KLSYN: A Formant Synthesis Program

D.H. Klatt

{Revised for the IBM-PC implementation by Keith Johnson}

The KLSYN speech synthesis program accepts user commands to create parametric data to control a digital speech synthesizer, and it produces an output waveform file with a user-specified name. {The IBM-PC version of KLSYN was first implemented by Keith Johnson & Yingyong Qi at Ohio State University in 1987. Since then several minor modifications have been added by Keith Johnson.}

The synthesizer is the same as the one documented in some detail in Klatt (1980), except that the voicing source has been augmented so as to permit a choice between two glottal waveforms. The new voicing source waveform is intended to be more flexible and thus be capable of producing more natural changes in voice quality over the duration of a sentence, if controlled properly. The theory of control, and the new control parameters are all described herein.

Introduction and overview

{This section describes procedures and tools for speech syntheis at the Research Laboratory of Electronics at MIT. Although the procedural details differ from one implementation to the next, this section provides a valuable insight into the speech synthesis strategies developed and used by Dennis Klatt. The IBM-PC version of the synthesizer produces Cspeech files, so recording, playback and spectral analysis can be accomplished on a PC using Cspeech.}

In a typical session of synthesis activity, the following steps would be taken, resulting in the production of an audio tape containing a listening test (identification, discrimination, or paired-comparison) of synthetic stimuli {or for use with an on-line speech perception setup.}:

Record. The program RECORD is used to analog-to-digital convert an audio recording of someone saying the utterance to be synthesized. Using the waveform display and editing features of RECORD, the digitized speech waveform should be trimmed to conform to the duration of the planned synthesis waveform, and then saved as a file. This file will be referred to as the "natural model" in the paragraphs below.

Specto. The program SPECTO should then be used to produce a pseudo-spectrogram of the natural model waveform file on the LXY printer/plotter. Information to be printed will also include estimates of formant frequencies, overall amplitude, and fundamental frequency versus time.

Klsyn. The program KLSYN is the synthesizer routine. It is used in a manner described later in this manual. The result is a synthesis waveform file which will be referred to as the "synthesis output file" in the paragraphs below.

Play. The program PLAY can be used to listen to the natural model and the synthesis output file. However, to make improvements in the synthesis, it isbest to rely on spectral comparisons using KLSPEC.

Klspec. The program KLSPEC permits direct spectral comparison of two (or more) waveform files. It is suggested that, for synthesis purposes, the spectral representation known as "spectrogram filtering" be used, although the "dft" spectrum is useful if one wishes to compare individual harmonics, and the "critical-band" spectrum is useful if one wishes to use filters with wide bandwidths at the higher frequencies to combat statistical fluctuations in noise spectra. Detailed notes should be taken as you step through the two waveform files, first noting differences in fundamental frequency and overall amplitude. Then, when these have been corrected, you should look for differences in spectral shape, and the speed of onsets and offsets.

To make corrections to the synthesis parameter time functions, you should return to KLSYN, recalling the synthesis parameters that were automatically saved as a file. Corrections are made to the parameter tracks, and a new synthetic waveform file is generated. This process should be iterated until the differences between synthesis and natural model are minimal.

At this point, one may wish to develop an ideal "end-point" synthetic stimulus for some other syllable or word, repeating the steps described above. Finally, one might generate a continuum between the endpoints, or a set of stimuli that differ along some acoustic dimension or dimensions. At that point, one will have available a set of synthetic waveform files with different first names.

Maketape. The program MAKETAPE is used to automatically randomize a set of waveform files and make (a) identification tests, (b) 4IAX discrimination tests, or (c) paired comparison tests. The audio signal for these tests comes out through the digital-to-analog converter of the VAX, and can be recorded on audio tape. {The system described in Johnson & Teheranizadeh (1992) can be used to create perception experiment running programs for online collection of listeners' responses and response times, etc.}

Klattalk. A software version of KLATTALK is available for observation of parameter values used in synthesis by rule. The standard KLATTALK program has been augmented so as to produce an output config/parameter file 'name.doc', appropriate for input to KLSYN, by typing the ‘KT -d8’ command. In this way, "first guess" parameter values for any English utterance may be obtained from text input to KLATTALK.

A block diagram of the speech synthesizer is shown in Figure 1. There are almost 50 parameters and constants which the user can change to determine aspects of the synthetic speech that it produces, as described in the following sections.

----

insert Figure 1 about here

Table I. Default configuration of KLSYN.

Sym	Name	V/C	Min	Max	Default
sr	Sampling rate	C	5000	20000	10000
du	Utterance duration	C	30	5000	500
nf	Number of cascade formants	C	1	8	5
ss	Source select	C	1	2	1
rs	Random seed	C	1	99	1
os	Output select	C	0	20	0
g0	Overall gain control	V	0	80	60
f0	Fundamental frequency	V	0	5000	1000
at	Amplitude of turbulence	V	0	80	0
oq	Glottal open quotient	V	10	80	50
tl	Glottal spectral tilt	V	0	34	0
sk	Glottal skew	V	0	100	0
dF	Delta F1 (at open glottis)	V	0	100	0
db	Delta B1 (at open glottis)	V	0	400	0
av	Amp cascade of voicing	V	0	80	60
ah	Amplitude of aspiration	V	0	80	0
F1	First formant frequency	V	180	1300	500
F2	Second formant frequency	V	550	3000	1500
F3	Third formant frequency	V	1200	4800	2500
F4	Fourth formant frequency	V	2400	4990	3250
F5	Fifth formant frequency	V	3000	4990	3700
f6	Sixth formant frequency	V	3000	4990	4990
fz	Nasal zero frequency	V	180	800	280
fp	Nasal pole frequency	V	180	500	280
b1	First formant bandwidth	V	30	1000	60
b2	Second formant bandwidth	V	40	1000	90
b3	Third formant bandwidth	V	60	1000	150
b4	Fourth formant bandwidth	V	100	1000	200
b5	Fifth formant bandwidth	V	100	1500	200

b6	Sixth formant bandwidth	V	100	4000	500
bz	Nasal zero bandwidth V	40	1000	90	
bp	Nasal pole bandwidth V	40	1000	90	
ap	Amp of parallel voicing	V	0	80	0
af	Amplitude of frication	V	0	80	0
ab	Amplitude of the bypass	V	0	80	0
a1	First formant amplitude	V	0	80	0
a2	Second formant amplitude	V	0	80	0
a3	Third formant amplitude	V	0	80	0
a4	Fourth formant amplitude	V	0	80	0
a5	Fifth formant amplitude	V	0	80	0
a6	Sixth formant amplitude	V	0	80	0
an	Nasal formant amplitude	V	0	80	0
p1	First formant bandwidth	V	30	1000	80
p2	Second formant bandwidth	V	40	1000	200
p3	Third formant bandwidth	V	60	1000	350
p4	Fourth formant bandwidth	V	100	1000	500
p5	Fifth formant bandwidth	V	100	1500	600
p6	Sixth formant bandwidth	V	100	4000	800

Starting KLSYN

The program is evoked simply by typing {from any directory}:

klsyn<CR>

The program begins by typing out a line identifying itself {and by reminding the user of the command line options}. It then reads in the default synthesizer configuration file "default.con". This file specifies the utterance duration and parameter update interval in ms among other things -- all of which can be changed by user commands. Finally, the program types the user prompt symbol ">".

The default configuration is listed in Table I, and is described below. Default values are assigned to all of the constants and variable parameters of synthesis as indicated in the table.

How to Begin with a Special Synthesizer Configuration

Alternatively, one can startup KLSYN with a previously prepared set of defaults. In particular, a command is available for changing the default values in the file default.con (the ‘c’ command). When a waveform is finally synthesized, the synthesizer configuration (set of default values) is saved in a file with the same first name as the waveform file. For example, if the waveform file were called "baa.snd", then the configuration file would be named "baa.doc". The next time you want to run the synthesizer using these special defaults as a starting point, simply type:

```
klsyn baa<CR>
```

and the configuration file baa.doc will be read instead of default.con. If the file baa.doc did not exist in the above example, the system will complain and abort.

Other Optional Arguments

The program KLSYN can take arguments of two types. The first type is an argument preceded by a minus sign, which establishes a mode of operation for the session. There are three (mutually compatible) options, the first of these identifies the user as a NOVICE. The second option requests that the file be synthesized in BATCH mode. In batch mode the synthesizer reads in a synthesizer configuration file (either default.con or a file specified on the command line) and immediately synthesizes the waveform. The last option requests that AUTOMATIC GAIN CONTROL be active during the waveform synthesis. In this mode, when a file is synthesized the waveform is calculated once to determine the peak output level, then the overall gain (G0) is adjusted so that the file is as loud as possible without peaking, and the waveform is calculated over again. {Batch mode and AGC mode were not included in the original KLSYN.}

```
klsyn<CR>           (default  KLSYN mode)
klsyn -n<CR>        (novice user print verbose messages)
klsyn -b<CR>        (batch mode)
klsyn -g<CR>(automatic gain control)
klsyn -b -g<CR>     (batch mode with automatic gain control)
```

The second type of argument, identified by the absence of a minus sign, is the name of a configuration file, as discussed above.

It is often the case that information such as the format of the arguments to KLSYN is needed, but this manual is not at hand. In almost all cases, typing '?' in place of an appropriate command or argument will get you a list of legal options at that point. In particular, the following command works: klsyn ?<CR>

Synthesis Commands

After startup initialization, the program types the prompt ">", indicating that it is ready to accept user commands from the terminal. The permitted commands (one-character typed symbols, not to be followed by a carriage return) are listed in Table II, and described more fully in the following paragraphs. {Note that commands (except '?') are lower-case letters.}

Table II. Synthesis commands in KLSYN.

Command	Action
?	Print a list of legal commands
q	Quit the program
p	Print the synthesis parameter default values
c	Change parameter default values
e	Enter synthesis parameter time function
s	Synthesize waveform

? HELP. At any point during synthesis, it is possible to type "?" and get a printout of the acceptable commands.

q QUIT Program. Any VAX program {and most PC programs} can be halted by typing '^C', i.e. control-c, but if this is done while the program is in graphics mode, it may leave the terminal in a very bad state. Therefore, the 'q' quit command is provided as the preferred method of aborting activity. There is no way to resume after a quit action, all parameter data is lost forever. Of course, normally, the program terminates gracefully by saving everything at the end of an 's' synthesize command.

p PRINT Synthesis Parameter Default Values. A complete list of synthesis constants and variable parameters, along with their default values and expected minimum and maximum values, is obtained by typing "p". If the list is too long for the terminal screen, one can halt the printout by hitting the "no-scroll" key {ctrl-s on PCs} to stop and {ctrl-q} to restart character transmission to the screen. After a waveform has been synthesized, it is possible to get a listing of this default information by typing the command:

```
print name.doc<CR>
```

c CHANGE Parameter Default Value. To change the default value of any synthesis constant or variable parameter, type "c". The system will respond:

```
Par:
```

and you should type the 2-character symbol for the parameter to be changed (or '?' to get a listing of acceptable 2-char names). If a legal 2-char name is typed, the system will then respond:

```
Change value of xx from yy to
        Value:
```

and you should type the desired value. If your requested value falls outside the range specified by the minimum and maximum values listed in the configuration, the system will ask whether you really want to use this value. Accepted responses are "y" (yes) or "n" (no). Do not respond "y" unless you are reasonably certain that such a request makes logical sense. Consider an example. To change the constant "duration of the stimulus" to 300 msec, one would type:

```
c
Par: du<CR>
Change value of du from 500 to
Val: 300<CR>
>
```

The default values for variable parameters are used throughout the synthesis unless a time function is specified for each parameter using the "d" command described immediately below.

e ENTER Synthesis Parameter Time Function. All variable parameters (those not followed by a "C" in the "V/C" column of Table I can be varied by specifying values for each time increment (5 msec default update interval), using the "e"command. The system responds with:

```
        Par:
```

and you should type the appropriate 2-character symbol. The program then requests (time,value) pairs for the synthesis parameter time funcion. It calculates (by linear interpolation) values for the time frames left unspecified while entering the functions.
So, for instance, if the user enters the pairs (0ms, 0dB); (50ms,60dB); (450ms,60dB); (495ms,0dB) to define an amplitde of voicing (AV) function, the value at time 100 ms will be set at 60 dB, while the value at time 5 ms will be 6dB. Time functions are entered interactively. When “time:” is the prompt, typing <CR> indicates that no more values are to be entered at the moment. Similarly, when “par:” is the prompt, typing <CR> indicates that no more variable parameters are to be modified. Values exceeding suggested limits are detected and the user is asked if this is really what was intended. A typical exchange (which puts a falling f0 trajectory on a 500ms syllable) looks like this:

```
>e
        par:f0
Empty line terminates input
                time:0
                value (was 1000):1400
                time:500
                value (was 1400):900
                time:<CR>
        par:<CR>
>
```

s SYNTHESIZE Waveform. When all of the variable parameters have been given appropriate default values or appropriate time functions, a waveform can be synthesized, using the "s" command. The first name of the waveform file is requested upon entering this command. Names should be descriptive of the synthesis, but are constrained to be no more than 8 characters in length, and composed of alphabetic characters and digits, freely mixed (the "-" and similar characters are not allowed).

After synthesis, the waveform is saved in a file which can be read into Cspeech for further analysis and comparison with a natural model. Relevant information to enable the user to resynthesize the identical utterance in the future is then saved in the "name.doc" file, and the program halts.

The peak output level is printed at the end of the synthesis in dB, where any number greater than zero dB indicates that the signal has exceeded the bits available to the digital-to-analog converter, and will have to be synthesized again with source amplitudes reduced (see the ‘G0’ configuration parameter in Table I). On the other hand, any level below about -12 dB will not use the two highest order bits of the d/a converter, and might be profitably resynthesized at a higher level, if this is consistent with the experiment. {The automatic gain control option avoids this problem.}To get back to a position where it is possible to modify the source amplitudes of the waveform stimulus called ‘name’, type:

```
klsyn name<CR>
```

The configuration/parameter file ‘name.doc’ is automatically reloaded. An overall gain control on the synthesis, "g0", is available so that one need not increase/decrease all the values in each active source amplitude parameter track to achieve a change in output signal level. Following synthesis of an important waveform, it is good practice to obtain a listing of the default parameter values and the variable parameter data for future reference. This is done with the command:

```
print name.doc<CR>
```

Parameters and Constants

A list of the constants and variable parameters that control the synthesizer is shown in Table I. The two-character symbols stand for full names given in column 2. Each control variable has been assigned a default value, which is indicated in the last column. It will be used during synthesis unless changed by the user. Parameters that must remain constant throughout an utterance are indicated by a ‘C’ in column 3, all others may bevaried using the ‘e’ enter command.

Minimum and maximum values are also indicated in columns 4 and 5. These are "soft" limits that can be over-ridden; they suggest normal range of variation, and help detect typing errors.

KLSYN synthesizer configuration at startup time

The following paragraphs define each of the constants and variable parameters of Table I. Though nominally variable, these parameters take on the default value for all time unless the user employs the ‘e’, "enter synthesis parameter time function", command to specify a parameter time function in the form of a sequence of straight-line segments.

sr The constant ‘sr’, "sampling rate", is the number of output samples computed per second of synthetic speech. It is suggested that the default value of 10,000 samples/sec not be changed unless the user understands the digital signal processing implications of such a change (for example, if only ‘sr’ is increased, the spectrum of the synthetic speech will tilt down). However, if a sampling rate of 16,000 samples/sec is desired, one can change ‘nf’, the number of formants in the cascade branch, to 8 and obtain synthesis that is nearly identical below 5 kHz to that generated at 10,000 samples/sec (see description below of parameter ‘nf’).

An antialiasing low-pass filter with a cutoff frequency of 4500 -4800 Hz must be used when playing out files created with the default 10,000 samples/sec setting for ‘sr’. If ‘sr’ is changed to any new value it is necessary to use a filter with a cutoff frequency appropriate for the sampling frequency.

du The constant ‘du’, "duration", of the utterance to be synthesized, is the number of msec from beginning to end of the current synthetic utterance, including at least 25 msec at the end to allow the waveform to decay naturally after you have turned off all the sound sources.

The current maximum value for ‘du’ is 1000 (one second). (Actually, the maximum utterance duration is 200 frames times ui). The specified value for ‘du’ will be rounded up to the nearest multiple of ‘ui’, the number of msec in a parameter update time interval.

ui The constant ‘ui’, "update interval", is the number of msec of waveform generated between times when parameter values are updated. The default value of 5 ms is frequent enough to mimic most rapid parameter changes that occur in speech (in fact, 10 ms updates may be often enough). Under special circumstances, a shorter update interval, e.g. 1 ms, might be desirable, but note the qualification given in the next paragraph.

Parameters involved in generating the glottal source waveform (‘f0’ ‘av’ ‘oq’ ‘tl’ ‘sk') are not changed at the exact time specified by the update interval. Instead, their change in value is delayed to the next waveform sample at which glottal opening occurs. For low values of fundamental frequency, this delay may be as much as 10 ms (the average delay is 5 ms when ‘f0’ is 100 Hz, and 2.5 ms when f0=200 Hz).

If this were not done, it would be as if spurious excitation occurred at the update rate, resulting in perceptible auditory distortion. (The fact that formant frequencies and bandwidths change at the update time means that small waveform distortions synchronized to the update rate are unavoidable.) Delaying changes to the voicing source control parameters in order to synchronize them with the time of primary excitation of the vocal tract both removes the update interval periodicity of the distortions, and better hides them under the signal.

nf The constant ‘nf’, "number of formants in cascade vocal tract", specifies how many formants, counting from F1 up to a maximum of F8, are actually in the cascade vocal tract. The default value is 5, which is an appropriate number if the sampling rate is 10,000 samples/sec and the speaker has a vocal tract length of 17 cm. (i.e. the average spacing between formants will then be 1000 Hz).

If the speaker that you are trying to model has a vocal tract length significantly different from 17 cm, or if the ‘sr’ sampling rate parameter has been changed, you may wish to modify ‘nf’. For example, to model a typical female voice with a vocal tract length about 20% shorter than the average male, one would set ‘nf’ to four.

If the sampling rate is changed to 16,000 samples/sec, then a male voice should have 8 formants in the frequency range from 0 to 8 kHz, and thus ‘nf’ should be set to 8. Only the lower 6 formant frequencies and bandwidths are settable by the user; the frequency and bandwidth of the seventh and eighth formants are fixed at F7=6500, B7=500, F8=7500, B8=600. The parallel vocal tract has only 6 formants, so that one would have to move F6 up in frequency to generate noise spectra with peaks above the default value of F6=4990 Hz when ‘sr’ is increased.

It should be clear that ‘nf’ only crudely approximates variations in vocal tract length. If, for example, a speaker had a vocal tract length 10% shorter than the typical male, one would have to use five formants in the cascade branch, setting the higher formants appropriately higher in frequency, and then use the ‘tl’ tilt parameter to achieve the correct general spectral tilt for this voice.

ss The constant ‘ss’, "source switch", is a switch that determines which of two voicing source waveforms is used for synthesis. The default value, 1, causes a low-pass filtered impulse train to be generated, while the value 2 causes a more natural waveform with a definite sharp closing time to be invoked. Each has its own set of advantages and disadvantages.

Impulse Train. A train of impulses is filtered by a critically damped second-order low-pass digital filter, resulting in an approximation to the glottal waveform such as is shown in Figure 2. The spectrum falls off at -12 dB per octave for low and mid frequencies and then flattens out. (Above 4 kHz, harmonics are further attenuated by a down-sampling low-pass filter, but this should have little effect on the perceived quality of a vowel.)

The primary advantage of the filtered impulse train is that the source spectrum is perfectly regular, with no ‘glottal zeros’. The 2-pole low-pass filter has a nominal cutoff frequency of zero Hz, and a bandwidth (which determines the width of the open portion before the time waveform asymptotically approaches zero) that is proportional to the synthesis parameter ‘oq' (the open portion of the voicing period expressed in percent). The spectrum of this source can be tilted down to simulate a mode of vibration where the vocal folds do not meet at the midline, using the ‘tl’ tilt-of-the-glottal-source parameter described below.

The disadvantage of this waveform is that primary excitation of the vocal tract occurs at glottal opening time, and there is no excitation at glottal closing time. Thus the phase of the source is incorrect, even though the source magnitude spectrum is probably to be preferred for its regularity, at least in some psychophysical tests. (Fortunately, the phase of the source spectrum is not of great perceptual importance, especially under listening conditions where room acoustics impose their own phase distortions on the sound reaching the ears.)

----
insert Figure 2 about here

Figure 2. Comparison of waveforms and spectra of the impulsive (‘ss’=1) and natural (‘ss‘=2) glottal sources, using default settings to all other parameters.

Natural Pulse Train. The advantages of the natural glottal source are that the glottal volume velocity waveform has well-defined open and closing times, with an asymmetrical shape such that closing velocity is more rapid than opening velocity. The voicing volume velocity waveform obeys the equation:

Ug(t) = at**2 - bt**3

during the open phase of the period, and is zero for the remainder of the period. [The choice of synthesis waveform shape is based on suggestions contained in Rothenberg (1971) and in Fant (1983).] The spectrum of the natural source is somewhat irregular, with a weak zero at about 600 Hz (assuming default settings to all of the glottal source parameters except ‘ss’, which is set to 2.) Waveforms and spectra for the impulsive and natural voicing sources are compared in Figure 2. The natural glottal waveform can also be modified so as to tilt the spectrum down, using either ‘oq’ or ‘tl’, in order to mimic the effects of incomplete glottal closure and the concomitant rounding of the corner of the waveform at closure.

The disadvantage of the natural source waveform is that the magnitude spectrum is somewhat irregular, so that a formant will be slightly attenuated as it approaches a frequency of about 600 Hz (the actual zero locations depend on ‘oq’, the percent of the period in the open phase). This formant amplitude variation seems to occur in natural speech, but may not be desirable for particular synthesis stimulus sets.

rs The constant ‘rs’, "random seed", is the seed value given to the random number generator routine. Any number from 0 to 99999 can be specified. For each, you will get a quite different random number sequence (different frication and aspiration noises from those used to generate the previous stimuli).

On the other hand, stimuli all generated with the same value for ‘rs’ will have identical frication source and aspiration source waveforms. This is sometimes desirable if stimuli on a continuum are not to differ due to random fluctuations in e.g. a burst of frication noise.

os The constant ‘os’, "output waveform selector", determines which waveform is saved in the output file. If ‘os’ has the default value of zero, the normal final output of synthesis is saved. Other output options are given in Table III. For example, if you wished to see and spectrally analyze the voicing source waveform of the synthesizer by itself for a particular synthetic utterance, you would set ‘os’ to four.

Note that the radiation characteristic is applied if ‘os’ is greater than 4, but not if ‘os’ is less than 4. (Due to computational considerations, the derivative of the voicing source is usually computed directly, so that the actualsource waveform that is displayed when requested is approximated by sending the computed source waveform through a leaky integrator.) Thus, setting’os’=4 results in the actual voicing source waveform being generated, while setting’os’=5 produces the (first difference) of the voicing source waveform that ordinarily is routed to the parallel vocal tract model.

Table III. KLSYN output waveform options using ‘os’.

os	WAVEFORM SAVED	
0.	Normal synthesis output	
1.	Voicing periodic component alone	
2.	Aspiration alone	
3.	Frication alone	
4.	Glottal source (voicing, turbulence, and aspiration)	
5.	Glottal source sent to parallel vocal tract (AP)	+ radiation char
6.	Cascade vocal tract, output of nasal zero resonator	"
7.	Cascade vocal tract, output of nasal pole resonator	"
8.	Cascade vocal tract, output of fifth formant	"
9.	Cascade vocal tract, output of fourth formant	"
10.	Cascade vocal tract, output of third formant	"
11.	Cascade vocal tract, output of second formant	"
12.	Cascade vocal tract, output of first formant	"
13.	Parallel vocal tract, output of sixth formant alone	"
14.	Parallel vocal tract, output of fifth formant alone	"
15.	Parallel vocal tract, output of fourth formant alone	"
16.	Parallel vocal tract, output of third formant alone	"
17.	Parallel vocal tract, output of second formant alone	"
18.	Parallel vocal tract, output of first formant alone	"
19.	Parallel vocal tract, output of nasal formant alone	"
20.	Parallel vocal tract, output of bypass path alone	"

f0 The variable ‘f0’, "fundamental frequency", is the rate at which the vocal folds are currently vibrating in Hz times 10. I.e. if a fundamental frequency of 100 Hz is desired, then’f0’ is set to 1000. The additional accuracy resulting from a specification of fundamental frequency to 0.1 Hz adds some naturalness to a slowly changing pitch glide.

A new fundamental period is computed each time the vocal folds begin to open. The value of ‘f0’ existing at that time instant is used to determine the new period. Several other parameters of the voicing source (‘av’, ‘no’, ‘tl’, ‘sk’) change value at this time rather than changing at the nominal update time -- otherwise discontinuities could occur in the voicing waveform.

The fundamental period is quantized in a digital speech synthesizer. In this simulation, the period (time between instants when glottal opening occurs) is quantized to increments of 1/40000 sec. (Patent pending by Digital Equipment Corporation). This means that at 100 Hz, ‘f0’ is effectively specified in 0.25 Hz steps (0.25% quantization error), while at 200 Hz, ‘f0’ is quantized in 0.5 Hz steps (still a 0.25% quantization error in ‘f0’). This accuracy is necessary to avoid perceptible "staircase pitch" problems for slowly gliding ‘f0’ in the higher pitch ranges; it is achieved by running the glottal source simulation at a sampling rate four times that specified by ‘sr’, and lowpass/downsampling this waveform before sending it to the vocal tract model.

av The variable ‘av’, "amplitude of voicing" is the amplitude in dB of the voicing source waveform sent through the cascade vocal tract. A value of 0 dB turns off (zeros) the signal. A value of about 60 dB produces a level for vowel synthesis that is close to the maximum non-overloading level; such values should be used to keep the signal in the higher-order bits of the digital-to-analog converter.

The synthesizer does not necessarily turn voicing on and off at exactly the time specified by the ‘av’ time function. The effect of a change in ‘av’ is delayed until the instant of the next glottal waveform opening. If the natural source, ‘ss’=2, is used, the primary excitation of the vocal tract actually begins even later, at glottal closure some ‘oq’ percent of the voicing period following the time of glottal opening.

If ‘av’ is suddenly turned off, no more glottal pulses will be issued, and the vocal tract response to the previous pulse will continue to die out, taking 10 to 20 msec to become totally inaudible.

If ‘av’ is suddenly turned ON, and you wish a glottal pulse to be issued at exactly that time, it is necessary to have set ‘f0’ to zero for a period of time prior to this event, and to turn ‘f0’ on simultaneous with the time that ‘av’ is turned on. This procedure should be followed in order to specify voice onset time for a plosive as an exact number of update intervals later than burst onset.

ah The variable ‘ah’, "amplitude of aspiration", is the amplitude in dB of the aspiration noise sound source that is combined with periodic voicing, if present (‘av’>0), to constitute the glottal sound source that is sent to the cascade vocal tract. (Voicing can be sent to the parallel vocal tract by making ‘ap’ non-zero, but aspiration cannot be sent to the parallel vocal tract. Instead, one would use ‘af’, the amplitude of frication noise.) A value of zero turns off the aspiration source, while a value of 60 results in an output aspirated speech sound with levels in formants above F1 roughly equal to the levels obtained by setting ‘av’ to 60.

The spectrum of the aspiration noise source is nearly flat, actually falling slightly with increasing frequency. To best approximate an aspirated speech sound, one should probably increase ‘b1’, the first formant bandwidth, to anywhere from 200 to 400 Hz, thus simulating the effect of additional low-frequency losses incurred when the glottis is partially open.

at The variable ‘at’, "amplitude of turbulence", is the amplitude in dB of turbulence noise generated at the glottis during the open phase of a glottal vibration. The noise is identical to aspiration except (1) the source is turned off during the closed phase of a glottal cycle, and (2) the output level rises and falls with changes to the variable ‘av’. Thus this breathiness dimension of voicing is zero when ‘av’ is set to zero, whereas aspiration noise is not influenced by the setting of ‘av’. Usually ‘ah’ is used to generate aspiration for voiceless aspirated plosives and [h] sounds, while ‘at’ is used to add a breathiness quality to the voicing source.

A value of 60 will make the voice quite breathy. To achieve a good match to natural breathiness, however, one should probably also tilt down the source spectrum, using ‘tl’, increase the open phase of a glottal cycle, ‘oq’, to a little more than half the period, and perhaps increase ‘b1’.

oq The spectrum of a voicing source pulse train can vary in two fairly distinct ways. The relative amplitude of the first harmonic can increase or decrease, or the general tilt of the spectrum can go up and down. To change primarily just the first harmonic amplitude, the ‘oq’ parameter is varied, while the parameter ‘tl’ affects the general spectral tilt (see Figure 3).

----
insert Figure 3 about here

Figure 3. The effect of changes in ‘oq’ (top panel) and changes in tl’ (bottom panel) on the waveform and spectrum of the voicing source.

The variable ‘oq’, "percent of voicing period with glottis open", is a nominal indicator of the width of the glottal pulse when using the default impulse train glottal source, and it is the exact number of samples in the open period when using the natural voicing source (‘ss’=2). A value of ‘oq’=50, the default value, corresponds to a 5 msec open portion of the fundamental period at the default sampling rate(10000 samples/sec) and default F0(100 Hz*10), see Figure 3.

There are many male speakers for whom the duration of the open portion of the fundamental period does not change as fundamental frequency changes over a fairly wide range. To simulate the behavior of this kind of speaker, one must adjust ‘oq’ to be inversely proportional to the fundamendal frequency parameter ‘f0’, which is rather a bother.

Other speakers tend to produce speech with the duration of the open portion of the cycle being a constant fraction of the total period, e.g. about half of the period. To synthesize this type of speech it is not necessary to change ‘oq’ during synthesis.

The effect of changes in ‘oq’ on the spectrum is illustrated in Figure 3. A narrow glottal pulse, as may occur in creaky voice, or when trying to speak loudly, results in a spectrum relatively rich in higher-frequency components, while a wider glottal pulse, as may occur in a breathy offset to speaking, results in a spectrum rich in energy below the first formant. Thus to match an observed strong first harmonic in the spectrum of a natural utterance, increase ‘oq’.

The synthesizer routine checks to see that ‘oq’ does not result in an open portion of the glottal pulse which exceeds the duration of the period, and truncates requests that exceed the duration of the current period and prints a warning to the user about the inappropriate value of ‘oq’.

tl The variable ‘tl’, "spectral tilt of voicing", is the (additional) downward tilt of the spectrum of the voicing source, in dB as realized by a soft one-pole low-pass filter. The effect of changes in ‘tl’ on the voicing source spectrum is illustrated in Figure 4. A value of zero has no effect on the source spectrum, while a value of 24 tilts the spectrum down gradually such that frequency components above about 3 kHz are attenuated by about 24 dB relative to a more normal source spectrum.

The tilt parameter is an attempt to simulate the spectral effect of a "rounding of the corner" at the time of closure in the glottal volume velocity waveform due either to an incomplete closure, as in breathiness, or an asynchronous closure such that the anterior portion of the vocal folds meet at the midline before the posterior portions come together.

The tilt parameter is also useful in simulating a voicebar, wherein only lower-frequency components are radiated from the closed vocal tract. For many speech synthesis situations, this would be the only use for ‘tl’. However, ‘tl’ is a good parameter to use in attempts at matching the spectral details of a particular natural utterance.

sk The variable ‘sk’, "skew to alternate periods", is the number of 25 microsecond increments to be added to and subtracted from successive fundamental period durations in order to simulate one aspect of vocal fry, the tendency for alternate periods to be more similar in duration than adjacent periods.

Such aperiodicities, when introduced, have fairly strong perceptual consequences. This kind of change to normal voicing occurs throughout speech for some voices, and at the initiation and cessation of voicing in a sentence for many others. There is no need to play with this parameter in most synthesis situations.

F1 F2 F3 F4 F5 f6 The "formant frequency" variables determine the frequency in Hz of up to six resonators of the cascade vocal tract model, and of the frequency in Hz of each of six additional parallel formant resonators. Normally, the cascade branch of ‘nf’=5 formants is used to generate voiced and aspirated sounds, while the parallel branches are used to generate fricatives and plosive bursts. Since formants are the natural resonant frequencies of the vocal tract, and frequency locations are independent of source location, the formant frequencies of cascade and corresponding parallel resonators must be identical.

Suggested values for formant frequencies of a number of English sounds were published in Klatt (1980). The tables are reproduced below as Table IV and Table V for easy reference, although it is recommended that synthesis parameter values be based on analysis, synthesis, and comparison of a real utterance, rather than just from theory and matches to the idiolect of D. Klatt.

Table IV. Formant frequency and bandwidth targets for selected English vowels. If two values are given the first occurs early in the vowel and the second occurs late in the vowel.

Vowel	F1	F2	F3	b1	b2	b3
[i]	310	2020	2960	45	200	400
	290	2070	2960	60	200	400
[ɪ]	400	1800	2570	50	100	140
	470	1600	2600	50	100	140
[eɪ]	480	1720	2520	70	100	200
	330	2020	2600	55	100	200
[ɛ]	530	1680	2500	60	90	200
	620	1530	2530	60	90	200
[æ]	620	1660	2430	70	150	320
	650	1490	2470	70	100	320
[ɑ]	700	1220	2600	130	70	160
[ɔ]	600	990	2570	90	100	80
	630	1040	2600	90	100	80
[ʌ]	620	1220	2550	80	50	140
[oʊ]	540	1100	2300	80	70	70
	450	900	2300	80	70	70
[ʊ]	450	1100	2350	80	100	80
	500	1180	2390	80	100	80
[u]	350	1250	2200	65	110	140
	320	900	2200	65	110	140
[ɹ]	470	1270	1540	100	60	110
	420	1310	1540	100	60	110
[aɪ]	660	1200	2550	100	70	200
	400	1880	2500	70	100	200
[aʊ]	640	1230	2550	80	70	140
	420	940	2350	80	70	80
[oɪ]	550	960	2400	80	50	130
	360	1820	2450	60	50	160

Formant frequencies generally move continuously and slowly in time (relative to the default 5 msec parameter update interval 'ui'). An exception is the closure and release of a stop consonant. During closure, the first formant 'F1' is typically at a frequency of about 180 Hz. (The first formant frequency does not go below about 180 Hz under any circumstances due to the mass and compliance of cavity walls and air trapped in the closed vocal tract.) Upon release, the first formant frequency may rise quite rapidly over the first 5 to 10 msec, giving the appearance of a discontinuous jump to a frequency to as high as 400 Hz at the time of the first visible glottal pulse following the burst in a syllable such as [ba].

Table V. Formant frequency and bandwidth targets for selected English consonants.

Son.	F1	F2	F3	b1	b2	b3					
[w]	290	610	2150	50	80	60					
[j]	260	2070	3020	40	250	500					
[r]	310	1060	1380	70	100	120					
[l]	310	1050	2880	50	100	280					
Fric.	F1	F2	F3	b1	b2	b3	a3	a4	a5	a6	ab
[f]	340	1100	2080	200	120	150	0	0	0	0	57
[v]	220	1100	2080	60	90	120	0	0	0	0	57
[θ]	320	1290	2540	200	90	200	0	0	0	28	48
[ð]	270	1290	2540	60	80	170	0	0	0	28	48
[s]	320	1390	2530	200	80	200	0	0	0	52	0
[z]	240	1390	2530	70	60	180	0	0	0	52	0
[ʃ]	300	1840	2750	200	100	300	57	48	48	46	0
Affr.											
[tʃ]	350	1800	2820	200	90	300	44	60	53	53	0
[dʒ]	260	1800	2820	60	80	270	44	60	53	53	0
Plos.											
[p]	400	1100	2150	300	150	220	0	0	0	0	63
[b]	200	1100	2150	60	110	130	0	0	0	0	63
[t]	400	1600	2600	300	120	250	30	45	57	63	0
[d]	200	1600	2600	60	100	170	47	60	62	60	0
[k]	300	1990	2850	250	160	330	53	43	45	45	0
[g]	200	1990	2850	60	150	280	53	43	45	45	0
Nas.	fp	fz	F1	F2	F3	b1	b2	b3			
[m]	270	450	480	1270	2130	40	200	200			
[n]	270	450	480	1340	2470	40	300	300			

b1 b2 b3 b4 b5 b6 The "formant bandwidth" variables determine the bandwidths of resonators in the cascade vocal tract model. Since formant bandwidths depend in part on source impedance, and turbulence sources contribute more losses, the synthesizer provides separate control of bandwidths ‘p1’ ‘p2’ ‘p3’ ‘p4’ ‘p5’ ‘p6’ for the parallel formants.

If the number of formants in the cascade branch is left at the default value of ‘nf’ = 5, then the ‘b6’ variable has no meaning and no effect on the synthetic waveform.

The resonator bandwidth variable has two effects on the frequency-domain shape of the vocal tract transfer function. An increase in bandwidth reduces the amplitude of the formant peak and simultaneously increases the width of the peak as measured 3 dB down from the peak. Perceptual experiments indicate that both of these changes have perceptual consequences, but that the change in peak height is much more audible than the width change.

In a cascade synthesizer, adjustments to formant peak heights in order to match the spectrum of a recorded voice can be achieved either by changing the general slope of the voicing source spectrum (using ‘tl’) or by changing individual formant bandwidths. Changing formant bandwidths is an effective way to mimic quite closely the voice quality of a speaker, but some guidelines are offered to help avoid the perceptual problems of aberrant bandwidth specification:

1. If a bandwidth is set to a value less than the soft limits given in Table 1, there is a danger that whistle-like harmonics will be heard when a harmonic of the fundamental sweeps past the formant frequency.

2. If the bandwidths of the lower formants are wider than the suggested guidelines of Table 1, the synthetic voice will begin to sound buzzy. In this case, all bandwidths should be reduced, and then ‘av’ can be reduced to get back to an appropriate overall spectral level.

fp fz The variable ‘fp’, "frequency nasal pole", in consort with the variable ‘fz’, "frequency nasal zero", can mimic the primary spectral effects of nasalization in vowel-like spectra. In a typical nasalized vowel, the first formant is split into peak-valley-peak (pole-zero-pole) such that ‘fp’ is at about 300 Hz, ‘F1’ is higher than it would be if the vowel were non-nasalized, and ‘fz’ is at a frequency approximately halfway between ‘fp’ and ‘F1’.

When returning to a non-nasalized vowel, ‘fz’ is moved down gradually to a frequency exactly the same as ‘fp’. The nasal pole and nasal zero then cancel each other out, and it is as if they were not present in the cascade vocal tract model.

bp bz The variables ‘bp’, "bandwidth nasal pole", and ‘bz’, "bandwidth nasal zero", are set to default values of 90 Hz. It is difficult to determine appropriate synthesis bandwidths for individual nasalized vowels, but, fortunately, one can achieve good synthesis results without changing these default values in most cases.

af The variable ‘af’, "amplitude frication", determines the level of frication noise sent to the various parallel formants and bypass path. The variable should be turned on gradually for fricatives (e.g. straight line from 0 to 60 dB in 90 msec), and abruptly to about 60 dB for plosive bursts.

a1 a2 a3 a4 a5 a6 ab The variables ‘a1’ ‘a2’ ‘a3’ ‘a4’ ‘a5’ ‘a6’ ‘ab’, "amplitudes parallel formants", determine the spectral shape of a fricative or plosive burst. If a formant is a front cavity resonance for a particular fricative articulation, one might set the formant amplitude to 60 dB as a first guess. Formants associated with the cavity in back of the constriction should have their amplitudes set to zero initially, (The amplitude of the first parallel formant,’a1’, is therefore zero for all English fricatives.) and then all parallel formant amplitudes should be adjusted on a trial-and-error basis, comparing synthesized frication spectra with a natural frication spectrum.

The bypass path amplitude is used when the vocal tract resonance effects are negligible because the cavity in front of the main fricative constriction is too short, as in [f], [v], [th], [dh], [p], [b].

p1 p2 p3 p4 p5 p6 The variables ‘p1’ ‘p2’ ‘p3’ ‘p4’ ‘p5’ ‘p6’, "bandwidths parallel formants" are set to default values that are wider than the bandwidths used in the cascade vocal tract model. It is difficult to measure formant bandwidths accurately in noise spectra, even when a fairly long sustained fricative is available for analysis. However, these default values can be used in most situations. The only adjustment is then made to the parallel formant amplitudes in order to match details in a natural frication spectrum.

All-Parallel Synthesis Using ‘ap’ and ‘an’ The variable ‘ap’, "amplitude voicing parallel", is the amplitude, in dB, of voiced excitation of the parallel vocal tract. Normally, this would be allowed to remain at the default value of zero since the cascade vocal tract would be used for generating the voicing component of all voiced sounds (even voicebars and voiced fricatives).

However, there are circumstances where a vowel with special characteristics (e.g. two-formant vowels) can only be generated using the greater flexibility (individual control of formant amplitudes) of the parallel vocal tract. A value of ‘ap’ = 60 would be a good choice to synthesize a typical vowel using the parallel vocal tract model. Of course, ‘av’, would be set to zero.

The parallel formant amplitude variables must then be adjusted to get the right spectral shape for the vowel. A good starting point is to set parallel formant amplitudes ‘a1’ ‘a2’ ‘a3’ ‘a4’ ‘a5’ to 60 dB. This will give exactly the right relative formant amplitudes for a non-nasalized vowel with formant frequencies at 500, 1500, 2500, 3500 and 4500 Hz. However, as formant frequencies are changed from these values (appropriate for a uniform tube), formant amplitudes can quickly diverge from those in a corresponding cascade vocal tract model. (The formant amplitude will increase/decrease as formant frequency is increased/decreased, but there is no automatic adjustment such that formants "riding on the skirt" of a lower-frequency formant are attenuated as this formant frequency is lowered.) Trial-and-error adjustment of parallel formant amplitudes is then necessary.

an The variable ‘an’, "amplitude parallel nasal formant", is normally not used. However, when employing the parallel vocal tract to synthesize vowels, as discussed above, ‘an’ can be used to simulate the effects of nasalization on vowels and nasal murmurs. To achieve nasalization, one would set ‘fp’ to about 280 Hz (the default value) and adjust both ‘an’ and ‘a1’ to levels matching a nasalized vowel spectrum.

g0 An overall gain control, ‘g0’, is included to permit the user to adjust the output level without having to modify each source amplitude time function. The nominal value is 60 dB. To increase the output by e.g. 3 dB, one would simply use the ‘c’ command to set ‘g0’ to 63. In unusual circumstances, it might be desirable to make ‘g0’ a variable, and control it as a function of time. This is permitted, although I can’t think of a very good example of when such a procedure would be advantageous.

References

Fant, G. (1983), “The Voice Source: Acoustic Modeling”, *Speech Transmission Laboratory,* **QPSR 4/1982**, Royal Institute of Technology, Stockholm, Sweden, 28-48.

Johnson, K. & Teheranizadeh, H. (1992), “Facilities for speech perception research at the UCLA phonetics lab”, UCLA Working Papers in Phonetics, **??**, ??-??.

Klatt (1980) "Software for a Cascade/Parallel Formant Synthesizer", *J. Acoust. Soc. Am.* **67**, 971-995.

Rothenberg, A. (1971), “Effect of Glottal Pulse Shape on the Quality of Natural Vowels”, *J. Acoust. Soc. Am*., **53**, 1632-1645.