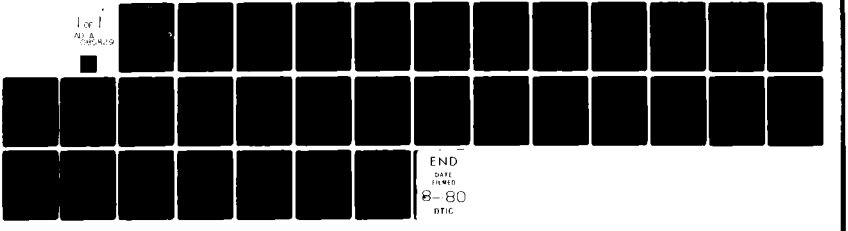
AD-A085 829 SPEECH COMMUNICATIONS RESEARCH LAB LOS ANGELES CA F/G 5/7
FINAL REPORT ON CONTRACT NUMBER N00014-76-C-0483,(U)
MAR 79 J E SHOUP, D H KLATT N00014-76-C-0483
UNCLASSIFIED NL



END
DATE FILMED
8-80
DTIC

(12)

FINAL REPORT

OFFICE OF NAVAL RESEARCH

(6) Final Report for CONTRACT NO. N00014-76-C-0483

Number

(11) 1 MAR 1979

(12) 32

(10) JUNE E. SHOUP

and

DENNIS H. KLATT

(15) N0014-76-C-0483

Speech Communications Research Laboratory, Inc.

806 West Adams Boulevard

Los Angeles, California 90007

ABSTRACT

Synthetic "set"-"sat" and [G AH G] - [G AA G] continua were constructed by the simultaneous manipulation of vowel duration and the first formant target frequency of the vowel. In a randomized identification test, a highly trained phonetician transcribed perceived vowel quality when instructed to ignore irrelevent charges in vowel duration. Results are compared with a groupof native English listeners who made phonemic judgments when presented with the same tape. It was found that there were large individual differences in the phonetic labels assigned to particular stimuli, so that cross-subject comparisons are not easily made. However, the one phonetically trained observer did show evidence ofcriterial shifts when only duration was changed. calling into question the commonly held notion that phonetically trained observers can estimate vowel quality independently of duration.

A second study of factors contributing to the perception of a natural voice quality in synthetic computer-generated speech has been initiated. A set of hypotheses to be investigated are described in Section II of this report. An additive harmonic speech synthesizer has been designed and implemented in software in order to carry out an evaluation of the hypotheses concerning the perceptual importance of various factors contributing to naturalness in synthetic speech. The harmonic synthesizer is described in Section III.

SECTION I

INFLUENCE OF VOWEL DURATION ON PHONETIC PERCEPTION

1.1 Introduction

The English vowels /AA/ and /AH/ differ primarily in two ways -- the first formant frequency of /AA/ is higher, and the duration of /AA/ is longer. If a syllable such as /g AA g/ is spoken rapidly, English listeners are likely to misperceive it as /g AH g/, even if it has formant trajectories appropriate for /g AA g/.

However, a phonetician is trained to ignore durational cues and to transcribe the vowel quality that was actually realized (Peterson and Shoup, 1966). If the midpoint formant values for the fast /G AA G/ are set equal to the target frequencies for [AA], the phonetician should be able to indicate this fact no matter what the speaking rate and duration of the vowel.

The purpose of the experiment to be described below was to determine the extent to which trained phoneticians meet this ideal by comparing their performance with that of untrained native English listeners. A set of synthetic consonant-vowel-consonant stimuli have been devised such that, when shortened, the vowel becomes ambiguous with another English vowel. Two vowel pairs have been studied -- /AA/ versus /AH/ and /AE/ versus /EH/. In particular, a shortened word "sat" should be heard as "set" by most English listeners and a shortened nonsense utterance /G AA G/ should sound like /G AH G/ to

untrained listeners, even if there is no vowel reduction concomitant with the shortening. For a trained phonetician, however, shortening not accompanied by vowel reduction should not change perceived vowel quality.

1.2 Stimuli

The Vowels /EH/ - /AE/. Broadband spectrograms of the words "set" and "sat" spoken in isolation by DHK are shown in Figure 1. Typical values for the vowel duration D and the first formant target F1 for "set" are D=190 ms and F1=510 Hz. For the word "sat", the values are D=280 ms and F1=630 Hz. These two parameters serve as the primary acoustic cues differentiating /EH/ from /AE/, although there may be small differences between the two vowels on other acoustic dimensions (Peterson and Barney, 1952).

Twenty nine stimuli along a "set" - "sat" continuum have been constructed by the simultaneous manipulation of vowel duration (from 160 ms to 340 ms) and first formant frequency of the vowel (from 480 Hz to 660 Hz). The stimulus inventory is defined in Figure 2. Since the two vowels are slightly diphthongized toward schwa before a postvocalic /D/, the first formant values refer to the initial vowel target. The first formant value near the end of the vowel is 60 Hz higher. All other parameter values are the same for all stimuli of the continuum. Synthesizer control parameter values were obtained by trial and error until a natural utterance was duplicated according to spectral criteria.

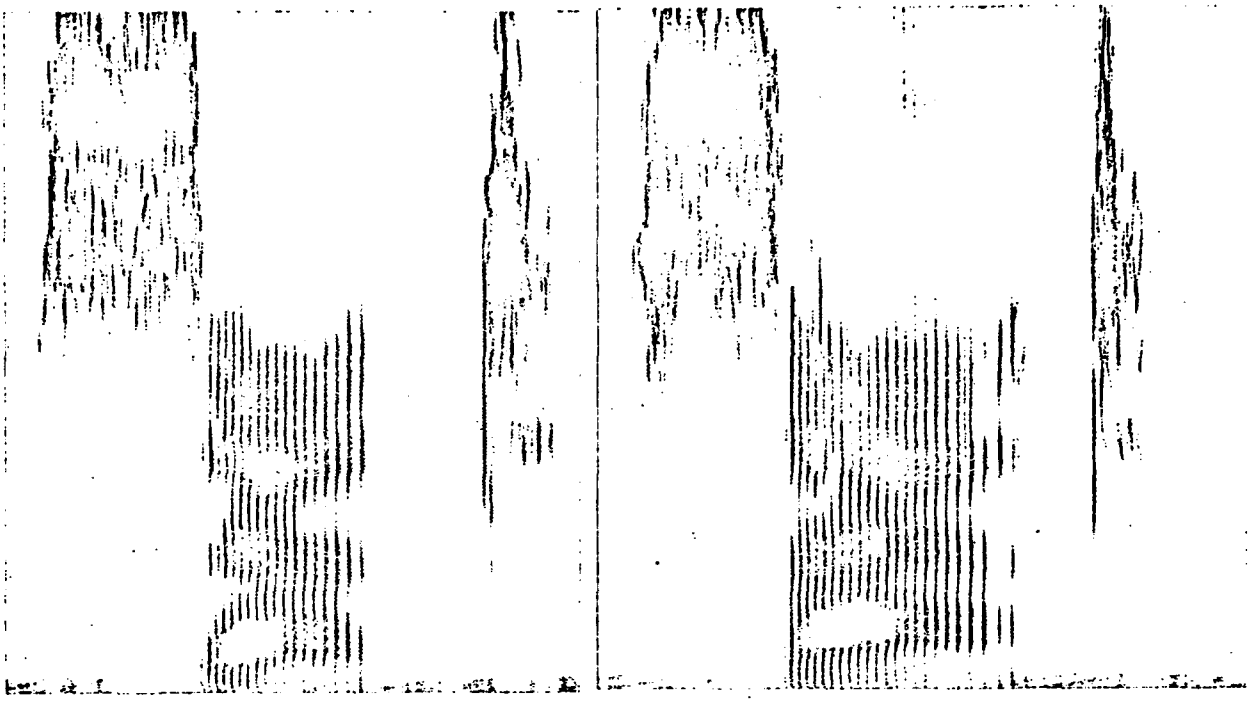


Figure 1. Broadband spectrograms of the words "set" and "sat".

The values for D and F1 span the expected phoneme boundary in such a way as to cover the range of expected interactions between these two acoustic cues. The endpoint stimuli 4 and 26 are perceived as clear natural versions of the two words.

The Vowels /AH/ - /AA/. There exists another vowel pair of English, /AH/-/AA/, for which the primary distinguishing acoustic characteristics are vowel duration and first formant target. A second set of stimuli involving these vowels have been included in the experiment to determine the perceptual interaction of duration and first formant target in another region of the vowel space.

Stimulus Number (1 - 29):

	First Formant Target F1 (Hz)						
D (ms)	480	510	540	570	600	630	660
160			8	13	18	23	27
190		4	9	14	19	24	28
230	1	5	10	15	20	25	29
280	2	6	11	16	21	26	
340	3	7	12	17	22		

Figure 2. Vowel Duration and first formant frequencies of 29 stimuli ranging from "set" to "sat". Also shown are broadband spectrograms of the four endpoint stimuli and a naturally spoken token of the words "set" and "sat".

Another reason to include a second vowel continuum has to do with the well-known contrast effect in vowel perception; there is often a pronounced shift in the responses of subjects to vowel stimuli as a function of the acoustic characteristics of the previous stimulus. If the previous stimulus was labeled as an /EH/, then the next stimulus from an /EH/-/AE/ continuum is less likely to be called an /EH/ than if the previous stimulus was labeled /AE/. In an attempt to overcome, in part, this contrast effect, we have assembled an identification test in which adjacent test stimuli come from alternate continua.

Typical values for vowel duration and first formant target for /AA/ in the environment of /G/ are D=320 ms and F1=740 Hz. For /AH/, the values are D=240 ms and F1=590 Hz. These two parameters serve as the primary acoustic cues differentiating

/AH/ from /AA/. Twenty nine stimuli along a [G AH G] - [G AA G] continuum have been constructed by the simultaneous manipulation of vowel duration (from 180 to 360 ms) and first formant frequency at vowel midpoint (560 to 740 Hz). The stimulus inventory is defined in Figure 3.

A randomized 300-item test tape was prepared from the 29 "set"-"sat" stimuli and the 29 [G AH G]-[G AA G] stimuli. The 58 different stimuli were played rapidly in succession at the beginning of the tape in order to familiarize the listeners with the range of variation in the vowels. There followed the 300-item identification test in which stimuli were separated by 3-second pauses, with an extra 5-second pause every tenth stimulus to ensure synchrony of responses with the answer sheet. Adjacent stimuli of the identification test came from different continua. Five randomizations of the 29 stimuli from the /EH/-/AE/ continuum and five different randomizations of the /AH/-/AA/ continuum were folded together to make a 290-item randomized sequence, and the first ten items were repeated at the end to complete the 300-item test. (The first 5 and last 5 items of the test were not scored.)

Instructions to the expert phoneticians were given on a piece of paper which read in part: "Transcribe each vowel that you hear for phonetic quality while disregarding the variations in duration that are present. You may use a broad or narrow phonetic transcription; we are only interested in your self consistency and ability to ignore spurious variation in vowel

Stimulus Number (1 - 29):

	First Formant Target F1 (Hz)						
D (ms)	560	590	620	650	680	710	740
180			8	13	18	23	27
210		4	9	14	19	24	28
250	1	5	10	15	20	25	29
300	2	6	11	16	21	26	
360	3	7	12	17	22		

Figure 3. Vowel Duration and first formant frequencies of 29 stimuli ranging from [G AH G] to [G AA G]. Also shown are broadband spectrograms of the four endpoint stimuli and a naturally spoken token of the nonsense syllables.

duration even though this variation is phonemic in your native language.

The instructions to the control subjects read in part: "Transcribe phonemically the vowel qualities that you hear. We expect most stimuli will be heard as /EH/, /AE/, /AH/, or /AA/, but there may be other vowels present, such as e.g. /IH/ or /AO/."

Further instructions given to both sets of subjects included a warning that the randomization happened to produce long sequences where one of the intended four vowels was unlikely to be heard, so please do not charge your criteria when you notice the prolonged absence of a particular response.

1.3 Results

The tape was played to one expert phonetician (JES) and to four control subjects familiar with phonemic notation. The results for the control subjects can be expressed as the first formant value at the phoneme boundary, as a function of stimulus duration. The phoneme boundary is defined as the F1 value at which the response pattern crosses 50 percent /AA/ responses. The boundary is at F1=625 Hz for a stimulus duration of 280 ms, i.e. a duration that is ambiguous between /AA/ and /AH/. The boundary shifts toward more /AA/ responses with increasing vowel duration according to the expected trading relation between the two cues. Changing vowel durations near this ambiguous region has an effect of shifting the phoneme boundary 5% (in terms of F1) for a 20% change in duration. Further from the maximally ambiguous stimulus duration, changes to vowel duration have less of an effect on the phoneme boundary, i.e. duration is not as strong a cue as first formant frequency.

This data was from the best two subjects, i.e. the subjects that had the least difficulty assigning labels to the vowels. Unfortunately, some of the remaining subjects, including the trained phonetician, had great difficulty in categorizing the vowel for some stimuli. This was manifested in the data by a good deal of variability in the response label assigned to the identical stimulus when it was repeated later in the test.

Thus our original objective, to compare trained phoneticians with untrained listeners, cannot be performed with statistical significance. Nevertheless, it is clear that both groups of subjects change their responses when only duration is manipulated. Unfortunately, the data from this kind of experiment do not tell us any more than this, i.e. it is not possible to quantify the role of duration in the perception of individual subjects.

This research was abandoned in favor of a more interesting theoretical question when it was discovered that a similar study had just been published (Mermelstein, 1978). The new topic is described in the next section.

SECTION II

PERCEPTION OF NATURAL VOICE QUALITY

2.1 Theoretical Considerations

The acoustic theory of vowel production that has been developed by Fant (1960), Stevens and House (1961) and others states that, in the frequency domain, the output spectrum for a vowel, P(f), is the product of (1) a source spectrum S(f), (2) a transfer function of the vocal tract T(f), and (3) a radiation characteristic R(f). The vocal tract transfer function and radiation characteristic describe a linear system that can be modeled by a product of poles and zeros.

The theory indicates that the output waveform for a vowel, p(t), can be synthesized by sending a quasi-periodic time waveform s(t) (that is analogous to a glottal volume velocity sound source) through a linear filter having a transfer function given by the product of T(f)*R(f). Assuming that the source waveform can be approximated by an analytic function, a vowel can also be synthesized by sending an impulse train through a linear filter having a transfer function given by the product of S(f)*T(f)*R(f).

This is the most common way that vowels are produced by formant synthesizers. Fundamental frequency and amplitude control parameters determine characteristics of an impulse train that excites a set of analog or digital formant resonators plus additional poles and zeros that approximate the desired transfer function.

A cascaded set of digital formant resonators can approximate the ideal vocal tract transfer function T(f) quite accurately for non-nasalized vowels, especially if the experimeter has dynamic control over both the frequencies and bandwidths of all of the formant resonators. Similarly, the radiation characteristic can be approximated very well by a first difference calculation (i.e. a zero at the origin in the z-plane).

A question that still remains, however, is how best to approximate the voicing source waveform and spectrum. It has often been said that formant synthesizers produce a somewhat mechanical, non-human sound, and that this is due to deficiencies in the representation of the voicing source. There are several hypotheses concerning specific deficiencies in the source representation, but there has been no systematic perceptual investigation of the relative importance of each. The purpose of the present investigation is to determine the relative perceptual importance of various factors that might improve the naturalness of the speech of formant synthesizers. The representational deficiency hypotheses fall into five general categories that are reviewed in the following paragraphs.

2.2 Hypotheses to be Tested

1. Stationarity

The first explanation that has been offered has to do with the perfect periodic regularity of synthetic waveforms that are produced when the synthesizer control parameters are held constant. It is argued that the auditory system is capable of detecting this regularity and thus distinguishing the synthesis from human speech which is never perfectly periodic for a number of reasons.

What aspect of the presumed variability in natural source waveforms is of greatest perceptual importance? There appear to be three candidates. The first concerns a possible random component in the times of opening and closing of the vocal folds during voicing. It has been claimed that the addition of perhaps a one percent random jitter to the period of the synthetic voicing waveform will have the desired perceptual effect of humanizing the voice (HYPOTHESIS 1).

A second possible random variation concerns the size and shape of each glottal pulse (HYPOTHESIS 2). Visual examination of high-speed motion pictures of the vibrating vocal folds suggests that the detailed vibration pattern does vary from cycle to cycle, with both regular (e.g. alternating periods more similar than adjacent periods) and irregular unpredictable variations being observed. Thus one might, for example, investigate the perceptual effect of adding a random component to the amplitude of each glottal pulse.

A third source of random variation is the generation of turbulence noise at certain times in the glottal opening/closing cycle (Rothenberg, 1974). If the airflow through the relatively narrow glottal opening exceeds a critical velocity, perceptible aspiration noise will be generated. Aspiration noise has a different spectrum from periodic voicing harmonics so that its presence would be perceived best at higher frequencies. Thus one might study the perceptual effect of adding aspiration noise to the voicing source waveform just before closure and immediately after opening of the glottis (HYPOTHESIS 3).

2. Phase Relations Among Harmonics

A second theory concerns the role of phase in vowel perception. During the production of a sustained vowel at constant pitch, the spectrum consists of a set of well-defined harmonics, each of which can be characterized by a magnitude (usually expressed in dB) and a phase angle (usually expressed as a delay relative to the timing of a positive peak in the waveform corresponding to the first harmonic). If all harmonics of the source waveform are synthesized to have the same (zero) phase, then there will be a time during the period when they all add together in phase and produce a large positive peak in the waveform. It has been suggested that this peakiness is an undesirable unnatural aspect of synthetic speech that should somehow be reduced by modifying the phase relations between

source harmonics. A synthesizer might try to achieve this goal in two ways -- either by adding a random component to the phase relations between source harmonics (HYPOTHESIS 4), or by trying to duplicate the detailed phase relations seen in real speech. For example, is it only necessary to produce phase relations such that the waveform appears to decay over a period (HYPOTHESIS 5). Or is it necessary to simulate the phase relations of the glottal source waveform, as seen in a plot of glottal area versus time (or inferred glottal volume velocity waveform) as well as the vocal tract transfer function induced waveform decay over time (HYPOTHESIS 6).

3. Pitch-Synchronous Changes to the Vocal Tract Transfer Function

The third theory questions the simplifying assumption that the vocal tract transfer function is independent of changes in glottal state over a single period. For example, recent calculations by Fant (personal communication) indicate that the bandwidth of the first formant of /a/ increases from about 50 Hz during the closed-glottis portion of the period to over 400 Hz when the glottis is open. Small shifts (on the order of 50 Hz) in formant frequency also occur for some vowels when the glottis is open. Pitch-synchronous changes in formant frequencies and bandwidths could be synthesized to investigate the perceptual effect of this detailed time variation, as compared with the use of suitable average formant frequency and bandwidth values (HYPOTHESIS 7).

4. Source Spectrum Irregularities

A fourth theory concerns the details of the voicing source spectrum. Typically, the source spectrum used in formant synthesis has a smooth monotonically decreasing envelope (harmonic amplitudes decrease at a rate of about -12 dB per octave of frequency increase). On the other hand, natural source spectra have a harmonic amplitude envelope that include notches at certain frequencies. The notches are an inherent consequence of a source waveform that includes a distinct closed portion of the period. If a formant frequency is changed such that the formant energy concentration moves into the region of one of these notches, its amplitude may be attenuated by as much as 10 to 20 dB. Perceptual consequences of the incorporation of notches at fixed frequency could be investigated by simple modifications to the source spectrum and waveform (HYPOTHESIS 8).

The perceptual effects of notches that change location as a function of fundamental frequency, formant position, or glottal state are also of interest (HYPOTHESIS 9). However, time-varying source irregularities are difficult to investigate since little is known about how natural source waveforms change as a function of these variables.

5. Changes Over the Duration of an Utterance

A fifth theory concerns the dynamic control of the voicing source. What changes to waveform and spectrum are seen

as fundamental frequency is varied or as lung pressure decreases toward the end of the utterance and the vocal folds are relaxed in preparation for breathing? How should sound intensity be varied over the utterance?

Typically, a synthetic waveform is constrained to retain a fixed spectral envelope as fundamental frequency is varied. Generally, no attempt is made to introduce a terminal breathy quality by adding aspiration and an associated boost in the amplitudes of the lowest frequency harmonics at the ends of utterances (HYPOTHESIS 9). Neither is there an attempt to simulate changes to spectral slope seen at high fundamental frequency (Monsen, Engebretson, and Vemula, 1978) (HYPOTHESIS 10), or to simulate the kinds of diplophonia (alternate peaks earlier and larger) that occur at low fundamental frequencies and at the end of utterances (HYPOTHESIS 11). It is likely that the "droning" quality characteristic of sentences generated with a fixed source spectral envelope is the most important defect to be removed by dynamic changes to the source model.

Specifically, over an utterance of several syllables duration, how should the source spectrum be changed as a function of fundamental frequency changes and position of the syllable in the breath group? Systematic investigation of natural speech, e.g. reiterent "mamama" speech would help to guide the search for relevant experimental variables.

2.3 Experimental Design

What is the perceptual importance of each of the above-mentioned differences between simplified vowel synthesis and observed characteristics of normal speech waveforms and spectra? This study will attempt to find answers to some of these questions through systematic examination and comparison of the effects of a number of stimulus variables. The general approach will be to obtain naturalness ratings and preference ratings for pairs of stimuli that differ along the appropriate physical dimensions. Simple vowels and longer vocalic sequences (synthesized or perhaps in part from a recording of an utterance) will be used as reference stimuli. In some cases, direct estimation of discriminability of certain physical dimensions may be obtained since it is not clear that a negative result (no systematic preference between a pair of stimuli) is due to a lack of discriminability or simply to an insignificant change in naturalness.

Most of the stimuli to be generated require straightforward modifications to an existing digital formant synthesizer that has been described previously (Klatt, 1980). However, in a few cases, it is desirable to have direct control over the amplitude and phase of each harmonic in the voicing source. Therefore, we have designed a second vowel synthesizer (known as the "additive harmonic synthesizer") that can generate the same waveform as a formant synthesizer, but by adding together sinusiods of the appropriate frequencies, amplitudes, and phases.

SECTION III

AN ADDITIVE HARMONIC SYNTHESIZER FOR PERCEPTUAL RESEARCH

3.1 Introduction

The output waveform for a stationary vowel can be represented by the sequence:

$$p(nT) = An*\cos((2\ PI\ n\ F0\ T) + PHn)$$

where n is an index of all harmonics below e.g. 5000 Hz, T is the sampling interval and equals 1/10000, F0 is the fundamental frequency in Hz, An is the amplitude of the nth harmonic, and PHn is the phase of the nth harmonic. Calculation of An and PHn for a given functional representation of the source waveform is complex, but equations are provided below.

Since the formant synthesizer produces an output spectrum $P(f)=S(f)*T(f)*R(f)$, it follows that the additive harmonic synthesizer should require that each harmonic of frequency n*F0 have an amplitude $An=|P(n*F0)|$ and a phase $PHn=/\underline{P(n*F0)}$. Analytic expressions for these quantities are given below.

An example of additive harmonic synthesis is shown in Figure 4. The figure indicates successive stages in the approximation of an impulse train by a sum of harmonics combined in cosine phase.

The magnitude of the transfer function of the vocal tract can be computed and used to adjust the amplitudes of harmonic

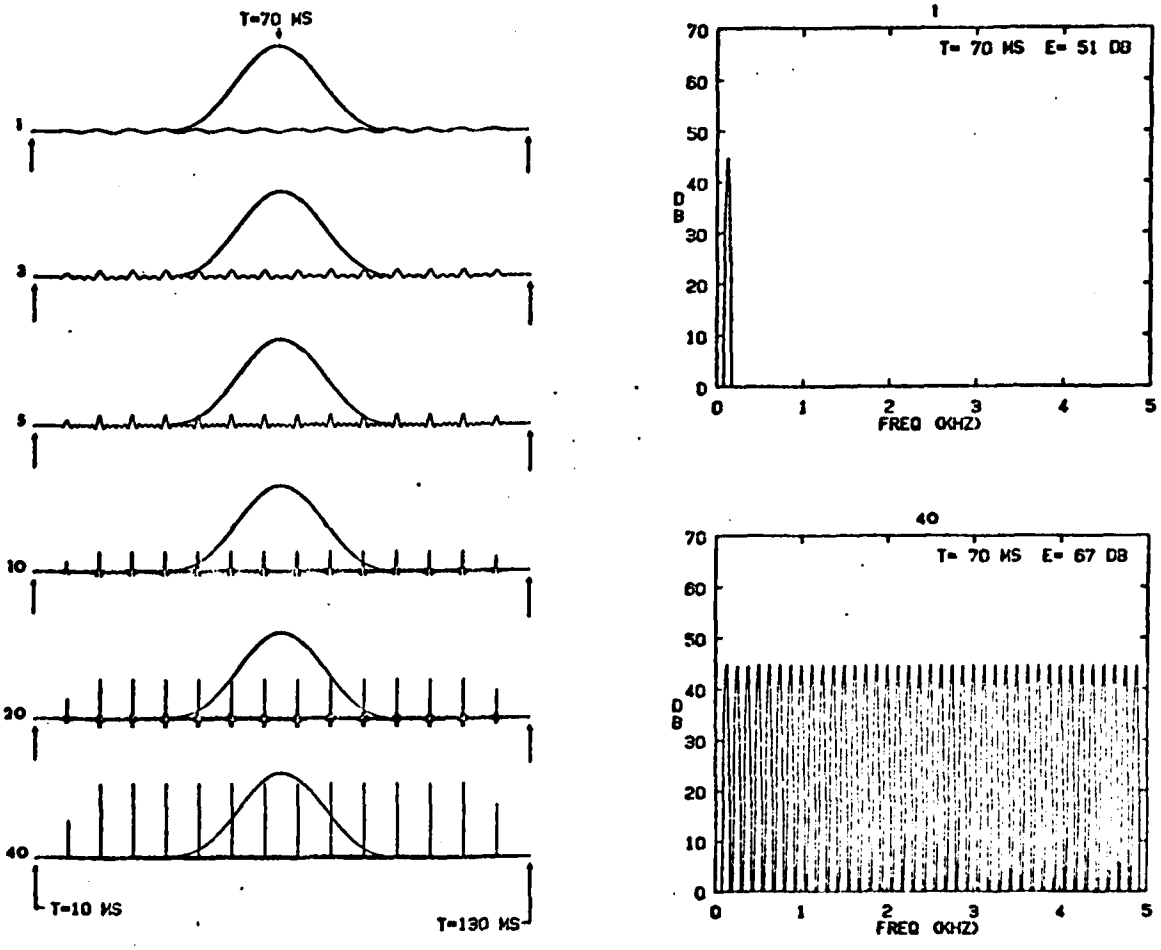


Figure 4. Generation of an impulse train by additive harmonic synthesis. As the number of harmonics increases, the approximation to an ideal impulse improves.

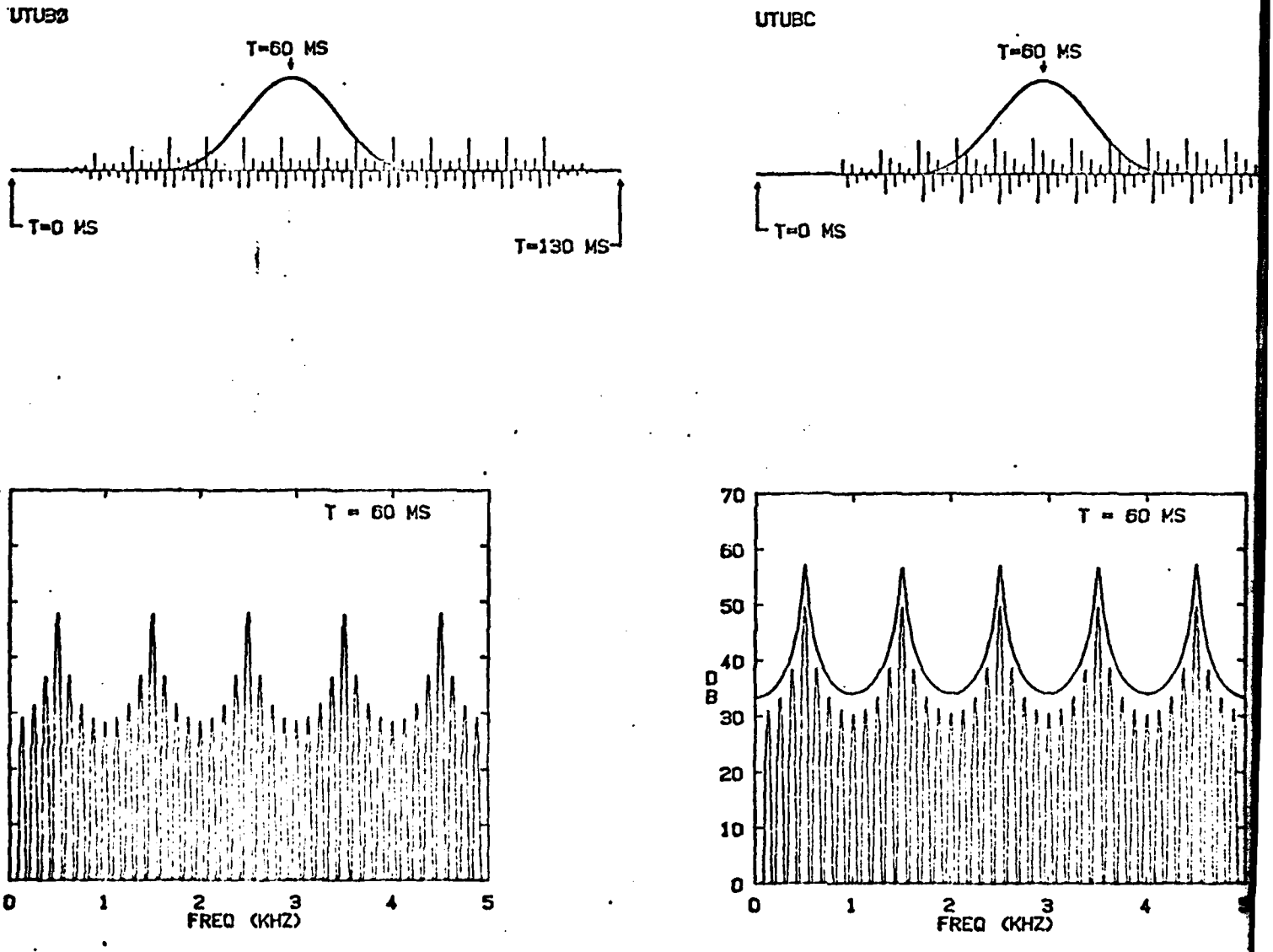


Figure 5. Synthetic waveform and spectrum corresponding to the excitation of a uniform vocal tract by an impulse train. In part a, the harmonics have been added in cosine phase, while in part b, the phase delay imposed by the vocal tract transfer function has been incorporated.

components making up the synthetic impulse train. The result is shown in Figure 5 for a transfer function analogous to a uniform tube closed at the glottis and open at the lips.

The lower panel of Figure 5 indicates the effects of the phase of the vocal tract transfer function. Inclusion of the phase causes the waveform to decay exponentially between impulse excitations of the vocal tract, although the magnitude spectrum is identical to the cosine phase waveform shown in Figure 5a. It takes about a half millisecond for the impulse excitation to travel from the glottis to the lips; one thus sees the consequences of a sequence of waves traveling back and forth in the uniform tube.

The net effect of imposing a -12 dB/octave roll-off to the source spectrum and a +6 dB/octave boost due to the radiation characteristic are shown in Figure 6a. The result is an additive harmonic synthesis of a neutral vowel. In Figure 6b, a formant resonance synthesizer has been used to generate a comarable waveform. As can be seen, the waveform and spectra are identical in the two cases.

The situation becomes more complex when synthesizer control parameters are allowed to vary so that the stationarity assumption is violated. The easiest way to maintain an approximate correspondence between the outputs of a formant synthesizer and the harmonic synthesizer is to change control parameter values pitch synchronously. In this case the signal is stationary during a glottal period, and harmonics can be added

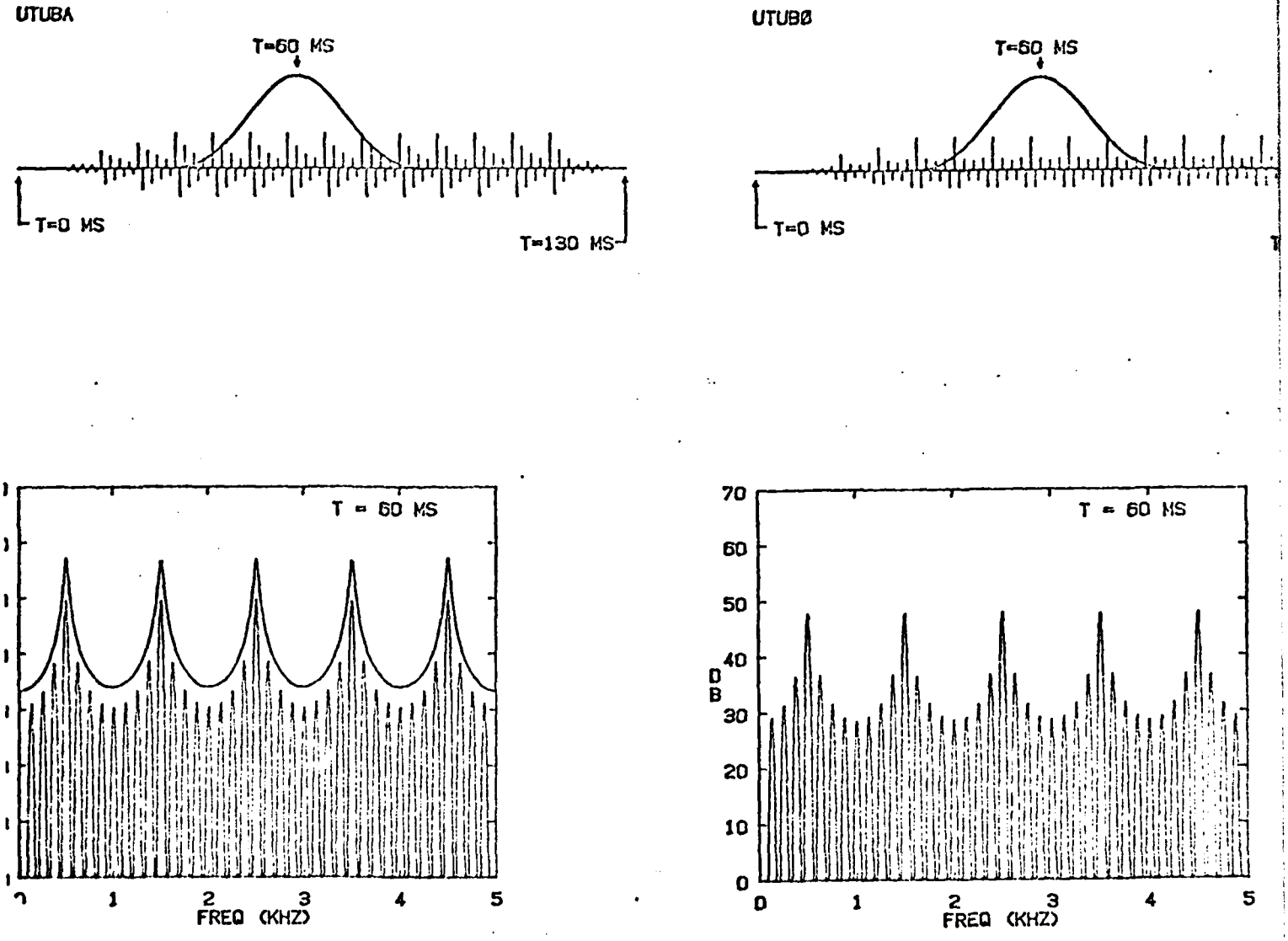


Figure 6. Comparison of additive synthesis and formant-resonator synthesis of a vowel having F0=100 Hz, -12 dB/octave source falloff, +6 dB/octave radiation characteristic, formant frequncies of 500, 1500, 2500, 3500, and 4500 Hz, and formant bandwidths of 70, 70, 70, 70, and 70 Hz.

together in the same way as before. The only issue is how to compute initial conditions such that the signal is the same as in a formant synthesizer. As a crude approximation, we have assumed that initial conditions are zero, i.e. the decaying waveform from the previous period can be ignored. Perceptual tests to be described below indicate whether this assumption can be justified.

3.2 Implementation Issues

The execute file HARSYN.EXE generates vowel-like synthetic speech sounds by adding together sinusiodal harmonics. To run the program, type <SP.KLATT.EXE>HARSYN(cr).

Synthesis control parameters are listed below. Default values are specified, as well as the range over which a parameter can be varied.

Name	Default	Function
AV	0	amplitude of voicing (0 to 70 dB)
F0	0	fundamental frequency (50 to 1000 Hz)
F1-F5	/EH/	formant frequencies (150 to 4999 Hz)
B1-B5	/EH/	formant bandwidths (40 to 1000 Hz)
DBO	-6	spectral slope in decibels per octave (-20 to +20)
DBK	0	spectral slope in decibels per kilohertz (-20 to +20)
PHR	0	degree to which phase lag of harmonic source is random (100 corresponds to random phase,

		0 to cosine phase).
PHT	100	degree to which phase lag of vocal tract transfer function is added to overall phase lag (full=100, none=0, negative full=-100)
GO	60	overall gain control in dB (0 - 100)
CSW	0	switch to engage a cascaded set of formants by the additive-harmonic source if set to 1.

Output waveform samples are computed according to the equation:

$$OUTPUT(mT) = \sum_{n=1}^{100} A(n) * COS(ARG(mT,n)) \qquad (1)$$

where $ARG(mT,n) = ARG(mT-T,n) + (2\ PI\ n\ F0\ T)$

and $ARG(1,n) = PH(n)$

If the fundamental frequency, F0, changes during the utterance, it is very important to compute output samples in the manner shown above rather than simply computing COS((2 PI nF0 mT) + PH(n)). The amplitude of a harmonic, A(n), is determined by the vocal tract transfer function (i.e. formant frequencies and bandwidths) as well as glottal source parameters. The phase parameter PHR(n) is used to control the phase of the glottal source harmonics, while PHT controls the amount of phase lag associated with the vocal tract transfer function. Details of how control parameters simulate the acoustic characteristics of the voicing source, the vocal tract transfer function, and the radiation characteristic are described in later sections.

Commands for control parameter specification by the user are the same as in the synthesis program NEWSYN.EXE and will be described in NEWSYN-DOCUMENTATION.MRN.

Source files used to compile and load the execute file HARSYN.EXE reside in <SP.KLATT.EXE>. The main calling program is HARSYN.FOR, and required subroutines include HARSY1.FOR and KOPLIB.REL. If charges are made to the Fortran routines, type LOAD @HARSYN to load a new version. Then type SAV HARSYN.EXE(cr) to obtain a new execute file.

3.3 Details of Control Parameter Specification

Control parameters are specified every 5 ms in the input data array. However, the synthesizer only updates parameters at the beginning of a period, using parameter values specifed at the time of the onset of the period in question. In this way, F0 and formant frequencies and bandwidths are constant throughout a period. The overall amplitude of the period is linearly interpolated from one end of the period to the next in order to minimize amplitude discontinuities at period boundaries.

Amplitude of Voicing AV

The amplitude of the voicing source, AV, is specified in decibels every 5 msec. A value of about 60 turns on the voicing source such that waveform peak amplitudes are nearly maximum, while a value of 0 turns off the voicing source. The value in dB is converted into a linear scale factor, AVOICE. If there is a

change in AVOICE from the previous frame, the actual amplitude of voicing for each of the waveform samples in the period is linearly interpolated from the amplitude of voicing specified in the previous period.

At the beginning of the utterance, it is assumed that the previous value of AV is zero. This results in a natural linear rise in amplitude of the output waveform. A similar linear fall at the end of the synthetic utterance should be generated by setting AV to 0 in the final frame in order to avoid "clicks" in the A/D output.

Fundamental Frequency F0

The fundamental frequency of vocal fold vibrations, F0, determines the frequencies of the harmonics that are present in the output. Fundamental frequency is constrained to stay between 50 and 1000 Hz (stimuli with F0 outside this range are not likely to be needed by the experimenter). Up to 100 harmonics of the fundamental frequency are generated, subject to the constraint that harmonics having a frequency that becomes greater than 4900 Hz are linearly interpolated to zero amplitude, and remain at zero amplitude until the frequency of the harmonic falls below 4901 Hz. As noted above, it is essential that Equation 1 be used to compute waveform samples if F0 is changed dynamically during the synthetic utterance.

The amplitude of each harmonic is determined by a multiplicative combination of (1) the amplitude of voicing, (2)

spectral shaping associated with the glottal source and radiation characteristic, and (3) the magnitude of the vocal tract transfer function. Control parameters that determine factors 2 and 3 are described next.

Spectral Slope in dB/octave, DBO

The combined effects of the voicing source and radiation characteristic on the magnitude spectrum are simulated by specifying the average spectral fall-off in dB per octave, using the control parameter DBO. A typical choice for DBO might be -6 because the voicing source typically falls off at -12 dB/oct while the radiation characteristic adds a +6 dB/octave increase to the spectrum. The control parameter DBO must be held constant over an utterance. The value assigned to DBO can be as high as +20 dB/octave, and as low as -20 dB/octave. In the additive harmonic synthesizer, harmonic amplitudes AHARM(NH) are adjusted according to the equation:

$$AHARM(NH) = NH^{**}(DBO/6.)$$

i.e. the harmonic number is raised to a power given by the real number DBO/6.

Spectral Slope in dB/kHz, DBK

An additional spectral slope in decibels per kiloHertz can be imposed on the synthesis through use of the constant control parameter DBK. The parameter has a default value of zero, i.e. it has no influence on the output. However, there are cases

where it is desired to manipulate spectral slope, and use of the control parameter DBO may result in too large an influence at low frequencies. For example, setting DBK to -2 will attenuate higher frequencies at a rate of -2 dB/kHz, i.e. formant amplitudes will fall at approximately 2 dB per formant (assuming an average formant spacing of 1000 Hz). The effect is implemented by the equation:

```
AHARM(NH) = AHARM(NH) * 2**(DBK*FHARM(NH)/6000.)
```

The first harmonic amplitude is not affected by these two glottal source slopes. This is achieved by a post normalization of the form:

```
AHARM(NH)=AHARM(NH)/AHARM(1)
```

Formant Frequencies F1, F2, F3, F4, F5

The effects of the vocal tract transfer function are determined by settings of five formant frequencies, F1-F5, (and five formant bandwidths, B1-B5). The magnitude of the vocal tract transfer function at each harmonic frequency FHARM(NH) is computed from this formant specification. The theoretical magnitude of the vocal tract transfer function is used to adjust harmonic amplitudes in the synthesis. Formant frequencies are not linearly interpolated over the period, but rather change discretely. The magnitude of the vocal tract transfer function at frequency F is given in Gold and Rabiner (1968),

and the effect of the vocal tract transfer function on a harmonic is given by:

AHARM(NH) = AHARM(NH)*T(FHARM(NH))

Formant Bandwidths B1, B2, B3, B4, B5

The formant bandwidths, B1-B5, also influence the magnitude of the vocal tract transfer function according to the equation given above. If there is a change in formant bandwidths from one 5 ms frame to the next, bandwidth values change discretely rather than being interpolated over the 50 samples of a frame.

Source Phase Lag, PHR

The phase spectrum of the output is determined by two phase paramters, PHR and PHT. Randomness can be imposed on source phase relations by increasing PHR from 0 to 100. If PHT=0 and PHR=0, harmonics will all be in cosine phase, and the waveform will have a maximal peak factor (ratio of peak amplitude over a period to rms amplitude). One way to reduce the peak factor toward a value more characteristic of naturally-spoken vowels is to add a random component to the phase relations PH(n) as defined in Equation 1. If RAN(n) is a random number ranging from -pi to pi, then the phase of a sourceharmonic is given by:

PH(n) = (PHR*RAN)/100.

Transfer Function Phase Lag, PHT

The phase relations between output harmonics are the summed effects of source harmonic relations, as governed by the control parameter PHR, and the phase lag imposed by the vocal tract transfer function. This latter phase lag is related to the tendency for the waveform to decay in amplitude over the interval of each period. If PHT is set to 100, a phase lag corresponding to this natural decay is computed. Leaving out this systematic phase lag (PHT=0), or making the waveform grow exponentially over each period (PHT=-100) can have important perceptual consequences. The parameter is presently included so as to permit psychophysical assessment of the importance of waveform decay.

The phase lag associated with the vocal tract transfer function is given by:

PHARM(NH) = sum [2*WT - ATAN(X1/Y) - ATAN(X2/Y)]

Overall Gain Control G0

A constant gain control G0 in dB can be used to raise or lower the overall level of the synthesis without having to change the detailed contour AV. A default value of 60 is assumed. To reduce the gain by 6 dB, one would set G0 to 54.

Cascade/Harmonic Vocal Tract Implementation CSW

The additive harmonic synthesizer normally simulates the vocal tract transfer function, T(f), by appropriate adjustments to the amplitudes and phases of source harmonics. This is exactly the same as sending the source function through a cascaded system of resonators having a transfer function T(f) as long as fundamental frequency remains constant and T(f) remains stationary. However, if F0 is not stationary, then two methods are not exactly the same because the decay of one excitation pulse into the following period is not considered in the additive harmonic verison.

In order to compare the perceptual consequences of this discrepency, we have provided a synthesis option such that the voicing source and radiation effects are synthesized by the additive harmonic technique, and this waveform is then passed through a digital filter consisting of five cascaded formants. If the control parameter CSW is set to 1, this digital filter is excited. Otherwise, the normal all-additive-harmonic technique is used.

References

Fant, C.G.M. (1960), Acoustic Theory of Speech Production. The Hague: Mouton & Co.

Gold, B. and L.R. Rabiner (1968), Analysis of Digital and Analog Formant Synthesizers, IEEE Transactions on Audio and Electroacoustics, AU-16, 81-94.

Klatt, D.H. (1980), Software for Cascade/Parallel Synthesizer, Journal of the Acoustical Society of America, Vol. 67, No. 3, 971-995.

Mermelstein, P. (1978), Difference Limens for Formant Frequencies of Steady-State and Consonant-Bound Vowels, Journal of the Acoustical Society of America, Vol. 63, No. 2, 572-580.

Monsen, R.B., A.M. Engebretson, and N.R. Vemula (1978), Indirect Assessment of the Contribution of Sub-Glottal Air Pressure and Vocal-Fold Tension to Changes of Fundamental Frequency in English, Journal of the Acoustical Society of America, Vol. 64, No. 1, 65-80.

Peterson, G.E. and H.L. Barney (1952), Control Methods Used in a Study of Vowels, Journal of the Acoustical Society of America, Vol. 24, 175-184.

Peterson, G.E. and J.E. Shoup (1966), The Elements of an Acoustic Phonetic Theory, Journal of Speech and Hearing Research, Vol. 9, No. 1, 68-99.

Rothenberg, M. (1974), Glottal Noise During Speech, Speech Transmission Laboratory - Quarterly Progress and Status Report, No. 2-3, 1-10.

Stevens, K.N. and A.S. House (1961), An Acoustical Theory of Vowel Production and Some of its Implications, Journal of Speech and Hearing Research, Vol. 4, 303-320.