Chapter 3

Description of the Cascade/Parallel Formant Synthesizer

The Klattalk system uses the KLSYN88 cascade-parallel formant synthesizer that was first described in Klatt and Klatt (1990). This speech synthesizer is based on an original design described in Klatt (1980). This chapter[1] will outline the acoustic theory upon which KLSYN88 is based and will motivate a number of design decisions and modifications that have been made since the publication of the original Klatt (1980) synthesizer. Specific synthesis equations for sound sources and vocal tract transfer functions of various types are presented in Section 3.2. The effect on the synthesis of each control parameter is described in detail in Section 3.3, and examples are given of synthesis values for several different types of speech sounds. A final section will discuss simplifications made to permit real-time hardware implementation of the synthesizer in Klattalk. Readers familiar with Klatt (1980) may wish to skim over the first two sections and concentrate on the control parameter definitions of Section 3.3.

3.1 Overview

The broadband spectrogram shown in the lower half of Figure 3.1 illustrates a method of analyzing speech to determine the frequencies at which energy is present as a function of time. Time is plotted along the horizontal axis, frequency along the vertical axis, and blackness at any point is monotonically related to the energy in a frequency band 300 Hz wide, as averaged over a time interval of 1 or 2 ms.

The waveform shown at the top of Figure 3.1 shows a short sample of periodic voicing from the /ɪ/ of "this" followed by aperiodic noise from the /s/. Voicing is generated by

[1]This version was printed November 6, 1990 from the file Vax [klatt.klattalkbook] ch3klsyn88.tex. For the most up to date versions, copy to DEC20 using CFTP program, and big-latex it. All figures are in the top left filing cabinet drawer of my office labeled "Manuscripts in preparation". Book title: KLATTALK. Subtitle: The Conversion of English Text to Speech.

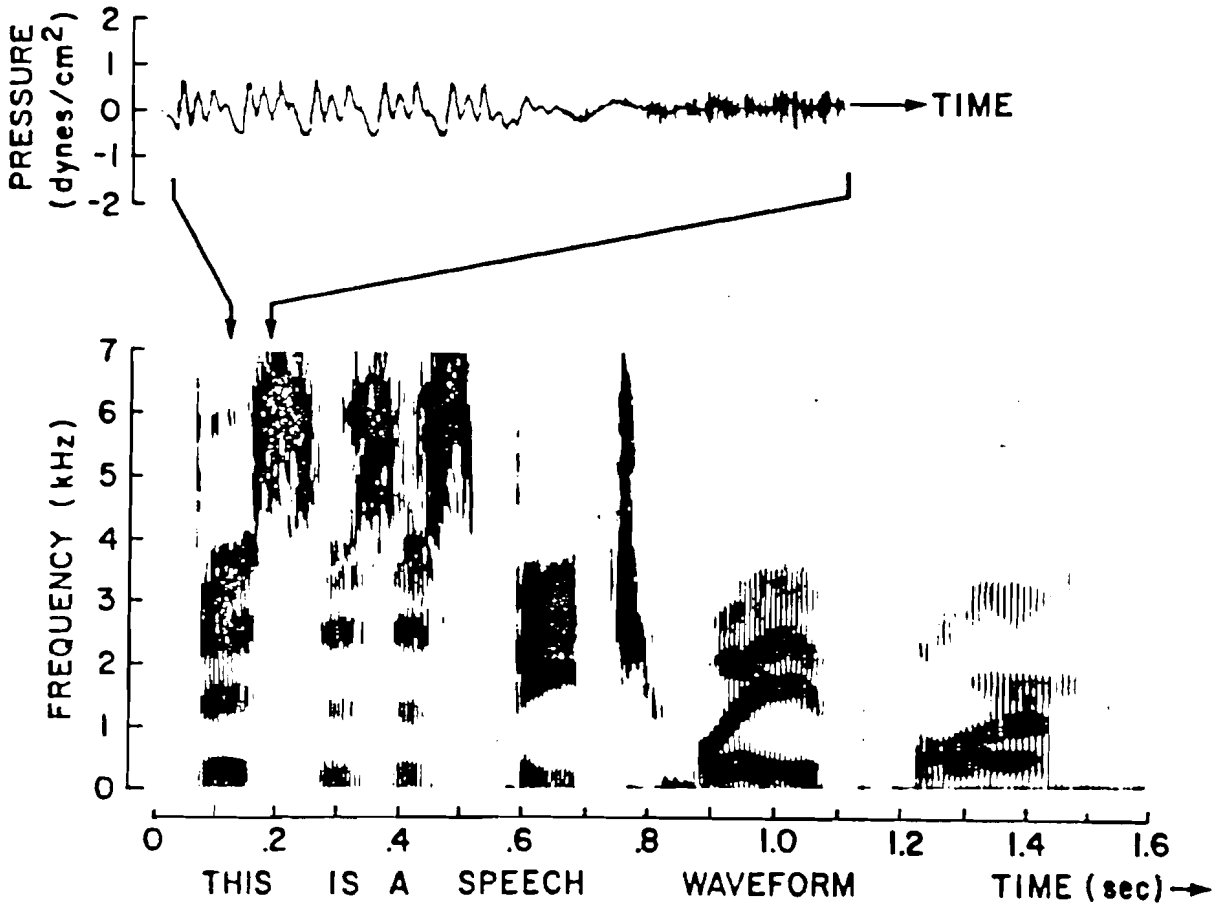


Figure 3.1: *Sound pressure waveform (top) and broadband sound spectrogram (bottom) of a sample of speech.*

vibrations of the vocal folds of the larynx. It shows up on the spectrogram as a series of vertical striations, each corresponding roughly to the excitation caused by the sudden termination of airflow as the vocal folds come together during vibration.

The horizontal dark bands seen during voicing are due to the resonances of the vocal tract which modify the sound produced by the voicing source. These resonances, which are called formants, move about in time in response to the motions of articulators such as the tongue, jaw, lips, and velum. Formant frequencies are the main acoustic evidence to indicate the articulation employed by the speaker during the production of many speech sounds.

During noise production, as in the /s/ of "this", a turbulence noise source is created at a constriction formed by the tongue tip. Only the "front cavity" (i.e., the portion of the vocal tract in front of the constriction) higher frequency natural resonant modes of the vocal tract are excited to form broad dark areas in the spectrogram. The lower formants, being back-cavity resonances, are usually not excited by the noise source.

In general, speech is produced by the sequential activation of one or more sound sources which then excite the resonances of the vocal tract, resulting in characteristic acoustic

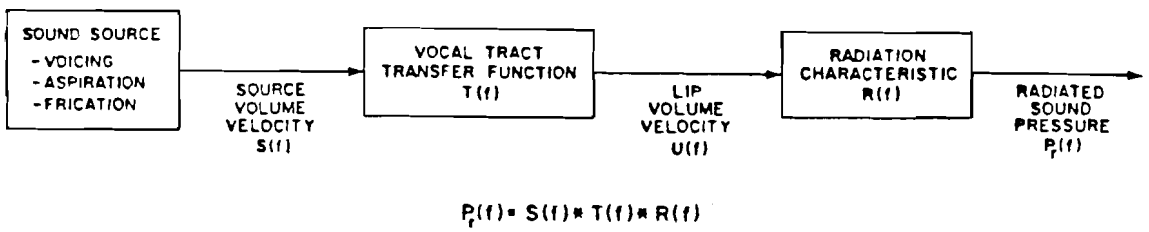


Figure 3.2: *The output spectrum of a speech sound, $P(f)$, can be represented in the frequency domain as a product of a source spectrum $S(f)$, a vocal tract transfer function, $T(f)$, and a radiation characteristic, $R(f)$.*

patterns of sound radiating from the lips and nose of the talker. Only two basic types of sound sources are activated to produce most of the sounds of the languages of the world: (1) quasi-periodic sources involving the vibration of some structure such as the vocal folds, tongue tip, lips or uvula, and (2) turbulence noise sources – either sustained, as in a fricative such as /s/ or an aspirated sound such as /h/, or a brief burst of noise, as at the release of a plosive such as /t/ and /d/. For some sounds, a transient source is generated by abruptly releasing a closure behind which a positive or negative pressure has been created.

Duplication of the acoustic patterns that appear on spectrograms serves as the end goal of efforts to produce synthetic speech. The ultimate criteria are, of course, perceptual: "Is the speech intelligible? Does it sound natural?" However, Holmes (1961; 1973) has shown that a well-designed formant-based speech synthesizer which can duplicate the pattern seen on a broadband spectrogram, i.e. the smoothed magnitude spectrum as it changes in time, is capable of producing speech that is indistinguishable from the original recording. In this sense, it is reasonable to base synthesizer performance on objective spectral comparisons, rather than informal subjective listening that is known to be strongly influenced by experimenter bias and expectations.

Historically, speech synthesizers fall into two broad categories: (1) articulatory synthesizers that attempt to model faithfully the mechanical motions of the articulators and the resulting distributions of volume velocity and sound pressure in the lungs, larynx, vocal tract and nasal tract, and (2) formant synthesizers which attempt to approximate directly the speech waveform and spectrum by a simpler model formulated in the acoustic domain. Klattalk employs a formant model of speech generation since current articulatory models are fairly primitive, and rules to control the muscles or shapes of the articulators in such a model are difficult to optimize.

The synthesizer design employed in Klattalk is based on an acoustic theory of speech production that was first presented in Fant (1960), and is summarized in Figure 3.2. According to this view, one or more sources of sound energy are activated by creation of a pressure drop across a constriction in the airway, usually by the build-up of lung pressure. Treating each sound source separately, we may characterize it in the frequency domain by

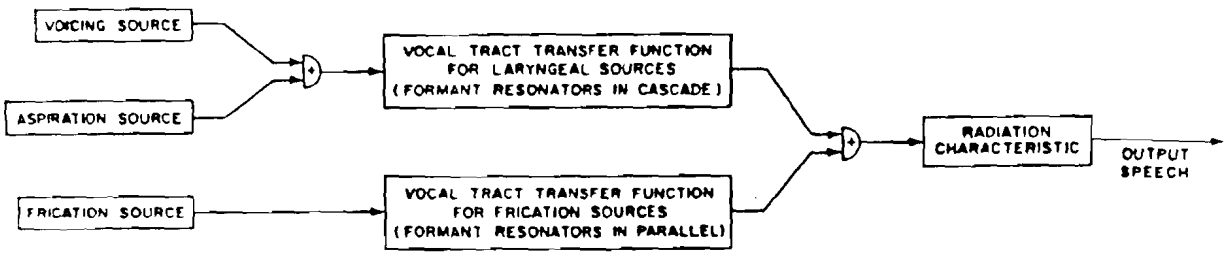


Figure 3.3: *Simplified block diagram of the synthesizer.*

a source spectrum, $S(f)$, where f is frequency in Hz. Each sound source excites the vocal tract, which acts as a resonating system analogous to an organ pipe. Since the vocal tract is a linear system, it can be described in the frequency domain by a linear transfer function, $T(f)$, which is a ratio of output lip-plus-nose volume velocity, $U_l(f)$, to source input, $S(f)$. Finally, sound is radiated from the lips and/or nose. The spectrum of the sound pressure that would be recorded some distance from the lips of the talker, $P_r(f)$, is related to lip-plus-nose volume velocity, $U_l(f)$, by a radiation characteristic, $R(f)$, that includes the effects of directional sound propagation from the head.

Each of the above relations can also be recast in the time (waveform) domain. This is the domain in which a waveform is actually generated in the computer. A sampled version of $P_r(t)$, denoted by $P_r(nT)$ consists of samples of the synthetic output waveform that are usually taken 10,000 times/second, i.e. every $T = 0.0001$ seconds, where n is an integer.

The synthesizer to be described includes components to simulate the generation of several different kinds of sound sources, components to simulate the vocal-tract transfer function, and a component to simulate sound radiation from the head. A simplified block diagram of the synthesizer is shown in Figure 3.3. The laryngeal sound sources—voicing and aspiration noise (as in /h/)—are combined into a glottal [2] volume velocity waveform $U_g(t)$ that excites the vocal tract. The vocal-tract model consists of digital formant resonators connected in cascade (the output of one serving as the input to the next). The output of the vocal-tract model is a lip volume velocity waveform, $U_l(t)$.[3] Radiation of this sound about the head results in a sound pressure waveform $P_r(t)$ that can be measured by a microphone placed about a fixed distance in front of the head.

There is a second model of the transfer function of the vocal tract when the sound source is not at the larynx, as for example in a fricative or plosive. In this latter case, a frication source generates a turbulence noise waveform that excites a set of digital formant

[2]The glottis is the space between the vocal folds of the larynx.

[3]When a nasal consonant is produced, sound propagates from the nasal tract, so that $U_l(t)$ should be thought of as including any volume velocity from the nares.

resonators that have to be connected in parallel for reasons to be explained shortly. The resulting output lip volume velocity is again transformed into radiated sound pressure by the radiation characteristic.

Waveform Sampling Rate

Most of the sound energy of speech is contained in frequencies between about 70 and 7000 Hz (Dunn and White, 1940). However, intelligibility tests of band-pass filtered speech indicate that intelligibility is not measurably degraded if the energy at frequencies above about 5600 Hz is removed (French and Steinberg, 1947). This bandwidth limitation would correspond to removing the portions of the acoustic patterns of Figure 3.1 that appear above 5.6 kHz. Speech low-pass filtered in this way sounds perfectly natural.

We have selected a 5000 Hz upper cutoff frequency as the default for KLSYN88, although the user is free to increase or decrease the constant control parameter **SR** (sampling rate) to change the limit. According to the sampling theorem, if an audio waveform contains no energy above 5 kHz, it can be sampled at a digital sampling rate of 10,000 waveform samples per second, stored in the computer as a sequence of numbers, and then converted back to the original audio waveform (using a digital-to-analog converter and a 5000 Hz external low-pass filter) with no degradation.

Bits per Sample

A 12-bit digital-to-analog (D/A) converter is used to transform a sequence of digital waveform samples $P_r(nT)$ into a time-varying voltage that can be played through a loudspeaker. Assuming that the synthetic speech waveform makes use of the full range of D/A bits, 12 bits is more than sufficient for high fidelity playback of synthetic speech with no noticeable quantization noise. The software version of KLSYN88 that runs on a general-purpose digital computer is written in C and uses floating point calculations to minimize accumulation of quantization noise.

Parameter Update Rate

The speech signal is the result of sound generation in a system with dynamically changing sources and vocal-tract configurations. As can be seen in Figure 3.1, some acoustic dimensions, such as formant frequencies, change gradually in time, while others, such as which sound source is active, change nearly instantaneously. It is not practical to change synthesis parameters continuously due to the computational cost. We therefore perform the same trick that is used in movies to fool the eye into seeing continuous motion; "frames" of speech are synthesized with control parameters held constant, and then the control parameter values are suddenly changed before generating the next frame (waveform chunk).

Control parameter values are normally updated in this manner every 5 milliseconds. A frame of speech is thus 50 samples of the output waveform at 10,000 samples/sec. This

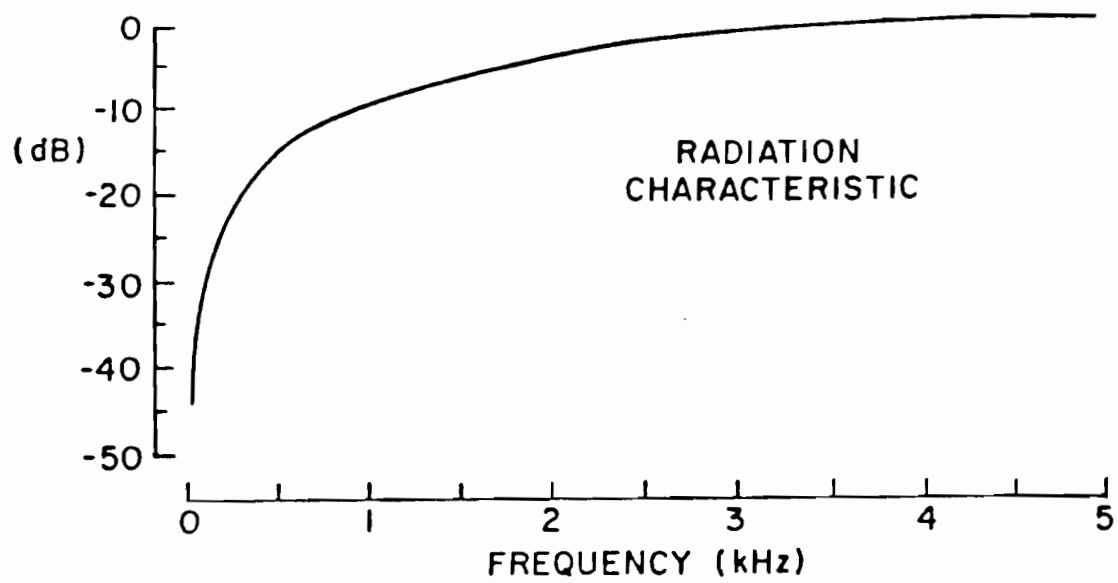


Figure 3.4: *Transfer function of the radiation characteristic.*

parameter update rate is frequent enough to fool the ear and mimic even the most rapid of formant transitions and brief plosive bursts. For example, the pattern shown in Figure 3.1, having a duration of 1.6 sec, would be synthesized with about 300 frames. The choice of frame duration does not appear to be very critical—many speech waveform encoding systems use frame rates more than double the present figure.

Folding the Radiation Characteristic into Source Calculations

The radiation characteristic, shown at the right in Figure 3.3, is not realized as a separate module in the synthesizer. Since it involves a simple time derivative, the transformation is instead incorporated as a part of the models of the sound sources. This method of accounting for the radiation characteristic permits simplification of the sound source calculations, as will be seen. The sound pressure measured directly in front of the lips is approximately proportional to the temporal derivative of the lip-plus-nose volume velocity, and inversely proportional to r, the distance from the lips (Fant, 1960). If the radiation characteristic were not folded into the sound source models, it could be approximated in the synthesizer by taking the first difference of lip-nose volume velocity:

$$P_r(nT) \propto \frac{dU_l(t)}{dt} \approx U_l(nT) - U_l(nT - T) \tag{3.1}$$

The radiation characteristic adds a gradual rise in the overall spectrum, as shown in Figure 3.4, and severely attenuates very low frequencies.

Representing the radiation characteristic as a derivative or as a simple first difference is only an approximation, since it assumes that the lip-plus-nose output acts as a simple nondirectional source. This assumption is valid at lower frequencies (below about 2.5 kHz),

where the wavelength is greater than the head size. At higher frequencies, the radiated sound becomes somewhat directional, and the sound pressure on the axis directly in front of the mouth is up to 5 dB greater than that for a nondirectional source (Morse, 1948).

If a microphone were placed on the chest below the lips, the effects of the radiation characteristic would be reduced, and less high-frequency sound would reach the microphone. The reduced energy at high frequencies may make a voice seem more pleasant, but it also has a slightly deleterious effect on the intelligibility of individual speech sounds because the low frequencies partially mask the audibility of the more important weaker high frequency components.

3.1.1 The Nature of the Voicing Source

The major difference between the cascade/parallel formant synthesizer originally described in Klatt (1980) and the new KLSYN88 design is in the voicing source. In this section we motivate the need for greater flexibility in voicing source control in order to simulate natural variations in voice quality related to laryngealization and breathiness, especially for female voices (Klatt and Klatt, 1990).

Some of the degrees of freedom of the larynx during voicing and the resulting acoustic output are shown in Figure 3.5. Voice quality variation associated with changes in glottal opening is illustrated in physiological terms in the row A of Figure 3.5, which shows a schematic view of the glottis from above. The positions of the arytenoid cartilages (triangles) and vocal processes (the points where the vocal folds insert into the arytenoid cartilages) are illustrated for laryngealized, modal, and breathy phonation.

The characteristics of a modal voice are illustrated in column 2 of Figure 3.5. The vocal folds are nearly approximated, leading to a typical volume velocity waveform (panel 2B) with an open quotient of about 50 to 60% of the period and a waveshape during the open phase that is slightly skewed (closure is more rapid than opening). The spectrum of the modal voicing source (panel 2C) has an average falloff of about -12 dB per octave of frequency increase, above a frequency of 300-400 Hz.

In preparation for laryngealized phonation (column 1 of Figure 3.5), the arytenoids are positioned so as to close off the glottis, and perhaps even apply some medial compression to the vocal processes. When lung pressure is applied to the system, the vocal folds vibrate, producing a glottal volume velocity waveform as shown in panel 1B of the figure. The glottal pulse is relatively narrow, i.e. the duration of the open portion of a fundamental period is relatively short. In addition, the fundamental frequency is substantially lowered during laryngealization, and there may be period-to-period irregularities in both the duration of the period and the amplitude of the glottal volume velocity pulse (Timke *et al.*, 1959). Possible perceptual cues to laryngealization (associated with changes to the source spectrum) are a reduction in the relative amplitude of the fundamental component in the source spectrum (panel 1C) and a lowered fundamental frequency contour.

The glottal configuration during a breathy vowel is shown in panel 3A of Figure 3.5.

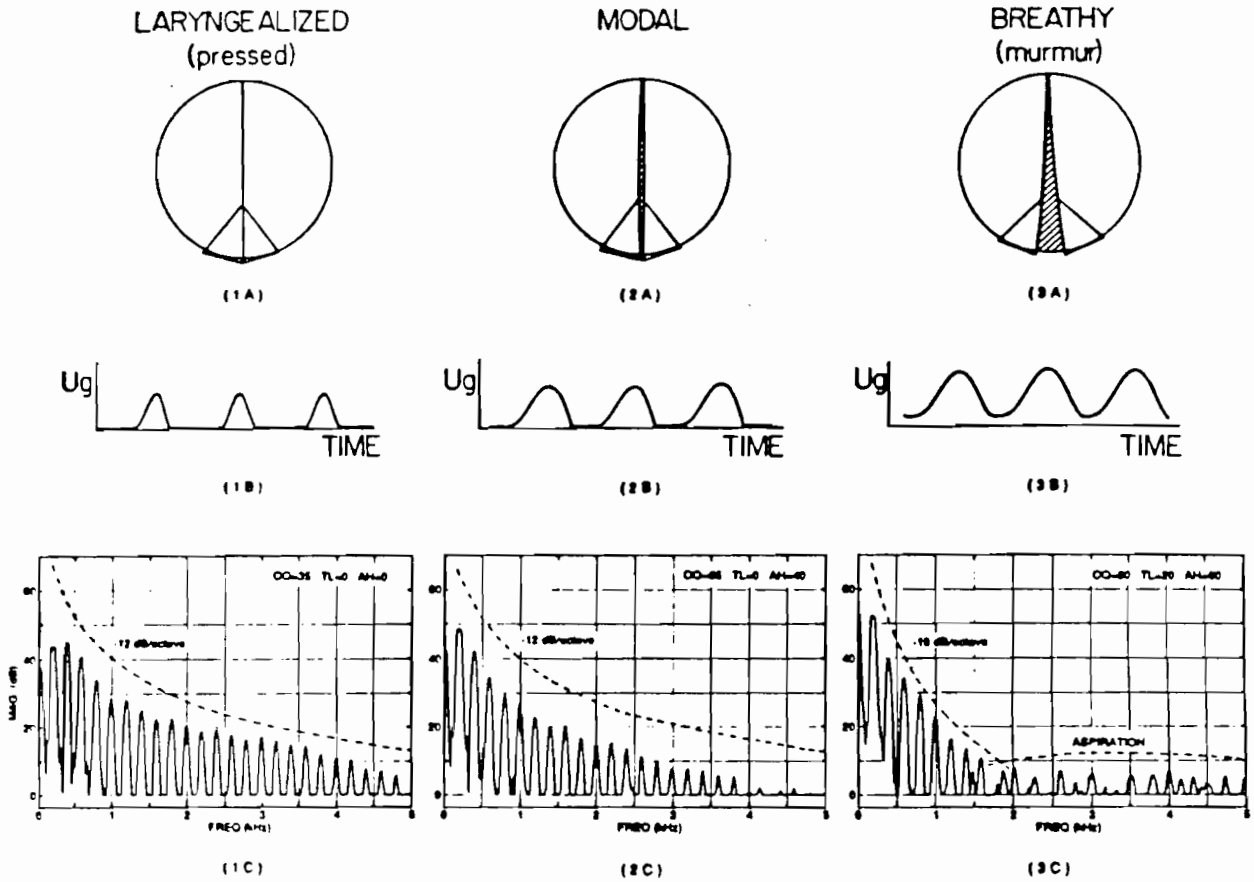


Figure 3.5: *Glottal configurations (row A) for (1) laryngealized, (2) modal and (3) breathy vowels. An increased opening at the arytenoids results in glottal volume velocity waveforms (row B) with a progressively longer duration open period, an increased dc flow, and a less abrupt closure event. The source spectra (row C) have a more intense fundamental component from left to right, and the breathy configuration results in a spectrum with weaker high-frequency harmonics being replaced by aspiration noise. Figure adapted from Stevens (1977).*

The arytenoid cartilages are well separated at the back, but the vocal processes are sufficiently approximated that the vocal folds vibrate when a lung pressure is applied to the system. Since the glottis is never completely closed at the back over the vibratory period, there is considerable dc airflow (panel 3B). This increased airflow results in the generation of turbulent aspiration noise, which is combined with the periodic voicing component to form a source spectrum consisting of both harmonics and random noise (panel 3C). Being relatively weak in amplitude, the aspiration noise might not be audible were it not for the fact that the vibratory behavior of the vocal folds is modified in a breathy vowel (Fant, 1980). Ordinarily, as illustrated in the middle column, the vocal folds close simultaneously along their length, leading to an abrupt cessation of airflow and relatively strong excitation of higher harmonics at the instant of closure. In a breathy vowel, however, the folds close first at the front, and then closure propagates posteriorly, leading to a volume velocity waveform with a rounded corner at closure (panel 3B). The implications of this behavior for the harmonic components

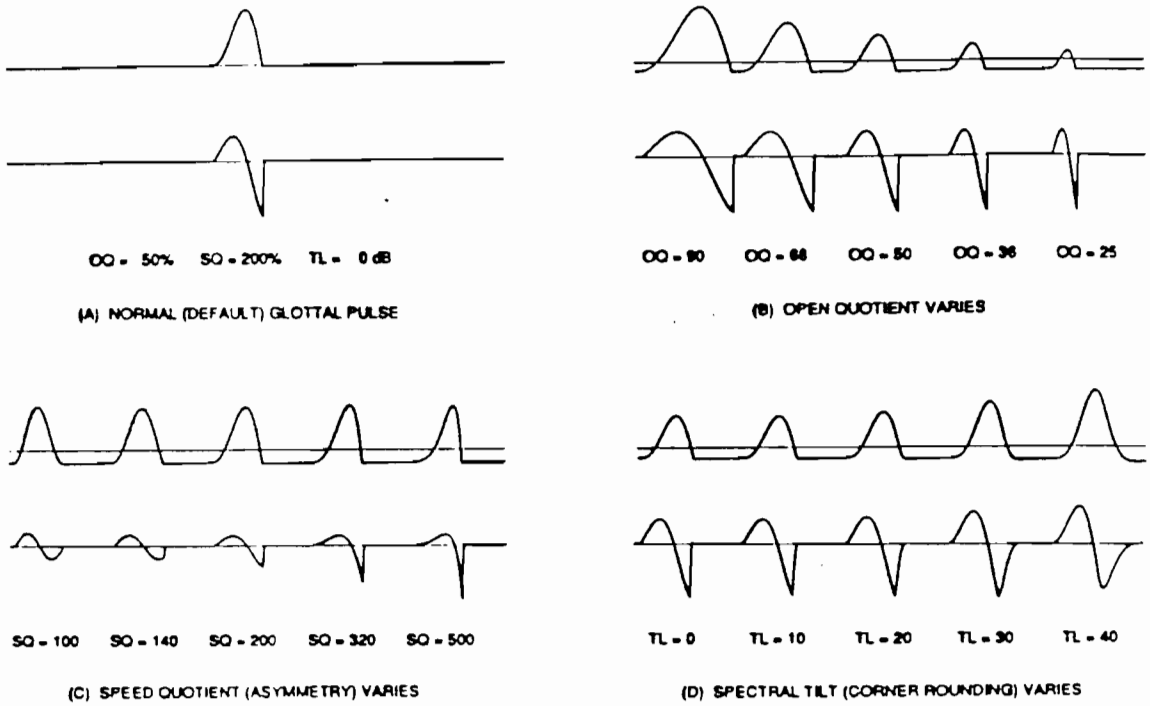


Figure 3.6: *A modified version of the Liljencrants and Fant (LF) model was employed to generate the glottal volume velocity waveforms $U_g(t)$ and their first derivatives $U_g'(t)$, which are shown as each of three control variables is varied. In parts (B) through (D), five 10-msec periods of the source waveform are shown as a selected control parameter takes on a set of values typical of its range of variation in speech.*

of the source spectrum are two-fold – the waveform is more nearly sinusoidal and thus has a very strong fundamental component, and the amplitudes of higher harmonics are attenuated substantially due to non-simultaneous closure (panel 3C). Possible perceptual cues to a breathy vowel are thus an increase in the relative amplitude of the fundamental component in the spectrum, and replacement of higher harmonics by aspiration noise.

Both laryngealization and breathy phonation are common occurrences in speech. Utterance-initial stressed vowels in English frequently begin with a laryngealized onset. Breathy phonation is common for many female speakers, and often occurs at the margins of voiceless consonants for all speakers (Klatt and Klatt, 1990). Utterances often end in a breathy mode of vibration, and unstressed syllables tend to be somewhat more breathy than stressed syllables. Thus there is a clear need for a synthesizer design that is capable of dynamic control over the parameters involved in this type of voice quality variation.

The formant synthesizer originally described by Klatt (1980) employs an impulsive voicing source model that is not capable of mimicking all of these behaviors. The impulsive source is retained as a synthesizer option, but two new models of the voicing source have been added. Recent efforts to characterize the essential features of the voicing source waveform $U_g(t)$ for different male and female voices have led to several new parametric models of glottal output (Rosenberg, 1971; Rothenberg *et al.*, 1975; Fant, 1979; 1982; Titze, 1984;

Ananthapadmanabha, 1984; Fant, Liljencrants and Lin, 1985; Allen and Strong, 1985; Fujisaki and Ljungqvist, 1986; Klatt, 1987). These models, which have many common features, have led to the creation of a new more natural default source model for KLSYN88 (called KLGLOTT88, Klatt and Klatt, 1990). A slightly modified version of the Liljencrants-Fant "LF" voicing source model (Fant, Liljencrants and Lin, 1985) is also provided as an option in KLSYN88.

The characteristics of the waveform can be described by conventional parameters such as **F0**, the fundamental frequency of voicing, and **AV**, the peak amplitude of the glottal pulse, as well as new parameters (1) **OQ**, the open quotient—or ratio of open time to total period duration, (2) **SQ**, speed quotient—or ratio of the duration of the rising portion to the duration of the falling portion of the glottal open phase, and (3) **TL**, spectral tilt—or the additional spectral change associated with "corner rounding" in which closure is non-simultaneous along the length of the vocal folds. Effects of each of these parameters on the waveform and spectrum of the voicing source are shown in figures 3.6 and 3.7 respectively. As can be seen from the figures, the open quotient determines the relative strength of the first harmonic, the speed quotient can introduce spectral zeros, and the tilt parameter attenuates high-frequency components (which are usually replaced by breathiness noise through use of the aspiration amplitude **ah** control parameter). The KLGLOTT88 model and the LF model will be fully defined mathematically in Section 3.2, and instructions for control will be given in Section 3.3.

Complications: Glottal pulse timing irregularities

The waveshapes of successive periods of $U_g(t)$ in a sustained vowel need not be identical. the literature includes terms such as "jitter", the period-to-period random fluctuations in period durations (Horii, 1979), "shimmer", the period-to-period random fluctuations in glottal pulse amplitude (Horii, 1980), and "diplophonic double-pulsing", the tendency for a voice to sometimes vibrate in a mode where pairs of glottal pulses move toward one another, with the first often being attenuated in amplitude (Timke *et al.*, 1959). We consider each of these deviations from perfect periodicity in turn.

Jitter and Shimmer. It is well known that a constant f_0 is to be avoided in speech synthesis because the result is a peculiarly mechanical sound quality. An example of an analysis of fundamental frequency of a female subject attempting to hold a constant pitch is shown in Figure 3.8a. while the wavering nature of the f_0 trace may in small part be due to analysis artifacts,[4] It is known that normal physiological mechanisms can impart these kinds of fluctuations. In an insightful correlational analysis of f_0 and EMG data, Baer (1978) was able to show that a single muscle fiber twitch in the cricothyroid causes a predictable not-insignificant local increase in f_0, and that normal statistical variations in fiber firing can be expected to produce fluctuations in f_0 not unlike those observed in the figure.

[4]The fundamental frequency extraction algorithm that was employed to produce this data is of the harmonic sieve type, which probably averages out some rapid period-to-period changes within the 25-ms analysis window rather than accentuating them.

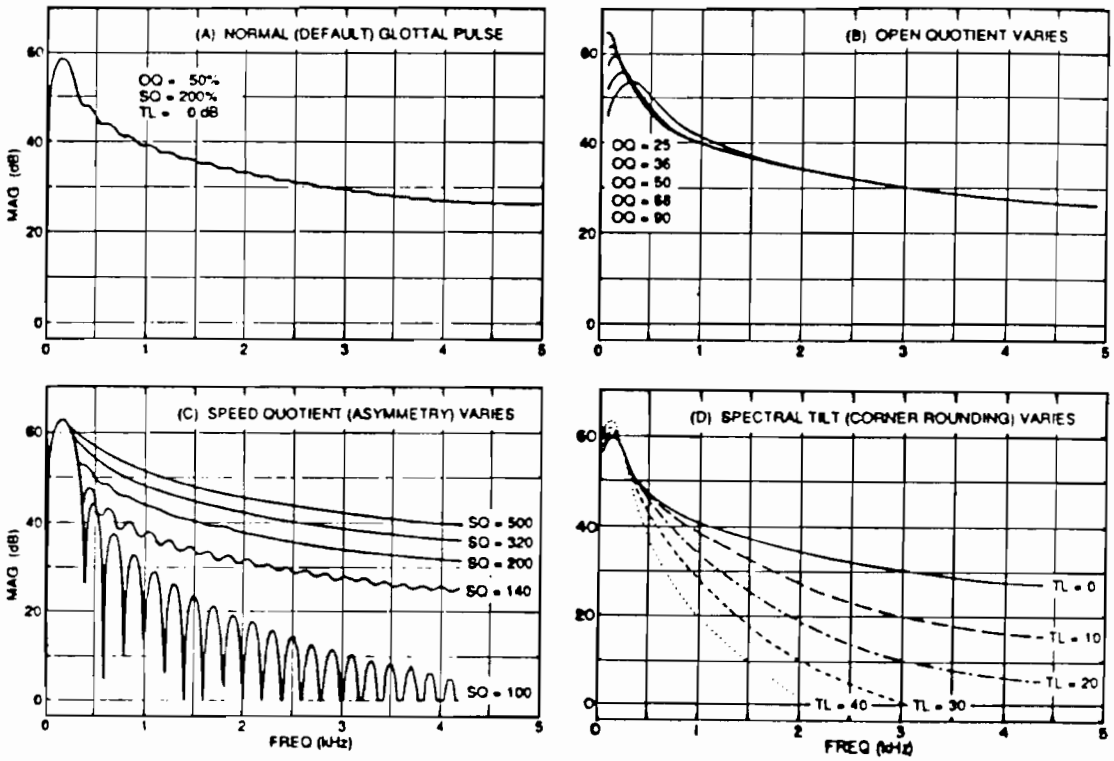
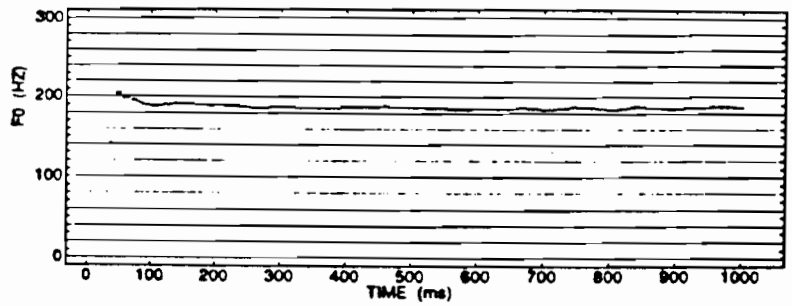


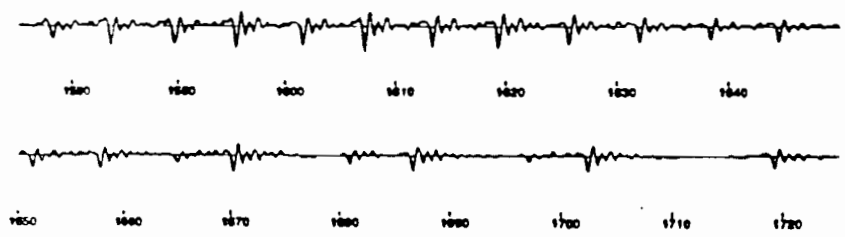
Figure 3.7: *DFT magnitude spectra are shown corresponding to the $U_g'(t)$ waveform of a single glottal pulse synthesized by the modified LF model at several values for each of three control parameters. Waveforms for each condition were displayed in Figure 3.6. These curves represent the spectral envelope of a harmonic spectrum that would result from synthesizing a train of such pulses.*

The mechanical quality of synthesis at constant f_0 can be reduced or eliminated simply by introducing a normal intonation contour to the synthesis (Rosenberg, 1968), but there are often time intervals where the f_0 is nearly constant, and some sort of simulation of the f_0 flutter or jitter seen in Figure 3.8 would be desirable. Jitter, defined as the period-to-period variability in f_0, has been measured in sustained vowels for both normal and pathological voices (Lieberman, 1961; 1963; Horii, 1979; 1980; Hollien *et al.*, 1973; Askenfelt and Hammarberg, 1986). If the appropriate parameter to characterize jitter and shimmer is the standard deviation of a presumed Gaussian distribution of periods or pulse amplitudes respectively, then normal voices sustaining the vowel /a/ contain a jitter of about 0.5% to 1.0% (Hollien *et al.*, 1973). This is slightly less than the detectability threshold—perceptual data indicate a detectability threshold for jitter of about 2% and for shimmer of about 10% or 1 db (Pollack, 1971)—calling into question the utility of adding this kind of gaussian jitter to synthesis. It is also likely that the jitter and especially shimmer measured by these techniques is in part a measurement artifact due to superposition effects (Milenkovic, 1987).

The nature of a better random component for the synthesis of jitter has been a subject of debate, since most efforts to introduce audible random jitter to the pitch period in synthesis have led to a harsh voice quality (Rozsypal and Millar, 1979). The new KLSYN88 voicing



(A) VOWEL SUSTAINED AT CONSTANT PITCH, NOTE F0 JITTER



(B) EXAMPLE OF DIPLOPHONIC DOUBLE PULSING

Figure 3.8: *Examples of deviations from perfect periodicity: (a) fundamental frequency contour of a female subject sustaining a vowel at constant pitch (note the waver, or inability to hold pitch constant), and (b) a speech waveform is shown in which various degrees of diplophonic double-pulsing are present; normal voicing (t=1580 to 1660 ms) suddenly changes to a vibration mode where the first of a pair of periods is delayed and reduced in amplitude (t=1660 to 1710) and the first pulse may disappear entirely (t=1710 to 1720).*

source model includes a mechanism for introducing a slow quasi-random drift or "flutter" to the f_0 contour (shimmer is not modelled). The term "flutter" has been adopted since jitter has a well-defined meaning that differs from the f_0 modification used in our synthesis strategy. Instead of using a random process to simulate jitter, we add to the nominal f_0 a quasi-random component which is in fact the sum of three slowly varying sine waves. The amount of flutter at any time is determined by the new **FL** control parameter. A value of FL=25% results in synthetic vowels with a quite realistic deviation from constant pitch. It is unlikely that this slowly varying flutter component is the only deviation from constant pitch in normal voicing, but it appears to be sufficient for synthesis purposes.

Diplophonic Double Pulsing. An example of diplophonic double pulsing is shown in Figure 3.8B. In the extreme, the alternate pulses may actually disappear, in which case the f_0 is halved. Obvious examples of double pulsing were observed sporadically, usually near the termination of an utterance, for more than a quarter of the speakers that were examined in a recent study (Klatt and Klatt, 1990). less extreme diplophonia may occur more often. The KLSYN88 voicing source model includes a mechanism for simulating double pulsing

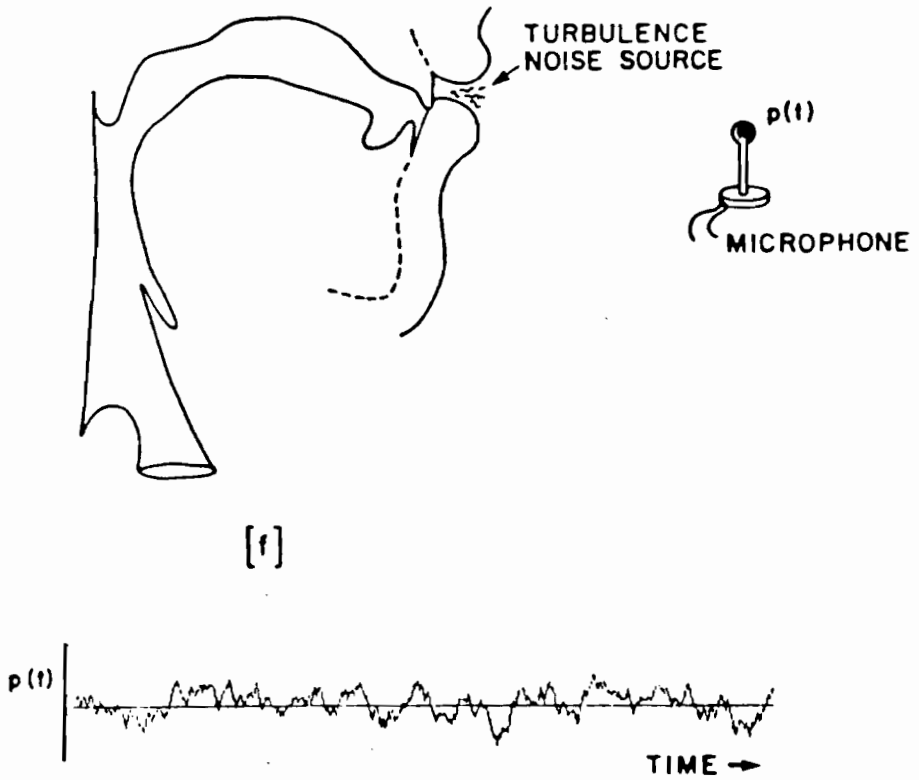


Figure 3.9: *Articulatory configuration for /f/, shown at the top, consists of a constriction between the upper teeth and lower lip which results in a turbulence noise sound pressure waveform, $p_n(t)$, shown below.*

using the DI (diplophonic double pulsing) control parameter. Alternate pulses are modified whenever DI is greater than zero. A modified pulse is delayed in time and attenuated in amplitude by an amount that is specified in terms of the maximum allowed modification in percent, where the maximum delay is such as to time the closure of the first pulse to be simultaneous with the opening of the next unaltered pulse, and the amplitude attenuation goes from one to zero on a linear scale as DI ranges from 0 to 100%. for example, if OQ is at 50%, setting DI to a value of 50% results in a first pulse of each pair that is delayed by a quarter of a period and is attenuated by half (-6 dB).

3.1.2 The Nature of the Noise Source

A second kind of sound source involves the generation of turbulence noise by the rapid flow of air past a narrow constriction. Turbulence noise is generated whenever a constriction is small enough, and the pressure drop across the constriction is large enough that airflow becomes turbulent (Stevens, 1971; 1977).

The articulatory configuration for /f/, shown in Figure 3.9, illustrates conditions under which turbulent airflow may be created. The vocal folds are spread apart so that the

only significant obstruction to flow occurs at the constriction between the upper teeth and lower lip. A turbulent jet is created such that an effective random sound-pressure source exists slightly downstream from the constriction (Stevens, 1971). An example of the noise waveform generated under these conditions is shown in the bottom part of Figure 3.9.

A noise-producing constriction can be created at several locations along the vocal tract, or in the larynx. The resulting noise is called aspiration if the constriction is located at the level of the vocal folds, as during the production of the sound /h/. If the constriction is located above the larynx, as during the production of sounds such as /f/, /θ/, /s/, /š/, the resulting noise is called frication noise. The explosion of a plosive release in /p/, /t/, /k/, /č/ also consists primarily of frication noise, although there is also the "step" response to the sudden release of the oral pressure (also known as the coherent response) that may have some perceptual importance. While the coherent response is not explicitly modeled, the user can reinitialize the random number generator to a special value that produces the same particularly flat spectrum burst over the first 5 to 10 ms, using control parameters **SB**, same burst, and **RS**, random seed.

The noise source is modeled by a pseudo-random number generator. The amplitude distribution of a random number sequence differs from that of normal turbulence noise. The amplitude distribution of such a sequence is flat between limits of -32768 and +32767, while turbulence noise generated by a physical process has a Gaussian amplitude distribution. The difference in amplitude distribution is barely perceptible under ideal listening conditions, and is not considered to be a serious flaw in the modeling of speech. If it were, one could add together 4 or perhaps 16 samples of pseudo-random noise to produce an excellent approximation to a Gaussian distribution. The computational cost was considered to be too great to be incorporated. In any event, there is usually some filtering of the noise during synthesis, and the output of the filtering tends to have a Gaussian distribution.

Spectrum of the Noise Source. The sequence of random numbers has a flat spectrum, as shown in Figure 3.10. The spectrum of a short windowed sample of synthetic noise (solid curve) fluctuates randomly about the expected long-term average noise spectrum (dashed line). Short samples of noise vary in their spectral properties due to the nature of random processes. The ear appears to be able to "average" over sufficiently long time intervals to estimate the relevant attributes of the spectrum from the output speech, even though short samples of noise vary so much in spectral detail.

Spectra of several naturally occurring noisy speech sounds are shown in Figure 3.11. These spectra are the result of a turbulence noise sound source exciting the vocal tract, and thus represent a combination of the spectrum of the sound source and the vocal tract transfer functions of /f/, /θ/, /s/, and /š/. However, since the vocal-tract transfer function for /f/ is nearly flat, the spectrum of /f/ in Figure 3.11 is approximately that of the turbulence noise source and radiation characteristic combined. Thus, in the frequency range from 0 to 5 kHz, the /f/ spectrum is approximately flat.

When voicing and turbulence noise generation co-exist, as in a voiced fricative such as /v/, /ð/, /z/, /ž/, both the voicing and noise sources are modified due to the activity of the other source. The voicing waveform is usually nearly sinusoidal due to an active

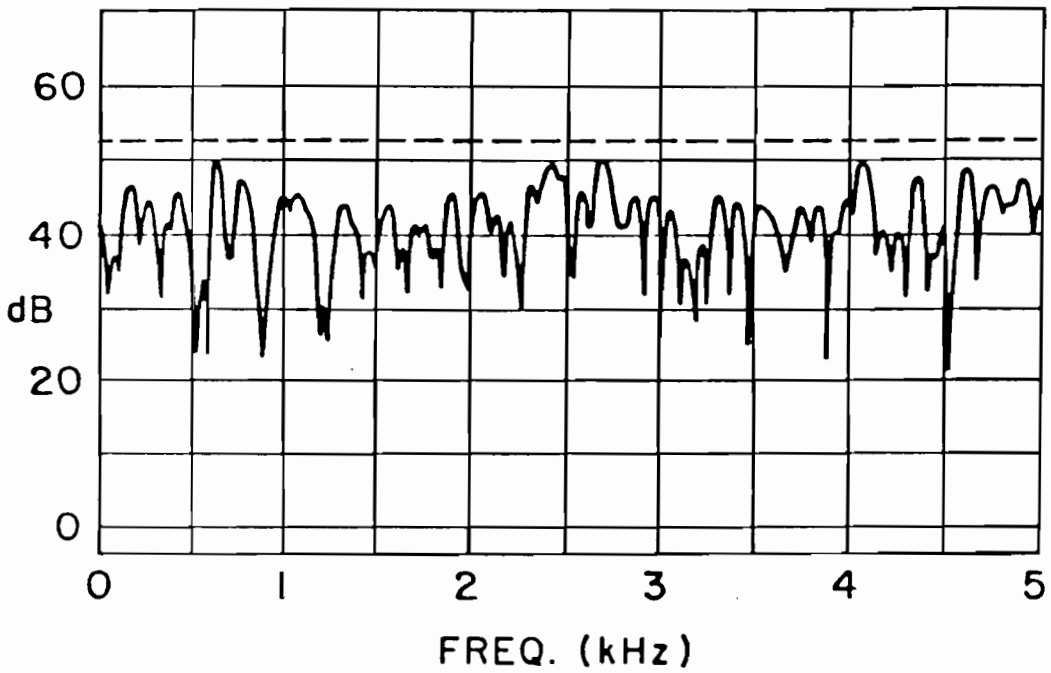


Figure 3.10: *The long-term average spectrum of a pseudo-random number generator (dashed line) is compared with a short 25-msec windowed chunk of synthetic noise. This flat spectrum is the composite spectrum of the synthesis noise source and radiation characteristic (see text). The average spectrum has been displaced upwards by a few dB relative to the short-time spectrum.*

partial opening of the glottis to permit greater airflow in order to support turbulence noise generation, and the noise source is amplitude-modulated periodically by the vibrations of the vocal folds. Noise intensity increases slightly during the open portion of each glottal vibration. Therefore the synthesizer should be capable of generating at least two types of frication waveforms—normal frication and amplitude-modulated frication.

Aspiration noise is essentially the same as frication noise in spectral characteristics. The same pseudo-random number generator and amplitude modulation algorithm are used to synthesize aspiration noise.

3.1.3 Types of Vocal Tract Transfer Functions

The acoustic characteristics of the vocal tract are determined by its shape as a function of distance from the larynx to the lips. Midsagittal views of the vocal tract of an adult male articulating /i/, /ɑ/, and /u/ are shown in the left column of Figure 3.12. Except at low frequencies, the vocal tract can be thought of as an (acoustically) hard-walled tube of varying cross-sectional area.

The fact that the tube has a bend in it is irrelevant—the tube can be straightened out without changing its sound-generating properties. Thus plots of cross-sectional area as

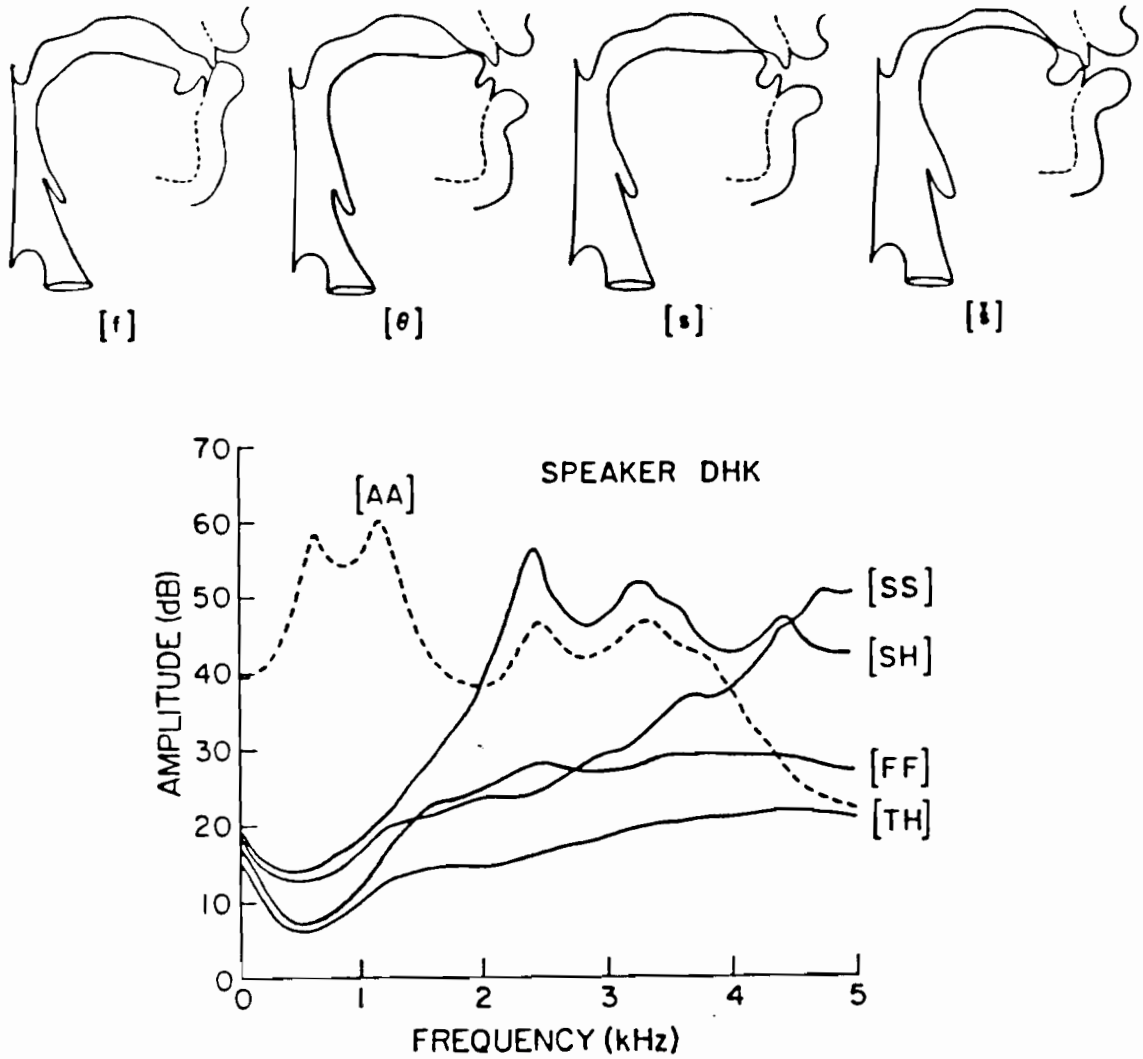


Figure 3.11: *Articulatory configurations and (preemphasized) DFT magnitude spectra of four English fricatives. Fricative spectra are compared with a typical /ɑ/ vowel spectrum. If it were not preemphasized, the /f/ spectrum would be approximately flat.*

a function of distance from the larynx, middle column of Figure 3.12, are all that is required to characterize the acoustic effects of the vocal tract on a sound source. Unfortunately, it is difficult to estimate accurately the cross-sectional area from an x-ray picture such as shown in Figure 3.12 since the cross dimension depends on aspects of articulation not seen in a lateral x-ray such as whether the tongue makes contact with the upper teeth (Fant, 1960; Ladefoged *et al.*, 1971).

The vocal tract forms a non-uniform transmission line whose behavior can be computed for frequencies below about 5 kHz by solving a one-dimensional wave equation (Fant, 1960). Above 5 kHz, three-dimensional modes would have to be considered. Solutions to the wave equation result in a transfer function $T(f)$ that relates glottal source volume velocity to output volume velocity at the lips:

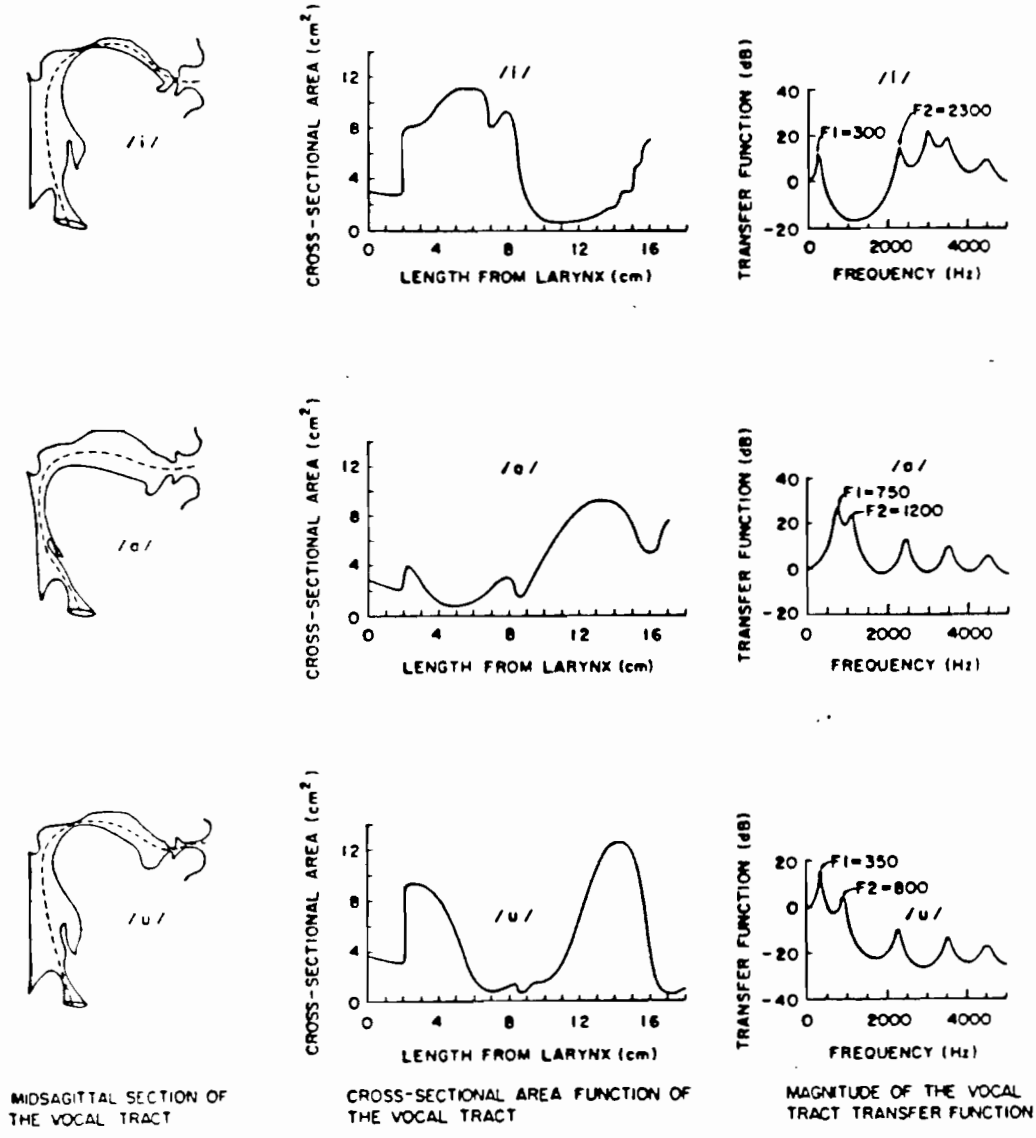


Figure 3.12: *Midsagittal views of the vocal tract (left), the acoustically relevant cross-sectional area of the vocal tract from larynx to lips (middle), and frequency domain vocal tract transfer functions (right) for the vowels /i/, /ɑ/, and /u/.*

$$T(f) = \frac{U_l(f)}{U_g(f)} \tag{3.2}$$

Vowels

The vocal-tract transfer function for vowels is the least complex of the various transfer functions for the sound types of English; it consists of simple transfer functions each characterized by a pair of poles or "resonances" that can be simulated by a set of digital resonators connected in cascade, as shown in the top part of Figure 3.13. The vocal tract acts as a complex

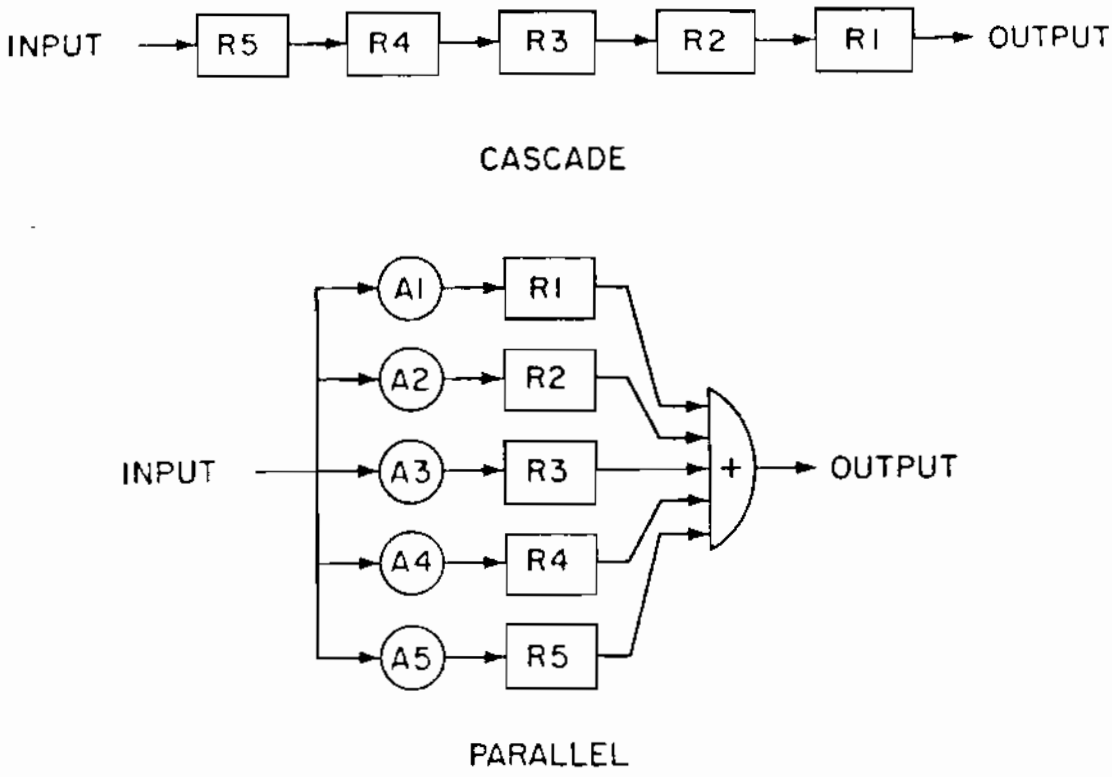


Figure 3.13: *Digital resonators can be configured in cascade with the output of one acting as input to the next (top), or in parallel, in which case each receives the same source input, whose gain is determined by an independent amplitude control, and outputs are summed algebraically (bottom).*

resonator, very much like an organ pipe, except that the locations of the resonant frequencies change as the shape of the tube is modified by motions of the jaw, tongue and lips. Locations of the five lowest resonant frequencies of the vocal tract for several vowels produced by an adult male are indicated by peaks in the transfer functions shown in Figure 3.12. During vowel production (sound source at the larynx, nasal tract closed off by a raised velum), the transfer function of the vocal tract, $T(f)$, conforms to the general form of an all-pole linear system (Fant, 1960):

$$T(f) = \prod_{n=1}^{\infty} \frac{(\pi BW_n)^2 + (2\pi F_n)^2}{(s + \pi BW_n + j2\pi F_n)(s + \pi BW_n - j2\pi F_n)} \tag{3.3}$$

where f is frequency in Hz, F_n is the nth formant frequency, BW_n is the nth formant bandwidth, s is the complex variable $(j2\pi f)$, π equals 3.14159..., and j represents the square root of minus one.

The equation can be separated into a product of two terms, one involving the lowest five formant resonators, and the second involving the remaining infinite number of terms, which fortunately converge to a more-or-less fixed spectral correction factor for frequencies below 5 kHz. Abbreviating by letting the constant S_n stand for $(\pi BW_n + j2\pi F_n)$ and by

letting S_n* represent the complex conjugate of S_n, the equation simplifies to:

$$T(f) = K(f) \times \prod_{n=1}^{5} \frac{S_n S_n*}{(s + S_n)(s + S_n*)} \tag{3.4}$$

where $K(f)$ is a fixed correction term that boosts frequencies near 5 kHz somewhat. It will turn out that $K(f)$ is not needed when the equation is converted to the digital domain of sampled difference equations, so its exact properties are not given here (see, for example, Fant, 1960). The constants in the numerator of Equation 3.4 insure that the transfer function has a value of unity at zero frequency, i.e., the dc airflow is unimpeded.

Each term of the product in Equation 3.4 defines the transfer function of a single formant resonator, consisting of a pair of poles. The transfer function conforms to an all-pole model because there are no side-branch resonators or multiple sound paths, and the source impedance is assumed to be infinite. Thus Equation 3.4 is good for non-nasalized vowels and most sonorant consonants. It will have to be modified if there is coupling to the nasal tract, or if the sound source is moved to a location above the larynx, as in a fricative such as /s/, as discussed below.

Five resonators are appropriate for simulating a vocal tract with a length of about 17 cm, the length of a typical male vocal tract, because the average spacing between formants is equal to about 1000 Hz. A typical female vocal tract is 15 to 20 percent shorter, suggesting that only four formant resonators be used to represent a female voice in a 5 kHz simulation (or that the simulation should be extended to about 6 kHz). The fifth formant (and even the fourth formant for a child's voice) can be effectively removed from the cascade vocal tract by a trick that is described in the section that presents equations for a digital resonator.

Ordinarily we are only interested in the magnitude of the transfer function, although the phase could also be plotted. Phase is a monotonically decreasing continuous function of frequency. The phase function has the special property that the value at each formant frequency F_n is given by $(n - \frac{1}{2})\pi$. The phase function for each individual formant resonator is such as to cause the impulse response of the resonator to be an exponentially damped sinusoid. The ear is not particularly sensitive to phase differences among speech sounds, and room acoustics distorts phase to the extent that it is not considered to be important in speech perception.

Formant Frequencies. Each of the five terms in Equation 3.4 introduces a peak in a magnitude spectrum of the type shown in Figure 3.12. The frequency locations of these resonant peaks are given by the formant frequency variables F_n. The lowest natural frequency of the vocal-tract transfer function is the first formant frequency, and is symbolized by F1. The next natural frequency, the second formant, is F2, etc. Formant frequencies are the primary parameters for distinguishing among vowels and sonorant consonants, with F1 and F2 being perceptually more important than the higher formants, which tend to vary less between phonemes. Typical formant values for vowels will be given in Chapter 4.

All of the formant frequencies decrease when the lips are closed. However, the first formant does not actually fall all the way to zero, as high-school physics would lead one to

believe. The approximation that the walls of the vocal tract are hard is not valid at low frequencies, so that the mass of the walls, coupled with the compliance of the air in the tube define a Helmholtz resonance that gives F1 a value of about 180 Hz (Fant, 1972). As a result, F1 never goes below about 180 Hz in any speech-related articulation.

Formant Bandwidths. The bandwidth of a formant resonance describes the width of the resonance as measured 3 dB down from the peak in the transfer function, as shown later in Figure 3.20. The bandwidth also affects the absolute height or amplitude of a spectral prominence associated with a formant, as discussed below. Formant bandwidths are a function of energy losses due to heat conduction, viscosity, cavity-wall motions, radiation of sound from the lips and the real part of the glottal source impedance. Increased losses widen the bandwidth and reduce the amplitude of a formant peak.

Bandwidths are difficult to deduce from spectral analyses of natural speech because the transfer function is only estimated at harmonic locations, and there are irregularities in the harmonic amplitudes of the glottal source spectrum to begin with. Bandwidths have been estimated by other techniques such as using a sinusoidal swept-tone sound source (Fujimura and Lindqvist, 1971). Results indicate that bandwidths of the first few formants for vowels vary from about 40 Hz to more than 300 Hz depending on the particular phonetic segment being spoken and on the degree of opening at the glottis and lips. Typical values for formant bandwidths for vowels, as deduced by an analysis-by-synthesis methodology, are presented in Chapter 4.

The ear is not very sensitive to changes in formant bandwidths. We are more sensitive to the effect such changes have on formant amplitudes than they have on the width of the resonance peak (Klatt, 1982). All else being equal, the height of a formant peak in the magnitude spectrum of the vocal-tract transfer function is proportional to 1/BW. Thus, for example, doubling the bandwidth BW of a formant causes a decrease of 6 dB in the peak spectrum amplitude of the prominence. If the formants are equally spaced with a spacing of S, then the ratio of the amplitudes of the peaks to the valleys in the magnitude of the transfer function is given by $2S/\pi BW$.

Figure 3.14 shows at the top the transfer function for a uniform tube, with the same bandwidth for each formant. The formant spacing is 1000 Hz, and the bandwidth is 100 Hz, giving a peak-to-valley ratio of 16 dB. The next panel illustrates how the amplitude of a particular formant changes as the bandwidth is manipulated. In the bottom panel of the figure, the frequency of F1 is halved, keeping the higher formant frequencies the same. The result is a reduction of the F1 peak in the transfer function by 6 dB and a decrease of 12 dB in the amplitudes of higher peaks. (This effect on formant amplitudes is discussed in Fant, 1960.)

Sonorant Consonants

Sonorant consonants such as /w/, /y/, /r/, and /l/ are similar to vowels in transfer function characteristics, and much of what has been said about vowels applies to these sounds. Certain formants for these consonants tend to have greater bandwidths than those normally found

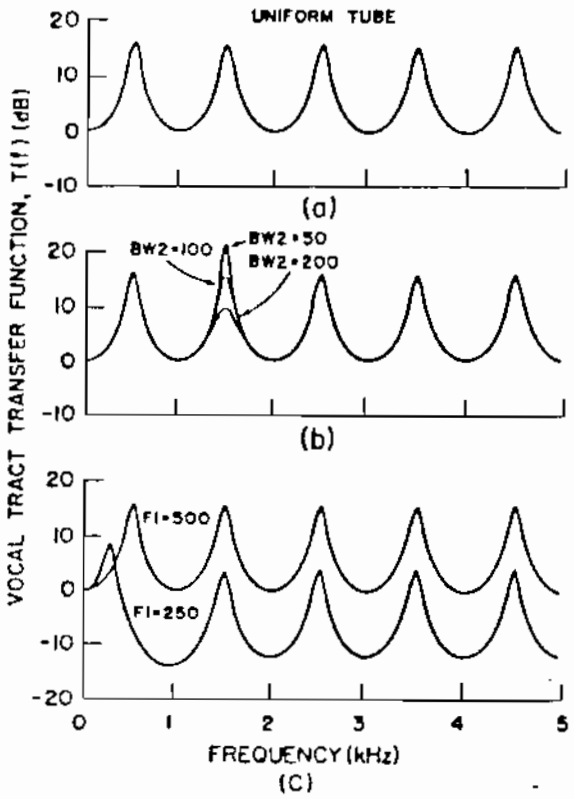


Figure 3.14: *Vocal-tract transfer functions illustrating changes in relative formant amplitudes as a function of selected manipulations to formant frequencies and bandwidths. Panel (b) illustrates that a doubling of formant bandwidth reduces a formant peak by about 6 dB Panel (c) shows that halving a formant frequency reduces the amplitude of the formant peak by 6 dB and also reduces the amplitudes of all higher frequency formant peaks by 12 dB.*

for vowels, because additional acoustic losses occur as a consequence of the more constricted vocal-tract constriction. In the case of /r/, and /l/ additional poles and zeros appear in the transfer function at higher frequencies (above about 2000 Hz) as a consequence of the more complex vocal-tract shape. These matters are discussed further in Chapter XX. The sonorant consonant (or "voiceless vowel") /h/ has the same type of vocal-tract transfer function as a vowel, but the voicing source is replaced by the turbulence noise of aspiration generated near the glottis.

Nasals

Nasal consonants /m/, /n/, and /ŋ/, as well as nasalized vowels, involve vocal tract shapes that are more complicated than a single tube (Figure 3.15). In a nasal consonant, the path from larynx to nasal opening (nares) forms one tube, while the oral cavity acts as an additional side-branch resonator. In a nasalized vowel, the nasal passages effectively act as a side-branch resonator (Fant, 1960; Hawkins and Stevens, 1985; Stevens *et al.*, 1987).

It is not possible to approximate nasal murmurs and the nasalization of vowels that

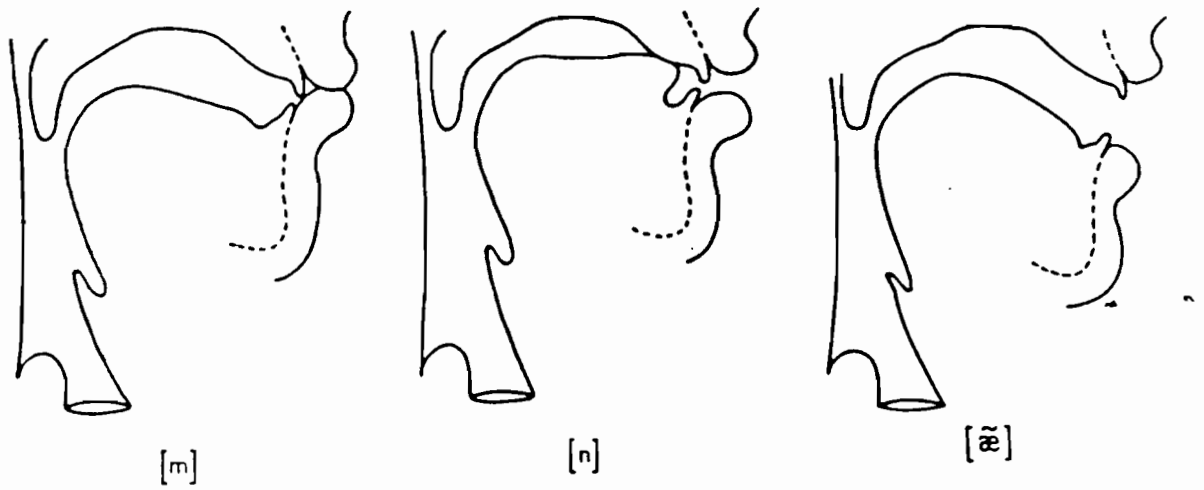


Figure 3.15: *Articulatory shapes of several nasal consonants and a nasalized vowel. The vocal tract consists of single tube plus a "side-branch" resonator.*

are adjacent to nasals with a cascade system of five resonators alone. More than five formants are often present in these sounds, and formant amplitudes do not conform to the relations inherent in a cascade configuration because of the presence of transfer-function zeros (Fujimura, 1960; 1962) A zero occurs at a frequency where a side-branch resonator causes an effective short circuit in the acoustic path from the glottis to the output. As a consequence, sound energy is reflected back to the source at this frequency, and does not appear at the output.

Typical spectra for a nasal murmur and for a nasalized /ɪ/ are shown in Figure 3.16. Nasalization of adjacent vowels is an important element in the synthesis of nasal consonants. The perceptually most important change brought on by the nasalization of a vowel is the reduction in amplitude and "splitting" of the first formant into a pole-zero-pole complex, which is caused by the presence of a nearby low-frequency nasal pole pair and zero pair. The first formant frequency also tends to shift slightly in nasalized vowels. Thus the auditory effects of nasalization can be approximated by the usual vowel cascade of formant resonators augmented by a low-frequency cascaded resonator and antiresonator.[5] The nasal pole-zero pair can be effectively removed from the vowel circuit to synthesize normal nonnasal vowels simply by setting the frequency of the zero equal to the frequency of the pole, as we will see below.

[5]The side-branch tube contributes more than a single pole/zero pair to the transfer function. Additional resonator/antiresonator pairs may be needed to adequately model nasalized vowels in a language such as French where nasalization of vowels is phonemic, in which case the tracheal pole-zero pair to be described below is available for "double-duty".

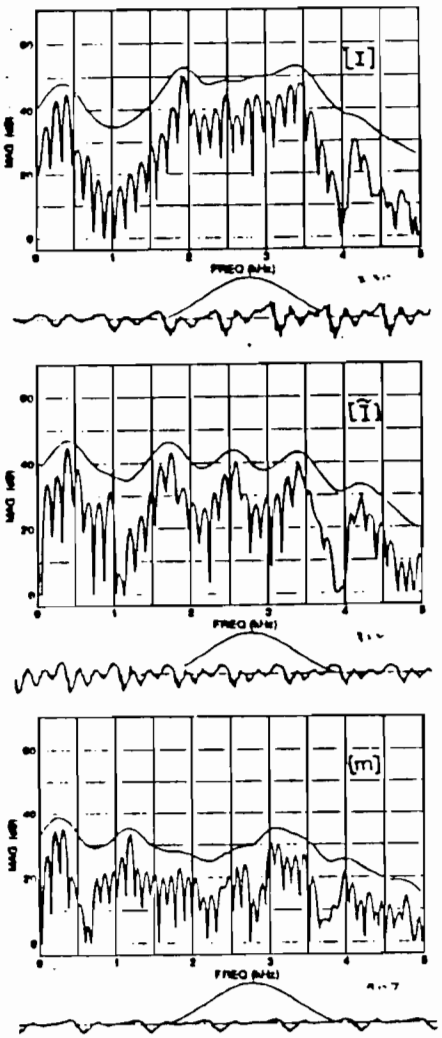


Figure 3.16: *DFT magnitude spectra and superimposed linear prediction spectra are compared for the vowel /I/, the same vowel when nasalized, and a nasal murmur /m/. All spectra were obtained from the recorded syllable "dim". The nasal murmur and the nasalized [I] have an extra pole pair and zero pair in the transfer function near F1, causing a broadened peak or a clear double peak.*

Obstruent Consonants

Articulatory shapes for selected obstruent consonants were shown in Figure 3.11. A constriction in the oral cavity induces turbulent airflow, and this noise source excites the vocal tract. Since the noise source is no longer at the larynx end of the vocal tract, the transfer function is not all-pole, but rather contains both poles and zeros.

The pole frequencies are temporally continuous with formant locations of adjacent phonetic segments because the poles are the natural frequencies of the entire vocal tract configuration, no matter where the source is located. Thus the use of vocalic formant frequency parameters to control the locations of frication maxima is theoretically well-motivated (and helpful in preventing the fricative noises from "dissociating" from the rest of the speech

signal).

The zeros in the transfer function for fricatives are the frequencies for which the impedance looking back toward the larynx from the position of the frication source is infinite, since the series-connected pressure source of turbulence noise cannot produce any output volume velocity under these conditions. The effect of transfer-function zeros is two-fold; they introduce notches in the spectrum and they modify the amplitudes of nearby formants. The perceptual importance of spectral notches is not great because masking effects of energy at adjacent frequencies limit the detectability of a spectral notch (Malme, 1959).

We have found that a satisfactory approximation to the vocal-tract transfer function for frication excitation can be achieved without explicitly introducing transfer-function zeros by employing a parallel set of digital formant resonators having amplitude controls, as shown in Figure 3.13. The presence of any transfer function zeros is accounted for by appropriate settings of the the formant amplitude controls. For example, if a formant is effectively cancelled by the presence of a nearby transfer-function zero, the amplitude of that formant resonator is set to zero.

A sixth formant has been added to the parallel branch specifically for the synthesis of very high frequency noise in /s/ and /z/. The main energy concentration in these alveolar fricatives is often centered on a frequency of about 6 kHz. This is above the highest frequency (5 kHz) that can be synthesized in a 10,000 sample/second simulation. However, in an /s/ there is gradually increasing frication noise in the frequencies below 6 kHz due to the low-frequency skirt of the 6 kHz formant resonance, and this noise spectrum can be approximated quite well by a broadly tuned resonator positioned at about 4900 Hz. We have found it better to include an extra resonator to simulate high-frequency noise than to move F5 up in frequency whenever an /s/ is to be synthesized because we thereby avoid clicks or false cues associated with moving energy concentrations.

Also included in the parallel vocal tract model is a bypass path. The bypass path with amplitude control AB is present because the transfer function contains no prominent resonant peaks during the production of /f/, /v/, /θ/, /ð/, and the synthesizer should include a means of bypassing all of the resonators to produce, in effect, a flat vocal-tract transfer function.

The frication-excited parallel formant resonators are assigned the same formant frequency values as are given to the corresponding formants in the cascade model. Bandwidths, however, are wider, due in part to the additional losses associated with increased radiation losses, increased losses at the constriction, and possible additional losses due to a widened glottal opening. This is the reason that separate bandwidth controls **B2F** through **B6F** are employed for the frication-excited parallel vocal tract model. There is no first formant resonator in the parallel model because the first formant is minimally excited during fricatives, primarily because an adjacent transfer function zero cancels the F1 pole.

Relatively simple rules for determination of ballpark settings for formant amplitudes can be derived from a quantal theory of speech production (Stevens, 1972; 1989). (See also Fant, 1960.) The theory states that only formants associated with the cavity in front of the oral constriction in Figure 3.11 are strongly excited by the noise source. The lowest front

cavity resonance for /f/, /θ/, /s/, and /š/ are approximately quarter-wavelength resonances of 8000, 7000, 6000, and 2500 Hz respectively. The value for /š/ is lower in part because partial lip rounding and protrusion tends to decrease the lowest front-cavity resonance. The configuration for /š/ also has a long narrow constriction behind the source, and the resonance of this constriction contributes to the spectrum.

Complications: Source-Tract Interactions

According to the original classical formulation of the acoustic theory of speech production (Fant, 1960; Flanagan, 1972), the voicing source is characterized as a "current source" because the volume velocity waveform $U_g(t)$ is said to depend very little on the shape or impedance of the vocal tract, at least for vowels. Similarly, the vocal-tract transfer function is assumed to be modeled well by a succession of time-invariant linear filters because the terminating impedance at the glottis, while varying over a period, is nonetheless high compared with the vocal-tract impedance. Recent work by Fant and his associates suggests that some of the original simplifying assumptions of the classical theory are not really valid. First of all, the presumed direct relationship between glottal area and glottal flow is perturbed by standing-wave pressure fluctuations in the pharynx, which invalidate an assumed constant transglottal pressure over a cycle. The pharyngeal pressure variations cause the glottal source flow waveform to take on ripple components at the frequency of F1, and may even be large enough to have a direct influence on the mechanical behavior of the vocal folds (Fant, 1985). Furthermore, as the glottis opens and closes, the vocal-tract transfer function undergoes rapid changes over a single period that may be of perceptual importance. Four phenomena not satisfactorily modeled by the old non-interactive theory can be identified from examination of the behavior of the interactive source-filter model; the first two affect the source waveform $U_g(t)$ and the second pair of phenomena affect the vocal-tract transfer function $T(f)$.

F1 Ripple in the Source Waveform. The transglottal pressure is an important variable in determining glottal volume velocity from the time variation in glottal area. However, transglottal pressure is not constant over a period as originally assumed, but rather varies due to pressure fluctuations associated with the F1 standing wave in the lower pharyngeal portion of the vocal tract (Fant, 1982; Fant, Lin and Gobl, 1985; Lin, 1990). The resulting glottal flow is nonlinear in that volume velocity through an orifice is proportional to the square root of the pressure drop (Stevens, 1971). Assuming a constant glottal area function from period to period, the non-interactive model predicts a succession of smooth identical volume velocity pulses, whereas the interactive model predicts a buildup of "F1 ripple" interaction as a standing wave is developed in the vocal tract. This effect, which Fant calls nonlinear superposition, result in an overall boost in the spectral amplitude of F1 relative to other formants because an F1 component is contained in the source waveform. The F1 amplitude boost is of unknown perceptual importance, but presumably could be crudely approximated by first formant bandwidth changes and perhaps an increase in the source spectral tilt using the new KLSYN88 voicing source model.

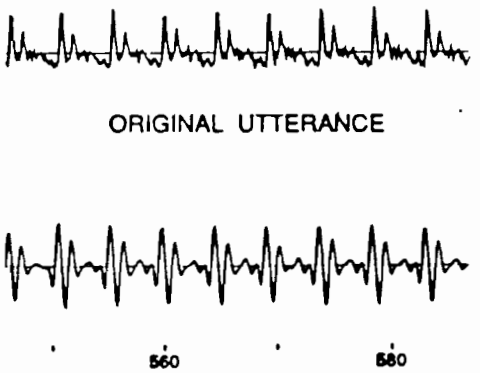


Figure 3.17: *The waveform of a breathy low vowel (top) has been bandpass filtered (bottom) to extract F1 and show how the bandwidth of the first formant can increase during the open part of a glottal period such that the waveform decays to zero much more rapidly during the second half of each cycle.*

Nonlinear F1 – f_0 Interaction. The pharyngeal pressure standing waves may actually influence the mechanical behavior of the vocal folds. One nonlinear effect that could be associated with acoustical-to-mechanical coupling is an increase in glottal source strength whenever F1 is near an integral multiple of f_0 (Fant and Martony, 1963; Fant and Ananthapadmanabha, 1982). Perhaps the pressure changes associated with the F1 standing wave induce a stronger closure if the phase is favorable, and this occurs whenever $F1 = n * f_0$. It should be possible to simulate the essential characteristics of this type of interaction by causing the synthesis parameter AV to increase whenever a harmonic is close to the frequency of F1. However, informal attempts to detect this effect in pitch glides produced by two speakers from our laboratory have been unsuccessful. Perhaps the interaction occurs only under some special glottal conditions and need not be modeled most of the time.

Truncation of the F1 Damped Sinusoid. The time-varying glottal impedance affects the vocal-tract transfer function primarily by causing losses at low frequencies to increase when the glottis is open. The first formant bandwidth may increase substantially, leading to a truncation of the damped sinusoid corresponding to F1 during the open portion of the period (Fant and Ananthapadmanabha, 1982), as illustrated in Figure 3.17. Effects of time varying formant bandwidths can be approximated in a formant synthesizer either by employing a perceptually equivalent (constant) increased bandwidth, or by varying bandwidth over a period. Perceptual data indicate that it is difficult but not impossible to hear the difference between a time-varying first formant bandwidth and an appropriately chosen

constant bandwidth (Nord *et al.*, 1984). Some time variation in formant frequencies may also be desirable; F1 has been observed to increase by as much as 10% during the open phase of a glottal cycle.

A method for changing first formant bandwidth and first formant frequency pitch-synchronously is included in the new KLSYN88 synthesizer. The variables **DF1**, "delta frequency of F1" the incremental increase in first formant frequency during the open portion of each period, and **DB1**, "delta bandwidth of F1", the incremental increase in first formant bandwidth during the open portion of each period, have been created in order to allow pitch-synchronous changes to **F1** and **B1**. For example, to have **F1** = 500 Hz during the closed phase and 550 Hz during the open phase of each period, one would set **F1** = 500 and **DF1** = 50. In a low vowel, the time variation in first formant bandwidth might be approximated by setting **B1** = 50 and **DB1** = 400. A perceptually nearly equivalent constant first formant bandwidth (equal spectral level of F1) corresponds to a first formant bandwidth setting of about 90 Hz. The default values for the **DF1** and **DB1** incremental parameters are set to zero because most users will not need to resort to this kind of detail during synthesis.

Tracheal Poles and Zeros. Tracheal resonances may show up as additional pole-zero pairs in the vocal-tract transfer function, especially for breathy phonation where the glottis is presumably open over its posterior portion throughout the glottal cycle (Fant *et al.*, 1972; Klatt and Klatt, 1990). An example is shown in Figure 3.18. Effects of tracheal coupling can be modeled in a formant synthesizer by adding one or more paired pole-zero resonators to the vocal-tract transfer function (Fant *et al.*, 1972; Ishizaka it et al., 1976; Cranen and Boves, 1987). Cranen and Boves (1987) measured values of the lowest three tracheal resonances of 510, 1350 and 2290 Hz from one male speaker. Slightly higher, but comparable values were observed in vowel spectra from ten female speakers (Klatt and Klatt, 1990).

A single tracheal pole-zero pair has been added to the cascade model of the vocal tract transfer function of the new KLSYN88 synthesizer in order to improve the synthesis of breathy vowels.[6] The variable **FTP**, "frequency of the tracheal pole", in consort with the variable **FTZ**, "frequency of the tracheal zero", can mimic the primary spectral effects of tracheal coupling in breathy vowels.

Tracheal resonances are often seen in breathy vowels at frequencies of about 550, 1300 and/or 2100 Hz (slightly higher for female voices). The best synthesis strategy is to pick the most prominent one for synthesis (or use the nasal pole-zero pair to simulate a second). Normally, the spectral dip or zero corresponding to the selected tracheal resonance is immediately below it in frequency, but in some cases it can be higher in frequency. Tracheal coupling usually begins and ends gradually as the glottis is opened or closed. This observation suggests a synthesis strategy in which both the tracheal pole and zero are moved together to the frequency location of an observed tracheal pole, and then the frequencies of the tracheal zero **FTZ** and the tracheal pole **FTP** are gradually moved together over perhaps 50 ms prior to glottal abduction to an appropriate value, as revealed by spectral analysis of the breathy

[6]The nasal pole-zero pair is also available much of the time for use in mimicking a second tracheal resonance if desired.

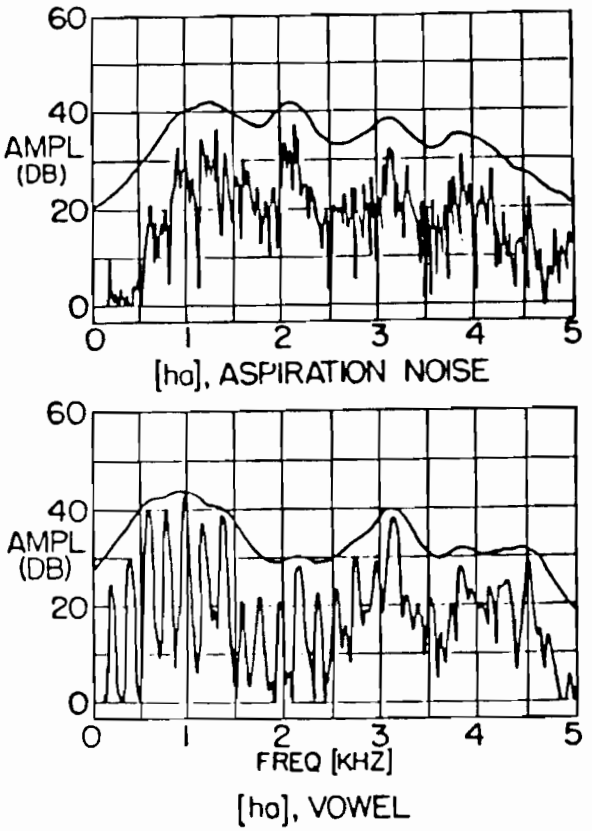


Figure 3.18: *Two examples of the intrusion of tracheal resonances in the spectra of speech sounds: in an /h/ noise, there is an extra resonance at about 2.2 kHz for this female speaker, and the same peak shows up in spectra of the following vowel, with additional evidence of an extra tracheal peak in the lower frequency F1-F2 region.*

interval.

The variables **BTP**, "bandwidth of the tracheal pole", and **BTZ**, "bandwidth of the tracheal zero", have default values of 180 Hz. It is difficult to determine appropriate synthesis bandwidths for individual tracheal resonances, but, fortunately, one can achieve good synthesis results without changing these default values in most cases.

If the location of a tracheal zero is not clear from analysis of a breathy vowel, one possible synthesis strategy is to leave the frequencies of the tracheal pole and zero overlapped, and simply increase the bandwidth of the the zero (and/or decrease the bandwidth of the pole) in order to reveal the presence of the tracheal pole as a resonance peak in the synthesis. Each doubling of the zero bandwidth will approximately increase the strength of the tracheal resonance by about 6 dB.

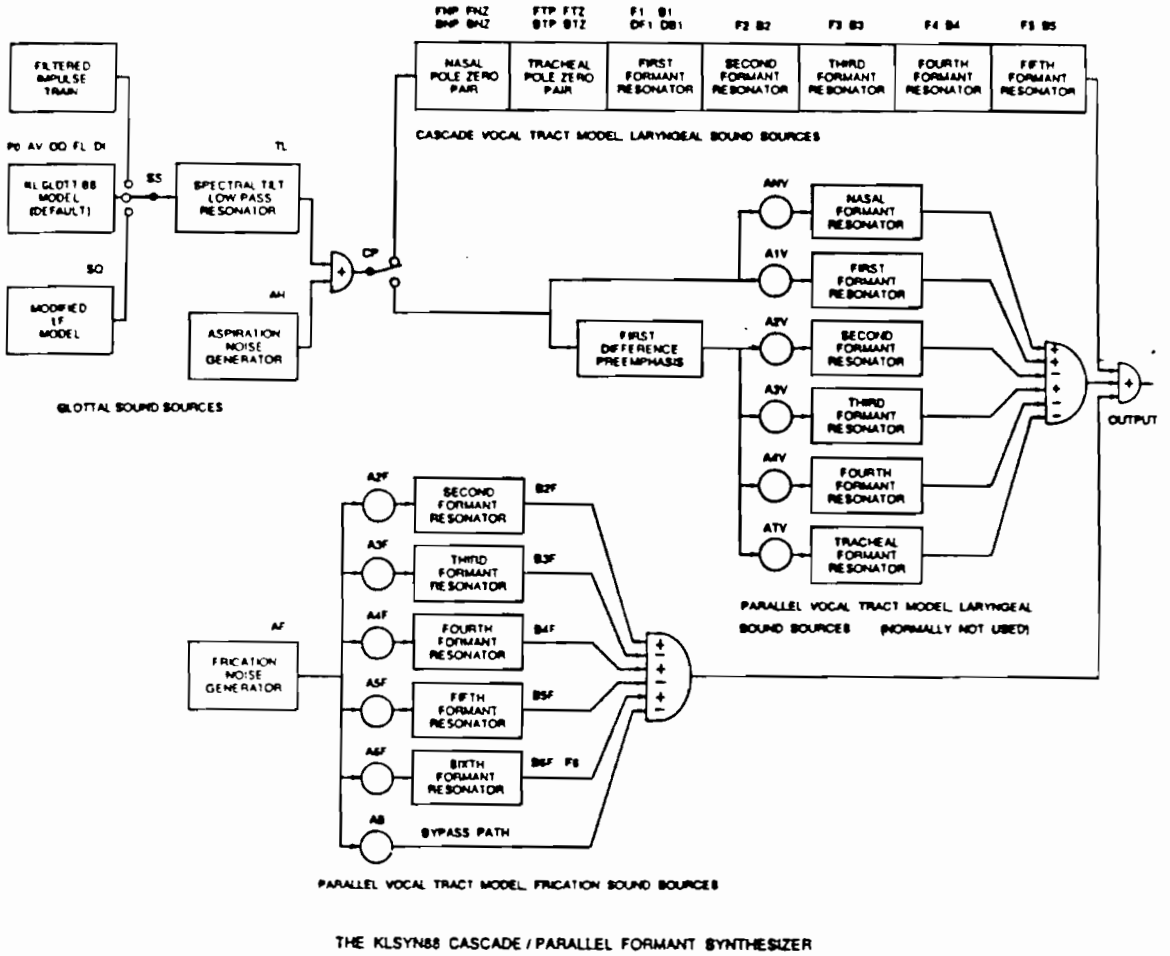


Figure 3.19: *Block diagram of the new KLSYN88 formant synthesizer. Three voicing source models are available: (1) the old Klatt-80 impulsive source, (2) the new KLGLOTT88 model (the default), and (3) the modified LF model. Also added are a tracheal pole-zero pair, and control parameters allowing the first formant frequency and bandwidth to vary over a fundamental period.*

In summary, the old cascade/parallel formant synthesizer (Klatt, 1980) has been modified to incorporate (1) a new voicing source model having flexible control of open quotient, spectral tilt, aspiration noise of breathiness, flutter to the timing of individual glottal pulses, and diplophonic double pulsing, (2) an ability to change the first formant bandwidth pitch-synchronously and (3) an extra pole-zero pair for simulating the introduction of a tracheal resonance in the vocal tract transfer function for a breathy vowel or /h/. This introductory section has outlined the various types of sound sources and vocal-tract transfer functions employed in the generation of English speech sounds. The next section will specify equations for the synthesis of each type of sound source and for the vocal-tract transfer function that is needed in a formant-based synthesis strategy.

3.2 Synthesis Algorithms

A complete detailed block diagram of KLSYN88 is shown in Figure 3.19. The SS source switch control parameter permits the choice between three voicing source algorithms. The CP cascade/parallel switch permits the choice between cascade or parallel synthesis of vowels and other larynx-excited sounds.[7] A frication noise source excites a set of resonators configured in parallel in order to model frication spectra and plosive bursts. The following subsections describe in detail the algorithms used in each component of the block diagram. Following that, the next section provides a users guide that describes the action of each synthesis control parameter.

3.2.1 Basic Synthesizer Building Block: A Digital Resonator

The basic building block of the synthesizer is a digital resonator having the properties illustrated in Figure 3.20. Two parameters are used to specify the input-output characteristics of a resonator: the resonant (formant) frequency F and the resonance bandwidth BW.

Samples of the output of a digital resonator, $y(nT)$, are computed from the input sequence, $x(nT)$, by the equation:

$$y(nT) = A\,x(nT) + B\,y(nT - T) + C\,y(nT - 2T) \tag{3.5}$$

where n is in integer counter of time, T is the reciprocal of the sampling rate of 10,000 samples/sec., and $y(nT - T)$ and $y(nT - 2T)$ are the previous two sample values of the output sequence $y(nT)$.

The constants A, B, and C are related to the resonant frequency F and the bandwidth BW of a resonator by the impulse-invariant transformation (Gold and Rabiner, 1968):

$$\begin{aligned} C &= -e^{(-2\pi BW\,T)} \\ B &= 2e^{(-\pi BW\,T)}\cos(2\pi F\,T) \\ A &= 1. - B - C \end{aligned} \tag{3.6}$$

The exponential and cosine functions are transcendental functions that are computed by calls to standard subroutines provided by VAX VMS in a computer simulation, but must be approximated (or obtained by table lookup) when using a limited compute-power synthesizer chip. Appropriate approximations for use in Klattalk will be given in Section 3.4.

A resonator can be effectively removed from the cascade synthesizer by setting the coefficients A, B, C to 1.0, 0.0, 0.0 respectively. This is necessary if the speaker one is trying

[7] The parallel model is rarely used except to generate certain pathological stimuli consisting of one or two formants. The larynx-excited parallel vocal tract model is not used at all in Klattalk.

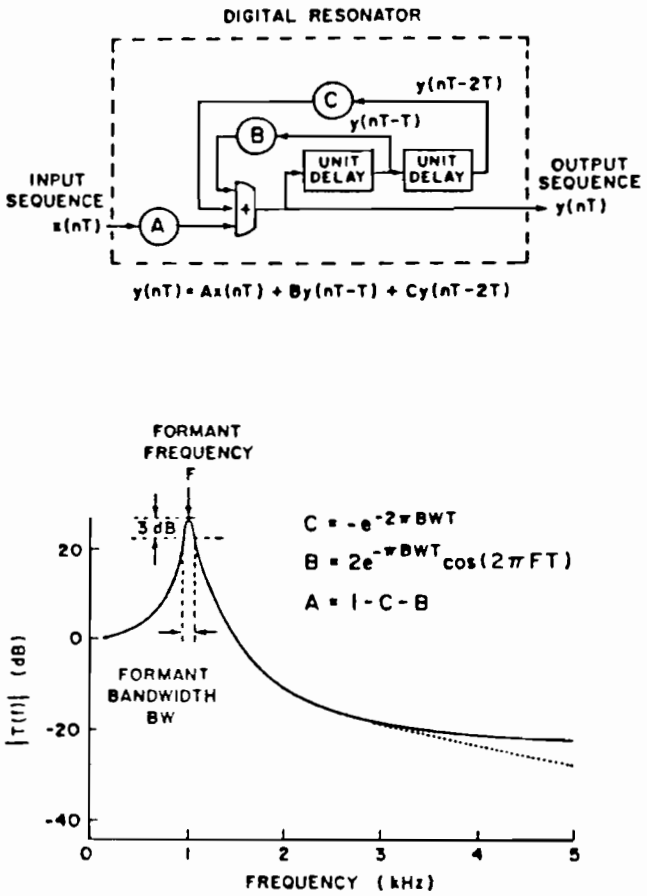


Figure 3.20: *The digital resonator shown in the form of a block diagram in the upper part of the figure has a transfer function, T(f), (magnitude of the ratio of output to input in the frequency domain) as shown. In this example, F=1000 Hz and BW=50 Hz. The transfer function of a corresponding analog resonator is shown by the dotted line.*

to imitate has a short vocal tract and correspondingly fewer formants in the frequency range up to 5 kHz.

Transfer Function of a Digital Resonator

A digital resonator is a second-order difference equation. The transfer function of a digital resonator has a sampled frequency response given by:

$$T(f) = \frac{A}{1. - B\,z^{-1} - C\,z^{-2}} \tag{3.7}$$

where $z = e^{(j2\pi fT)}$, and f is frequency in Hz, ranging from 0 to 5 kHz. The transfer function has a (sampled) impulse response identical to a corresponding analog resonator circuit at sample times nT (Gold and Rabiner, 1968), but the frequency responses of an analog and

digital resonator are not exactly the same, as can be seen in Figure 3.20. The digital resonator does not fall off in magnitude near 5 kHz because a digital resonator defined by Equation 3.7 actually includes an infinite set of higher-frequency poles. The transfer function is periodic, repeating every 10,000 Hz. For example, a resonator having a formant frequency of 500 Hz also has poles at 9500, 10500, 19500, 20500, etc. Hz. This is the reason why there is no need for the higher-pole correction term $K(f)$ in a digital realization of Equation 3.4.

Waveform Irregularities Caused by Sudden Changes to F and BW

The values of the resonator control parameters F and BW are normally updated every 5 ms, causing the difference equation constants to change discretely as an utterance is synthesized. In a physical system, such as the vocal tract, the resonance frequency and bandwidth change continuously. The changes are usually slow relative to the 5 ms update interval of the synthesizer, so that the discrete changes made to F and BW at the frame rate are incrementally too small to cause any problems. However, a rapid articulatory change, as in a stop release, can cause moderately large formant frequency changes within one 5 ms frame. Large changes to F or BW may introduce clicks and burps in the synthesizer output because the variables $y(nT-T)$ and $y(nT-2T)$ contain values based on the previous settings for F and BW.

The problem of rapidly changing the frequency of a digital resonator has been addressed by Fujisaki and Azami (1971). They show that it is possible to change frequency very rapidly without introducing undesirable transients, but at a significant computational cost. The history variables $y(nT-T)$ and $y(nT-2T)$ of Equation 3.5 must be modified every time that the frequency changes, using the previous sample of the A coefficient in the following way. Assume that the current formant frequency F results in A, B, C, and that a change to formant frequency F_{new} results in A', B', C'. All that is necessary is to change $y(nT-T)$ and $y(nT-2T)$ at the same time that F or BW are changed:

$$\begin{aligned} y(nT-T)' &= y(nT-T) \times \sqrt{A'/A} \\ y(nT-2T)' &= y(nT-2T) \times \sqrt{A'/A} \end{aligned} \tag{3.8}$$

where n is the number of the current output time sample. The correction guarantees that a change to the new desired frequency will not produce a discontinuous change in output waveform level. The correction is used every time that **F1**, **F2** and **F3** are set.

3.2.2 Voicing Source Models

The three choices of voicing source model are (1) the impulsive source that was used in Klatt (1980), (2) the KLGLOTT88 source model that serves as the default, and (3) a slightly modified version of the Liljencrants-Fant LF model. The default model will be described first. Control parameters having an effect on voicing source characteristics for all three source models include:

- AV, amplitude of voicing, in dB
- F0, voicing fundamental frequency, in tenths of a Hz
- OQ, open quotient of the glottal waveform, in percent of a full period
- TL, tilt of the voicing source spectrum, in dB down at 3 kHz
- FL, period-to-period flutter (quasi-random fluctuations) in f_0, in percent of maximum
- DI, degree of diplophonic double pulsing irregularity in f_0, in percent of maximum
- AH, amplitude of aspiration (breathiness) noise, in dB

The KLGLOTT88 Voicing Source Model (SS = 2)

The KLGLOTT88 volume velocity waveform has been parameterized to obey a relationship first proposed by Rosenberg (1971). During the open portion of the voicing waveform, the wave shape of the natural voicing source follows the equation:

$$U_g(t) = a\ (t^2 - \frac{t^3}{(\frac{OQ}{100}) * T0}) \tag{3.9}$$

where T0 is the duration of the fundamental period ($1/f_0$), **OQ** is the percent of the period T0 during which the glottis is open, and a is a constant chosen so that the waveform has the desired peak amplitude. The rather simple equation that has been selected to describe the glottal volume velocity during the open portion of a glottal cycle nonetheless simulates many of the characteristics of a natural glottal waveform, such as the tendency of the flow to decrease more rapidly than it increases.

Because the radiation characteristic has the effect of taking the derivative of the output from the lips, it is not necessary to implement Equation 3.9. Instead, we simplify by incorporating the radiation characteristic as part of the voicing source, i.e. by taking the derivative as a part of the source calculation (in an ideal linear system, the order of operations does not affect the output):

$$U'_g(t) = \frac{dU_g(t)}{dt} = a\ (2t - \frac{3t^2}{(\frac{OQ}{100}) * T0}) \tag{3.10}$$

The waveform and spectrum of the new natural voicing source are shown in Figure 3.21. The spectrum falls off at a rate of about -12 dB per octave of frequency increase. Also shown in the figure is the derivative of the source volume velocity, $U'_g(t)$, which has a spectrum that falls at about -6 dB/oct.

OQ, The Open Quotient. The only place that the open quotient control parameter appears in the source equations is in Equation 3.10, where it determines the duration of the

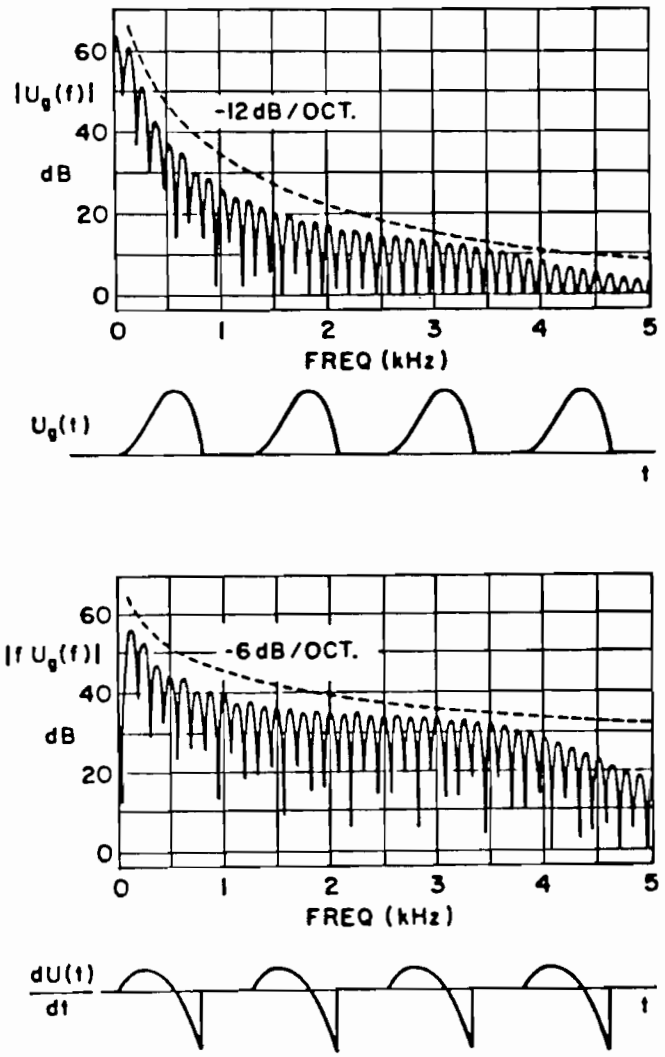


Figure 3.21: *Waveforms and spectra of the new KLGLOTT88 voicing source model with default values assigned to all control parameters. The upper panels represent the spectrum and waveform of $U_g(t)$ and the lower panels are for $U'_g(t)$. The dashed lines indicate the shape by an ideal spectrum that decreases at -12 or -6 dB/oct.*

open portion of a period T0. Open quotient is specified in percent of a full period. The spectral consequences of varying open quotient are shown in Figure 3.22B.

T0 The Number of Samples in the Fundamental Period. The fundamental frequency changes over the course of a sentence, often starting at a relatively high value, and falling at the end. Natural voicing is thus called "quasi-periodic" to indicate that the waveform is nearly repeatable, but changes slightly from period to period, either due to changes in fundamental frequency, or to changes in other parameters to be discussed below.

The synthetic waveform for which the spectrum is shown in Figure 3.22 is perfectly periodic: it repeats every 10 msec, and thus has a fundamental frequency of 100 Hz. The dynamic synthesizer control parameter **F0**, "fundamental frequency", specifies the rate at

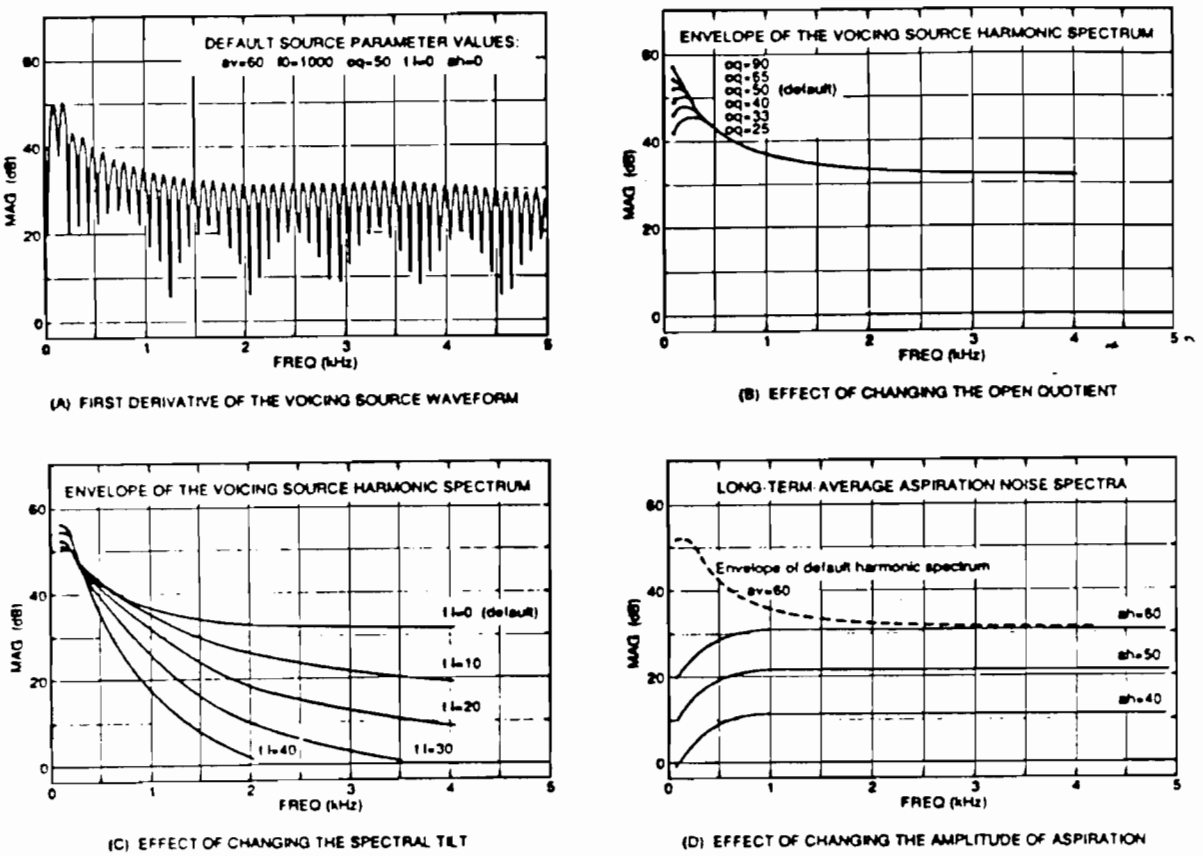


Figure 3.22: *DFT magnitude spectra are shown of $U_g'(t)$ as synthesized by the new KLGLOTT88 voicing source model with several values for each of three control parameters. Part (A) shows the spectrum of a train of pulses with $f_0 = 100Hz$, while the curves of parts (B) and (C) represent the spectral envelope of a harmonic spectrum that would result from synthesizing a train of such pulses as* **OQ** *and* **TL** *are varied. Part (D) gives the spectrum of the aspiration noise when* **AH** *is varied.*

which the vocal folds are currently vibrating in Hz. Actually, **F0** is specified internally in units of Hz times 10. If a fundamental frequency of 100 Hz is desired, then **F0** is set to 1000. The additional accuracy (reduced quantization) resulting from a specification of fundamental frequency to within a tenth of a Hz adds naturalness to a slowly changing pitch glide.

The period T0, in samples at 10,000 samples/sec, is given by:

$$T0 = integer\ part\ of\ \left(\frac{100,000}{F0}\right) \tag{3.11}$$

Period Quantization. A listener is very sensitive to changes in fundamental frequency (Flanagan and Saslow, 1958; Klatt, 1973). This sensitivity creates a problem in a digital simulation of the voicing sound source, because the period T0 is normally quantized to be an integer number of samples. For example, if the sampling rate is 10,000 samples per second, as it is in Klattalk, then a period of 8 ms corresponds to 80 samples. An increase

of one sample to the period would change fundamental frequency, f_0, from 125 Hz to 123.46 Hz. If only integral numbers of samples were available, the fundamental frequency of voicing would thereby be quantized to approximately a one or two percent step size for a male voice, and to about a 2 or 3 percent step size over the range of a female voice. This step size is large enough to be heard, especially if f_0 changes slowly and/or the pitch is high. The auditory impression, something like a staircase pattern, is unnatural.

Therefore, the original synthesizer described by Klatt (1980) has been modified so as to reduce period quantization. In effect, all of the computations involved in generating a voicing waveform are performed at 4 times the sampling rate of the rest of the synthesizer, i.e. 40,000 samples per second. This over-sampled waveform is then digitally low-pass filtered and downsampled to 10,000 samples/sec, using standard signal processing techniques. The f_0 quantization is reduced by a factor of four. The remaining quantization is less than the just-noticeable difference for a change in f_0 under normal time-varying listening conditions (Klatt, 1973).

The period T0, in samples at 40,000 samples/sec, is given by:

$$T0 = integer\,part\,of\left(\frac{400,000}{F0}\right) \tag{3.12}$$

This is the value of T0 that will be assumed in all of the equations involving the various voicing sources described below.

Downsampling Low-Pass Filter. Theoretically, the voicing waveform, which contains 40,000 samples/sec and thus reproduces the spectrum energy up to 20,000 Hz, should be sharply filtered to remove all energy above 5 kHz before downsampling to 10,000 samples/sec. Otherwise, the resulting spectrum can vary unpredictably depending on the details of which of the four samples are retained during downsampling. As a practical matter, a two-pole resonator, acting as a low-pass filter, appears to be adequate for the purpose. The two-pole filter has a nominal center frequency of 3600 Hz and bandwidth of 2400 Hz.

Flutter. The new KLSYN88 voicing source model includes a mechanism for introducing a slow quasi-random drift to the f_0 contour through the **FL** flutter control parameter. Instead of using a random process to simulate jitter, we add to the nominal f_0 a quasi-random component which is in fact the sum of three slowly varying sine waves:

$$\Delta f_0 = \frac{FL}{50} * \frac{F0}{100} * [sin(2\pi\,12.7t) + sin(2\pi\,7.1t) + sin(2\pi\,4.7t)]\ Hz \tag{3.13}$$

Sine wave frequencies of 12.7, 7.1, and 4.7 Hz were chosen so as to ensure a long period before repetition of the perturbation that is introduced. Δf_0 is added to the value of the **F0** control parameter before computing the period T0.

Diplophonia. If the **DI** diplophonic double pulsing control parameter is non-zero, a specified fraction of the closed portion of a glottal pulse is determined and either added or subtracted from the total period of alternate periods. The duration of this increment/decrement to the period is given by:

$$\Delta T0 = \frac{DI}{100} * T0 * (1.0 - \frac{OQ}{100}) \tag{3.14}$$

where T0 was given earlier in Equation 3.12. The closed part of a period comes at the end of each period, so if $\Delta T0$ is added to T0, the next pulse will be delayed. A delayed pulse is also attenuated in amplitude according to the equation:

$$AV_{lin} = AV_{lin} * (1.0 - \frac{DI}{100}) \tag{3.15}$$

where AV_{lin} is a linearized version of the **AV** control parameter that is specified in dB, see next paragraph.

Amplitude of Voicing. The amplitude of the voicing source, **AV**, is a scale factor, specified in dB and converted to a linear scale factor AV_{lin} by table lookup, that determines actual values of the voicing source amplitude by a simple multiplication:

$$voice = vsource \times AV_{lin} \tag{3.16}$$

where *vsource* is the nominal amplitude of the source, and *voice* is the amplitude of the excitation of the filters.

The synthesizer does not necessarily turn voicing on and off at exactly the time specified by the **AV** dynamic control parameter. The effect of a change in **AV** is delayed until the instant of the next glottal waveform opening. The primary excitation of the vocal tract actually occurs somewhat later, at the next glottal closure.

If **AV** is suddenly turned off, no more glottal pulses will be issued, and the vocal tract response to the previous pulse will die out, taking at least 20 msec to become totally inaudible (assuming there are no limit cycles, see discussion below of resonator characteristics).

The amplitude of voicing is scaled by an overall gain control parameter **GV**, in dB, according to the equation:

$$AV_{lin} = AV_{lin} * GV_{lin} \tag{3.17}$$

where GV_{lin} is a linearized version of of **GV**. There is also a hidden arbitrary scale factor to make sure that synthesis of a vowel having default values for all control parameters results in a waveform that has sufficient amplitude to occupy all of the bits of the D/A converter.

Spectral Tilt. The **TL** spectral tilt control parameter ranges in value from 0 to 41 dB. Tilt is realized by a critically damped digital resonator which serves effectively as a low-pass filter. The bandwidth of the resonator takes on a value which is specified by the conversion listed in Table 3.1. Values in the table have been adjusted to give a spectral tilt at 3 kHz equal to the control parameter value, as shown in Figure 3.22C. The frequency setting of the tilt digital resonator in order to achieve critical damping is $F_{tl} = 0.375\ BW_{tl}$.

Desired TL	BW_{tl}	Desired TL	BW_{tl}
0	5000	21	1071
1	4350	22	1009
2	3790	23	947
3	3330	24	885
4	2930	25	833
5	2700	26	781
6	2580	27	729
7	2468	28	677
8	2364	29	625
9	2260	30	599
10	2157	31	573
11	2045	32	547
12	1925	33	521
13	1806	34	495
14	1687	35	469
15	1568	36	442
16	1449	37	416
17	1350	38	390
18	1272	39	364
19	1199	40	338
20	1133	41	312

Table 3.1: *Conversion from the desired tilt, synthesizer control parameter* **TL** *in dB, to the bandwidth* BW_{tl} *in Hz of the low-pass digital resonator used to create varying amounts of spectral tilt to the voicing source.*

Normally, a digital resonator has unity gain at $f = 0$ Hz, but we desire a behavior pattern such that the gain near F1 is nearly constant with changes to **TL**. An approximation is used to make the gain at 300 Hz nearly constant; the digital resonator gain constant A_{tl} is adjusted whenever **TL**> 10 according to the formula:

$$A_{tl} = A_{tl}\ (1.0 + \frac{(TL - 10)^2}{1000}\ ) \tag{3.18}$$

Impulsive Source (SS = 1)

Many early speech synthesizers used a filtered impulse (Figure 3.23 top) to simulate the shape of the voicing waveform (see Klatt, 1980 and references cited therein). The phase of this waveform is wrong, since primary excitation of the vocal tract occurs at a time corresponding to instant the folds open. Furthermore, the spectrum envelope is perfectly regular (i.e., monotonically decreasing, in contrast with evidence characterizing aspects of normal voicing waveforms and spectra (Flanagan, 1958; Miller, 1959; Mathews, Miller and

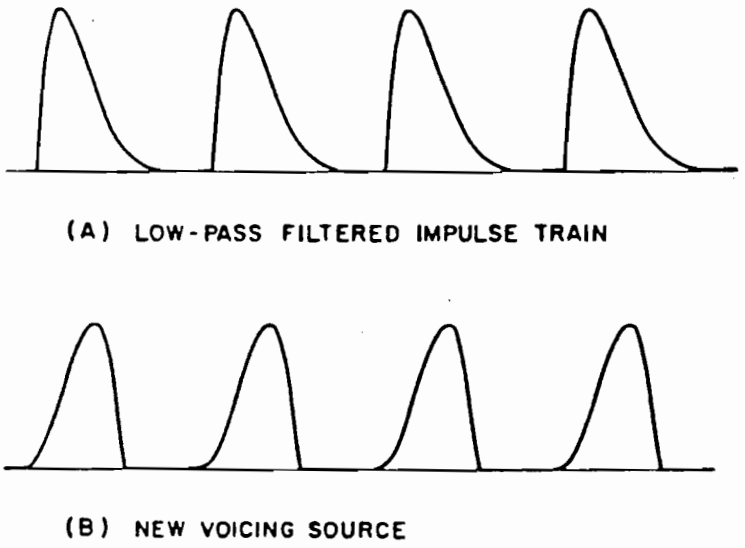


Figure 3.23: *Comparison of synthetic voicing source waveforms based on (a) filtering an impulse and (b) computing a new more natural wave shape.*

David, 1961; Monsen and Engebretsson, 1977; Fant, 1979; Sundberg and Gauffin, 1979; Ananthapadmanabha, 1984)). Nonetheless, the impulsive source is useful in many situations where simple regular source spectra are desired.

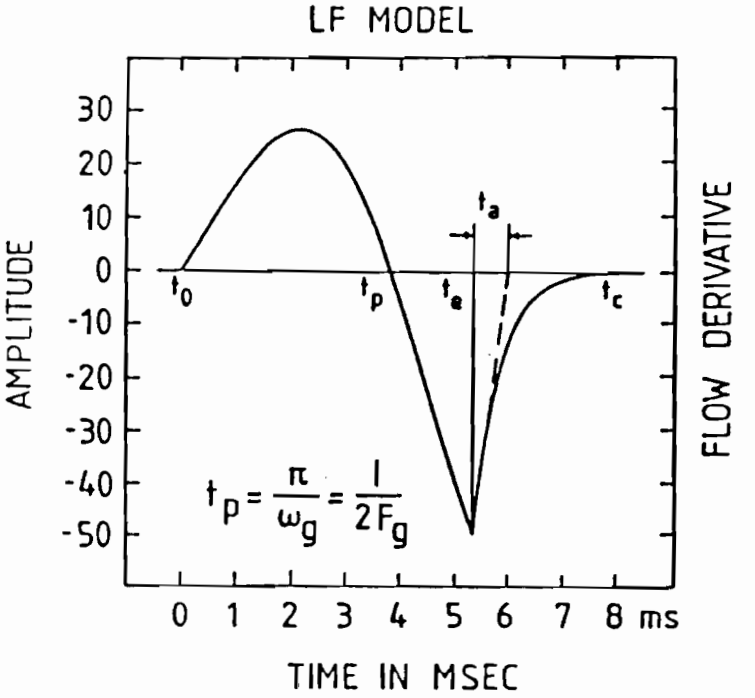
The impulsive source is realized by sending an impulse train through a low-pass digital resonator. The low-pass behavior is accomplished by setting the pole on the real axis, i.e., by setting $F_{imp} = 0$. The resonator bandwidth BW_{imp} determines the effective width of pulse, an example of which is shown in Figure 3.23. There is no well-defined closing time, but it is still possible to have the nominal pulse width change as a function of values of the **OQ** open quotient control parameter according to the equation:

$$BW_{imp} = \frac{10000}{T0 * (\frac{OQ}{100})} \tag{3.19}$$

where T0 was given earlier in Equation 3.12. The effect of a change in open quotient on the source spectrum should be a local increase/decrease of the amplitude of the first harmonic, but unfortunately there is no simple way to obtain this behavior for the impulsive source, and essentially the change to **OQ** causes a change in several harmonics in the low-frequency range.

Normally, a digital resonator has unity gain at $f = 0$ Hz, but we desire a behavior pattern such that the gain near F1 is nearly constant with changes to **OQ**. An approximation is used to make the gain at 300 Hz nearly constant; the digital resonator gain constant A_{imp} is adjusted according to the formula:

$$A_{imp} = A_{imp} * [1.0 + (0.00833 * T0 * \frac{OQ}{100})^2] \tag{3.20}$$



$$\text{Segment 1.} \quad E(t) = E_e e^{\alpha t} \sin \omega_g t \qquad (t_0 \le t \le t_e)$$

$$\text{Segment 2.} \quad E(t) = \frac{-E_e}{\epsilon t_a} \cdot \left[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)} \right] \qquad (t_e \le t \le t_c)$$

Figure 3.24: *The Liljencrants and Fant (LF) model is defined in terms of a set of waveform events, after Gobl (1988).*

where T0 was given earlier in Equation 3.12. All other voicing source control parameters affect the characteristics of the impulsive source in exactly the same way as they affect the natural source described above.

The Modified LF Model (SS = 3)

The Liljencrants-Fant (LF) model (Fant *et al.*, 1985) was originally formulated in terms of a set of times of waveform events, as illustrated in Figure 3.24, but it can easily be recast in terms of familiar parameters **AV** (amplitude of voicing), **F0** (fundamental frequency), **OQ** (open quotient), **SQ** (speed quotient), and **TL** (spectral tilt). The waveform of the derivative of $U_g(t)$ during the open phase is characterized as a portion of a single cycle of an underdamped sinusoid:

$$U_g'(t) = E_0 \, e^{\alpha t} \, sin(2\pi F_1 t) \qquad (3.21)$$

where the open phase corresponds to $(0 < t < (OQ * T0))$, $U_g'(t) = 0$ during the remainder of the period T0, and this equation expresses the output before computing the effect of any corner rounding (spectral tilt) due to nonsimultaneous closure. The equation can be simulated by the impulse response of an underdamped resonator; E_0 is a scale factor to make the amplitude of the negative peak in the flow derivative independent of other parameters; alpha is the parameter that is adjusted to make $U_g(t)$ return to zero flow at closure; and the sinusoidal frequency F_{lf} is equal to the reciprocal of the duration of the rising portion of the glottal flow waveform, which we re-express as:

$$F_{lf} = \frac{F0}{(\frac{OQ}{100} * 8)} * \frac{(SQ + 100)}{SQ} \tag{3.22}$$

If we assume **SS** = 3, all other control parameters are at their default values, and $T0$ is given by Equation 3.12, then F_{lf} = 375 Hz. Actual modifications that have been imposed in the LF model include a change in what aspect of the waveform is held constant as each control parameter is varied (recall Figure 3.6, and how corner rounding is achieved).

The bandwidth of the digital resonator used to generate the LF glottal output BW_{lf} is inversely proportional to the open period, and is complexly related to alpha in Equation 3.21. The relationship to alpha is computed by table lookup, using the speed quotient **SQ** as an argument to a 41-element array (100 <= **SQ** <= 500):

$$BW_{lf} = bwlftab[\frac{SQ}{10}] * \frac{200}{T0 * (\frac{OQ}{100})} \tag{3.23}$$

where interpolation via the remainder of $SQ/10$ is used for greater accuracy. Assuming default values for all control parameters except **SS** = 3, and $T0$ given by Equation 3.12, the computed value of $BW_{lf} = -20.1$ Hz.

The gain adjustment E_0 to the digital resonator is proportional to the open period, and is also complexly related to alpha. The relationship to alpha is computed by table lookup, using the speed quotient **SQ** as an argument to a 41-element array (100 <= **SQ** <= 500):

$$E_0 = gainlftab[\frac{SQ}{10}] * \frac{T0 * (\frac{OQ}{100})}{200} \tag{3.24}$$

where interpolation is again used for values not in the array. Assuming default values to all control parameters except **SS** = 3, and $T0$ given by Equation 3.12, the computed value of $E_0 = 16.1$. The "A" coefficient of the underdamped resonator is multiplied by E_0 every time that the resonator frequency and bandwidth are reset, i.e. at the beginning of each period.

Values for both arrays are presented in Table 3.2. Note that negative bandwidths imply an unstable filter, but the digital resonator is being used to generate less than one full period of its response before being reset to initial conditions at the beginning of each fundamental period.

Desired SQ	BW_{lf}	E_0	Desired SQ	BW_{lf}	E_0
100	0.0	27.4	310	-44.1	6.05
110	-0.6	26.3	320	-46.2	5.46
120	-2.0	25.3	330	-48.3	4.92
130	-4.0	24.3	340	-50.4	4.41
140	-6.0	23.2	350	-52.4	3.94
150	-8.0	22.1	360	-54.5	3.58
160	-10.4	21.0	370	-56.6	3.14
170	-12.7	20.0	380	-57.8	2.83
180	-15.3	18.8	390	-60.8	2.49
190	-17.8	17.6	400	-62.7	2.24
200	-20.1	16.1	410	-64.5	2.03
210	-22.4	14.9	420	-66.3	1.83
220	-24.7	13.8	430	-68.1	1.63
230	-27.0	12.8	440	-69.9	1.48
240	-29.2	11.7	450	-71.6	1.32
250	-31.4	10.6	460	-73.3	1.19
260	-33.6	9.81	470	-75.0	1.08
270	-35.8	9.00	480	-76.6	.982
280	-37.9	8.12	490	-78.2	.902
290	-40.0	7.36	500	-79.6	.832
300	-42.1	6.60			

Table 3.2: *Conversion from speed quotient* **SQ** *to digital resonator bandwidth* BW_{lf} *and gain factor* E_0 *in the LF voicing source model. Interpolation is performed for* **SQ** *values between those presented in the table.*

3.2.3 Equations for the Noise Source

The turbulence noise of frication and aspiration is simulated by a pseudo-random number generator using 16-bit arithmetic (Knuth, 1981). A new value for "ran", a 16-bit variable, is computed based on the previous value, according to the equation:

$$ran = 16 \; low \; order \; bits \; of \; [(ran * 20077) + 12345] \tag{3.25}$$

The variable *ran* cycles through all possible integers from $\pm 2^{15}$ before repeating. The noise source waveform consists of a sequence of 10,000 of these pseudo-random numbers per second. The spectrum is flat (recall Figure 3.10), and represents the combined spectrum of the noise source and the radiation characteristic.

The random number generator is reinitialized to a special value that produces a particularly flat spectrum over the next 5 to 10 ms, using control parameters **SB** (same burst), and **RS** (random seed). If **SB** = 1, then the random number generator is reset to the value of RS on every frame for which both frication amplitude and aspiration amplitude

control parameters are zero. This strategy will result in the same burst noise sequence every time **AF** is suddenly turned on to produce a burst of frication noise.

Amplitude Modulation. The output of the random number generator is amplitude modulated whenever the amplitude of voicing AV is greater than zero. Voiceless sounds (AV=0) are not amplitude modulated because the vocal folds are spread and stiffened, and do not vibrate to modulate the airflow. The degree of amplitude modulation is fixed at 50 percent in the synthesizer, and the modulation envelope is a square wave with a period equal to the fundamental period. This is accomplished by multiplying the variable *ran* by 0.5 during the nominal closed phase of a voicing period if **AV** is on, i.e. **AV** > 0. Experience has shown that it is not necessary to vary the degree of amplitude modulation over the course of a sentence, but only to ensure that it is present in voiced fricatives and voiced aspirated sounds.

Aspiration Amplitude. The amplitude of the aspiration noise source, **AH**, is a scale factor, specified in dB and converted to a linear scale factor AH_{lin} by table lookup, that determines actual values of the aspiration noise source amplitude by a simple multiplication:

$$asp = ran \times AH_{lin} \tag{3.26}$$

After scaling by the constant gain factor control parameter GH, the aspiration source has been calibrated such that a value of **AH**=60 dB will generate aspiration with a level appropriate for /h/, i.e. with a spectrum level in the F3 region that is a few dB less than that in the comparable vowel (assuming **AV** is set to 60 dB), while a value of 0 turns off the aspiration source.

Frication Amplitude. The amplitude of the frication noise source, **AF**, is a scale factor, specified in dB and converted to a linear scale factor AF_{lin} by table lookup, that determines actual values of the frication noise source amplitude by a simple multiplication:

$$fric = ran \times AF_{lin} \tag{3.27}$$

After scaling by the constant gain factor control parameter GF, the frication noise source has been calibrated such that a value of **AF**=60 dB will generate frication noise that excites each parallel formant (that has its amplitude control set to 60) to a spectrum level that is a few dB more intense than the comparable vowel (with **AV** set to 60 dB), while a value of 0 turns off the frication source.

3.2.4 Vocal Tract Models for Synthesis

The synthesizer configuration shown earlier in Figure 3.19 includes components to realize two different types of vocal-tract transfer function. The first, a cascade configuration of digital resonators shown at the top in Figure 3.13, models the resonant properties of the vocal tract whenever the source of sound is within the larynx. The second, a parallel configuration of digital resonators and amplitude controls shown at the bottom in Figure 3.13, models the

resonant properties of the vocal tract during the production of frication noise. In addition, there is a larynx-excited parallel configuration of resonators (shown at the middle-right in Figure 3.19) that is normally not used, but can be activated in order to generate certain types of pathological stimuli such as one or two-formant vowels.

Cascade Vocal-Tract Model for Sonorants

Five digital formant resonators connected in cascade (the output of one feeding the input of the next) are used to model the vocal-tract transfer function for vowels, liquids and glides, as shown in Figure 3.19. In addition, a nasal resonator and antiresonator are cascaded in order to approximate the transfer function for nasals and nasalized segments, and a tracheal pole-zero pair has been added to the cascade in order to approximate the effects of tracheal coupling on the vocal-tract transfer function. The transfer function $T(f)$ is completely specified by the frequencies and bandwidths of the five digital formant resonators plus the frequencies and bandwidths of the nasal and tracheal pole-zero pairs.

All-Pole Transfer Function. When the nasal pole-zero pair and the tracheal pole-zero pair have frequency and bandwidth settings that cause them to cancel one another out, the result is a transfer function consisting only of poles. Such a cascade of five digital formant resonators has a volume velocity transfer function that can be represented in the frequency domain as a product:

$$T(f) = \prod_{n=1}^{5} \frac{A_n}{1. - B_n z^{-1} - C_n z^{-2}} \tag{3.28}$$

where z is an abbreviation for the z-transform variable $e^{j2\pi fT}$, and the constants A_n, B_n, and C_n are determined by the values of the nth formant frequency F_n and nth formant bandwidth BW_n by the relations given earlier in Equation 3.6. Examples of $T(f)$ were given in Figure 3.12.

Pitch-Synchronous Change in First Formant Bandwidth and Frequency. The changes to first formant frequency **DF1** and bandwidth **DB1** occur in "square-wave" fashion, increasing at the instant of glottal opening, and decreasing at the instant of glottal closure, as determined by the open quotient. The change is accomplished by calling a subroutine to reset the digital resonator constants *A, B, and C* for the first formant at each glottal open and closing time.

Nasal Pole-Zero. Nasal murmurs and vowel nasalization are approximated by the insertion of an additional resonator and anti-resonator into the cascade vocal-tract model. An anti-resonance (also called an anti-formant or transfer-function zero pair) can be realized by slight modifications to the equations for a digital resonator. The frequency response of an anti-resonator is the mirror image of the response plotted in Figure 3.20 (i.e., replace dB by -dB). An anti-resonator is used in the synthesizer to simulate the effects of nasalization in the cascade model of the vocal-tract transfer function.

The output of an anti-formant resonator, $y(nT)$, is related to the input $x(nT)$ by the

equation:

$$y(nT) = A'x(nT) + B'x(nT - T) + C'x(nT - 2T) \tag{3.29}$$

where $x(nT - T)$ and $x(nT - 2T)$ are the previous two samples of the input, $x(nT)$, and the constants A', B' and C' are defined by the equations:

$$\begin{aligned} A' &= \frac{1.0}{A} \\ B' &= \frac{-B}{A} \\ C' &= \frac{-C}{A} \end{aligned} \tag{3.30}$$

where A, B, and C are obtained by inserting the anti-resonance center frequency F and bandwidth BW into Equations 3.6.

The default frequency of the nasal pole **FNP** and the nasal zero **FNZ** is 280 Hz. In a nasalized vowel or a nasal consonant, which are produced with a lowered velum, these parameters are shifted, and the bandwidths **BNP** and **BNZ** may be changed. The effect on the transfer function of /ɪ/ was illustrated earlier in Figure 3.16.

The nasal pole-zero pair is effectively removed from the cascade circuit during the synthesis of non-nasalized speech sounds if **FNP=FNZ** because the transfer function of a zero pair is the mirror image of the transfer function of a pole pair having the same frequency and bandwidth settings. The bandwidths of the nasal resonator and antiresonator should therefore be equal; a good value is about 80 Hz.

Tracheal Pole-Zero. A cascaded pole-zero pair is provided in the cascade branch of the synthesizer to mimic the addition of any "spurious" resonant peak due to tracheal coupling interaction, for example in a breathy vowel. Control parameters include the frequency **FTP** and bandwidth **BTP** of the tracheal pole, and the frequency **FTZ** and bandwidth **BTZ** of the tracheal zero.

Parallel Vocal Tract Model for Frication Sources

There are five formant resonators and a bypass path in the parallel configuration of Figure 3.19. Each is supplied with an independent amplitude control, and each receives, as input, the same sequence of samples from the frication noise source. Outputs are added together with alternating sign to best approximate the theoretically correct phase response of the transfer function (Klatt, 1980).

Careful adjustments to the amplitude controls **A2F**, **A3F**, **A4F**, **A5F**, **A6F**, and **AB**, permit specification of the spectral shape of the output frication noise for each obstruent of English. The resonator amplitude controls have been calibrated such that if one is set to 60 dB, and **AF** = 60, the peak spectral level is a few dB more intense than the level

in the corresponding vowel when **AV** is set to 60 dB. The **AB** amplitude control has been calibrated to give about the right level to an /f/ noise when set to 60 dB.

Parallel Vocal Tract Model for Laryngeal Sources (CP = 1)

The larynx-excited parallel vocal-tract model in Figure 3.19 is normally not used, but can be engaged by setting the cascade-parallel switch **CP** to 1. It consists of the four lowest formants plus the nasal pole pair and tracheal pole pair, all configured in parallel with amplitude controls **A1V** through **A4V**, **ANV** and **ATV**. Through suitable adjustments to these amplitude controls, it is possible to closely approximate the vocal-tract transfer function of any vowel.

Some of the these resonators are excited by the first difference of the source waveform (see Figure 3.19), in order to avoid generating a spectrum with too much energy at low frequencies and/or a zero in the low-frequency region of the spectrum. The frequency and bandwidth control values are identical to those for the cascade vocal tract model.

3.3 Synthesizer Control Parameters

The KLSYN88 synthesizer, summarized in Figure 3.19, consists of circuits to generate voicing, aspiration and/or frication, and circuits to approximate the sound source filtering performed by the vocal tract. The radiation characteristic has been folded into the sound sources for computational efficiency. There is a cascade formant model of the vocal tract transfer function for laryngeal sound sources, and a parallel formant model with formant amplitude controls for frication excitation. A third vocal-tract model in which the vocal-tract transfer function for laryngeal sound sources is approximated by formants configured in parallel is useful for some pathological synthesis applications, but is normally not used.

Control parameters are identified above each block in Figure 3.19. Some control parameter names have been changed slightly from Klatt (1980) in order to accommodate the new components and to be more mnemonic. A complete list of synthesizer control parameters is identified in Tables 3.3 and 3.4. The following paragraphs, which are numbered according to the control parameter ordering in the tables, detail the operation of the synthesizer in response to changes in each control parameter available to the user.

1. DU: Utterance Duration

The constant **DU**, "duration", is the total duration in msec of the waveform to be synthesized. The default value is 500 ms, which is sufficient for most simple single-syllable utterances.

Plan to include about 20 ms at the end of the utterance with all source amplitudes set to zero to allow the waveform to decay naturally to zero. This step avoids the generation

	SYM	MIN	VAL	MAX	DESCRIPTION
1.	DU	30	500	5000	Duration of the utterance, in msec
2.	UI	1	5	20	Update interval for parameter reset, in msec
3.	SR	5000	10000	20000	Output sampling rate, in samples/sec
4.	NF	1	5	6	Number of formants in cascade branch
5.	SS	1	2	3	Source switch (1=impulse, 2=natural, 3=LF model)
6.	RS	1	8	8191	Random seed (initial value of random number generator)
7.	SB	0	1	1	Same noise burst, reset RS if AF=0 and AH=0 (0=no,1=yes)
8.	CP	0	0	1	0 implies Cascade, 1 implies parallel tract excitation by AV
9.	OS	0	0	20	Output selector (0=normal,1=voicing source,...)
10.	GV	0	60	80	Overall gain scale factor for AV, in dB
11.	GH	0	60	80	Overall gain scale factor for AH, in dB
12.	GF	0	60	80	Overall gain scale factor for AF, in dB

Table 3.3: *Constant control parameters for the KLSYN88 synthesizer configuration. Each control parameter is assigned a two-letter name, a minimum value, a default value that applies if the user makes no changes, a maximum value, and an English description of its effect on the synthesis.*

of an audible click associated with sudden termination of a non-zero waveform.

The buffer can be made as large as 5 seconds (5000 msec) to accommodate relatively long sentences. Longer utterances are possible (the Vax is a virtual machine with essentially unlimited array sizes) but for visualization of the data given the unsophisticated plotting routines provided with the KLSYN88 software package, it is better to break up long utterances into segments, synthesize each, and then concatenate the resulting waveforms into a single long waveform using the program CONCAT. For example, to combine three waveform files into a fourth called 'outname.wav', one would type the command:

$ concat name1 name2 name3 outname⇓

2. UI: Update Interval

The constant **UI**, "update interval", is the number of msec of waveform generated each time parameters 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 to mimic very rapid formant transitions at plosive release.

Parameters involved in generating the voicing source waveform (**F0**, **AV**, **OQ**, **TL**, **DI**, **FL**) 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.

	SYM	MIN	VAL	MAX	DESCRIPTION
13.	F0	0	1000	5000	Fundamental frequency, in tenths of a Hz
14.	AV	0	60	80	Amplitude of voicing, in dB
15.	OQ	10	50	99	Open quotient (voicing open-time/period), in %
16.	SQ	100	200	500	Speed quotient (rise/fall time of open period, LF model), in %
17.	TL	0	0	41	Extra tilt of voicing spectrum, dB down @ 3 kHz
18.	FL	0	0	100	Flutter (random fluct in f0), in % of maximum
19.	DI	0	0	100	Diplophonia (pairs of periods migrate together), in % of max
20.	AH	0	0	80	Amplitude of aspiration, in dB
21.	AF	0	0	80	Amplitude of frication, in dB
22.	F1	180	500	1300	Frequency of the 1st formant, in Hz
23.	B1	30	60	1000	Bandwidth of the 1st formant, in Hz
24.	DF1	0	0	100	Change in F1 during open portion of a period, in Hz
25.	DB1	0	0	400	Change in B1 during open portion of a period, in Hz
26.	F2	550	1500	3000	Frequency of the 2nd formant, in Hz
27.	B2	40	90	1000	Bandwidth of the 2nd formant, in Hz
28.	F3	1200	2500	4800	Frequency of the 3rd formant, in Hz
29.	B3	60	150	1000	Bandwidth of the 3rd formant, in Hz
30.	F4	2400	3250	4990	Frequency of the 4th formant, in Hz
31.	B4	100	200	1000	Bandwidth of the 4th formant, in Hz
32.	F5	3000	3700	4990	Frequency of the 5th formant, in Hz
33.	B5	100	200	1500	Bandwidth of the 5th formant, in Hz
34.	F6	3000	4990	4990	Frequency of the 6th formant, in Hz (frication or if NF=6)
35.	B6	100	500	4000	Bandwidth of the 6th formant in Hz (only applies if NF=6)
36.	FNP	180	280	500	Frequency of the nasal pole, in Hz
37.	BNP	40	90	1000	Bandwidth of the nasal pole, in Hz
38.	FNZ	180	280	800	Frequency of the nasal zero, in Hz
39.	BNZ	40	90	1000	Bandwidth of the nasal zero, in Hz
40.	FTP	300	2150	3000	Frequency of the tracheal pole, in Hz
41.	BTP	40	180	1000	Bandwidth of the tracheal pole, in Hz
42.	FTZ	300	2150	3000	Frequency of the tracheal zero, in Hz
43.	BTZ	40	180	2000	Bandwidth of the tracheal zero, in Hz
44.	A2F	0	0	80	Amplitude of frication-excited parallel 2nd formant, in dB
45.	A3F	0	0	80	Amplitude of frication-excited parallel 3rd formant, in dB
46.	A4F	0	0	80	Amplitude of frication-excited parallel 4th formant, in dB
47.	A5F	0	0	80	Amplitude of frication-excited parallel 5th formant, in dB
48.	A6F	0	0	80	Amplitude of frication-excited parallel 6th formant, in dB
49.	AB	0	0	80	Amplitude of frication-excited parallel bypass path, in dB
50.	B2F	40	250	1000	Bandwidth of frication-excited parallel 2nd formant, in Hz
51.	B3F	60	320	1000	Bandwidth of frication-excited parallel 3rd formant, in Hz
52.	B4F	100	350	1000	Bandwidth of frication-excited parallel 4th formant, in Hz
53.	B5F	100	500	1500	Bandwidth of frication-excited parallel 5th formant, in Hz
54.	B6F	100	1500	4000	Bandwidth of frication-excited parallel 6th formant, in Hz
55.	ANV	0	0	80	Amplitude of voicing-excited parallel nasal formant, in dB
56.	A1V	0	60	80	Amplitude of voicing-excited parallel 1st formant, in dB
57.	A2V	0	60	80	Amplitude of voicing-excited parallel 2nd formant, in dB
58.	A3V	0	60	80	Amplitude of voicing-excited parallel 3rd formant, in dB
59.	A4V	0	60	80	Amplitude of voicing-excited parallel 4th formant, in dB
60.	ATV	0	0	80	Amplitude of voicing-excited parallel tracheal formant, in dB

Table 3.4: *Control parameters that can be varied over time in the KLSYN88 synthesizer configuration. Each control parameter is assigned a two- or three-symbol name, a minimum value, a default value that applies if the user makes no changes, a maximum value, and an English description of its effect on the synthesis.*

If this were not done, it would be as if spurious excitation occurred at the update rate, resulting in perceptible auditory distortion and perhaps a pitch sensation at the update rate (200 Hz if UI=5 ms). [8] Delaying changes to the voicing source control parameters in order to synchronize them with the time of glottal opening both removes the update interval periodicity of the distortions, and better hides any artifacts associated with parameter changes under the signal because the waveform has been zero prior to the parameter change.

3. SR: Sampling Rate

The constant **SR**, "sampling rate", is the number of output samples computed per second of synthetic speech. The default value of 10,000 samples/sec produces an output waveform in which frequency components up to 5000 Hz are generated. It is suggested that the default value 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, as shown in Figure 3.25B. It would be necessary to add a formant to the cascade branch, Figure 3.25C, in order to compensate and restore the desired spectral tilt.

The low-pass filter at the D/A converter output must be set to about 4800 Hz when using the default 10,000 samples/sec setting of **SR**. (Signal components at 5000 Hz are attenuated by 5 dB, while alias components above 6 kHz are attenuated by more than 60 dB by the high-performance filter that is used in a current version of the system.)

If a higher sampling rate of e.g. 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**). The filter switch should be changed to select the fixed 7500 Hz low-pass filter if a sampling rate of 16000 samples/sec is employed. If **SR** is changed to any new value other than 16000 samples/sec, a suitable fixed low-pass filter is not available at the patch panel in the current version, and it may be necessary to reconnect phono-connector cables at the back of the patch panel so as to replace the two fixed filters with a variable low-pass filter.

4. NF: Number of Cascaded Formants

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

[8] The fact that formant frequencies and bandwidths change at the update time means that small waveform distortions synchronized to the update rate are likely to be produced in spite of the care we have taken to avoid their generation in the voicing source calculations. Experience shows that these do not create any perceptual problems.

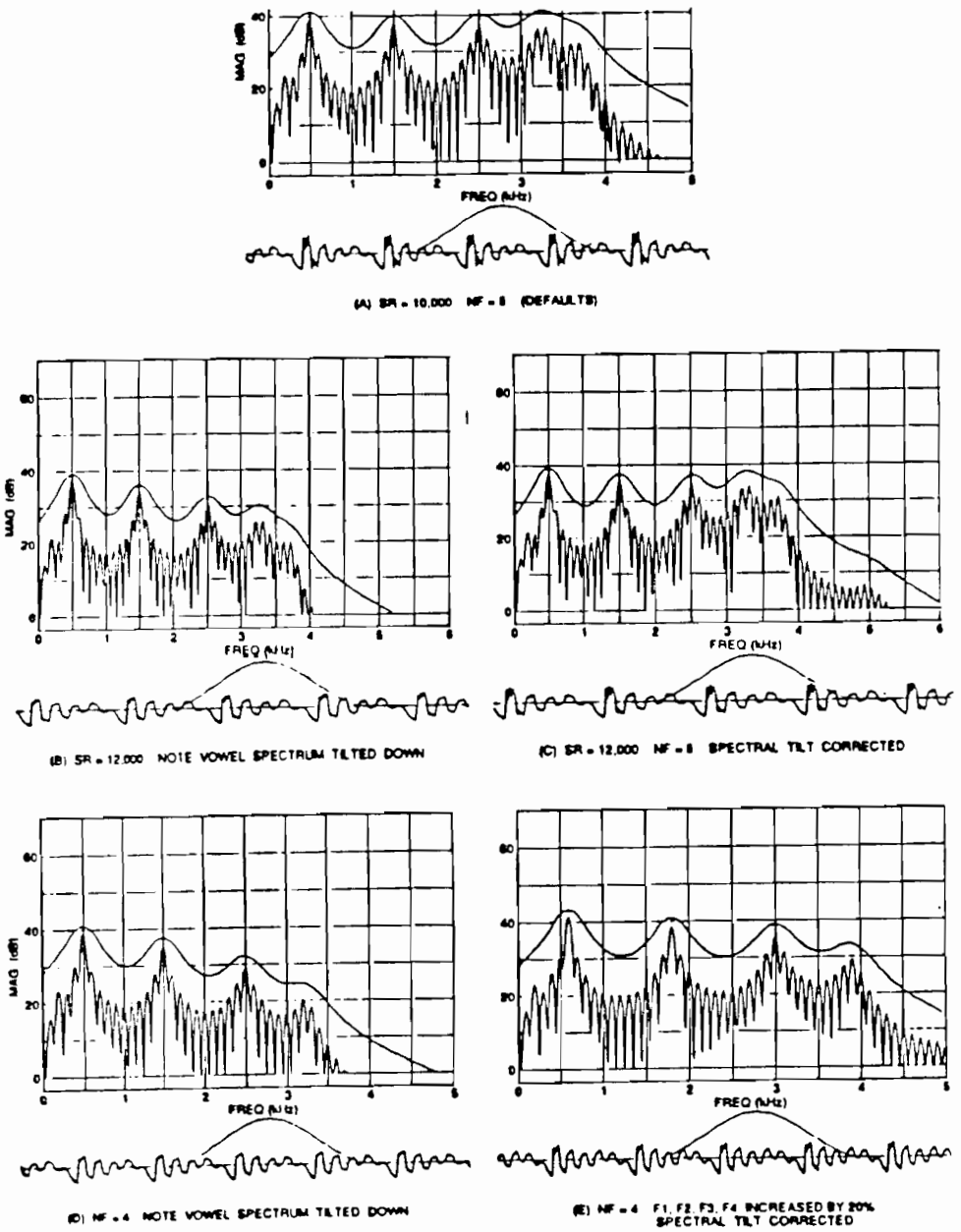


Figure 3.25: *Spectrum of first difference of the waveform of a synthetic vowel (a) using default values for all control parameters, (b) increasing the sampling rate* SR *to 12,000 samples per second, (c) setting* SR=*12,000 and increasing the number of formants in the cascade vocal tract model to* (NF=*6), (d) reducing* NF *to 4 formants in the cascade vocal tract model, and (e) reducing* NF *to 4 while simultaneously raising the frequencies of all formants by 20% to approximate a female with a smaller head.*

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, as illustrated in the lower panels of Figure 3.25.

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 frication-excited 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 if **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.

5. SS: Source Switch

The constant **SS**, "source switch", is a switch that determines which of three voicing source waveforms is used for synthesis. A value of 1 causes a low-pass filtered impulse train to be generated, while the (default) value 2 causes a new more natural waveform with a definite sharp closing time to be generated. Each has its own set of advantages and disadvantages, as discussed below. Finally, a modified version of the Liljencrants-Fant (LF) model (Fant, Lin and Gobl, 1985) is activated if a value of 3 is selected. This model is similar in many ways to the **SS** = 2 natural waveform, but allows the use of one additional control parameter **SQ**, speed quotient, to modify waveform symmetry.

Impulse Train, $SS = 1$. 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 3.26A. The spectrum falls off monotonically at -12 dB per octave for low and mid frequencies and then flattens out.

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 "center frequency" of zero Hz, and a bandwidth (which controls the width of the open portion of the glottal pulse) that is set to be proportional to the synthesis parameter open quotient, **OQ**. The spectrum of this source can also be tilted down to simulate a mode of vibration where the vocal folds do not meet at the midline, using the **TL** 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, [9] even though the source magnitude spectrum is probably to be

[9]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.

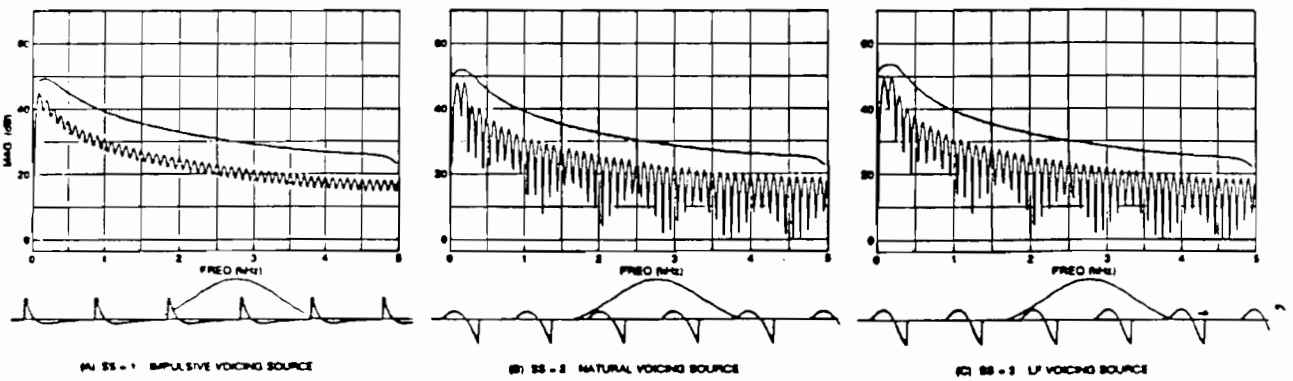


Figure 3.26: *Comparison of $U_g'(t)$ (obtained by setting* **OS**=*4) waveforms and spectra of (A) the impulsive source (*SS=*1), (B) the default natural glottal source (*SS=*2), and (C) the LF model (*SS=*3), using default settings to all other parameters.*

preferred for its regularity, at least in some psychophysical experiments.

KLGLOTT88 Pulse Train, $SS = 2$. The new KLSYN88 model has a voicing volume velocity waveform that obeys the equation:

$$U_g(t) = at^2 - bt^3$$

during the open phase, and is zero for the remainder of the period. The spectrum of this KLGLOTT88 source is somewhat irregular, with a weak zero at about 600 Hz (assuming default settings to all of the glottal source parameters.) Waveforms and spectra for the impulsive and KLGLOTT88 voicing sources are compared in Figures 3.26A and 3.26B. The KLGLOTT88 glottal waveform can also be modified so as to increase/decrease the first harmonic amplitude, using **OQ**, or to tilt the spectrum down, using **TL**, in order to mimic the effects of incomplete glottal closure and the concomitant rounding of the corner of the voicing source waveform at closure.

The disadvantage of the KLGLOTT88 source waveform is that the magnitude spectrum is somewhat irregular, so that a formant will be slightly attenuated as it approaches a frequency of any glottal zeros (the actual zero locations depend on **OQ**, the number of samples in the open phase). This formant amplitude variation seems to occur in natural speech, but may not be desirable for synthesis of particular stimulus sets.

Liljencrants-Fant LF Model, $SS = 3$. We have also included as an option (SS=3) the use of a slightly modified version of the Liljencrants-Fant (LF) model (Fant *et al.*, 1985). The model was originally formulated in terms of a set of times of waveform events, but it can easily be recast in terms of familiar parameters **AV** (amplitude of voicing), **F0** (fundamental frequency), **OQ** (open quotient), **SQ** (speed quotient), and **TL** (spectral tilt). The waveform of the derivative of $U_g(t)$ during the open phase is characterized as a portion of a single cycle of an underdamped sinusoid:

$$U'_g(t) = E_0\ e^{\alpha t} sin(2\pi F_g t) \qquad 0 < t < (OQ * T0)$$

$$U'_g(t) = 0 \qquad (OQ * T0) < t < T0$$

where this equation expresses the output before computing the effect of any spectral tilt associated with non-simultaneous closure. The equation can be simulated by the impulse response of an underdamped resonator. As shown in Figure 3.26C, the default pulse shape and spectrum of the LF model is very similar to that of the natural source output (**SS**=2). The effects of the control parameters **OQ** and **TL** on the waveform and spectrum (Figure 3.7), are essentially the same as for the KLGLOTT88 source, but the additional speed quotient **SQ** control parameter permits greater flexibility in the control of the depth of source spectral zeros, as shown in Figure 3.7C.

6. RS: Random Seed

The constant **RS**, "random seed", is the initial seed value given to the random number generator routine. Any number from 0 to 8191 can be specified. For each consecutive integer, you will get a quite different random number sequence (statistically different frication and aspiration noises from those used to generate the previous stimuli). For example, to generate a series of 9 stimuli, each with a different burst spectrum, one might set **RS** to 1,2,...,9.

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, for example, a burst of frication noise.

In a short burst of noise, the spectrum varies depending on the random nature of the numbers drawn in sequence. Usually such a spectrum has a lack of energy at certain frequencies, and it is difficult to create a spectral peak in such a frequency region. Therefore, it is sometimes desirable to select a sample of noise that, in the short run, happens to have a fairly uniform spectrum representative of the long-term average spectrum of the noise generation process. Examination of spectra associated with a number of random seeds revealed two seed values that generate quite good spectra for the synthesis of 5-ms or 10-ms bursts. (See Figure 3.27.) To obtain these frication source spectra for a burst, the **SB** parameter (described next) must remain at its default value of 1, see below; otherwise the noise source is free-running and the burst spectrum that is obtained will vary with the temporal location of the burst.

7. SB: Same Burst

The constant **SB**, "same burst", if set to 1, causes the random number generator to be reset to the initial seed value described above whenever **AF** and **AH** are both zero (as is typically the case during closure before a plosive release). The net effect is to get exactly the same

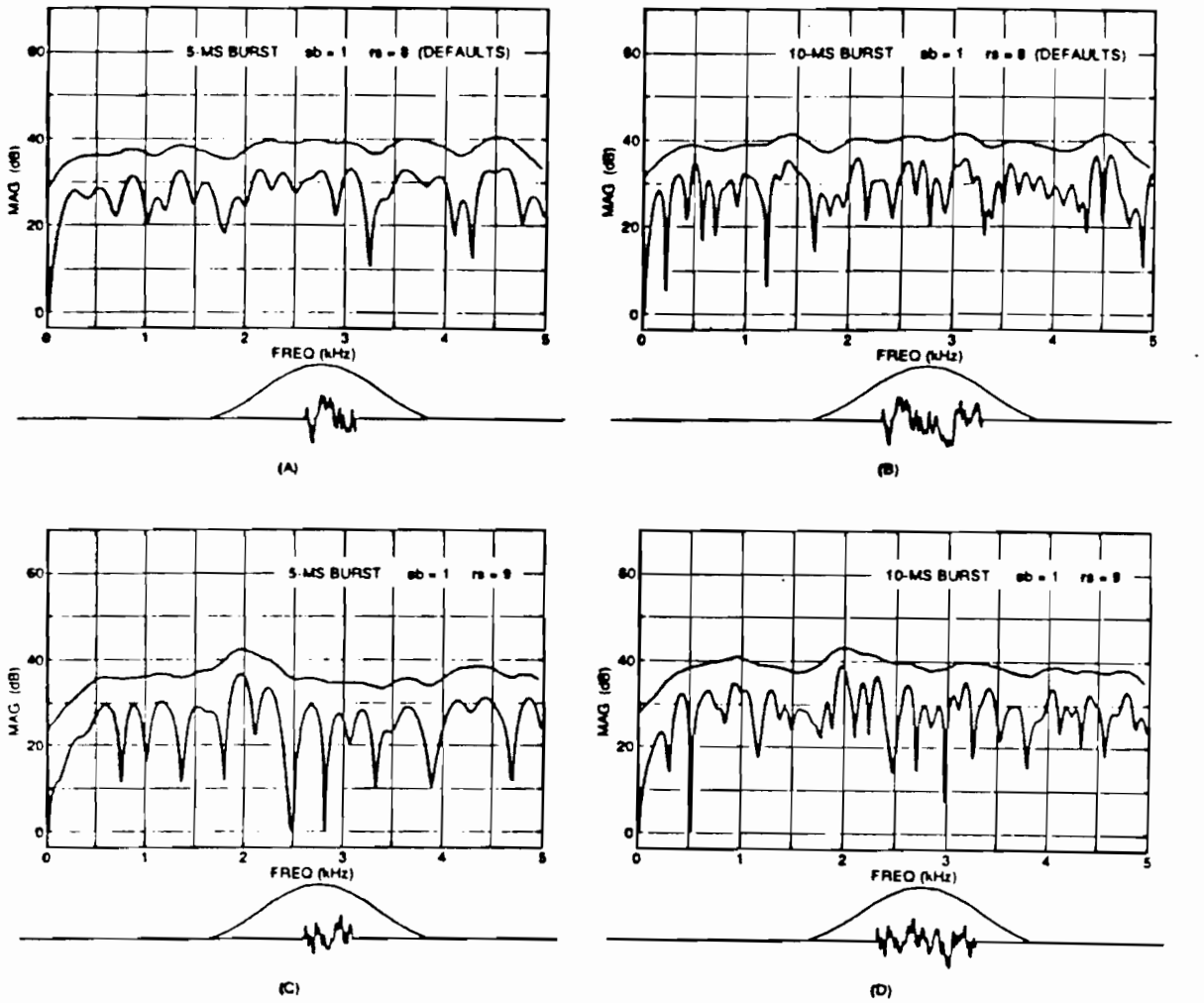


Figure 3.27: *Spectra of several brief noise bursts produced by the frication source with* SB *"same burst" set to 1 and the* RS *"random seed" set to two carefully selected values. These spectra are quite close to the average spectrum produced by the statistical process, making* RS*=8 or 9 a good choice for many purposes.*

noise sequence at the onset of each plosive burst and frication noise in an utterance if SB=1. With a good choice of seed value, a short burst with a reasonably flat noise source spectrum is obtained every time, as seen in Figure 3.27. Of course the random fluctuations that occur from token to token in natural speech are missing in such a strategy, but intelligibility is under greater control. The strategy is also consistent with the idea that part of the burst release is a "coherent" step response that is in fact the same for each repetition of a plosive (Maeda, 1987; Stevens, in preparation).

8. CP: Cascade/Parallel Switch

The constant CP, "cascade/parallel switch," determines whether vowels are generated by the cascade branch of the synthesizer (CP=0), or whether the synthesizer is reconfigured

so that formants for vowel synthesis are in parallel (**CP**=1). The default is cascade vowel synthesis.

When it is desired to generate vowels by parallel synthesis, the system is configured so that **A1V**, **A2V**, **A3V**, **A4V**, **ANV**, **ATV** control the amplitudes of the four lowest formants and the amplitude of a nasal formant and a tracheal formant that can be excited under special circumstances. A strategy for synthesizing a vowel using the parallel model is described in the section below that discusses these amplitude controls.

9. OS: Output Switch

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 3.5. For example, if you wish 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 set to four or greater. [10]

10. GV, GH, GF: Source Gains

Overall constant gain controls for each sound source, **GV**, "gain of voicing," **GH**, "gain of aspiration," and **GF**, "gain of frication," [11] are included to permit the user to adjust the output level without having to modify each source amplitude time function. This is especially useful when the final output level of a waveform is significantly greater or less than the optimum of -2 to -6 dB. The nominal default value for the **GV**, **GH** and **GF** control constants is 60 dB. To increase the output by e.g. 3 dB, one would simply use the 'c' command to set **GV GH GF** to 63.

Three gain controls are provided to permit easy adjustment of the relative levels of voicing, aspiration and frication in experiments where these might be the variables of interest across a set of stimuli.

13. F0: Fundamental Frequency

The variable **F0**, "fundamental frequency", is the rate at which the vocal folds are currently vibrating in Hz times 10. For example, if a fundamental frequency of 100 Hz is desired, then **F0** is set to 1000. The additional accuracy resulting from a specification of fundamental

[10] Due to computational simplifications, the radiation characteristic has been folded into the source calculations. In order to plot a source waveform without showing the effect of the radiation characteristic, the source waveform that is displayed for **OS** less than 4 is modified by sending it through a (slightly leaky, k=0.99) integrator.

[11] These three gain constants replace the old control parameter 'g0'.

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

Table 3.5: *KLSYN output waveform options using* **OS**.

frequency to 0.1 Hz adds some naturalness to a slowly changing pitch glide because listeners are very sensitive to small pitch changes.

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**, **OQ**, **TL**, **DI**, **FL**) change value at this time rather than changing at the nominal update time – otherwise discontinuities could occur in the voicing waveform.

If it is desired to initiate a glottal pulse at a specified time, the procedure to follow is to set **F0** to zero over the previous update interval. The instant that **F0** becomes non-zero will cause a glottal pulse to be issued of amplitude **AV**. Note, however, that the first glottal event in a period is the open gesture, so that primary excitation of the vocal tract using the natural source will not occur until glottal closure, a few milliseconds after the specified time (depending on the value of **F0** and **OQ**).

14. AV: Amplitude of Voicing

The variable **AV**, "amplitude of voicing" is the amplitude in dB of the voicing source waveform. 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. In this way, there are no waveform discontinuities due to a sudden change in **AV** in the middle of the open portion of a period.

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, normally taking about 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 at least one update interval prior to this event, and to turn **F0** on simultaneous with the time that **AV** is turned on (see discussion in F0 section above). 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. The result is indicated by the sample parameter values and output synthetic waveform in Figure 3.28.

Figure 3.28 shows waveforms of the glottal source and the output when the parameters are adjusted to give a voice onset time of 40 ms. For the impulsive source (SS=1), the voice onset time is exactly 40 ms, whereas with the KLGLOTT88 source (**SS**=2) the effective VOT is longer since the primary glottal excitation occurs after the glottal waveform as initiated.

15. OQ: Open Quotient

The spectrum of a voicing source pulse train can be made to 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 3.22B and 3.22C).

The variable **OQ**, "open quotient", is the ratio of the open time of a glottal cycle to the total duration of the period, in percent. The open quotient is a nominal indicator of the width of the glottal pulse when using the default impulse train glottal source, and it determines the exact duration of the open period when using the natural voicing source (SS=2) and the LF model (SS=3). A value of **OQ**=50%, the default value, results in a 5 msec open portion at a 100-Hz fundamental frequency (10-msec period).

There are many speakers for whom the open quotient does not change over an utterance, even if fundamental frequency changes over a fairly wide range. Thus, it is not necessary to change **OQ** during synthesis when generating most simple short speech stimuli. Of course, if there is a laryngealized vowel onset, open quotient should be reduced to, for example, about 30 at onset, and a smooth transition to a value of about 50 might occur over the first 50 ms of voicing. The open quotient **OQ** should be increased to 70 or 80 in a breathy voice or a voicing interval at the margins of a vowel adjacent to a voiceless consonant. Transitions involving increased open quotient typically take about 50 to 70 ms

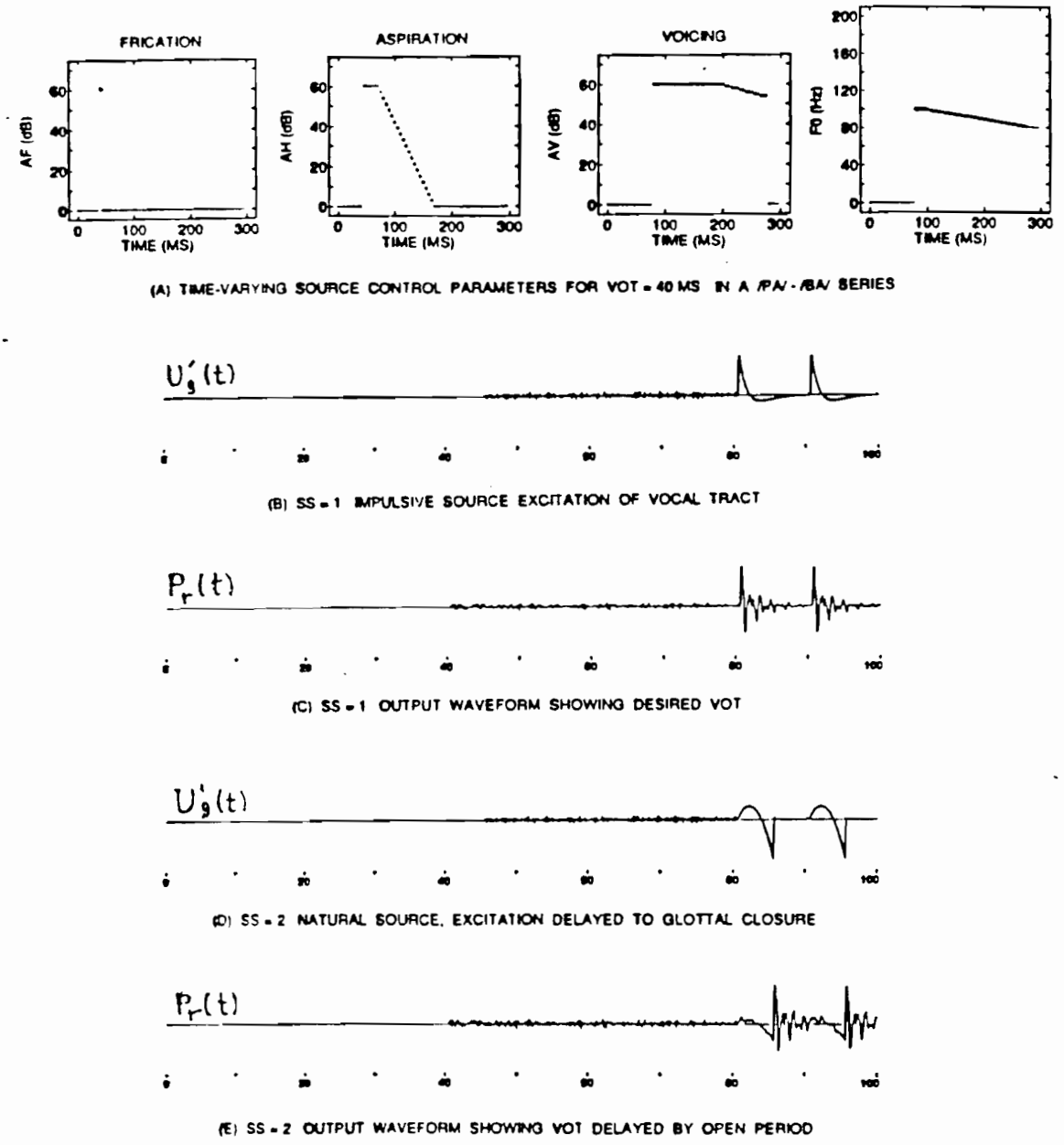


Figure 3.28: *Control parameter values intended to produce a plosive burst at t=40 ms followed by an interval of aspiration and then voicing onset at t=80 ms are shown in part (A). Due to the algorithm for generation of a glottal pulse, the desired voice onset time is realized only if* SS=1 *(see parts (B)and (C)). When using the KLGLOTT88 glottal source, the effective VOT is longer by a duration equal to the open period (about 5 ms in this example) because the glottal closure event that provides primary excitation of the vocal tract occurs later in the period, as shown in parts (D) and (E).*

or more.

The effects of changes in **OQ** on the spectrum were illustrated in Figure 3.22B.[12] A narrow glottal pulse, as may occur in creaky voice, or when trying to speak loudly, results in a spectrum with a relatively weak fundamental component, but the source spectrum remains essentially unchanged in mid and high-frequency regions. A wider glottal pulse, as may occur in a breathy vowel, results in a spectrum rich in energy below the first formant. As can be seen in Figure 3.22B, the change in relative amplitude of the first harmonic is more than 12 dB for physiologically plausible variations to the open quotient parameter. Thus to match an observed strong first harmonic in the spectrum of a natural utterance, one would increase **OQ**.

16. SQ: Speed Quotient

The variable **SQ**, "speed quotient of the LF voicing source", determines the ratio of the duration of opening of the glottis to the duration of closing. In a normal glottal pulse, this ratio is about 2.0, or 200%, the default. It does not appear that **SQ** varies over time for a given speaker, but there may be differences between speakers, with females tending to have slightly smaller speed quotients. The effect on the spectrum of reducing the speed quotient is to tilt the spectrum down at high frequencies and to introduce more pronounced spectral dips (recall Figure 3.7C). Thus if a voice appears to have source irregularities, one method of attempting to match the spectrum would be to decrease the speed quotient. Speed quotient has no effect unless one uses the LF model (**SS**=3).

17. TL: Spectral tilt

The variable **TL**, "spectral tilt of the voicing source", is the (additional) downward tilt of the spectrum of the voicing source, in dB, as realized by a soft two-pole low-pass filter. The spectral effects of changes in **TL** on all source models are illustrated in Figure 3.22C. A value of zero has no effect on the source spectrum, while a value of 20 tilts the spectrum down gradually such that frequency components at 3 kHz are attenuated by 20 dB relative to a more normal source spectrum. A 20 to 30 dB setting to the tilt parameter has the effect of tilting the spectrum down an additional -6 dB per octave, which is about the increase in tilt seen for many breathy vowels.

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 associated with breathy phonation, where the anterior portion of the vocal folds meet at the midline before the posterior portions come together. Spectral tilt, **TL**, is a good parameter to use in attempts at matching the spectral details of a particular natural utterance. It seems to be the case that voicing spectra requiring significant tilt are "breathy" and also contain

[12]This plot does not apply to the impulsive source. The spectral effects of changes in open quotient for the impulsive source are more similar to a change in spectral tilt, Figure 3.22C.

significant amounts of aspiration noise at mid and high frequencies (Klatt and Klatt, 1990). (See discussion under control parameter **AH** below.)

The tilt parameter is also useful in simulating a voicebar, wherein only lower-frequency components are radiated from the closed vocal tract. One would set **F1**=180 Hz (the closed-cavity lowest natural frequency of the vocal tract) and **TL**=40 to obtain the appropriate spectral shape for a voicebar, and adjust **AV** to get the correct level as a function of time. For speech synthesis situations employing stylized speech (as opposed to copying a natural model in detail), **TL** is normally not varied except perhaps at margins with voiceless consonants.

18. FL: Flutter

The variable **FL**, "flutter," introduces a quasi-random fluctuation in fundamental frequency that can be introduced to prevent synthesis of a waveform with perfectly constant pitch. It simulates the natural tendency of humans to drift or waver slightly when attempting to produce a constant pitch. Natural flutter is approximated in the synthesizer by adding together three slowly varying sine waves of amplitudes specified by the flutter **FL** parameter. The amount of flutter is calibrated in percent of a maximum. A value of 15 or 25% approximates normal variation. The default is set to zero in order to permit the user to obtain exactly the fundamental frequency value specified by **F0** at all times, but some flutter will make sentences and long sustained vowels sound more natural.

19. DI: Diplophonic Double Pulsing

The variable **DI**, "diplophonic double pulsing", is the degree (in percent) to which the first of every pair of glottal pulses migrates toward the second and is attenuated in amplitude (recall Figure 3.8). Normally, **DI** is set to zero, and the waveform is perfectly periodic, but there are voices that become diplophonic under some conditions and to varying degrees. A value of 50% would cause the first member of each pulse pair to be attenuated by 50% and to move half-way toward the second pulse. A value of 100% would attenuate the first pulse to zero, effectively halving the fundamental frequency.

Such aperiodicities, when introduced, have fairly strong perceptual consequences. This kind of modification of normal voicing occurs throughout speech for a few voices, and at the initiation and especially at cessation of voicing in a sentence for many others. However, there is no need to utilize this parameter in most synthesis situations.

20. AH: Amplitude of Aspiration

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, to constitute the glottal sound source that is sent to the vocal tract. 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, as shown in Figure 3.22D. The variable **AH** should be turned on and off gradually for an initial [h] (0 to 60 dB in about 90 ms), and more abruptly following the burst of an aspirated [p,t,k], as has been illustrated in Figure 3.28.

The spectrum of the aspiration noise source is nearly flat above 1 kHz (after including the radiation characteristic). 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. It may also be necessary to add one or more pole zero pairs to simulate tracheal resonances, using the tracheal pole-zero pair control parameters **FTP**, **BTP**, **FTZ**, **BTZ** (and perhaps utilize the nasal pole-zero pair as well if more than one extra resonance is to be mimicked).

Aspiration noise simultaneous with voicing is quite common for speech intervals in which the voice is breathy. If voicing and aspiration are both present, then the amplitude of the aspiration noise waveform is modulated by a 50% duty-cycle square-wave amplitude modulation that increases the noise during the most open part of the cycle. A value of **AH** of 50 dB or more will make the voice quite breathy. To achieve a good match to natural breathiness, however, one should probably also tilt down the source spectrum 20 to 30 dB, using **TL**, increase the open phase of a glottal cycle **OQ** to 60 to 70%, and increase **B1**. The effects of these changes on the source spectrum were summarized in Figure 3.22.

21. AF: Amplitude of Frication

The variable **AF**, "amplitude of frication", determines the level of frication noise sent to the various parallel formant resonators and to the bypass path. After folding in the effect of the radiation characteristic, the frication source spectrum in the synthesizer is essentially flat.

The variable **AF** should be turned on gradually for fricatives (0 to 60 dB in about 90 msec), and abruptly to about 60 dB for plosive bursts. A value of 60 will excite a parallel formant resonator with about the same spectral level as in the voicing source at that frequency. To match particular frication and burst spectra, see the discussion of parallel formant amplitudes **A2F**—**A6F** below.

The **AF** control parameter is logically redundant in that one could make the same dynamic changes to each of the amplitude controls of the frication-excited parallel formant vocal tract model instead. However, it seems easier and more straightforward to make all dynamic changes to **AF**, and to set the parallel formant amplitudes to constants that can then be adjusted on a trial-and-error basis until the spectrum of the noise is as desired.

22. F1, F2, F3, F4, F5, F6: Formant Frequencies

The variables **F1**, **F2**, **F3**, **F4**, **F5**, **F6**, "formant frequencies," determine the frequencies in Hz of up to six resonators of the cascade vocal tract model, and the frequency in Hz of each

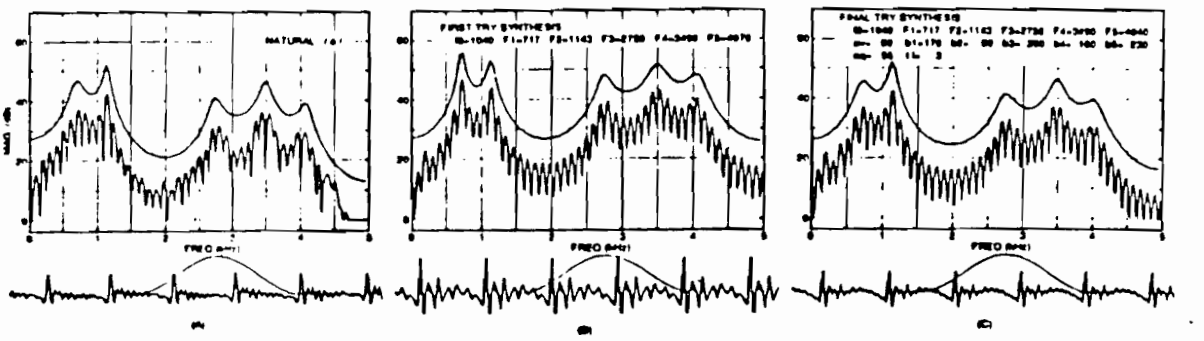


Figure 3.29: *Spectra of (a) a natural [ɑ] vowel, (b) synthesis in cascade mode (CP=0) with formant frequency values obtained from linear-prediction analysis of the natural model, (c) synthesis after trial-and-error adjustment of formant frequencies, bandwidths, and glottal source parameters. See text.*

of five additional wide-bandwidth parallel formant resonators. Normally, the cascade branch of NF=5 formants is used to generate voiced and aspirated sounds, while a set of wider-bandwidth resonators configured in parallel is 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.

An example of the process of adjusting the formant parameters to match a synthesized vowel to a naturally produced vowel is shown in Figure 3.29. A spectrum of natural speech, from the vowel [ɑ], is to be synthesized. In this case, the natural vowel is essentially harmonic throughout the spectrum, and conforms to the ideal theory, with no pronounced extra poles, zeros, or harmonic irregularities. The KLSPEC linear prediction analysis shown in part (A) of the figure resulted in estimated values for fundamental frequency **F0** and formant frequencies that were used in a first attempt to synthesize the vowel. These values, and the result of this first try, are shown in part (B) of the figure. Linear prediction did remarkably well at estimating most formant frequency values, but default formant bandwidths are not right, and the source spectral tilt may have to be adjusted.

Next, trial-and-error tuning of formant frequencies and bandwidths was performed. The criterion for detailed adjustment of a formant frequency is usually based on the relative amplitude of the strongest 2 or 3 harmonics in the formant complex. The formant frequency is adjusted in small steps until the harmonics have the proper relative amplitudes. In some cases, it may be desirable to consult temporally adjacent spectra so as to ensure that formant frequency estimates are continuously and slowly changing in time rather than jumping around semi-randomly from frame to frame. Formant bandwidths are adjusted on a trial-and-error basis in order to get relative formant amplitudes within about 1 to 2 dB of the natural model, and to get the depth of the valleys between formants about right. In order to satisfy both criteria, it is usually necessary to simultaneously adjust the source spectral tilt parameter **TL** and the amplitude of voicing **AV**. It is often also necessary to modify the voicing source open quotient **OQ** to adjust the relative amplitude of the first harmonic. Only a small

change from the default was needed here, however. The final result of these fine-tuning efforts, shown in part (C) of the figure, is a very good match to the original spectrum. Even the waveform of the synthesis shows a remarkable resemblance to the natural waveform. Not all natural speech spectra can be matched with this degree of fidelity, but as long as the amplitudes of the formant peaks are matched within 1 or 2 dB, and the spectral valleys are of generally similar depth to the original, a remarkably good perceptual result is usually obtained.

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. [13] 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 as high as 400 Hz at the time of the first visible glottal pulse following the burst in a syllable such as [ba]. Based on a suggestion by Fujisaki and Azami (1971), special precautions have been taken to adjust the resonator memory variables whenever a first formant frequency jump of this magnitude occurs so as to avoid audible clicks and pops in the synthesis when F1 is suddenly changed. Therefore, the user need not worry about making sudden changes to F1 as required in a plosive transition.

23. B1, B2, B3, B4, B5, B6, Formant Bandwidths

The variables **B1**, **B2**, **B3**, **B4**, **B5**, **B6**, "formant bandwidths", determine the bandwidths of resonators in the cascade vocal-tract model. 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 has no effect on the synthetic waveform.

Formant bandwidths depend in part on the source impedance. Since turbulence sources contribute more losses, the synthesizer provides separate control of bandwidths **B2F**, **B3F**, **B4F**, **B5F**, **B6F** for the frication-excited parallel formants, see below.

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 the change in peak height (change in relative intensity) is much more audible than the width change. Changing peak heights by bandwidth adjustment can also be thought of as changing the depth of the valleys between formant peaks; peak-to-valley ratios are a useful secondary spectral cue to the correct value for a formant bandwidth.

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

[13]The first formant frequency does not normally go below about 180 Hz due to the mass of the cavity walls and the compliance of the air trapped in the closed vocal tract.

a speaker, as demonstrated by the example shown in Figure 3.29. The following guidelines are offered to help avoid the perceptual problems of aberrant bandwidth specification:

- If a bandwidth is set to a value less than the soft limits given in Table 3.4, there is a danger that whistle-like harmonics may be heard when a harmonic of the fundamental sweeps past the formant frequency.
- If the bandwidths of the lower formants approach the suggested maxima of Table 3.4, the synthetic voice will begin to sound buzzy. In this case, offending bandwidths should be reduced, and the spectral tilt **TL** parameter should be employed to solve the problem of spectral balance between the amplitudes of the higher and lower formants.

24. DF1, DB1: Delta F1 and Delta B1 during Open Phase of Period

Fant and Ananthapadmanabha (1982) have noted that the first formant frequency and bandwidth are not necessarily constant over the duration of a period. When the glottis is open, the first formant frequency may increase by as much as 10% and glottal losses may increase the first formant bandwidth significantly, especially for low vowels. The perceptual importance of this effect is unclear. In most situations one can pick appropriate average values for **F1** and **B1** that match the spectrum averaged over several periods rather than making **F1** and **B1** vary pitch synchronously. However, the variables **DF1**, "delta F1," the incremental increase in **F1** during the open portion of each period, and **DB1**, "delta b1", the incremental increase in **B1** during the open portion of each period, have been created in order to allow pitch-synchronous changes to **F1** and **B1** if desired.

The change to first formant frequency and bandwidth occurs in "square-wave" fashion, increasing at the instant of glottal opening, and decreasing at the instant of glottal closure, as determined by the open quotient **OQ**. For example, to have **F1**=500 Hz during the closed phase and 550 Hz during the open phase of each period, one would set **F1**=500 and **DF1**=50.

In a low vowel, the time variation in first formant bandwidth might be approximated by setting **B1**=50 and **DB1**=400. An example of a one-formant vowel synthesized with these values for time variation of first formant frequency and bandwidth is shown in Figure 3.30. A perceptually nearly equivalent constant first formant bandwidth (equal spectral level of F1) corresponds to a first formant bandwidth setting of about 90 Hz, as shown in the third panel of the figure.

The default values for the **DF1** and **DB1** incremental parameters are set to zero because most users will not need to resort to this kind of detail during synthesis.

36. FNP, FNZ, BNP, BNZ: Nasal Pole-Zero Pair

The variable **FNP**, "frequency of the nasal pole", in consort with the variable **FNZ**, "frequency of the nasal zero", can mimic the extra pole and zero in nasal murmur spectra, as well

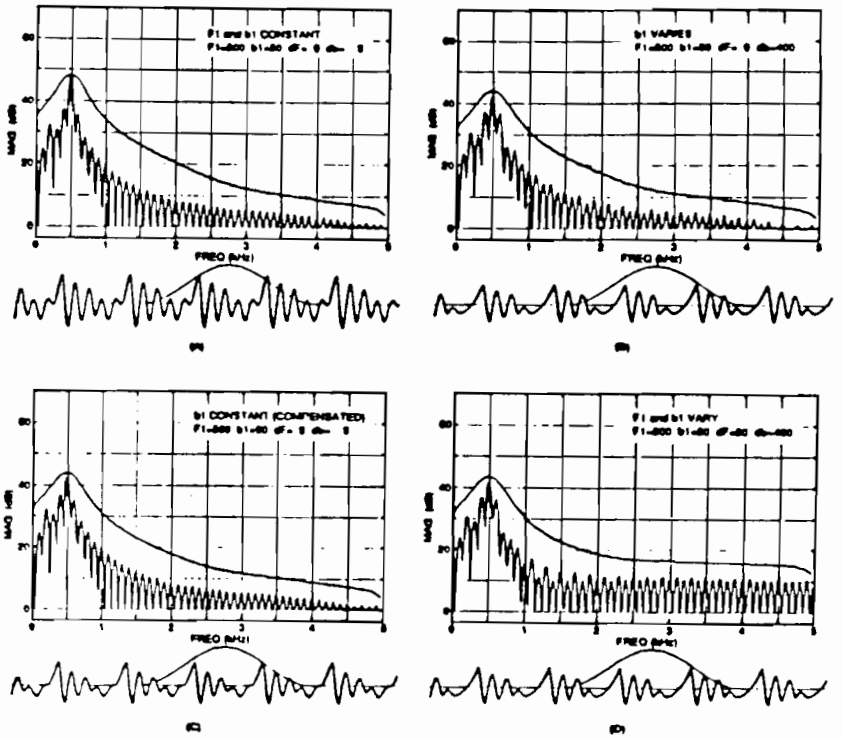


Figure 3.30: *Waveform and spectrum of a one-formant vowel synthesized with (a) constant formant frequency and bandwidth, (b) time-varying bandwidth, (c) an equivalent constant bandwidth, and (d) time-varying formant frequency and bandwidth.*

as the primary spectral effects of nasalization in vowel-like spectra. In a typical nasalized vowel, the first formant region of the spectrum is split into peak-valley-peak (pole-zero-pole). The additional pole-zero pair, **FNP** and **FNZ**, is usually above **F1**, except for vowels having a high **F1**, in which case the additional pair could be below **F1**.[14]

The variables **BNP**, "bandwidth of the nasal pole", and **BNZ**, "bandwidth of the 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.

A stylized version of the /ni/ syllable can be synthesized by the strategy shown in Figure 3.31. In the nasal murmur, the frequency of the nasal zero has been moved up in frequency so as to tilt the overall spectrum down. At the instant of release, the zero moves down quickly to a position between F1 and the nasal pole, simulating a pole-zero-pole configuration characteristic of a partially nasalized vowel. As the velum closes down over the course of the vowel, the strategy is to gradually adjust the frequency of a nasal zero to approach that of the corresponding nasal pole. The nasal pole and nasal zero then cancel each other out (at 500 Hz in this example), and it is as if they were not present in the cascade vocal-tract model, resulting in synthesis of a normal vowel at the end of the syllable.

[14]It does not make any difference to the synthesis strategy or spectral results which peak is assigned to F1 and which is assigned to FNP.

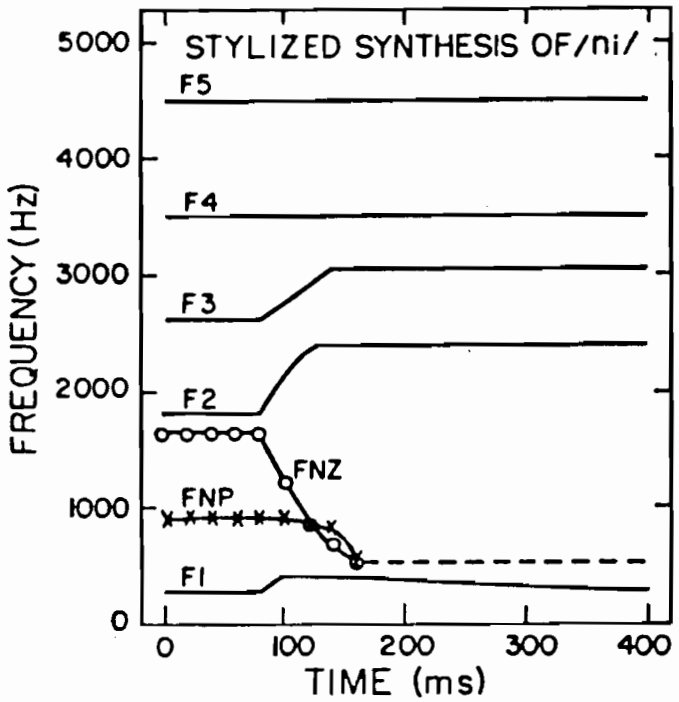


Figure 3.31: *Time course of formant frequency parameters used to produce a stylized synthesis of the syllable /ni/. The frequency of the nasal zero jumps rather abruptly from 1400 to 500 Hz following the tongue-tip release of the /n/, but formant poles must move more continuously even at this event.*

Figure 3.32 illustrates the mechanics of matching a natural nasal-vowel syllable. In a transition from a nasal consonant to a vowel, one can often see evidence of an extra low-frequency nasal pole-zero pair, as well as an additional pole-zero pair that is sometimes seen at higher frequencies. To mimic the "extra" resonance at 710 Hz in the murmur shown in part (b) of the figure, the nasal pole frequency **FNP** is set to 320 Hz during the [n], **F1** is set to the higher 710-Hz frequency of the pole-zero-pole complex, and the zero is located on the basis of evidence of a spectral dip, as well as how this choice influences the general tilt of the spectrum. The zero **FNZ** has been placed at 640 Hz, near **F1** and increased its bandwidth in order to reveal the adjacent pole. [15]

There is a second weak nasal pole at about 1130 Hz in the murmur. Ordinarily, one would not bother to match this detail, but the mechanism is available to do so if one so desires. Simply adjust the tracheal pole and zero frequencies and bandwidths, as shown in part (c) of the figure. As the bandwidth of the zero is increased relative to the bandwidth of the pole, the amplitude of the pole increases; this is a good method for adjusting the height of a nasal pole if the zero frequency cannot be moved because it would change the spectral

[15]While there does seem to be a zero near this harmonic, the overall result of synthesis was too much high frequency energy, which was corrected by using the **TL** parameter. More likely, the correct solution would be to place **FNZ** much higher in frequency in order to tilt the spectrum down without resorting to a somewhat implausible value for this voicing source variable. Of course, this adjustment would result in a synthetic spectrum where the harmonic at 640 Hz would be too strong, but the weakness of that harmonic amplitude might alternatively be due to a source zero that has not been modeled.

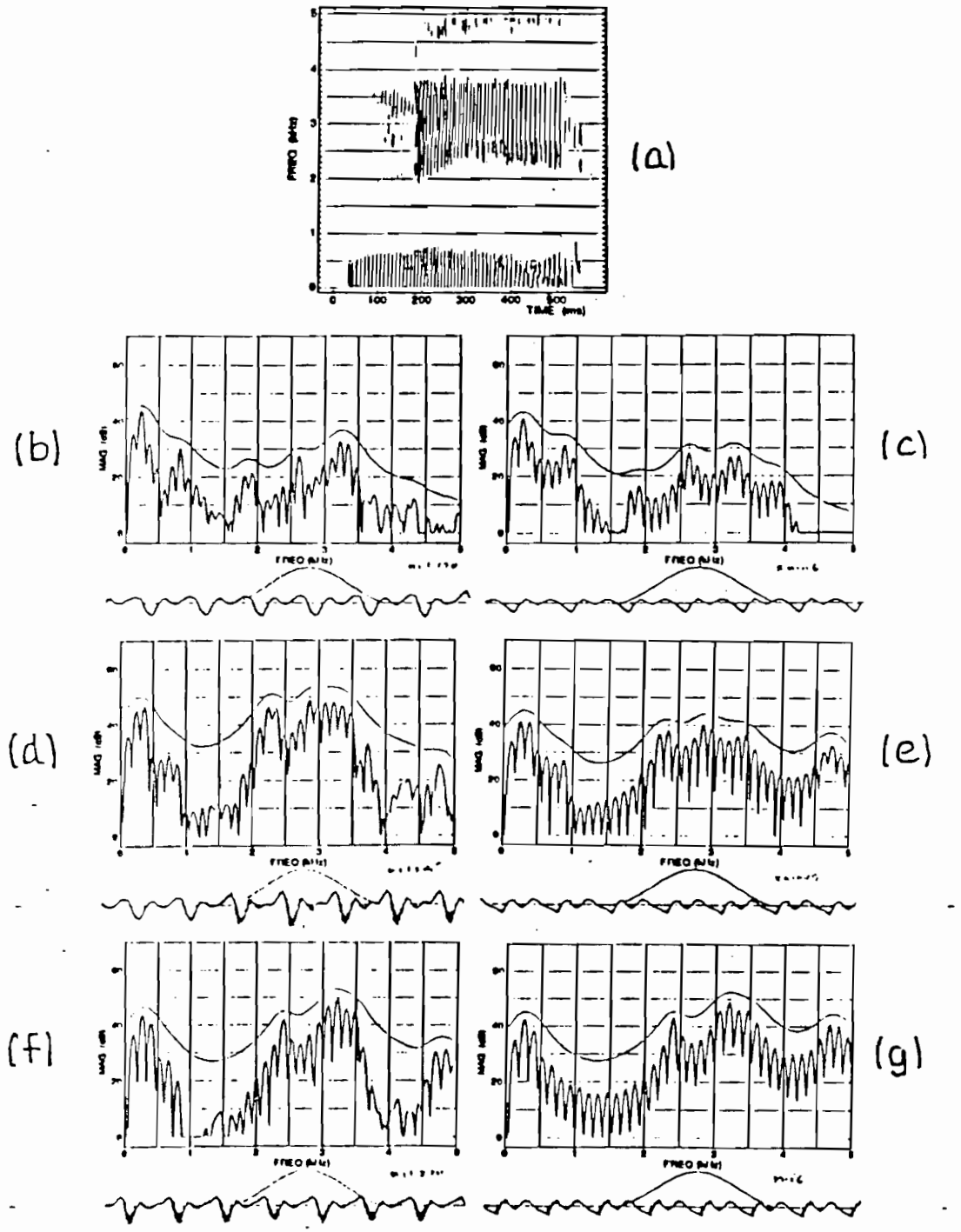


Figure 3.32: *A broadband spectrogram of (a) a natural [ni] syllable has been sampled at three time points to produce spectra of (b) the [n] nasal murmur, (c) a synthetic imitation, (d) nasalization at the onset of the vowel [i], (e) a synthetic imitation, (f) normal voicing near the middle of the [i], and (g) a synthetic imitation. Note the extra nasal poles at about 710 Hz and at 1130 Hz in the [n] murmur and in the nasalized onset of the following vowel.*

balance in the wrong way. Part (d) of Figure 3.32 illustrates how the nasal poles persist into the initial portion of the vowel due to a delay in velar closing with respect to stop release. Synthesis values for a nasalized and non-nasalized /i/ are shown in parts (e) and (g) of the figure.

40. FTP, FTZ, BTP, BTZ: Tracheal Pole-Zero Pair

The variable **FTP**, "frequency of the tracheal pole", in consort with the variable **FTZ**, "frequency of the tracheal zero", can mimic the primary spectral effects of tracheal coupling in breathy vowels. A cascaded pole-zero pair is provided in the cascade branch of the synthesizer to mimic the addition of a "spurious" resonant peak due to this tracheal coupling interaction. Tracheal resonances are often seen in breathy vowels at frequencies of about 550, 1300 and/or 2100 Hz (slightly higher for female voices). The best synthesis strategy is to pick the most prominent one for synthesis (or use the nasal pole-zero pair to simulate a second).

Normally, the spectral dip or zero corresponding to the selected tracheal resonance is close to the pole, and can be above it or below it in frequency. Tracheal coupling usually begins and ends gradually as the glottis is opened or closed. A suggested synthesis strategy, then, is to move both the tracheal pole and zero to the frequency location of the pole, and then to move the frequency of the tracheal pole **FTP** gradually up or down over perhaps 50 ms to an appropriate value, as revealed by spectral analysis of the transition.

The variables **BTP**, "bandwidth of the tracheal pole", and **BTZ**, "bandwidth of the tracheal zero", are set to default values of 180 Hz. It is difficult to determine appropriate synthesis bandwidths for individual tracheal resonances, but, fortunately, one can achieve good synthesis results without changing these default values in most cases. However, if the location of a tracheal zero is not clear from analysis of a breathy vowel, one possible synthesis strategy is to set **FTP** equal to **FTZ** at the frequency of the observed most prominent tracheal resonance, and simply increase the bandwidth of the the zero **BTZ** (or decrease **BTP**) in order to reveal the presence of the resonance peak in the synthesis. Each factor of two change in bandwidth will approximately increase the strength of the tracheal resonance by about 6 dB.

In a transition from a voiceless consonant to a vowel, illustrated in Figure 3.33, one can often see evidence of tracheal coupling. To mimic the tracheal resonance at 2100 Hz in part (b) of this example, the tracheal pole frequency **FTP** is set to 2100 Hz during the [h], and the tracheal zero frequency **FTZ** and bandwidth **BTZ** are adjusted to best match a nearby dip in the spectrum. Since two prominent extra tracheal pole-zero pairs are seen in the breathy voicing of part (d), both the tracheal pole-zero and the nasal pole-zero are employed to match the details of the spectrum. As the glottis closes down at the back over the course of the vowel, the synthesis strategy would be to gradually adjust the frequency and bandwidth of each zero to approach that of the corresponding tracheal pole. The tracheal pole and tracheal zero then cancel each other out, and it is as if they were not present in the cascade vocal tract model, resulting in synthesis of a normal vowel, as shown in part (g) of

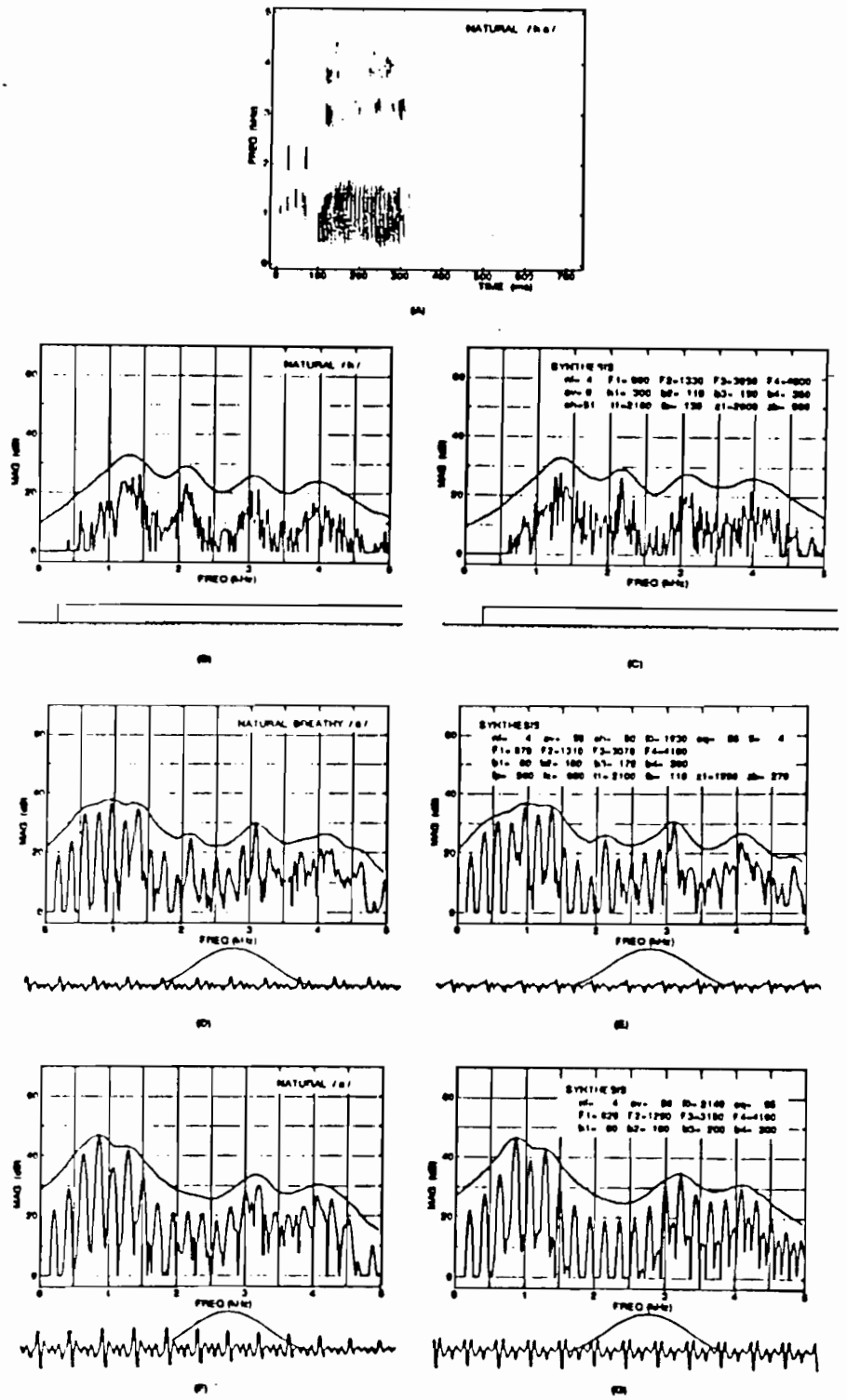


Figure 3.33: *A broadband spectrogram of (a) a natural [ha] syllable spoken by a female subject has been sampled at three time points to produce spectra of (b) [h] aspiration noise, (c) a synthetic imitation, (d) breathy voicing at the onset of the vowel [a], (e) a synthetic imitation, (f) normal voicing near the middle of the [a], and (g) a synthetic imitation. Note the extra tracheal formant peak at about 2100 Hz (between F2 and F3) in the [h] noise, and evidence of two tracheal poles at about 560 and 2100 Hz in the breathy voicing.*

the figure.

44. A2F, A3F, A4F, A5F, A6F, AB: Amplitudes of Frication-Excited Formants

The variables **A2F**, **A3F**, **A4F**, **A5F**, **A6F**, **AB**, "amplitudes of the parallel formants", determine the spectral shape of a fricative or plosive burst. The effect of turning each amplitude control on separately is shown in Figure 3.34. The spectral level of the frication peak corresponding to a control parameter setting of 60 dB is a few dB greater than that generated by the same formant of a vowel using default settings to all control parameter values and setting **AF** = 60 dB. 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 behind the constriction should have their amplitudes set to zero initially.[16] Then all parallel formant amplitudes should be adjusted on a trial-and-error basis, comparing synthesized frication spectra with a natural frication spectrum, as illustrated in Figure 3.35. In this example, a [k] burst is shown, together with an initial attempt at synthesis of this burst, with **A2F** and **A4F** set 60 dB, and a final match after further adjustment of A2F and **A4F**. Of course, frication bursts are short samples of random noise and a perfect match over the entire spectrum is usually impossible; in this case, it is best to ensure a match to the locations and amplitudes of formant energy concentrations.

The bypass path amplitude **AB** 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], [θ], [ð], [p], [b].

50. B2F, B3F, B4F, B5F, B6F: Bandwidths of Frication-Excited Formants

The variables **B2F**, **B3F**, **B4F**, **B5F**, **B6F**, "bandwidths of the frication-excited parallel formants" are set to default values that are typically wider than the bandwidths used in the cascade vocal tract model. Additional losses are due to acoustic losses at the constriction, as well as a smaller volume for storage of acoustic energy for a front-cavity resonance, leading to increased bandwidth.

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, or adjusted slightly as in the example of Figure 3.35. The only remaining adjustment is to the parallel formant amplitudes, in order to match details in a natural frication spectrum.

[16]The amplitude of the first parallel frication-excited formant, is therefore zero for all English fricatives. In fact, there is no wide-bandwidth parallel first formant in the synthesizer.

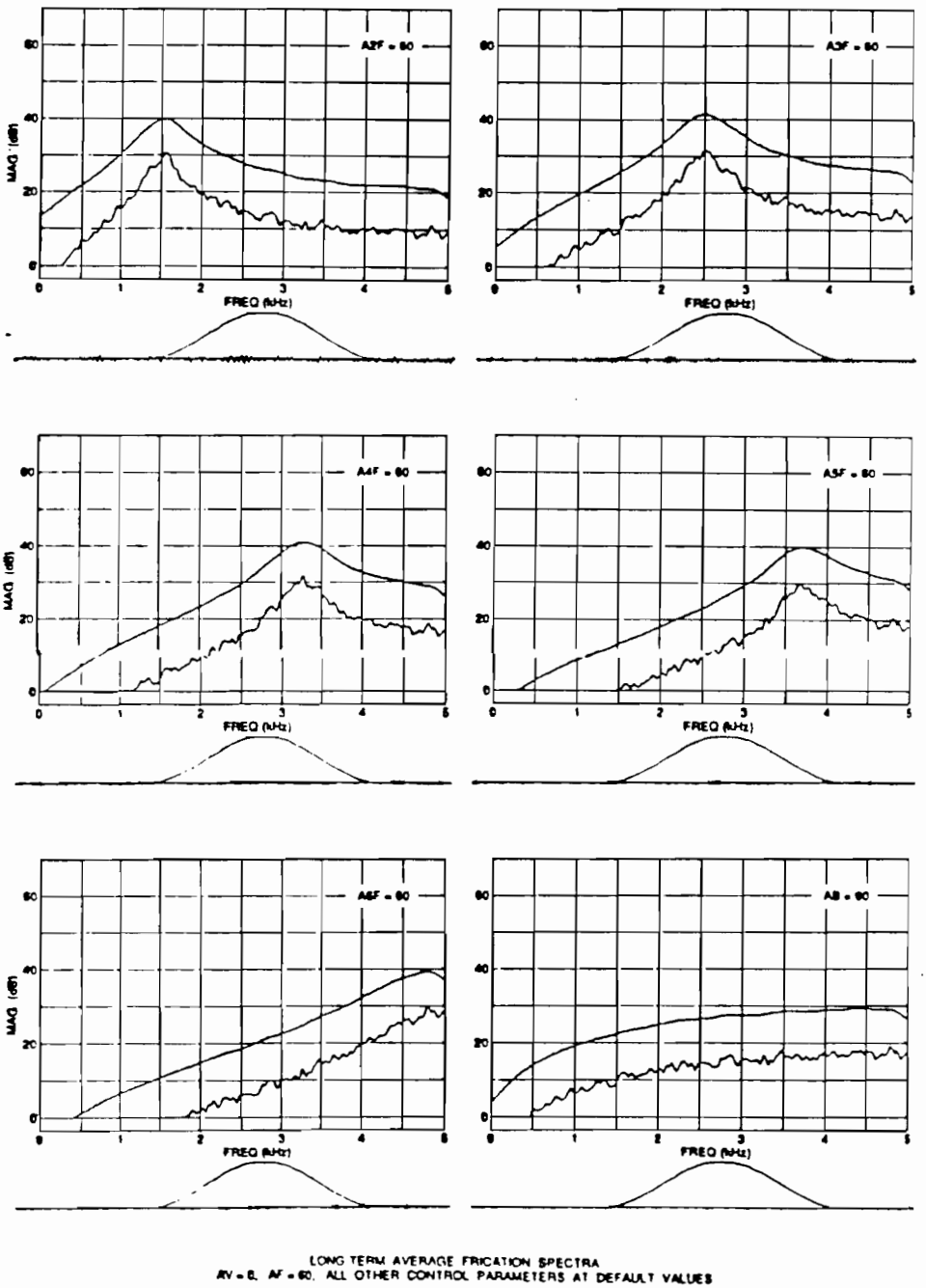


Figure 3.34: *Long-term average DFT spectra of the frication noise that is generated when* AF = *60 and each of the parallel frication amplitude controls is turned on individually.*

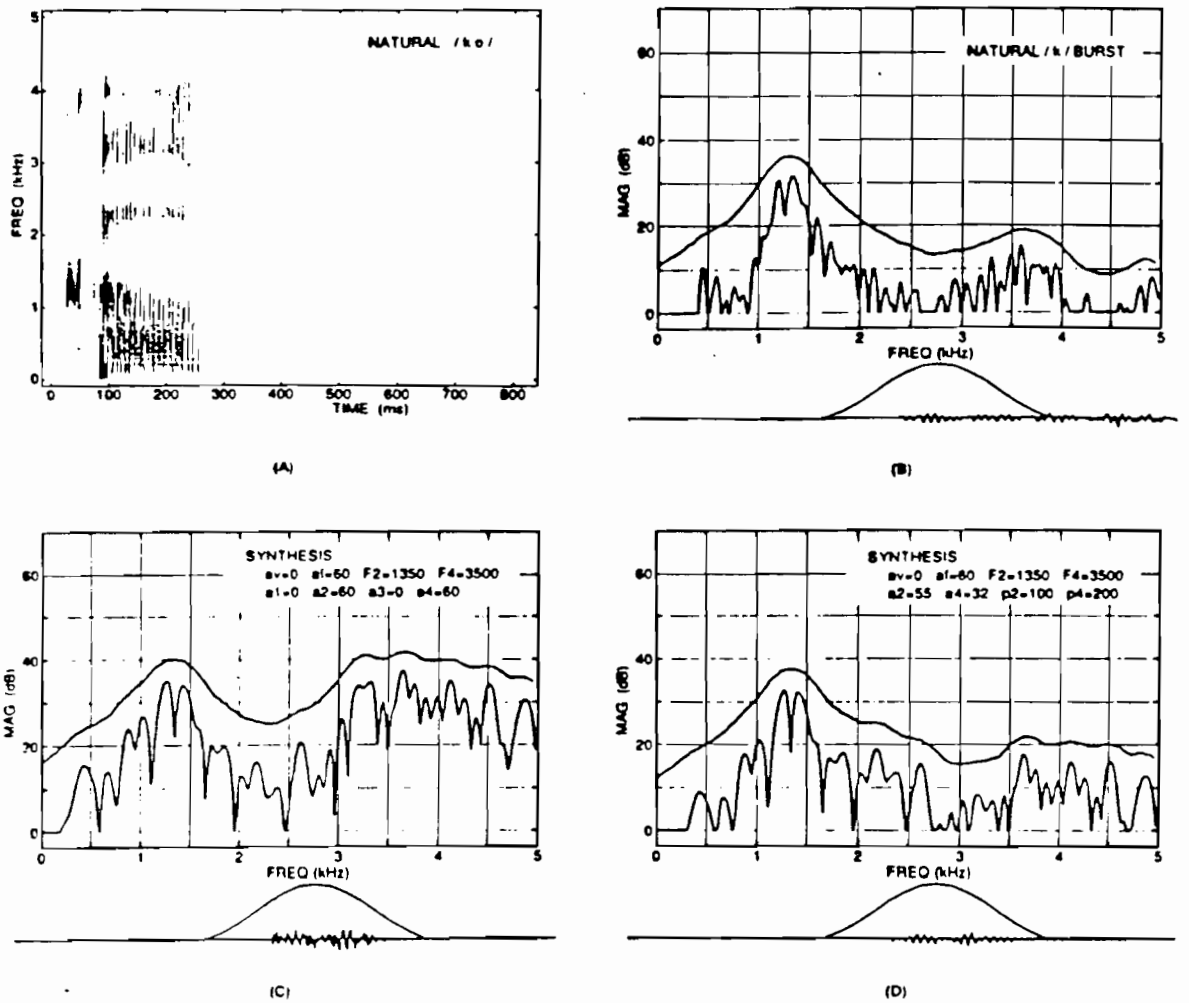


Figure 3.35: *(a) A broadband spectrum of the syllable [kɑ], (b) spectrum sampled at the release of the [k] burst, (c) synthesis of a 10-ms burst with front-cavity formant amplitudes* A2F *and* A4F *set to an initial guess of 60 dB and all other (back-cavity) formant amplitudes set to zero, and (d) synthesis after trial-and-error adjustment of formant amplitudes and parallel formant bandwidths.*

55. ANV: Extra Nasal Pole in Parallel Vowel Synthesis

The variable **ANV**, "amplitude of the parallel nasal formant", is normally not used. However, when employing the parallel vocal tract to synthesize vowels, as discussed below, **ANV** can be used to simulate the effects of nasalization on vowels and nasal murmurs. To achieve nasalization, one would set **FNP** to an appropriate value and adjust both **ANV** and **A1V** to levels matching a nasalized vowel spectrum.

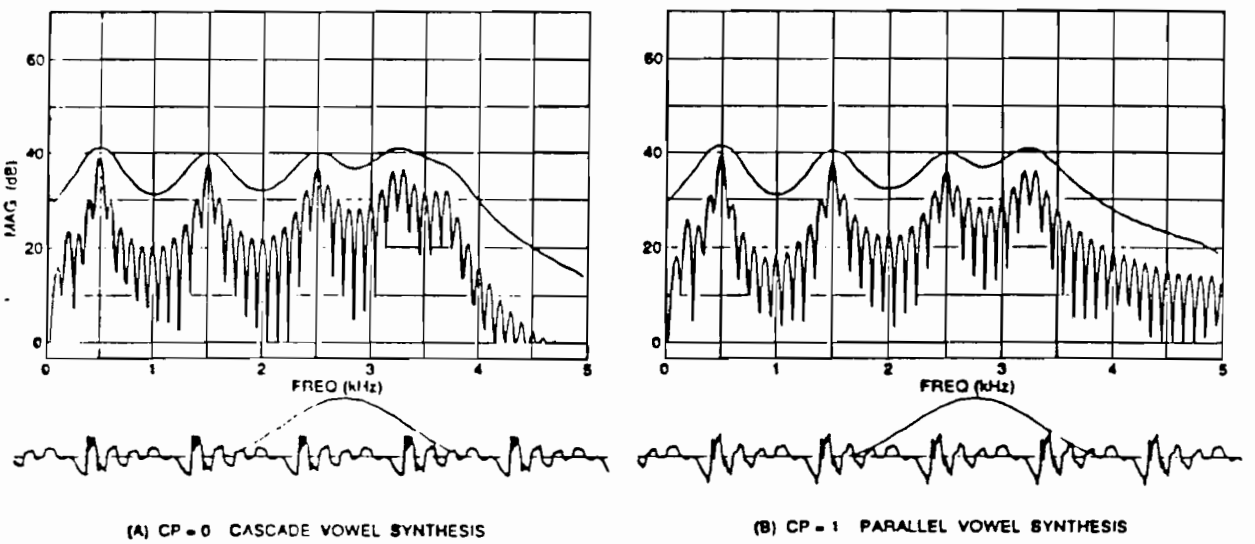


Figure 3.36: *Spectra of (a) a vowel produced by the cascade vocal tract model (CP=0), and (b) the same vowel produced by the parallel vocal tract model (CP=1). Both vowels were synthesized with default settings of all other control parameters.*

56. A1V, A2V, A3V, A4V: Amplitudes of Formants in Parallel Vowel Synthesis

In parallel mode, outputs are added with alternating sign, and the formants F2, F3, F4 and FT (frequency of tracheal pole) are excited by a first difference of the source waveform in order to maximally mimic the net effect of the ideal vocal tract transfer function (see Holmes, 1973, and Klatt, 1980, for justification).

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 larynx-excited parallel vocal tract. In this case, the control constant CP, "cascade/parallel switch", should be set to one. If **A1V**, **A2V**, **A3V**, **A4V** are set to about 60 dB (their default values), the levels of each formant are about right—at least for the default neutral-shaped vocal tract. (See Figure 3.36.) However, as formant frequencies change, complex interactions require that parallel formant amplitude controls be adjusted in order to match the theoretical transfer function of the vocal tract.

The parallel formant amplitude variables must then be adjusted to get the right spectral shape for the vowel. A good starting point when attempting to mimic a natural utterance is to first match the natural vowel in cascade synthesis mode, using formant bandwidth settings to get the depth of the valleys between formants about right, using **OQ** to set the relative amplitude of the first harmonic, and using **TL** to get the general spectral tilt about right. Then change over to parallel vowel synthesis (**CP=1**) and set parallel formant amplitudes **A1V**, **A2V**, **A3V**, **A4V** to initial guesses of 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 val-

ues (appropriate for a uniform tube), formant amplitudes can quickly diverge from those in a corresponding cascade vocal tract model.[17] Trial-and-error adjustment of parallel formant amplitudes is then necessary, as illustrated in Figure 3.37. This figure shows the spectrum of a naturally produced /i/, together with the spectrum of the same vowel produced with cascade synthesis. The lower panels show a first attempt at synthesis of the vowel with no adjustments of the gains of individual resonators (left panel), and the spectrum after the gains have been adjusted. The cascade and final parallel versions of the synthetic /i/ vowel shown in the figure are perceptually indistinguishable.

60. ATV: Extra Tracheal Pole in Parallel Vowel Synthesis

The variable **ATV**, "amplitude of the parallel tracheal formant", is normally not used. However, when employing the parallel vocal tract to synthesize vowels, as discussed above, **ATV** can be used to simulate the effects of tracheal coupling in breathy vowels. To add a resonance peak associated with tracheal coupling, one would set **FTP**, the frequency of the tracheal pole, to the frequency of the observed spectral peak and adjust **ATV** to match the observed spectral amplitude of the tracheal resonance, perhaps starting with a value of ATV=50 dB.

3.4 References

Allen, D.R. and Strong, W.J. (1985), "A Model for the Synthesis of Natural Sounding Vowels", *J. Acoust. Soc. Am. 78*, 58-69.

Ananthapadmanabha, T.V. (1984), "Acoustic Analysis of Voice Source Dynamics", *Speech Transmission Labs QPSR 2-3*, Royal Institute of Technology, Stockholm, 1-24.

Askenfelt, A. and Hammarberg, B. (1986), "Speech Waveform Perturbation Analysis: a Perceptual-Acoustical Comparison of Seven Measures,", *J. Speech and Hearing Res. 29*, 50-64.

Baer, T. (1978), "Effect of Single-Motor-Unit Firings on Fundamental Frequency of Phonation," *J. Acoust. Soc. Am. 64*, Suppl. 1, S90 (A).

Cranen, B. and Boves, L. (1987), "On Subglottal Formant Analysis," *J. Acoust. Soc. Am. 81*, 734-746.

Dunn, H.K. and White, S.D. (1940), "Statistical Measurements on Converstional Speech," *J. Acoust. Soc. Am. 11*, 278-288.

Fant, G. (1960), *Acoustic Theory of Speech Production*, 's-Gravenhage: Mouton and Co.

Fant, G. (1972), "Vocal Tract Wall Effects, Losses, and Resonance Bandwidths," *Speech Transmission Labs QPSR 2-3*, Royal Institute of Technology, Stockholm, Sweden, 28-52.

[17]The formants "riding on the skirt" of a lower-frequency formant are not automatically attenuated as this formant frequency is lowered in frequency, as is the case in a cascade configuration.

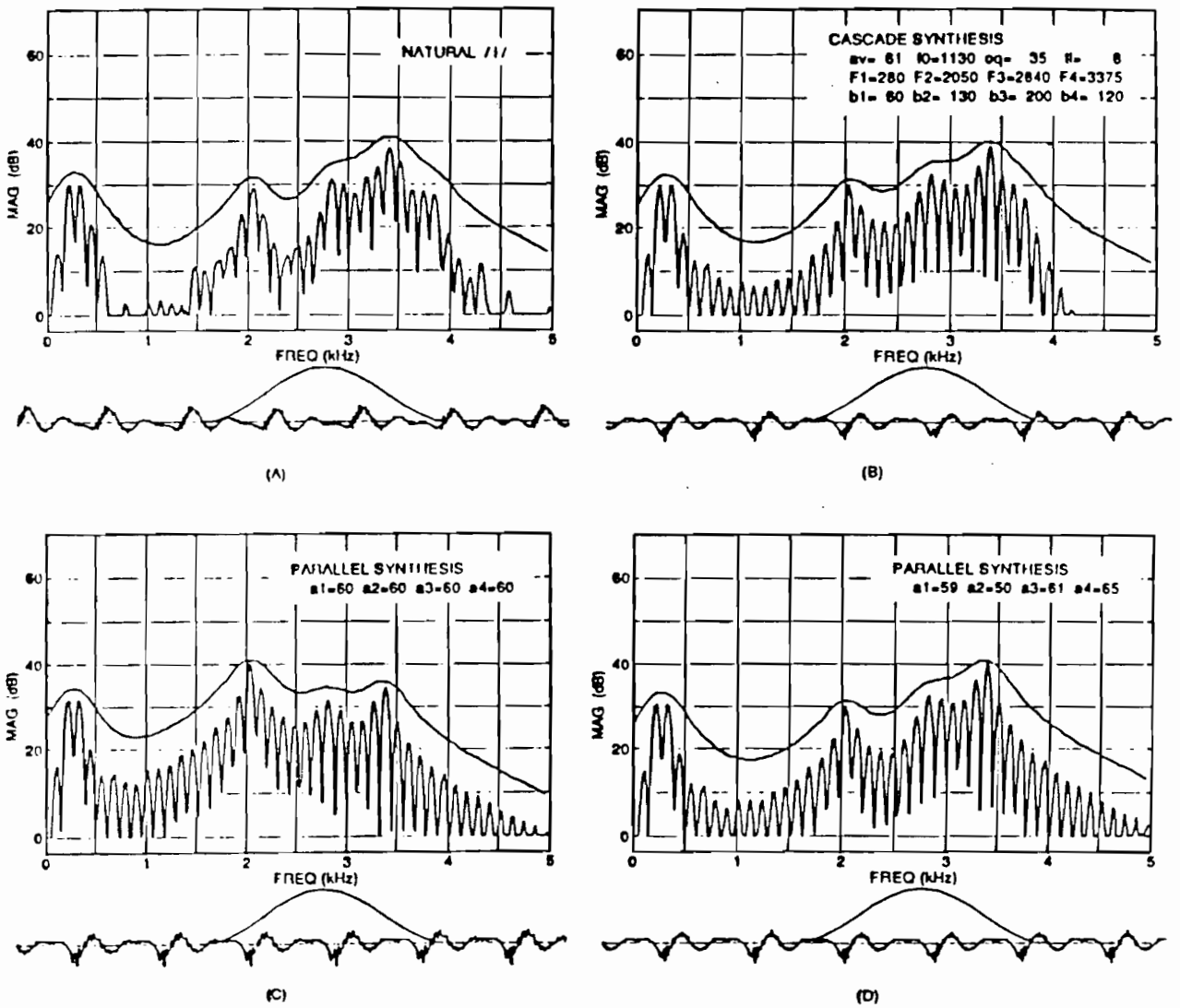


Figure 3.37: *(a) Spectrum of a naturally produced vowel /i/; (b) spectrum of a synthesized version of /i/ with a cascade connection of resonators; (c) spectrum resulting from attempt to synthesize /i/ with a parallel connection of resonators and with equal gain for each resonator; (d) spectrum of vowel synthesized with a parallel connection of resonators after adjusting individual gains to approximate spectrum in (a).*

Fant, G. (1979), "Glottal Source and Excitation Analysis", *Speech Transmission Labs QPSR 1*, Royal Institute of Technology, Stockholm, Sweden, 85-107.

Fant, G. (1980), "Voice Source Dynamics," *Speech Transmission Labs QPSR 2-3*, Royal Institute of Technology, Stockholm, Sweden, 17-37.

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

Fant, G. (1985), "The Voice Source: Theory and Acoustic Modeling", in I.R. Titze and R.C. Scherer (Eds.), *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control*, 453-464.

Fant, G. and Ananthapadmanabha, T.V. (1982), "Truncation and Superposition", *Speech Transmission Labs QPSR 2-3*, Royal Institute of Technology, Stockholm, Sweden, 1-17.

Fant, G. Ishizaka, K., Lindqvist, J., and Sundberg, J. (1972), "Subglottal Formants," *Speech Transmission Labs QPSR 1*, Royal Institute of Technology, Stockholm, Sweden, 85-107.

Fant, G., Liljencrants, J., and Lin, Q.G. (1985), "A Four-Parameter Model of Glottal Flow", *Speech Transmission Labs QPSR 4*, Royal Institute of Technology, Stockholm, 1-13.

Fant, G., Lin, Q.G., and Gobl, C. (1985), "Notes on Glottal Flow Interaction," *Speech Transmission Labs QPSR 2*, Royal Institute of Technology, Stockholm, 18-24.

Fant, G. and Mártony, J. (1963), "Speech Analysis," *Speech Transmission Labs QPSR 1*, Royal Institute of Technology, Stockholm, 1-5.

Flanagan, J.L. (1958), "Some Properties of the Glottal Sound Source," *J. Speech and Hearing Res. 1*, 99-116.

Flanagan, J.L. (1972), *Speech Analysis, Synthesis and Perception*, New York: Springer-Verlag (Second Edition).

Flanagan, J.L. and Saslow, M.G. (1958), "Pitch Discrimination for Synthetic Vowels," *J. Acoust. Soc. Am. 30*, 435-442.

French, N.R. and Steinberg, J.C. (1947), "Factors Governing the Intelligibility of Speech Sounds," *J. Acoust. Soc. Am.19*, 90-119.

Fujimura, O. (1960), "Spectra of Nasalized Vowels," *Research Laboratory of Electronics QPR 62*, MIT, Cambridge, MA, 214-218.

Fujimura, O. (1962), "Analysis of Nasal Consonants," *J. Acoust. Soc. Am. 34*, 1865-1875.

Fujimura, O. and Lindqvist, J. (1971), "Sweep-Tone Measurements of Vocal-Tract Characteristics," *J. Acoust. Soc. Am. 49*, 541-558.

Fujisaki, H. and Azami, S. (1971), 'Characteristics of Digital Pole Circuits for Simulating Time-Varying Vocal Tract Transfer Function", *Annual Report Engineering Res. Inst. 30*, 89-94, Univ. Tokyo, Tokyo Japan.

Fujisaki, H. and Ljungqvist, M. (1986), "Proposal and Evaluation of Models for the Glottal Source Waveform," *ICASSP-86*, 1605-1608.

Gobl, C. (1988), "Voice Source Dynamics in Connected Speech," *Speech Transmission Labs. QPSR 1,* Royal Institute of Technology, Stockholm, Sweden, 123-159.

Gold, B. and Rabiner, L.R. (1968), "Analysis of Digital and Analog Formant Synthesizers," *IEEE Trans. Audio Electroacoust. AU-16*, 81-94.

Hawkins, S. and Stevens, K.N. (1985), "Acoustic and Perceptual Correlates of the Non-Nasal/Nasal Distinction for Vowels," *J. Acoust. Soc. Am. 77*, 1560-1575.

Hollien, H., Michel, J., and Doherty, E.T. (1973), "A Method for Analyzing Vocal Jitter in Sustained Phonation," *J. Phonetics 1*, 85-91.

Holmes, J.N. (1961), "Research on Speech Synthesis", *Joint Speech Research Unit Report JU 11-4*, British Post Office, Eastcote, England

Holmes, J.N. (1973), "Influence of Glottal Waveform on the Naturalness of Speech from a Parallel Formant Synthesizer", *IEEE AU-21*, 298-305.

Horii, Y. (1979), "Fundamental Frequency Perturbation Observed in Sustained Phonation", *J. Speech and Hear. Res. 22*, 5-19.

Horii, Y. (1980), "Vocal Shimmer in Sustained Phonation," *J. Speech and Hearing Res. 23*, 202-209.

Ishizaka, K., Matsudaira, M., and Kaneko, T. (1976), "Input Acoustic Impedance Measurements of the Subglottal System," *J. Acoust. Soc. Am. 60*, 190-197.

Klatt, D.H. (1973), "Discrimination of Fundamental Frequency Contours in Synthetic Speech: Implications for Models of Pitch Perception", *J. Acoust. Soc. Am. 53*, 8-16.

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

Klatt, D.H. (1982), "Prediction of Perceived Phonetic Distance from Critical-Band Spectra: A First Step", *ICASSP-82*, 1278-1281.

Klatt, D.H. (1987), "Review of Text-to-Speech Conversion for English," *J. Acoust. Soc. Am. 82*, 737-793.

Klatt, D.H. and Klatt, L.C. (1990), "Analysis, Synthesis and Perception of Voice Quality Variations among Female and Male Talkers," *J. Acoust. Soc. Am. 87*, 820-857.

Knuth, D.E. (1981), "The Art of Computer Programming," *Volume 2 Seminumerical Algorithms*, Ch. 3, Reading, Mass: Addison Wesley, Second Edition, 9-24.

Ladefoged, P., Anthony, J., and Riley, C. (1971), "Direct Measurement of the Vocal Tract," *UCLA Working Papers in Phonetics 19*, 4-13. Also in *J. Acoust. Soc. Am. 49*, Suppl. 1., 104.

Lieberman, P. (1961), "Perturbation in Vocal Pitch", *J. Acoust. Soc. Am. 33*, 597-603.

Lieberman, P. (1963), "Some Acoustic Measures of the Fundamental Periodicity of Normal and Pathologic Larynges", *J. Acoust. Soc. Am. 35*, 344-353.

Lin, Q.G. (1990), "Speech Production Theory and Articulatory Speech Synthesis," *D.Sc. Thesis*, Royal Institute of Technology, Stockholm, Sweden.

Maeda, S. (1987), "On the Generation of Sounds in Stop Consonants," *Speech Communication Group Working Papers V*, Mass. Inst. of Technology, Cambridge, MA., 1-14.

Malme, C.I. (1959), "Detectability of Small Irregularities in a Broadband Noise Spectrum," *Research Laboratory of Electronics QPR 52*, MIT, Cambridge, MA, 139-141.

Matthews, M.V., Miller, J.E., and David, E.E., Jr. (1961) "Pitch Synchronous Analysis of Voiced Sounds," *J. Acoust. Soc. Am. 33*, 179-186.

Milenkovic, P. (1987), "Least Mean Square Measures of Voice Perturbation," *J. Speech and Hearing Res. 30*, 529-538.

Miller, R.L. (1959), "Nature of the Vocal Cord Wave," *J. Acoust. Soc. Am. 31*, 667-677.

Monsen, R.B. and Engebretson, A.M. (1977), "Study of Variations in the Male and Female Glottal Wave," *J. Acoust. Soc. Am. 62*, 981-993.

Morse, P.M. (1948), *Vibration and Sound*, New York: McGraw-Hill.

Nord, L., Ananthapadmanabha, T.V., and Fant, G. (1984), "Signal Analysis and Perceptual Tests of Vowel Responses with an Interactive Source-Filter Model", *Speech Transmission Labs QPSR 2-3*, Royal Institute of Technology, Stockholm, 25-52. Reprinted in *J. Phonetics 14*, 401-404.

Pollack, I. (1971), "Amplitude and Time Jitter Thresholds for Rectangular Wave Trains," *J. Acoust. Soc. Am. 50*, 1133-1142.

Rosenberg, A. (1968), "Effect of Pitch Averaging on the Quality of Natural Vowels," *J. Acoust. Soc. Am. 44*, 1592-1595.

Rosenberg, A. (1971), "Effect of Glottal Pulse Shape on the Quality of Natural Vowels", *J. Acoust. Soc. Am. 49*, 583-590.

Rothenberg, M., Carlson, R., Granstrom, B., and Lindqvist-Gauffin, J. (1975), "A Three-Parameter Voice Source for Speech Synthesis", in G. Fant (Ed.), *Speech Communication*, Uppsala: Almqvist and Wiksell, Vol. 2, 235-243.

Rozsypal, A.J. and Millar, B.F. (1979), "Perception of Jitter and Shimmer in Synthetic Vowels", *J. Phonetics* 7, 343-355.

Stevens, K.N. (1971) "Airflow and Turbulence Noise for Fricative and Stop Consonants", *J. Acoust. Soc. Am. 50*, 1180-1192.

Stevens, K.N. (1972), "The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data," in E.E. David and P.B. Denes (Eds.), *Human Communication: A Unified View*, New York: McGraw-Hill, 51-66.

Stevens, K.N. (1977), "Physics of Larynx Behavior and Larynx Modes", *Phonetica 34*, 264-279.

Stevens, K.N. (1989), "On the Quantal Nature of Speech," *J. Phonetics 17*, 3-45.

Stevens, K.N. (In preparation) *Acoustic Phonetics.*

Stevens, K.N., Fant, G., and Hawkins, S. (1987), "Some Acoustical and Perceptual Correlates of Nasal Vowels," in R. Channon and L. Shockey (Eds.), *Festschrift for Ilse Lehiste*, Dordrecht, the Netherlands: Foris, 241-254.

Sundberg, J. and Gauffin, J. (1979), "Waveform and Spectrum of the Glottal Voice Source," in B. Lindblom and S. Ohman (Eds.), *Frontiers of Speech Communication Research*, New York: Academic, 301-322.

Timke, R., von Leden, H., and Moore, P. (1959), "Laryngeal Vibrations: Measurements of the Glottic Wave. Part II: Physiological Considerations," *A.M.A. Arch. Otolaryng. 69*, 438-444.

Titze, I.R. (1984), "Parameterization of the Glottal Area, Glottal Flow, and Vocal Fold Contact Areas", *J. Acoust. Soc. Am. 75*, 570-580.