Csound version of the Klatt speech synthesizer: A University of California, Davis studio report
Wayne Lawson
Department of Music, University of California, Davis
Davis, CA 95616

ABSTRACT
A version of the KL SYN88 speech synthesizer, designed by Dennis Klatt, has been implemented as a Csound orchestra at the Computer and Electronic Music Studio at the University of California, Davis. The synthesizer, a terminal-analog type, includes controls over both cascade and parallel filter systems, a vocal source (whose controls include the open quotient of the vocal period, diplophonic pulsing, slow jitter, and vibrato), an aspiration source, and an affrication source. The Csound orchestra is used to synthesize music in which sound color is controlled independently.

1. COMPUTER MUSIC AT DAVIS

Speech-like sounds have been at the center of my work in computer music over many years. Making musical structures of such sounds is one of the main emphases in the Computer and Electronic Music Studio that I direct at the University of California, Davis. This paper is a report on our relatively new studio and a description of our work on synthesis of speech-like sounds using a new synthesizer written as a Csound orchestra.

The studio at Davis is equipped with a range of up-to-date workstations, including a Silicon Graphics Indigo, Sun Sparcstations, a NeXT, and a Mac IIx. We have high quality digital recording equipment, about 4 MB of disk storage for waveforms, and a small number of MIDI-controlled synthesizers and a keyboard. For mainly pedagogic purposes the studio also includes analog recording devices and an old Buchla synthesizer. We offer undergraduate courses in electronic and computer music that are open to both music majors and non-majors. The new PhD program in music at Davis is flexible and student-centered. Composition and Theory is among the concentrations available, and within that concentration students may choose to emphasize computer music composition and research. The main thrust of our work is software synthesis of composed music; we tend to be less concerned with interactive performance and issues of real-time synthesis. On the other hand, much of our work is highly computation-intensive, so we require the power of modern workstations for effective development of sonic structures and for efficient polishing of synthesized music. We use Csound (Vernon 1992) and an approach to score specification using C macros that we find highly flexible and musically natural. The emphasis is on composition projects even at the beginning levels, and a number of good pieces have been composed and realized in the studio.

2. THE KL SYN88 SYNTHESIZER

Our use of Csound as a computer-synthesis program and my own continued interest in vocal sounds motivated us to attempt to implement as a Csound orchestra a version of the well-known and widely-used speech synthesizer program written by Dennis Klatt, of the Research Laboratory of Electronics at MIT. Our orchestra is based on a recent version of the synthesizer, called KL SYN88, as described in articles by Professors Klatt and his daughter (Klatt 1980, Klatt and Klatt 1990), the last of which was completed shortly before his death from throat cancer in December 1988.

Klatt's synthesizers are of the "terminal analog" type. These kinds of synthesizers are properly regarded as functional models of the vocal apparatus, rather than models of the physical apparatus itself. The latter
include vocal-tract analog designs in which the cross-sectional area as a function of distance from the glottis (or the lips) is modeled directly. Although such models are likely in the future to improve on terminal analog designs, the details of how to control them are not understood well enough at present to meet the goal of synthesizing natural-sounding speech.

In terminal-analog designs the vocal tract is treated as a single complex filter system that modifies the acoustic energy reaching it from various sources. The source and filter system are assumed to be "weakly coupled"; that is to say, the effects of the filter system back onto the sources are taken to be negligibly small. Much acoustic phonetic research over several decades has dealt directly with the vocal tract as a filter system, so the control of terminal-analog synthesizers is well understood and highly natural-sounding speech has been produced with them.

The primary innovations in KLSYN88 are: (1) a new model of the glottal source for voiced sounds and (2) features that mimic certain observed interactions between the sources and the vocal filters. The cascade and parallel filters that model the vocal-tract transfer function are widely used and quite straightforward to implement. Along with frequency and gain, the Klatt's new glottal model, called KLGLOTT88, provides control over the proportion of the voicing period during which the vocal folds are open—the "open quotient", a spectral tilt parameter, a special kind of slow flutter in the timing of the glottal pulses, and a means of mimicking diplophonia, a delay and reduction in amplitude of every other glottal pulse. Extra resonances and antiresonances are provided to simulate the effects of tracheal coupling to the vocal tract, as are vocal pitch-synchronous controls of the first formant to simulate one kind of source-filter interaction. An aspiration noise source is used, among other things, to simulate breathy speech. Many of the new features in KLGLOTT88 were motivated by the goal of Klatt and Klatt to extend highly natural-sounding speech synthesis-by-machine to female speech.

3. KLSYN88 IN CSOUND

Our Csound orchestra eliminates certain optional features in the original program, but otherwise it provides some form of nearly all the parameters of KLSYN88 and a few in addition: quasi-periodic modulation of the voicing frequency and an increased number of resonances in the cascade branch of the filter to provide control over the higher frequencies usually called for in music, but not in speech. The sound generation portions of our synthesizer—our version of KLGLOTT88, separate pseudo-random sources for aspiration and affricative noises, and both cascade and parallel filter systems—are all combined into a single large Csound instrument, which is turned on at the beginning of any segment of sound and kept on throughout its duration. We have implemented an interesting detail of KLGLOTT88 by limiting changes in such parameters as voice frequency and gain to the closed portion of the glottal period. By doing so we can accommodate large changes in such parameters as voice frequency and amplitude without causing clicks and pops in the output signal.

The time-varying control parameters for the large sound-generating instrument are global control-rate variables generated by a number of small Csound instruments. The score consists almost entirely of calls to these instruments. Their parameters—p-fields, that is—specify single values of the control signals they generate and a duration. The instruments then generate a single segment of a piecewise linear, or in some cases piecewise exponential, global control signal, interpreting the duration parameter as the time between the last value and the present one.

Allocation of sound generation to a single instrument on all the time and allocation of the generation of control signals for that instrument to an number of other instruments has several advantages. The separation of control generation and sound generation provides a rough functional parallel to the physiological distinction between neuromuscular processes and their effects on the organs whose mechanical motions produce the sounds of speech. One could say that the system involving the lungs, the vocal folds, the throat, the tongue, and the lips—represented by the sound-generation instrument—is always "on" in
the sense that speech-like sounds will emanate from it whenever it is properly controlled by the nerves and muscles—represented by the control-generator instruments. An important practical advantage of this design is the temporal independence of the control-generator instruments. This can be illustrated in the synthesis of a breathy, vowel-like sound. The aspiration noise should be turned on at the beginning of the sound and kept at about the same amplitude throughout. The voicing, on the other hand, should rise in amplitude more slowly and die away more quickly than the aspiration. And of course the filter system may be changing quite out of synchrony with the sources, as in a sibilant.

Even the relatively high computational burden that the overall design entails is ameliorated by provisions we have built into the sound-generation instrument to turn off computation when output signals from various stages of the synthesizer fall below the level of audibility. The design seems implicit in Kliatt's own approach and has proven both practical and flexible for a number of composers who have used the orchestra.

4. C MACROS IN SCORE SPECIFICATIONS

There remains a large number of control signals—up to 42—that must be specified in the sound files that reference this orchestra. To avoid having to specify all these parameters for each sound event, we generate our numerical scores by means of a preprocessor that employs C macros. Something of the flavor of this procedure can be seen in the following examples of sound specifications from a recent piece of mine.

hcppvh(20, AN, DN, MP)
tcvvh(Qu, UI, CN, PN, PPN)
p2v(Qu, OE, EH, NW3, MF, Z)

Parameters that correspond are lined up vertically. The pitch and loudness specifications are semantically clear; duration is specified either by a number or, for commonly used durations, symbols—Qu for, example, means a quarter-note triplet. The AN, OE, etc., are specifications of sound color—an aspect of timbre derived from rowels, which I treat in this, and a number of recent compositions, as musically independent (Sawson 1985, 1986). Where there are two pitches or colors, grace-note-like slides in pitch or color, respectively, are to be generated at the beginnings of the events. The names—the hcppvh, for example—identify the type of the sound event; I like to think of the names as "containing" all the details not specified in their associated parameters. Each of the lines in these examples results in a sonic event of some complexity. Underlying each line is a hierarchy of macro definitions that at bottom generates numerical score statements in the form of instrument calls readable directly by Cosound. One might say that the hierarchical macro definitions in this preprocessor stage is an exfoliation of the two-level hierarchy in the design of the orchestra; that is to say they reflect in some sense my own compositional process and those of others who have tried this method.

5. USE OF THE SYNTHESIZER IN SOUND-COLOR MUSIC

The use of a synthesizer like KLSYN88 (in our Cosound version) for composing and realizing music raises an issue with both a pragmatic and an aesthetic side. Is it necessary to use a synthesizer for music composition that has all the features required for the generation of natural speech? The computational load of the full range of these features (even with bypasses at low amplitudes) is considerable, so a composer of reasonably complex music should consider carefully whether the wait for interim results on an inexpensive computer or the cost of a powerful workstation to speed up computation is justified.

The issue turns, of course, on the composer's own compositional interests and goals. If timbral complexity—including, perhaps, flexible timbral modulations within sound events—is not a high priority, then such a computationally demanding synthesis system is not necessary. Even in cases like my own,
where timbral structure across a span of a composition and timbral dynamics within single sound objects are important compositional goals, a simpler synthesizer may be sufficient. On the other hand, I find in the full range of vocal sounds—used sometimes in quite un-speech-like ways—a welcome variety of ways for differentiating and punctuating my pitch and sound-color structures.

Another of the attractions of the Klatt synthesizer for me is put in these strictly compositional issues, but rather in the more general associations the sounds seem to suggest. Even when the sounds I make with our version of KLSYN88 are quite unlike speech—in fact, some of them are clearly not producible at all by human talkers—attentive and sympathetic listeners may find themselves moving their jaws, tongues, and lips in a subliminal way, “talking along” to the music. Or listeners less inclined toward this kind of active involvement with the sounds may be led nevertheless to imagine some kind of anomalous human agent behind them.

This latter effect was pointed out to me by my collaborator in a recent project combining music with images. I synthesized the musical portion of a videotape, the images for which were drawn from a single photograph by Harvey Himmelblau of a burned-over pine forest in Yosemite National Park. The work, entitled Grave Trunks, is the first outcome of our project in which we sought specific perceptually satisfying analogies between dimensions of sound color on the one hand and dimensions of the photographic image on the other. We associated acute sound colors—those derived from the vowels in words like “heat”, “hate”, and “bat”—with portions of the photograph that were relatively light in tone: non-acute or grave sound colors—as in “meat”, “maat”, and “maul”—with dark tones in the photograph. The non-open sound colors—in the rounded vowels of certain European languages, Swedish “fy” and German “eise”, for example—were associated with portions of the photograph that appeared to have little depth of field or had a high edge density. Open sound colors—as in “bat” and “hot”—with the appearance of marked depth of field or low edge density. Among our goals in this project was to build on these fundamental analogies in the hope that the combined work would enhance each of our separate contributions. In a number of ways that we anticipated, this goal was met by Grave Trunks. An unexpected result was Himmelblau’s sense that the speech-derived musical sound provided a kind of concrete personification of human involvement that is otherwise missing in the usual silent viewing of a photograph.

Grave Trunks was well received at its premiere in May 1993, and we hope to improve the technical quality of our work by joining computer image processing with computer music in future projects.

6. REFERENCES


