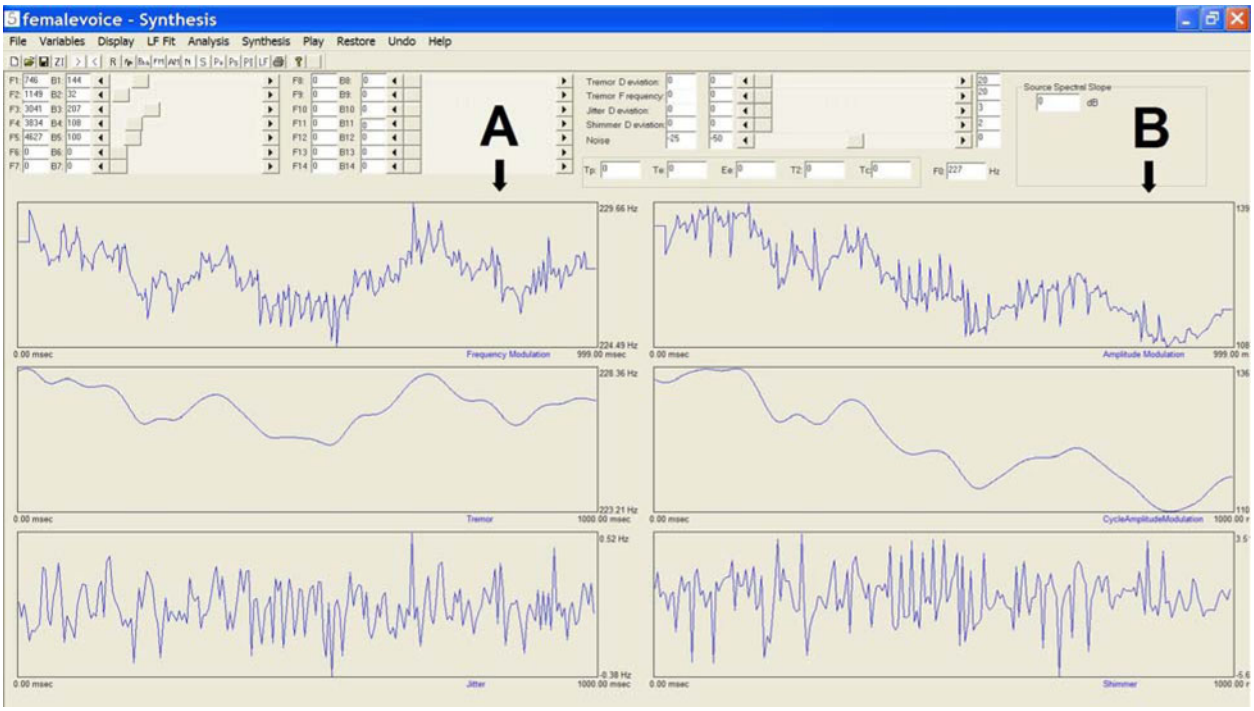


Figure 4. Modeling *f*0 ("A") and amplitude ("B") contours. The top frames show the unsmoothed output of the pitch and amplitude trackers, the middle frames show the smoothed pitch and amplitude tracks, and the bottom frame shows the residuals.

Modifying the Vocal Source

The synthesizer allows the source to be manipulated in both the time and spectral domains. Time-domain manipulations are accomplished by clicking and dragging the different LF parameters, which are plotted with dots on the flow-derivative pulses in the top right frame in the synthesizer window (labeled "A" in Figure 2A). However, it is often desirable to edit the harmonic voice source in the spectral domain as well as (or instead of) in the time domain. Although theoretical correspondences between LF model parameters and details of source spectral shape have been investigated (Fant, 1995; Ní Chasaide & Gobl, 1997), in practice, it is quite difficult to manipulate combinations of LF parameters in the time domain to achieve a specific desired spectral change. Further, although the LF model can accommodate a wide variety of source functions, it is not so flexible that it can model the complete range of glottal pulse shapes that occur even among normal speakers (Henrich et al., 2001; Shue, Kreiman, & Alwan, 2009). For these reasons, the synthesizer permits the source to be modified in the frequency (spectral) domain as well as in the time domain. This increased flexibility with respect to source shapes greatly improves accuracy and ease of modeling voices, particularly when pathology is present, and is a major feature that differentiates this synthesizer from others that are currently available.

Figure 5 shows the procedure for manipulating the voice source in the spectral domain. The amplitude or slope of an individual harmonic or a group of harmonics is manipulated by first selecting the desired harmonics, which are then connected by a line segment (indicated by an arrow in panel A of Figure 5). Users then click and drag the line segment to raise or lower either endpoint, thus changing the slope of that group of harmonics. The resulting increase in the amplitude of the first harmonic in this example is indicated by an arrow in panel B. The slope of the line segment (in dB) will appear in the upper right corner of the synthesis window (labeled "C" in Figure 5). Alternatively, it is possible to change the slope of a segment to obtain a specific target value by typing the desired value into cell C and then clicking the line segment. Users can also raise or lower the segment as a whole to change the absolute amplitudes of all of the selected harmonics, while preserving their relative amplitudes. The same procedures apply when manipulating any number of harmonics or the entire spectral slope at once. Any number of seg-

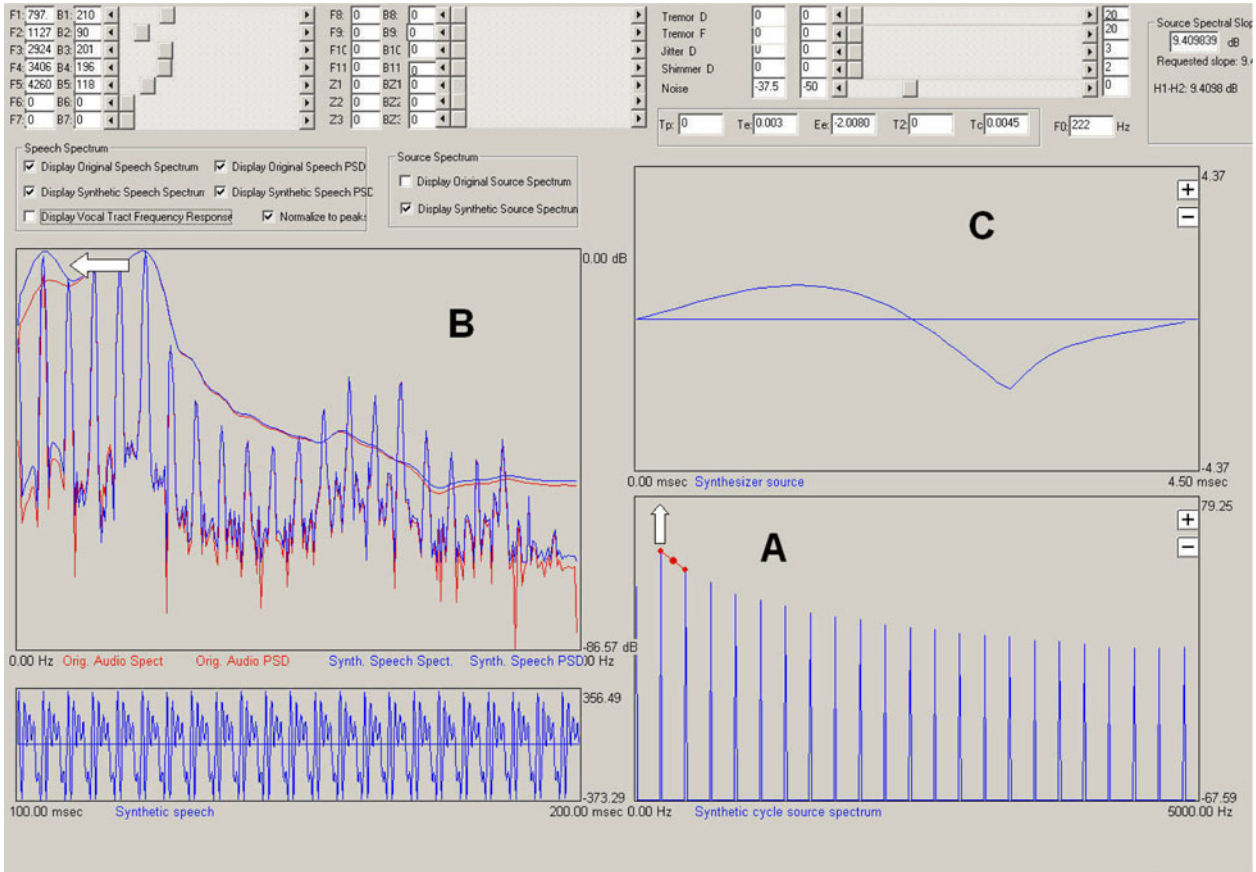


Figure 5. Manipulating the voice source in the spectral domain. In this example, the first two harmonics have been selected, as indicated by the line segment in panel A, and the amplitude of the first harmonic has been increased, as shown by the arrow. The increase in harmonic amplitude is indicated by a second arrow in panel B. Panel C shows the derivative of the glottal pulse corresponding to the spectrum in A. The slope of the line segment (in this case, 9.409 dB) appears in the cell labeled "Source Spectral Slope" in the uppermost right corner of the synthesis window.

ments can be created, including any number of harmonics, allowing precise control over the spectral details of the source excitation. To our knowledge, other synthesizers do not allow such precise modification of source spectra. For example, the TL parameter in KLSYN alters the overall spectral slope in units of dB down from 3 kHz.

Modifying the Inharmonic Voice Source

Adjustments can be made to the noise-to-signal ratio, jitter level, and shimmer level, either by sliding the appropriate cursor or by typing the desired value into the appropriate cell (labeled "D" in Figure 2A). It is also possible to change the amount of smoothing applied to the noise spectrum by adjusting the number of segments used in the piecewise approximation fit to the noise spectrum (from 1, which gives a flat noise spectrum, to an arbitrarily large number; the default value of 25 provides a perceptually good match to the target voice in most cases).

Modifying the Vocal Tract Filter

Vocal tract resonances or spectral zeros in the synthetic voice can be added, deleted, or moved, using the techniques described for the inverse filter. Bandwidths can be adjusted by dragging the appropriate sliding cursor or by typing the desired value into the appropriate cell.

Playing and Saving the Synthetic Tokens

At any point in the synthesis process, the user can play the synthetic speech, the original (target) voice sample, or the two together in the sequence original–original–synthetic–synthetic, by clicking the appropriate icon (Figure 2B). Synthesizer parameters and/or synthetic voice samples can be saved jointly or individually. All files are in ASCII format and can be imported into graphics, statistics, or word processor programs for manipulation or visualization. They can also be reopened in the synthesizer for additional editing, as desired, and they can be converted from ASCII to .wav format by using the integrated voice analysis package. This process is illustrated in detail in the software manual.

Application Examples

Example 1: Testing hypotheses about voice quality perception. As noted in the introduction, most evidence concerning relationships between acoustic parameters and perceived voice quality is correlational in nature. Lack of a theory describing how listeners perceive voice quality is partially to blame for this situation, but lack of a method for assessing the perceptual salience of specific acoustic parameters independent of other parameters has also impeded theoretical development. Because the voice synthesizer allows manipulation of parameters in discrete steps of a specific size, it is ideal for creating stimuli for use in studies examining the perceptual importance of specific acoustic parameters, thus addressing both of these concerns. It is impossible for speakers to spontaneously produce such stimuli, because they cannot vary individual vocal parameters with any precision or without covarying something else; for example, untrained speakers are generally unable to modify pitch without changing loudness at the same time, or modify loudness without changing pitch.

The voice synthesizer provides two options for studying voice quality experimentally, both of which can provide direct evidence of listeners' sensitivity to any parameters under study. First, users can create stimuli by systematically varying a parameter or parameters, saving the resulting audio files, and then using these sounds to test listeners' sensitivity, to compare the perceptibility of an attribute in different contexts, and so on. Eight sets of such stimuli (four representing a male talker, four representing a female talker) are included in the online supplementary materials database. Details of these stimuli are given in Table 1. For each sex, the first series demonstrates changes in relative amplitudes of the first two harmonics (H_1–H_2; four steps); the second demonstrates changes in the $f0$ contour (four different contours); the third demonstrates changes in the overall slope of the harmonic source spectrum (four steps); and the fourth demonstrates changes in the noise-to-harmonics ratio (four steps).

We have undertaken several experiments using this approach. For example, we asked native speakers of Gujarati (an Indo-Aryan language that uses such phonatory contrasts to signal different word meanings; Fischer-Jörgensen, 1967), Thai (which contrasts lexical tones, but not phonation types), Mandarin Chinese (which contrasts tones that are also characterized by allophonic changes in phonation; Belotel-Grenié & Grenié, 2004), and English (which uses neither contrast) to distinguish stimuli that differed only in H_1–H_2 (Kreiman, Gerratt, & Khan, 2010). We found that Gujarati and Mandarin speakers are significantly more sensitive to these contrasts than are either English or Thai speakers, indicating that learning during language acquisition affects perceptual sensitivity with re-

Table 1
Synthetic Voice Stimuli Created by Varying Individual Synthesizer Parameters

Stimulus Series	File Names	Steps
H_1–H_2	H1h2male.wav H1h2female.wav	0, 5, 10, 15 dB
$f0$ contour	F0male.wav F0female.wav	natural contour, flat contour, sine wave modulation, irregular modulation
Spectral slope	Slopemale.wav Slopefemale.wav	12, 9, 6, 3 dB/octave
Noise-to-harmonics ratio	NHRmale.wav NHRfemale.wav	−40, −30, −25, −20 dB

Note—All parameters except those indicated have been held constant across all stimuli.

spect to voice quality. However, extra attention to *f*0 alone (as occurs in Thai) is not sufficient to confer such benefit, despite the fact that *f*0 is the time-domain equivalent of H_1, and speaking a tone language has been shown to affect the central and peripheral processing of *f*0 (e.g., Krishnan & Gandour, 2009).

Second, the synthesizer can be used to examine quality using a method of adjustment task. In this application, listeners can manipulate a given parameter directly (e.g., by moving a sliding cursor) and can then immediately compare the synthetic token to the target voice (which can be synthetic or natural), after which they may readjust the parameter until tokens match or until they just hear a difference between the tokens. In a study using this approach, we were able to demonstrate that listeners' assessment of the extent of noise in a voice depends solely on the noise-to-harmonics ratio and not on levels of jitter and shimmer, which do not appear to be perceptually salient, apart from the overall noise level (Kreiman & Gerratt, 2005).

Example 2: Validating acoustic measurements using analysis by synthesis. A second application of the voice analysis–synthesis system is in assessing the accuracy and validity of acoustic measures of voice. By synthesizing a voice using a set of measured parameters, it is possible to determine the accuracy of those measurements by comparing the quality of the synthetic and target natural tokens. For example, when *f*0 is high and the frequency of *F*1 is low (as in the vowel /i/ spoken by a woman or a child), many automatic formant-tracking algorithms will mistakenly identify the first harmonic (H_1) as *F*1, leading to large errors in formant frequency measurements. Such errors are also common when the voice source is quasi-sinusoidal or when the glottis fails to close completely during phonation, because these factors increase the prominence of the first harmonic in the voice spectrum (Shue et al., 2009). Because vocal tract resonances affect the amplitudes of harmonics near a resonance peak, errors in formant estimation can also have an impact on measurement of harmonic amplitudes, thus adding error to parameters such as H_1^*–H_2^* (the relative amplitudes of the first two harmonics, corrected for the influence of vocal tract resonances; Hanson, 1997; Iseli & Alwan, 2004) or H_1^*–A_3 (the corrected amplitude of H_1, relative to the amplitude of the third formant). These measures are widely used as indices of phonemic voice quality contrasts (e.g., Andruski & Ratliff, 2000; Gordon & Ladefoged, 2001; Huffman, 1987) and in studies of prosody (Strik & Boves, 1992), so issues of measurement accuracy have significant cross-disciplinary importance.

After inverse filtering, the vowel token of interest to generate initial parameter estimates, the perceptual validation process begins with adjustment of formant frequencies in the synthesizer, so that the quality of the synthetic vowel perceptually matches that of the target token. Because JNDs for formant frequencies are small (on the order of 10–30 Hz; Stevens, 1998), large errors like those that occur when H_1 is mistakenly identified as *F*1 are obvious and easy to detect by ear—the synthetic copy sounds like the wrong vowel. After formant frequencies have been corrected, bandwidths can be adjusted. Selection of the correct bandwidth is an important step, because bandwidths directly influence the amplitude of harmonics near the associated resonance. Because of this interdependency between bandwidths and harmonic amplitudes in the synthesis process, it is impossible to unambiguously determine values for either parameter, unless one is constrained. Fortunately, empirical data exist describing the relationships across *f*0, formant frequencies, and bandwidths (Hawks & Miller, 1995). Values for bandwidths are calculated by using the procedure described by Hawks and Miller, after which harmonic amplitudes are manipulated in the synthesizer until both synthetic and natural spectra and voice qualities match precisely. At this point, harmonic amplitudes are both perceptually and acoustically correct and can be measured directly from the synthetic source spectrum, as described above. (See Kreiman, Gerratt, Iseli, et al., 2008, for more discussion.) In our experience, perceptually corrected values for H_1–H_2 can differ by as much as 27 dB from those calculated without perceptual validation. These values greatly exceed estimated JNDs for H_1–H_2 of about 2.1–3.9 dB (Kreiman & Gerratt, 2010).

Limitations and Conclusions

Synthesizer development is ongoing, and new releases are planned to address current limitations. The version described here is available in the Psychonomic Society supplemental archive, and the most current release version can be downloaded from the software Web site. The software currently does not allow users to vary formant frequencies over time, which limits its usefulness for modeling connected speech or phonation contrasts in natural languages. However, interactive time-variant vocal tract modeling is under development and will be included in the next release version. The sampling rate is also currently restricted to 10 kHz. Although this is adequate to capture most important voice-quality details, future implementations that can accommodate higher sampling rates will produce an even better quality synthesis. Current plans also include developing a strategy for synthesizing period-doubled (subharmonic) phonation. This strategy will be based on acoustic modeling of a large sample of such phonation, and we hope it will provide a better match to the quality of these voices than does the DI parameter in KLSYN.

Despite present limitations, the UCLA voice analysis and synthesis programs provide experimenters with a useful tool for creating and modeling voice signals. Natural voice samples can be examined by using the inverse-filtering software; supplied sample voices can be modified slightly or drastically in the synthesizer; and parameter files can be exported, manipulated, and reimported, or they can be created completely from scratch. A user-friendly interface makes it possible to create and export stimuli or to implement method-of-adjustment tasks. Because users can quantify the vocal source by using as many or as few parameters as they desire, the synthesizer is not subject to limitations with respect to the glottal

pulse shapes it can mimic, a significant innovation. Both source and vocal tract parameters can be modeled individually or in concert with other parameters to examine interactions between acoustic attributes in determining perceived quality. Finally, because the software is available as open-source freeware, users can add or modify any features in any manner they desire. As a result of its power and flexibility, the synthesizer package offers the opportunity to test many hypotheses about the acoustic basis of voice quality perception, using experimental, rather than correlational, approaches.

AUTHOR NOTE

The UCLA voice synthesizer and the inverse-filtering programs were written by N.A.-B., with additional programming support from Brian Gabelman and Diane Budzik. Development of this software was supported by Grant DC01797 from the National Institute on Deafness and Other Communication Disorders. Address correspondence concerning this article to J. Kreiman, Division of Head/Neck Surgery, UCLA School of Medicine, 31-24 Rehab Center, Los Angeles, CA 90095-1794 (e-mail: jkreiman@ucla.edu).

REFERENCES

AIRAS, M. (2008). TKK Aparat: An environment for voice inverse filtering and parameterization. *Logopedics Phoniatrics Vocology*, **33**, 49-64. doi:10.1080/14015430701855333

ANANTHAPADMANABHA, T. V. (1984). Acoustic analysis of voice source dynamics. *Speech Transmission Laboratory Quarterly Progress & Status Report*, **25**(2-3), 1-24.

ANDRUSKI, J., & RATLIFF, M. (2000). Phonation types in production of phonological tone: The case of Green Mong. *Journal of the International Phonetic Association*, **30**, 37-61.

BAKEN, R. J. (1987). *Clinical measurement of speech and voice*. Boston: College Hill.

BANGAYAN, P., LONG, C., ALWAN, A. A., KREIMAN, J., & GERRATT, B. R. (1997). Analysis by synthesis of pathological voices using the Klatt synthesizer. *Speech Communication*, **22**, 343-368. doi:10.1016/S0167-6393(97)00032-0

BELOTEL-GRENIÉ, A., & GRENIÉ, M. (2004). The creaky voice phonation and the organisation of Chinese discourse. *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages* (ISCA, Beijing, China), pp. 5-8.

COLTON, R. H., CASPER, J. K., & LEONARD, R. (2005). *Understanding voice problems: A physiological perspective for diagnosis and treatment*. Baltimore: Lippincott Williams & Wilkins.

CUTLER, A., DAHAN, D., & VAN DONSELAAR, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language & Speech*, **40**, 141-201.

DE KROM, G. (1993). A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech & Hearing Research*, **36**, 254-266.

FANT, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton.

FANT, G. (1979). Glottal source and excitation analysis. *Speech Transmission Laboratory Quarterly Progress & Status Report*, **20**(1), 85-107.

FANT, G. (1995). The LF-model revisited. Transformations and frequency domain analysis. *Speech Transmission Laboratory Quarterly Progress & Status Report*, **36**(2-3), 119-156.

FANT, G., LILJENCRANTS, J., & LIN, Q. (1985). A four-parameter model of glottal flow. *Speech Transmission Laboratory Quarterly Progress & Status Report*, **26**(4), 1-13.

FISCHER-JÖRGENSEN, E. (1967). Phonetic analysis of breathy (murmured) vowels in Gujarati. *Indian Linguistics*, **28**, 71-139.

FITCH, W. T., NEUBAUER, J., & HERZEL, H. (2002). Calls out of chaos: The adaptive significance of nonlinear phenomena in mammalian vocal production. *Animal Behaviour*, **63**, 407-418. doi:10.1006/anbe.2001.1912

FUJISAKI, H., & LJUNGQVIST, M. (1986). Proposal and evaluation of models for the glottal source waveform. *Proceedings of the IEEE International Conference on Acoustical Speech Signal Processing*, 1605-1608.

GERRATT, B. R., & KREIMAN, J. (2001). Measuring vocal quality with speech synthesis. *Journal of the Acoustical Society of America*, **110**, 2560-2566. doi:10.1121/1.1409969

GERRATT, B. R., TILL, J., ROSENBEK, J. C., WERTZ, R. T., & BOYSEN, A. E. (1991). Use and perceived value of perceptual and instrumental measures in dysarthria management. In C. A. Moore, K. M. Yorkston, & D. R. Beukelman (Eds.), *Dysarthria and apraxia of speech: Perspectives on management* (pp. 77-93). Baltimore: Brookes.

GOLDINGER, S. D., PISONI, D. B., & LOGAN, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **17**, 152-162. doi:10.1037/0278-7393.17.1.152

GORDON, M., & LADEFOGED, P. (2001). Phonation types: A cross-linguistic overview. *Journal of Phonetics*, **29**, 383-406.

HANSON, H. M. (1997). Glottal characteristics of female speakers: Acoustic correlates. *Journal of the Acoustical Society of America*, **101**, 466-481. doi:10.1121/1.417991

HAWKS, J. W., & MILLER, J. D. (1995). A formant bandwidth estimation procedure for vowel synthesis. *Journal of the Acoustical Society of America*, **97**, 1343-1344.

HENRICH, N., D'ALESSANDRO, C., & DOVAL, B. (2001). Spectral correlates of voice open quotient and glottal flow asymmetry: Theory, limits and experimental data. *EUROSPEECH 2001*, 47-51.

HUFFMAN, M. K. (1987). Measures of phonation type in Hmong. *Journal of the Acoustical Society of America*, **81**, 495-504.

ISELI, M., & ALWAN, A. (2004). An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation. In *Acoustics, Speech, and Signal Processing (ICASSP 2004 Proceedings)*, pp. 669-672.

JAVKIN, H. R., ANTOÑANZAS-BARROSO, N., & MADDIESON, I. (1987). Digital inverse filtering for linguistic research. *Journal of Speech & Hearing Research*, **30**, 122-129.

KARLSSON, I. (1991). Female voices in speech synthesis. *Journal of Phonetics*, **19**, 111-120.

KAWAHARA, H. (1997). Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited. *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, **2**, 1303-1306.

KEMPSTER, G. B., GERRATT, B. R., VERDOLINI ABBOTT, K., BARKMEIER-KRAEMER, J., & HILLMAN, R. (2009). Consensus auditory–perceptual evaluation of voice: Development of a standardized clinical protocol. *American Journal of Speech–Language Pathology*, **18**, 124-132. doi:10.1044/1058-0360(2008/08-0017)

KLATT, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, **67**, 971-975.

KLATT, D. H., & KLATT, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, **87**, 820-857. doi:10.1121/1.398894

KREIMAN, J., & GERRATT, B. R. (2005). Perception of aperiodicity in pathological voice. *Journal of the Acoustical Society of America*, **117**, 2201-2211.

KREIMAN, J., & GERRATT, B. R. (2010). Perceptual sensitivity to first harmonic amplitude in the voice source. *Journal of the Acoustical Society of America*, **128**, 2085-2089.

KREIMAN, J., GERRATT, B. R., ISELI, M., NEUBAUER, J., SHUE, Y.-L., & ALWAN, A. (2008, August). The relationship between open quotient and H_1^*–H_2^*. In *Proceedings of the 6th International Conference on Voice Physiology & Biomechanics*. Tampere, Finland.

KREIMAN, J., GERRATT, B. R., KEMPSTER, G. B., ERMAN, A., & BERKE, G. S. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech & Hearing Research*, **36**, 21-40.

KREIMAN, J., GERRATT, B. R., & KHAN, S. D. (2010). Effects of native language on perception of voice quality. *Journal of Phonetics*, **38**, 588-593. doi:10.1016/j.wocn.2010.08.004

KRISHNAN, A., & GANDOUR, J. T. (2009). The role of the auditory brainstem in processing linguistically relevant pitch patterns. *Brain & Language*, **110**, 135-148. doi:10.1016/j.bandl.2009.03.005

MICHAELIS, D., GRAMSS, T., & STRUBE, H. W. (1997). Glottal-to-noise

excitation ratio: A new measure for describing pathological voices. *Acustica*, **83**, 700-706.

NÍ CHASAIDE, A., & GOBL, C. (1997). Voice source variation. In W. J. Hardcastle & J. Laver (Eds.), *The handbook of phonetic sciences* (pp. 427-461). Oxford: Blackwell.

NYGAARD, L. C., & PISONI, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, **60**, 355-376.

QI, Y., & HILLMAN, R. E. (1997). Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals. *Journal of the Acoustical Society of America*, **102**, 537-543.

ROTHENBERG, M. (1973). A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. *Journal of the Acoustical Society of America*, **53**, 1632-1645. doi:10.1121/1.1913513

ROTHENBERG, M. (1977). Measurement of airflow in speech. *Journal of Speech & Hearing Research*, **20**, 155-176.

SHRIVASTAV, R. (2003). The use of an auditory model in predicting perceptual ratings of breathy voice quality. *Journal of Voice*, **17**, 502-512. doi:10.1067/S0892-1997(03)00077-8

SHUE, Y.-L., KREIMAN, J., & ALWAN, A. (2009). A novel codebook search technique for estimating the open quotient. *Interspeech 2009*, 2895-2898.

STEVENS, K. N. (1998). *Acoustic phonetics*. Cambridge, MA: MIT Press.

STRIK, H. (1998). Automatic parameterization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses. *Journal of the Acoustical Society of America*, **103**, 2659-2669.

STRIK, H., & BOVES, L. (1992). On the relation between voice source parameters and prosodic features in connected speech. *Speech Communication*, **11**, 167-174.

VOLODINA, E. V., VOLODIN, I. A., ISAEVA, I. V., & UNCK, C. (2006). Biphonation may function to enhance individual recognition in the dhole, Cuon alpinus II. *Ethology,* **112**, 815-825. doi:10.1111/j.1439-0310.2006.01231.x

YUMOTO, E., GOULD, W. J., & BAER, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *Journal of the Acoustical Society of America*, **71**, 1544-1550.

NOTE

1. The LF model is composed of two segments: the product of a growing exponential and a sinusoid, followed by a decaying exponential. In our modifications, a linear term was added to the second (exponential) segment of the model to force the flow back to 0 at the end of each glottal cycle. This has the effect of flattening out this segment of the model somewhat relative to the original LF model, but improves fit to many pathological voices. Second, the present implementation of the LF model abandons the original constraint that the areas under each segment of the model be equal. The instant of glottal closure (model point t_c) is not constrained to equal the beginning instant for the following cycle (point t_0) in our implementation, so that glottal closure is formally modeled in our implementation. (The original LF model saves a parameter by assuming that the flow returns to 0 at the end of each pulse, which does not routinely occur when modeling pathological phonation.) Finally, point t_a (which controls the manner in which the pulse returns to 0 during closure) was replaced by point t_2, defined as the time increment to 50% decay during glottal closure.

SUPPLEMENTAL MATERIALS

C++ code for the voice synthesizer and inverse filter, the user's manual, and audio files may be downloaded from http://brm.psychonomic-journals.org/content/supplemental. Future updates may be downloaded from www.surgery.medsch.ucla.edu/glottalaffairs/.

(Manuscript received February 23, 2010;
revision accepted for publication June 23, 2010.)