CENTER FOR COMPUTER RESEARCH IN MUSIC AND ACOUSTICS
AUGUST 1989

Department of Music
Report No. STAN-M-57

SYNTHESIS OF THE SINGING VOICE USING A PHYSICALLY PARAMETERIZED MODEL OF THE HUMAN VOCAL TRACT

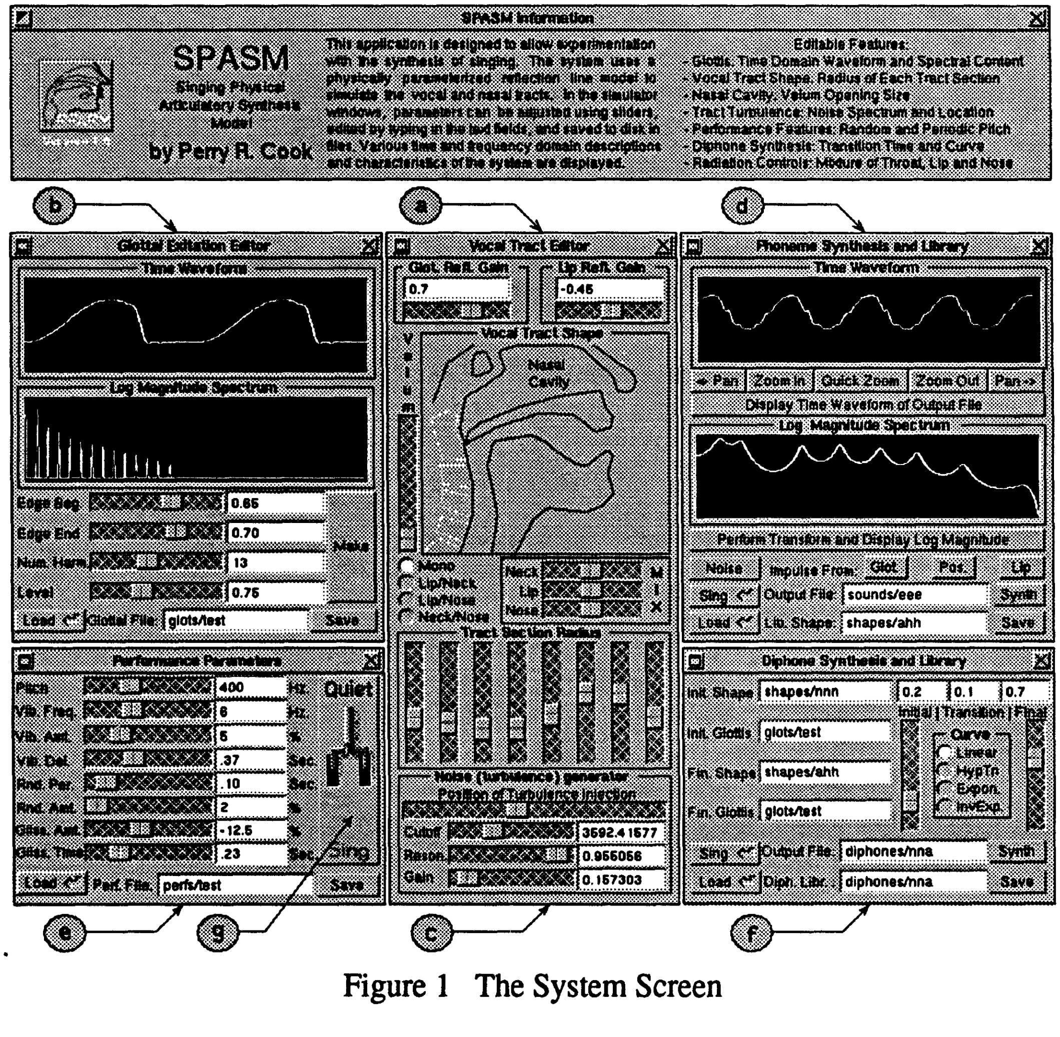
Perry R. Cook

CCRMA
DEPARTMENT OF MUSIC
Stanford University
Stanford, California 94305

SYNTHESIS OF THE SINGING VOICE USING A PHYSICALLY PARAMETERIZED MODEL OF THE HUMAN VOCAL TRACT

Perry R. Cook (PRC@CCRMA@SAIL.STANFORD.EDU)
Center for Computer Research in Music and Acoustics (CCRMA), Stanford Univ.

Using a physically parameterized model of the human vocal tract, a system has been constructed which allows graphical interactive experimentation with various control parameters. The vocal tract is parameterized by its shape, which is graphically displayed by a cross section of a human head. Sliders on an editor window control the radius of each vocal tract segment and the size of the velum opening into the nasal tract. Impulse responses can be obtained from the glottis, the lip, or an arbitrary point within the tract, and frequency domain transfer functions can be computed and displayed. The glottal pulse is additively synthesized from Fourier coefficients controlled by simple parameters in the editor, or from a library file of coefficients derived from analysis data. T simulate the turbulences of fricatives and other consonants, a filtered noise source can be made arbitrarily resonant at one frequency, and can be placed at any point within the vocal tract. Diphones can be constructed by specifying initial and final sets of parameters and a speed and interpolation curve for transition.



This application is designed to allow experimentation with the synthesis of singing. The system uses a physically parameterized reflection line model to simulate the vocal and nasal tracts. In the simulator windows, parameters can be adjusted using sliders, edited by typing in the text fields, and saved to disk in files. Various time and frequency domain descriptions and characteristics of the system are displayed.

Editable Features:
- Glottis: Time Domain Waveform and Spectral Content
- Vocal Tract Shape: Radius of Each Tract Section
- Nasal Cavity: Velum Opening Size
- Tract Turbulence: Noise Spectrum and Location
- Performance Features: Random and Periodic Pitch
- Diphone Synthesis: Transition Time and Curve
- Radiation Controls: Mixture of Throat, Lip and Nose

Edge Beg. 0.65
Edge End 0.70
Num. Harm. 13
Level 0.75
Glottal File: glots/test

Glot. Refl. Gain 0.7
Lip Refl. Gain -0.45
● Mono
○ Lip/Neck
○ Lip/Nose
○ Neck/Nose
Neck
Lip
Nose
Cutoff 3592.41577
Reson. 0.955056
Gain 0.157303

Impulse From: Glot. Pos. Lip
Output File: sounds/eee
Lib. Shape: shapes/ahh

Pitch 400 Hz.
Vib. Freq. 6 Hz.
Vib. Amt. 5 %
Vib. Del. .37 Sec.
Rnd. Per. .10 Sec.
Rnd. Amt. 2 %
Gliss. Amt. -12.5 %
Gliss. Time .23 Sec.
Perf. File: perfs/test

Init. Shape shapes/nnn 0.2 0.1 0.7
Init. Glottis glots/test
Fin. Shape shapes/ahh
Fin. Glottis glots/test
Curve
● Linear
○ HypTn
○ Expon.
○ InvExp.
Output File: diphones/nna
Diph. Libr.: diphones/nna

Figure 1 The System Screen

INTRODUCTION

The synthesis of singing has been investigated from within the frameworks of most music synthesis methods. Primitive speech systems use time domain methods, such as playback o digitally recorded phonemes or words, but such systems are inappropriate for music synthesis. Other systems are based on the final spectrum (McAulay & Quatieri 1986) (Chowning 1989). Still others, specifically source/filter systems such as linear predictive coding, are more closely related to the physics of the vocal tract and may include parameter for controlling formants or other spectral features directly (Rodet 1984). The mapping of such pseudo-physical systems onto the actual physical components of the vocal mechanism usually not sufficient to allow direct synthesis (rather than by analysis of recordings) based on notions from vocal pedagogy and speech physiology such as tongue position and glottal effort. The purpose of this project was to provide parameterizations of the vocal source, the vocal tract, and various control functions which are as physically correct as possible within certain computational criteria. This model of the human voice organ was then placed in an application program which allows graphical interactive experimentation with the control parameters. The speed and ease of interactive software design on the NeXT machine, its special purpose signal processing hardware, Display PostScript graphics, and high quality digital audio capabilities make the NeXT an ideal workstation for this application.

VOCAL TRACT SHAPE

The vocal tract is modelled by a spatial filter, or lattice filter (Markel & Gray 1976)(Kelly & Lochbaum 1962). The tract is divided into a number of sections, each of the same length determined by the sampling rate and the speed of sound. The filter reflection coefficients a computed from the relationship between the characteristic impedances of adjacent sections. Since the characteristic impedance is a function of the cross-sectional area of a section, and thus the radius, the coefficients can be computed entirely from the tract section radius measurements. The nasal tract is simulated with another lattice filter, and the two tracts are coupled at a bifurcation point located at the velum opening with a three-way scattering junction. Figure 1a shows the **Vocal Tract Shape** editor and display. Shapes can be saved or loaded from a shape library. Sliders in the graphical editor window control the radius of each segment of the tract. The path through the nasal airway is controlled by a velum position slider. A graphical cross-section of a human head provides immediate feedback to the user about the vocal tract shape as controlled by the sliders, or as loaded from a file. A additional text window showing the radii in centimeters allows the user to enter parameters with greater accuracy. Impulse responses can be obtained from the glottis, the lip, or an arbitrary point within the throat, and the log-magnitude transfer function can be computed and displayed. Switches and sliders control and mix the lip, nose, and throat radiation outputs.

GLOTTAL SOURCE

Rather than the impulse source used by many LPC speech and singing systems, the glottal source used in this system copies the time-domain and spectral properties of the pressure waveform of the human glottis. The glottal source waveform is additively synthesized fro

Fourier coefficients controlled by simple parameters entered in the editor, or from a library file of coefficients derived from analysis data. Figure 1b shows the **Glottal Exitation** editor The simple parameter editor controls operate principally on the time-domain glottal waveform. Parameters include the number of harmonics to be used for synthesis (primarily to prevent aliasing), the overall amplitude, and the position and slope of the falling edge of the glottal pulse, which has been shown to be an important feature when describing vocal effort (Sundberg 1987). Graphical displays of the log-magnitude spectrum and the time-domain waveform are provided. Parameters may be saved to and loaded from disk files.

THE NOISE SOURCE

Fricative consonants behave spectrally much like filtered noise, exhibiting one or two resonant peaks (Heinz & Stevens 1961). To simulate the turbulences of fricatives and othe consonants, a noise source can be placed at any point in the throat path of the vocal tract. The output of the noise source can be made arbitrarily resonant at one frequency by a two pole filter. This allows a tuned source of local turbulence to be injected at a point of constriction. Noise source parameters are saved as part of shape files. Figure 1c shows the **Noise Generator** editor.

PHONEME SYNTHESIS AND PERFORMANCE CONTROL

Once the glottal source, the noise source, and the vocal tract shape have been established, synthesis of a short musical 'performance' can be accomplished by mouse clicking the synthesis button, which is located in the **Phoneme Synthesis and Library** window shown i Figure 1d. A default set of natural amplitude and frequency control functions are available for synthesis, or the user can edit the performance features using sliders or text fields. Performance controls affect frequency features and are located in the **Performance Parameters** window, which is shown in Figure 1e. Once the synthesis is complete, the res is heard via the computer's internal digital to analog converters. The file can be played bac repeatedly by mouse clicking the **Sing** button. By typing any file name into the **Output Fil** field and pressing the carriage return, the file is played back (if it exists), and synthesis to different files allows speedy A/B comparison of sound examples. The time domain waveform and the log magnitude spectrum of the phoneme can be displayed.

DIPHONE SYNTHESIS

Diphthongs can be constructed by specifying initial and final sets of parameters, an interpolation curve, and the time in seconds of the initial and final steady state segments. Figure 1f shows the **Diphone Synthesis and Library** window. Since the synthesis yields one second of sound, specifying the duration of the initial and final states also specifies the transition time. Curves available for interpolation include linear, hyperbolic tangent, and exponentials. In the case of the glottis, the interpolation is carried out between initial and final wave tables. For the noise generator, the filter pole parameters are interpolated in the plane. For speed, the vocal tract and nasal tract scattering relationships are interpolated in the reflection coefficient space.

REAL TIME DSP SYNTHESIS

By clicking the **Sing** switch in the **Performance Features** window, the system begins to synthesize in real time. Performance features which are active are pitch, vibrato speed and amount, and random vibrato amount. The tract section and velum sliders make the appropriate adjustments in the real time model and sound, and new glottis wavetables may t synthesized and down-loaded to the DSP chip.

OTHER EDITORS AND CONTROLS

Other windows for controlling the system may be activated from the main menu. One such window is the **Shape Space Interpolator**. This window allows the user to enter a number c shape filenames into text fields. There are regions in which these shapes are active, and the user may control the current vocal tract parameters by moving a cursor about the window, thus determining the "mix" of shapes. Another window , the **Formant Editor**, allows the user to edit in the formant domain (the one in which the ear perceives speech sounds). Whe this window is activated, the system impulse response is obtained. A log-magnitude transform is computed and displayed, and peaks are located and marked. Each of the first few (selectable) peaks is associated with a text-field/slider control, and the user may move t markers to new locations. By depressing the **Doit** button, the system adaptively moves the formant peaks to the desired locations, modifying the vocal tract in a least squares perturbation fashion. The path taken during the adaptive modification can be written to disl as a trajectory file.

REFERENCES

McAulay, R.J., & T.F. Quatieri. 1986. "Speech Analysis/Synthesis Based on a Sinusoidal Representation." *IEEE Trans. Acoust. Speech and Sig. Proc.* ASSP-34(4):744-754.

Chowning, J.M. 1989. "Frequency Modulation Synthesis of the Singing Voice." In *Some Current Directions in Computer Music Research.*Cambridge MA.: MIT Press: 57-63.

Rodet, X. 1984. "Time-Domain Formant Wave-Function Synthesis." *Computer Music Journal* 8(3): 9-14.

Markel, J. D., & A. H. Gray. 1976. *Linear Prediction of Speech.* New York: Springer-Verlang.

Kelly, J. L., & C. C. Lochbaum. 1962. Proc . Fourth Intern. Congr. Acoust. Paper G42: 1-4

Sundberg, J. 1987. *The Science of The Singing Voice,* Dekalb Il.: Northern Illinois Universi Press: 83-85.

Heinz, J. M., & K. N. Stevens. 1961. "On the Properties of Voiceless Fricative Consonants. *Journal of the Acoustical Society of America* 33: 589-596.