Behavior Research Methods, Instruments, & Computers
1994, 26 (4), 431-436

Track-Draw: A graphical interface for controlling the parameters of a speech synthesizer

PETER ASSMANN, WILL BALLARD, LAURIE BORNSTEIN, and DWAYNE PASCHALL
University of Texas at Dallas, Richardson, Texas

In this report we describe a graphical interface for generating voiced speech using a frequency-domain implementation of the Klatt (1980) cascade formant synthesizer. The input to the synthesizer is a set of parameter vectors, called *tracks*, which specify the overall amplitude, fundamental frequency, formant frequencies, and formant bandwidths at specified time intervals. Tracks are drawn with the aid of a computer mouse that can be used either in *point-draw* mode, which selects a parameter value for a single time frame, or in *line-draw* mode, which uses piecewise linear interpolation to connect two user-selected endpoints. Three versions of the program are described: (1) SYNTH draws tracks on an empty time-frequency grid, (2) SPECSYNTH creates a spectrogram of a recorded signal upon which tracks can be superimposed, and (3) SWSYNTH is similar to SPECSYNTH, except that it generates *sine-wave speech* (Remez, Rubin, Pisoni, & Carrell, 1981) using a set of time-varying sinusoids rather than cascaded formants. The program is written for MATLAB, an interactive computing environment for matrix computation. Track-Draw provides a useful tool for investigating the perceptually salient properties of voiced speech and other sounds.

Digital formant synthesizers provide an important tool for speech researchers concerned with identifying the information-bearing components of speech sounds and evaluating models of speech perception. In this paper we describe a new graphical interface for synthesizing voiced speech and speech-like sounds. The synthesis procedure is based on a frequency-domain implementation of the cascade formant model of speech, described by Klatt (1980). At present it is restricted to voiced speech; that is, it can generate approximations to speech sounds produced with quasi-periodic vibrations of the vocal folds (e.g., "ee"). It does not have the capability to generate voiceless sounds, such as fricatives (e.g., "sh").

Several versions of Klatt's (1980) cascade/parallel formant speech synthesizer have been developed and are available commercially (e.g., Bickley, Stevens, & Williams, 1992; Jamieson, Ramji, Kheirallah, & Nearey, 1993). The Track-Draw package provides three complementary features. First, it provides a simple graphical technique for entering a set of parameter vectors (tracks) that serve as input to the synthesizer. Second, it has a *copy synthesis* procedure, which enables the user to draw a set of formant tracks on the spectrogram of a recorded speech sample, synthesize the pattern, and compare the synthesized version with the original. Third, it includes a *sine-wave synthesis* procedure, which replaces the time-varying formant pattern of natural speech with a set of frequency-modulated sinusoids (Remez, Rubin, Pisoni, & Carrell, 1981).

The Track-Draw package contains three programs dedicated to specific tasks in speech synthesis. The programs are interactive, providing options to (1) draw a new track, (2) alter an existing track, (3) synthesize the signal, or (4) play back the synthesized signal for evaluation. The user can cycle through these options in any order to modify and update synthesis parameters until a satisfactory approximation to the desired signal is obtained.

The main program is called SYNTH. This program can be used to synthesize voiced speech and other complex sounds with the help of a very straightforward drawing tool. In addition to the basic drawing/synthesis mode provided by SYNTH, there are two programs for copy synthesis that allow parameter tracks to be superimposed on the spectrogram of a sample of recorded speech or other sounds. The first is called SPECSYNTH. Like SYNTH, it uses a frequency-domain implementation of a cascade formant synthesizer. The second, SWSYNTH, uses sine-wave synthesis. Each of the programs is described in detail below.

SYNTHESIS PROCEDURE

The synthesis method used by SYNTH and SPECSYNTH is a frequency-domain implementation of the Klatt cascade formant synthesizer (Klatt, 1980, 1982). The synthesizer generates voiced speech by digital summation of a set of harmonics whose frequencies are integer multiples of the fundamental frequency (F_0), and whose amplitudes and phases are computed according to

The contributions of Terry Nearey and Chris Darwin to earlier versions of these programs are gratefully acknowledged. We thank Alice O'Toole and two reviewers for their comments on an earlier version of the manuscript. Address correspondence to P. Assmann, School of Human Development, The University of Texas at Dallas, Box 830688, Richardson, TX 75083-0688 (phone: 214-690-2435; e-mail: assmann@utdallas.edu).

a linear transfer function that simulates the glottal source spectrum, vocal tract resonators, and lip radiation. The combined transfer function is the product of the source spectrum, $G(f)$, vocal tract transfer function, $V(f)$, and radiation function, $R(f)$:

$$S(f) = G(f) \cdot V(f) \cdot R(f) \quad (1)$$

The glottal source spectrum $G(f)$ is approximated by a second-order digital resonator with a frequency of zero and a bandwidth of 100 Hz. This results in a spectral slope of approximately −12 dB/octave toward higher frequencies. For convenience, the glottal source component is combined with the lip radiation function $R(f)$, which imparts a further +6-dB boost to the spectrum. The output is then lowpass filtered by a digital antiresonance, with a frequency of 1.5 kHz and a bandwidth of 6 kHz, to better approximate the source spectrum of normal voiced speech (Klatt, 1980).

The vocal tract resonances, called *formants*, are modeled using a cascade of second-order digital resonators. The vocal tract transfer function, $V(f)$, is the product of five resonances that represent the formants $F1$–$F5$ (see Klatt, 1980, Equations 2 and 6).

The synthesizer generates voiced speech by using a form of additive harmonic synthesis (Klatt, 1982). The amplitude (α_k) and phase (ϕ_k) of the kth harmonic are obtained by computing the magnitude and angle of the complex vector $S(f)$ at the frequency of the harmonic, kF_0. The synthesized waveform is computed as follows:

$$x(i) = \sum_{k=1}^{m} \alpha_k \cos\left[\theta(i,k)\right], \quad (2)$$

where

$$\theta(i,k) = \theta(i-1,k) + (2\pi k F_0 T) \quad (3)$$

and

$$\theta(1,k) = \phi_k.$$

In this formulation, m is the number of harmonics, F_0 is the fundamental frequency in hertz, and T is the reciprocal of the sampling frequency (by default, 8 kHz). Voiced speech generated by this synthesis model is very similar to that generated by the time-domain implementation described by Klatt (1980).

The SWSYNTH program generates *sine-wave speech*. Sine-wave speech replaces the time-varying formant structure of natural speech with a set of modulated sinusoids whose frequencies coincide with the center frequencies of the formants. SWSYNTH uses Equations 2 and 3 for synthesis, but replaces the m harmonic components (with frequencies kF_0) with sinusoidal frequency components, one for each formant. The frequency of the sinusoid is specified by the corresponding formant track

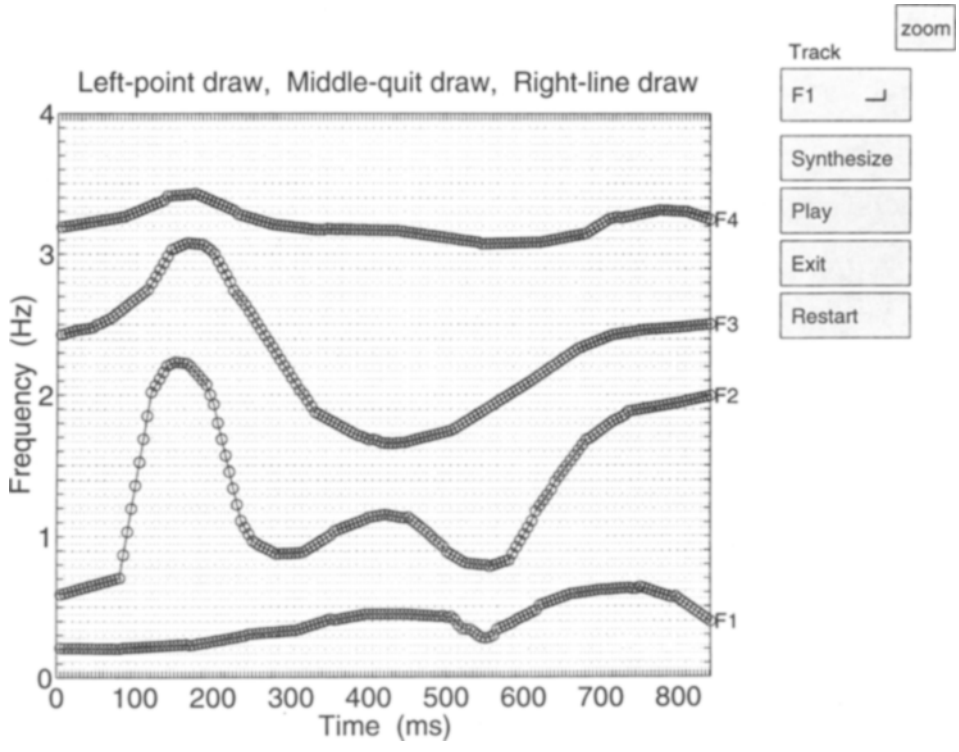

Figure 1. Main figure window generated by SYNTH. The display shows the time course of the frequencies of the formants. Program functions are controlled by pushbuttons on the right side of the display. To modify a parameter, the user selects a parameter code from the menu list, which appears when the button labeled "Track" is pressed.

($F1$–$F5$). The amplitude of the sinusoid is given by the transfer function $S(f)$, evaluated at the center frequency of each formant. Up to five frequency and amplitude-modulated sinusoids can be generated. If any frame along the frequency track is set to a value less than 30 Hz, it is effectively dropped from synthesis.

PROGRAM DESCRIPTION

SYNTH

The SYNTH program provides graphical control of the time-varying formant frequencies ($F1$–$F5$), fundamental frequency (F_0), overall amplitude (AV), and formant bandwidths ($B1$–$B5$). The graphical display consists of three figure windows. The main window panel displays the frequencies of the formants ($F1$–$F5$) as a function of time. On the right side of the panel there are several labeled pushbuttons that control track selection, synthesis, playback, display reset, and program termination.

Track drawing. Track drawing is initiated by selecting one of the options from the parameter list ($F1$–$F5$, F_0, AV, or $B1$–$B5$) given by the top button, labeled "Track." Data are entered for each parameter as a series of time-frequency coordinates (or time-amplitude coordinates, in the case of AV). These coordinates represent signal characteristics over a specified time interval, or *frame*. The frame duration is specified by the variable *nmspf*. Its default value of 5 msec can be reset as a command-line option.

Tracks can be modified by using one of two drawing modes. Coordinate values can be set individually (*point-draw* mode), or a group of frames can be assigned values by linear interpolation between two specified endpoints (*line-draw* mode). To use the point-draw mode, the cursor (by default, a small cross) is positioned at the desired location and the left mouse button is pressed. To use line-draw mode, the cursor is positioned at one of the desired endpoints of the line segment and the right button is pressed and held down. The cursor is then repositioned at the other endpoint and the right button is released. Each time a change is made using either point-draw or line-draw modes, the track is replotted to provide immediate feedback for the user. Track modification is terminated by pressing the middle button on a three-button mouse (or by pressing the escape key, if using a two-button mouse).

Synthesis. Synthesis is initiated by pushbutton control. When the synthesis button is pressed, a new figure window appears in the foreground. It displays a "schematic spectrogram" to indicate the progress of the synthesis on a frame-by-frame basis. Upon completion, the waveform of the synthesized sound is plotted beneath the schematic spectrogram.

Playback. Sound playback is initiated by pressing a dedicated pushbutton. The MATLAB function *sound.m* provides machine-specific audio playback.

Restart and exit. Control is provided to restart the program in midsession, to refresh the working variables

Figure 2. Second figure window generated by SYNTH. This window appears when the AV (overall amplitude) parameter code is selected from the "Track" menu list in Figure 1. Similar figure windows appear when the fundamental frequency (F_0) and formant bandwidth ($B1$–$B5$) parameter codes are selected.

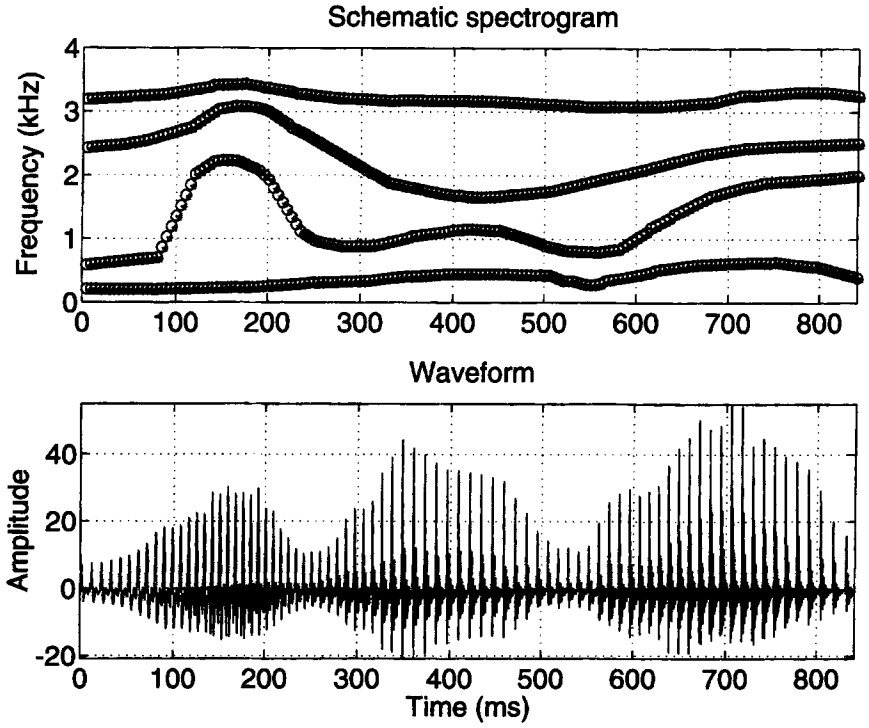

Figure 3. The synthesis window appears after the "Synthesis" button in the main figure window has been pressed (see Figure 1). The "Schematic Spectrogram" (upper panel) redisplays the formant frequency tracks created in the main figure window. The synthesized waveform (lower panel) is plotted when synthesis is completed.

used by SYNTH and to redisplay the figure windows. The exit button terminates the program and eliminates the figures generated by SYNTH, but maintains all workspace variables in program memory (including parameter tracks and synthesized waveform) for optional storage to disk.

Command-line mode. Parameters can also be specified directly on the command line, if desired. With this feature, it is straightforward to modify parameter tracks by applying an additive or multiplicative constant (e.g., to scale the frequencies of the formants, to shift the fundamental frequency up or down by one octave, or to alter the level of a previously synthesized sound). In addition, the time scale of the signal can be modified by resetting the duration parameter ("dur") and selecting the restart button. The parameter tracks are then stretched or compressed along the time axis using cubic spline interpolation.

Parameter tracks are maintained as MATLAB workspace variables and may be saved individually or collectively to disk for later retrieval or input into other programs. Workspace variables can be stored in MATLAB full-precision binary files (MAT-files) using the built-in *save* function, and later retrieved using the *load* command. MATLAB can also store and retrieve variables in standard ASCII format. This provides for easy transfer of data (including the synthesized waveform as well as the parameter tracks) between MATLAB and other software packages.

Figure 1 shows an example of the main figure window for the SYNTH program. The formant tracks are those used to generate a synthetic version of the sentence "We were away," appropriate for a male voice. In this example, the sample rate was 8 kHz, and only four formants are displayed in the range of 0–4 kHz.

Figure 2 shows the second figure window, which provides control of the overall amplitude (AV). This window appears when the AV track is selected from the track menu in the main figure window (Figure 1). Similar windows appear when the fundamental frequency (F_0) or formant bandwidth ($B1$–$B5$) tracks are selected.

Figure 3 displays the synthesis window. The upper panel is labeled "schematic spectrogram"; its main function is to show the progress of the synthesizer. As each frame is synthesized, a small asterisk appears in the unfilled circles that represent the current frame. The waveform of the synthesized utterance is displayed in the lower panel when synthesis is completed.

SPECSYNTH

SPECSYNTH creates a spectrogram from a digital sound waveform,[1] allowing copy synthesis through a tracing method. The synthesis procedure is the same as that used by SYNTH. When the program SPECSYNTH is started, it searches the contents of the working-memory buffer for a variable, y, which is assumed to represent the

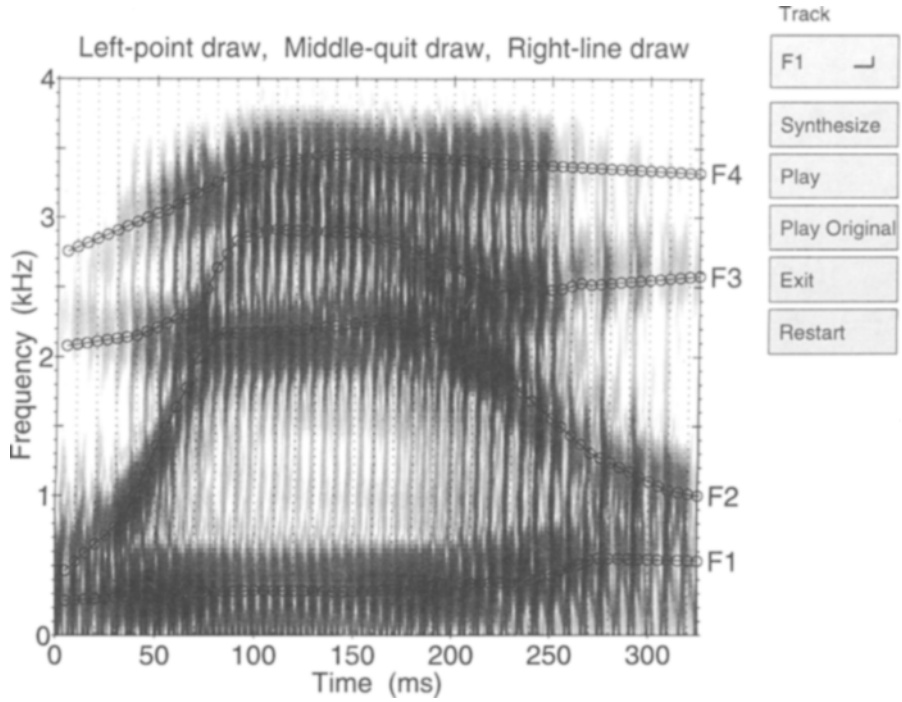

Figure 4. Main figure window generated by the SPECSYNTH program. The display shows the spectrogram of a user-defined waveform vector. Superimposed on the spectrogram are a set of formant frequency tracks that can be adjusted to match the formant pattern of the recorded voice. The spectrogram shown here was computed from a digital recording of a male voice producing the word "wheel."

input sound waveform. If y does not exist, it is generated as a 200-msec segment of a "neutral" vowel, similar to schwa.

Entering data/track drawing. Upon starting the program, a spectrogram is generated from the user-defined waveform vector. The spectrogram is generated with a fast Fourier transform (FFT) method. The lines representing the formant tracks are superimposed on the spectrogram. The fundamental frequency (F_0) track is estimated from the waveform by a sliding window cepstral analysis (Rabiner & Schafer, 1978). These initial fundamental frequency estimates can be refined or smoothed with the aid of the drawing functions. The overall amplitude (AV) is estimated from the waveform by computing the root-mean square energy in a sliding rectangular window. Editing and track drawing are mouse controlled, as described for the SYNTH program.

Saving the parameter tracks. Once the analysis of the speech signal is completed, the parameter tracks can be saved either in a MATLAB binary file or in a standard ASCII file. Thus, SPECSYNTH provides a useful tool for interactive measurements of the fundamental and formant frequencies of recorded speech. It is also possible to import formant and fundamental frequency measurements generated by other speech analysis programs for evaluation. Formant estimation accuracy can be assessed by visual comparison of spectrograms and superimposed formant tracks, as well as by synthesis and playback.

Playback. SPECSYNTH has two playback options. The program can play the synthesized utterance, or it can play the original sound for comparison purposes.

Other functions. The track selection, exit, and restart functions are identical to those used by SYNTH.

Figure 4 shows an example of the main figure window for the SPECSYNTH program. The spectrogram displays the word "wheel," spoken by an adult male. The formant tracks superimposed on the spectrogram generate an acceptable approximation to the spoken utterance.

SWSYNTH

The third program uses sine-wave synthesis rather than cascade formant synthesis. Like SPECSYNTH, the SWSYNTH program searches the contents of the working-memory buffer for a variable, y, which is assumed to represent the input sound waveform. If y has not been previously defined, the program generates a 200-msec segment of a "four-tone approximation" to the neutral vowel generated by SPECSYNTH. In other respects, SWSYNTH is similar to SPECSYNTH. A useful feature of SWSYNTH is its ability to generate individual frequency-modulated sinusoids (glissandi). A single modulated sinusoid is

generated when all but one of the frequency tracks are set to zero.

PROGRAM SPECIFICS

The software components described above are written for use with MATLAB Version 4.x. MATLAB is an interactive computing environment for matrix computation. It has several features that make it a desirable environment for speech analysis and synthesis. First, its matrix operations are well suited for digital signal processing applications, and many essential functions, such as convolution and Fourier transforms, are built-in. Unlike most conventional programming languages, it is interactive; no separate stages of compilation and linking are involved. MATLAB programs, called *M-files*, are ASCII text files and are highly portable across different platforms. The MATLAB language is written in a compact format. A powerful object-oriented graphics environment has been added in Version 4.0. MATLAB is available for many platforms, including Sun SPARCstation, VAX/VMS, 386- and 486-based PCs using the MS-Windows operating system, and Macintosh. To date, Track-Draw has been tested with Sun SPARCstations 1, 2, and 10, Macintosh IIcx, and with MATLAB for Windows using a PC/486 environment.[2]

Program Availability

The M-files which implement the Track-Draw package, along with a user's manual, can be obtained via anonymous ftp (file transfer protocol) or by floppy disk. Send requests by e-mail (assmann@utdallas.edu) or by regular mail to the first author.

REFERENCES

BICKLEY, C. A., STEVENS, K. N., & WILLIAMS, D. R. (1992). Control of Klatt synthesizer with high-level parameters. *Journal of the Acoustical Society of America*, **91**, 2442. (Abstract)

JAMIESON, D. G., RAMJI, K., KHEIRALLAH, I., & NEAREY, T. M. (1993). CSRE: A speech research environment. *Journal of the Acoustical Society of America*, **93**, 2394. (Abstract)

KLATT, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, **67**, 971-995.

KLATT, D. H. (1982, June). Harsyn: An additive harmonic synthesizer. *MIT Research Laboratory of Electronics: Speech Communication Group Working Papers*, pp. 47-60.

RABINER, L. R., & SCHAFER, R. W. (1978). *Digital processing of speech signals*. Englewood Cliffs, NJ: Prentice-Hall.

REMEZ, R. E., RUBIN, P. E., PISONI, D. B., & CARRELL, T. D. (1981). Speech perception without traditional speech cues. *Science*, **212**, 947-950.

NOTES

1. The sound waveform is a vector of sampled-data values. This vector can be loaded from disk (and would normally be derived from a recording of natural speech), or it may be obtained from a previously synthesized sample. MATLAB has several routines for sound file conversion and storage that can be used with the SPECSYNTH and SWSYNTH programs.

2. MATLAB is a registered trademark of The MathWorks, Inc. VAX/VMS is a trademark of Digital Equipment Corporation. Macintosh is a trademark of Apple Computers, Inc. PC is a trademark of International Business Machines Corporation. Windows is a trademark of Microsoft Corporation. SPARCstation is a trademark of Sun Microsystems, Inc.

(Manuscript received September 7, 1993;
revision accepted for publication March 22, 1994.)