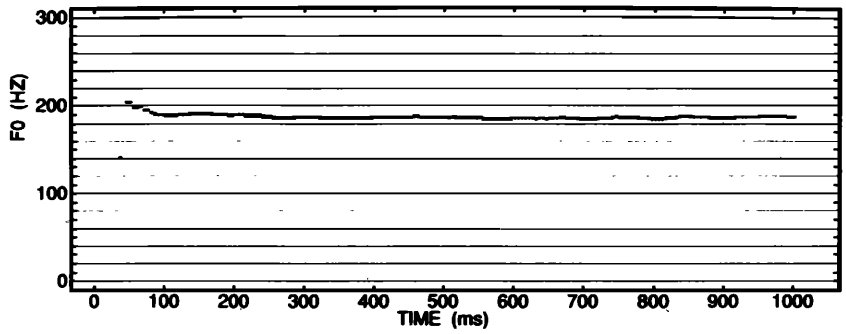
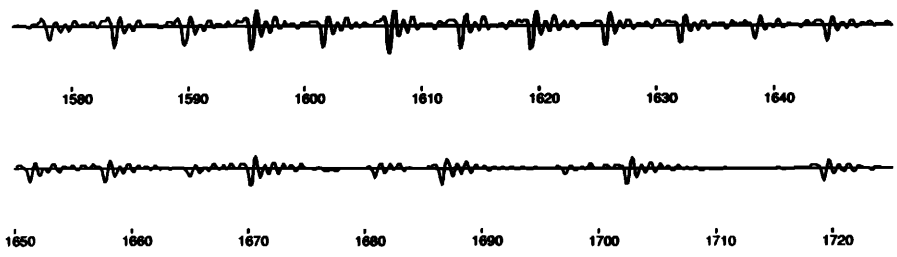


(a) VOWEL SUSTAINED AT CONSTANT PITCH, NOTE F0 JITTER



(b) EXAMPLE OF DIPLOPHONIC DOUBLE PULSING

FIG. 12. 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 in which various degrees of diplophonic double pulsing are present [normal voicing ($t = 1580$ to 1660) 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)]. (Diplophonic example extracted from the final syllable of female speaker LK [hɑ] imitation of "Steve eats candy cane.")

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. 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.

b. Diplophonic double pulsing. An example of diplophonic double pulsing is shown in Fig. 12(b). In the extreme, the alternate pulses may actually disappear, in which case 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 we have examined. Less extreme diplophonia may occur more often. The KLGLOTT88 voicing source model includes a mechanism for simulating double pulsing 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 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%–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).

2. Complications II: Source–tract interactions

According to the original classical formulation of the acoustic theory of speech production (Fant, 1960; Flanagan, 1972), the voicing source can be characterized as a "current source" because the volume velocity waveform $U_g(t)$ was said to depend very little on the shape or impedance of the vocal tract, at least for vowels. Similarly, the vocal-tract transfer function was assumed to be modeled well by a (succession of) time-invariant linear filter(s) because the terminating impedance at the glottis, while varying over a period, is nonetheless high compared with the vocal-tract impedance. These assumptions are illustrated in Fig. 13(a).

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. The essential characteristics of an "interactive source-filter model" that takes into account these complications are shown in Fig. 13(b). Four phenomena not satisfac-

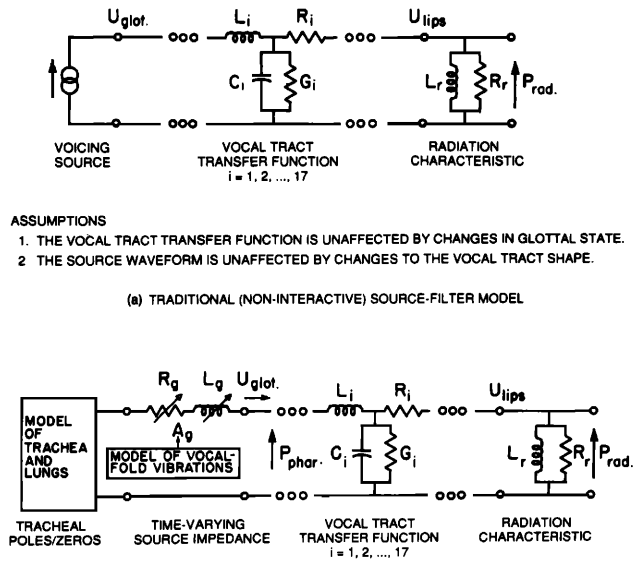


FIG. 13. The simplifications implicit in the classical acoustic theory of speech production, which constitute a noninteractive source-filter model shown in the top part of the figure, are contrasted with an interactive model in which voicing source glottal volume velocity $U_g(t)$ is influenced by pressure fluctuations above the glottis, and the vocal-tract transfer function changes over a period due to the time-varying glottal impedance.

torily modeled by conventional formant synthesizers can be identified from examination 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.

a. 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 was 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, 1982b; Fant *et al.*, 1985). The interaction is nonlinear in that volume velocity through an orifice is proportional to square root of the pressure drop (Stevens, 1971). Assuming a constant glottal area function from period to period, the noninteractive 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. The resulting $U_g(t)$ waveform includes significant "ripples" associated with the $F1$ standing wave, and the source waveform actually changes over the first two to three periods of voicing onset (Fant and Lin, 1987).

These effects, which Fant calls nonlinear superposition effects, 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. They are 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 KLGLOTT88 voicing source model. As is the case with most synthesis parameters, it is the change in parameter over time rather than its static value that has the greatest perceptual importance for improving naturalness. Thus, to mimic the buildup of an $F1$ ripple over the first few periods at voicing onset, a simultaneous decrease of **B1**, and increase in spectral tilt **TL**, might be performed. Such a change in relative formant amplitudes has been observed for laryngealized voicing onsets (Klatt, 1986b), but, in a transition between a voiceless consonant and a following vowel, the observed change is typically in the opposite direction, calling into question the generality and/or perceptual importance of this effect.

b. 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 Mártony, 1963; Fant and Ananthapadmanabha, 1982). Perhaps the pressure changes associated with the $F1$ standing wave induce a stronger closure if the phase is favorable (Rothenberg, 1985), and this occurs whenever $F1 = n \times f_0$. It should be possible to simulate the essential characteristics of this type of interaction by causing **AV** to increase whenever a harmonic is close to the frequency of $F1$. However, informal attempts to replicate this phenomenon with two speakers have been unsuccessful. Perhaps the interaction occurs only under some glottal conditions and not others.

In addition to the $F1$ standing wave induced in the vocal tract, it is possible that the increased impedance of a constricted vocal tract could influence source characteristics. The effect of a vocal-tract constriction on the vibratory behavior of the larynx has been studied by Bickley and Stevens (1986). Using both spectral analysis of natural speech produced under conditions of various suddenly applied oral constrictions and a subsequent modeling simulation, the authors found little change in source spectral characteristics until the constriction size became comparable to or smaller than that of a typical fricative. In this case, the glottal open time increased slightly, as did the bandwidth of the first formant, both being evidence of an increased average opening at the glottis. The authors conclude that, during the production of vowels and sonorant consonants, the effect of changes to oral constriction size on the mode of vibration of the larynx is probably negligible.

c. 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). Effects of time-varying formant bandwidths can be approximated in a formant synthesizer either by employing a perceptually equivalent constant 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.*, 1986). 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 KLSYN88.

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** in KLGLOTT88. 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. 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.

d. 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, 1986b). An example has been shown in Fig. 5. 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 *et al.*, 1976; Cranen and Boves, 1987). Berg (1960) originally observed a lowest tracheal resonance of 300 Hz. Ishizaka *et al.* (1976) found a much higher value of 640 Hz. Cranen and Boves (1987) measured values of the lowest three tracheal resonances of 510, 1350, and 2290 Hz from one male speaker. Fant *et al.* (1972) concluded from a modeling study that the latter estimates were more reasonable. The recent modeling work of Ananthapadmanabha and Fant (1982) and Rothenberg (1985) indicates that the actual effect on the vocal-tract transfer function of tracheal resonances is a complex function of the glottal configuration over time. A 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.

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[16]).

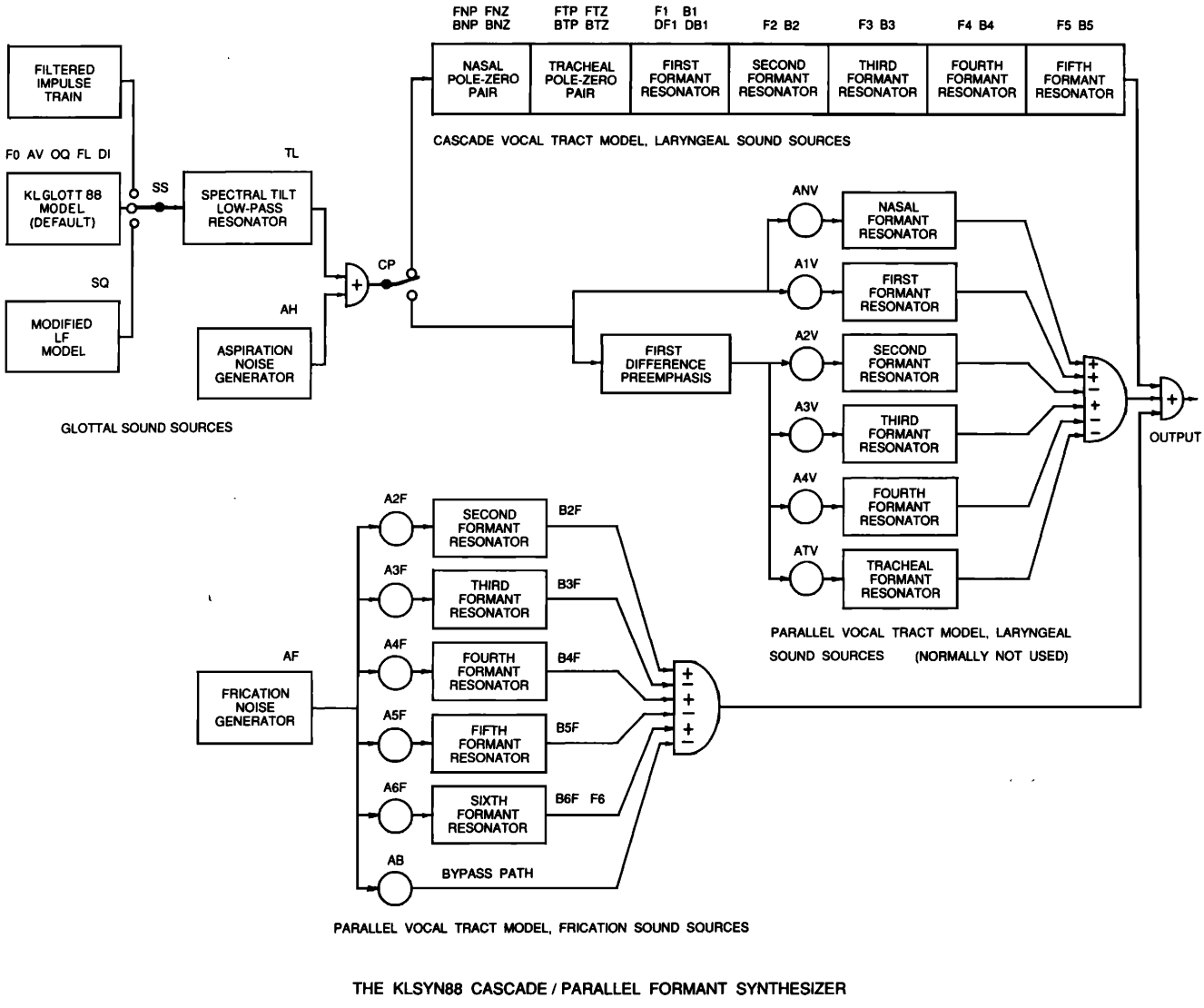
Normally, the spectral dip or zero corresponding to the selected tracheal resonance is immediately below it in frequency. Tracheal coupling usually begins and ends gradually as the glottis is opened or closed, which suggests a synthesis strategy in which both the tracheal pole and zero are usually moved together to the frequency location of an observed tracheal pole, and then the frequency of the tracheal zero **FTZ** is gradually moved down over perhaps 50 ms prior to glottal abduction to an appropriate value, as revealed by spectral analysis of the breathy 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 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 zero bandwidth will increase the strength of the tracheal resonance by about 6 dB.

C. Summary

In summary, the old cascade/parallel formant synthesizer, KLSYN (Klatt, 1980), has been modified to incorporate: (1) a version of the Liljencrants–Fant (LF) model of the glottal source; (2) 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; (3) an extra pole-zero pair for simulating the introduction of a tracheal resonance in the vocal-tract transfer function; and (4) an ability to change the first-formant bandwidth pitch-synchronously to simulate one of the interactions between source and vocal tract identified by Fant.

The resulting KLSYN88 synthesizer, summarized in Fig. 14, 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 specialized synthesis applications, but is normally not used. As was the case in the original formant synthesizer, the aspiration and frication noise sources are amplitude modulated, to simulate the effect of vocal-fold vibration, if **AV** is nonzero.



THE KLSYN88 CASCADE / PARALLEL FORMANT SYNTHESIZER

FIG. 14. Block diagram of the KLSYN88 formant synthesizer. Three voicing source models are available: (1) the old KLSYN impulsive source, (2) the 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.

Control parameters are identified above each block in Fig. 14. Some control parameter names have been changed slightly from Klatt (1980) in order to accommodate the new components and to be more mnemonic. There are, in addition, several constants that the user can modify; a complete list of synthesizer control parameters is identified in Tables XI and XII. New constant control parameters **RS** and **SB** permit the selection of a particular noise sample with a maximally flat spectrum, and duplication of that spectrum at every noise onset if desired. The parameters **GV**, **GH**, and **GF** are used to set scale factors for the individual sources, as indicated.

In the synthesis and perception experiments to be described below, some sentence-length reiterant utterances were generated using the KLSYN88 synthesizer. Various parameters of the synthesizer were manipulated to produce the stimuli, and the KLGLOTT88 source model was used in all of these experiments.

III. SYNTHESIS OF REITERANT UTTERANCES

One way to determine whether the acoustic correlates of breathiness identified in this study are perceptually important and sufficient cues to signal various voice qualities is to define a speech synthesizer that can manipulate these acoustic variables, and then attempt to mimic in detail some of the voices observed. Reiterant speech provides a significant advantage in these circumstances because one does not have to spend an inordinate amount of time attempting to deduce variations in the vocal-tract transfer function over each sentence to be synthesized. A tape recording played at the fall 1987 Meeting of the Acoustical Society of America demonstrated our success in this endeavor in that the synthesized and natural versions of [ʔɑ] and [hɑ] reiterant sentences from several female speakers are virtually indistinguishable.

In this section, strategies are presented for selecting optimal values over time for each glottal source synthesizer

TABLE XI. Constant control parameters for the KLSYN88 synthesizer configuration. Each control parameter is assigned a two-letter name, an indication of whether it is a constant or can be made to vary over time, 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.

SYM	V/C	MIN	VAL	MAX	Description
DU	C	30	500	5000	duration of the utterance, in ms
UI	C	1	5	20	update interval for parameter reset, in ms
SR	C	5000	10000	20000	output sampling rate, in samples/s
NF	C	1	5	6	number of formants in cascade branch
SS	C	1	2	3	source switch (1 = impulse, 2 = natural, 3 = LF model)
RS	C	1	8	8191	random seed (initial value of random number generator)
SB	C	0	1	1	same noise burst, reset RS if AF = 0 and AH = 0 (0 = no, 1 = yes)
CP	C	0	0	1	0 implies Cascade, 1 implies Parallel tract excitation by AV
OS	C	0	0	20	output selector (0 = normal, 1 = voicing source, ...)
GV	C	0	60	80	overall gain scale factor for AV, in dB
GH	C	0	60	80	overall gain scale factor for AH, in dB
GF	C	0	60	80	overall gain scale factor for AF, in dB

control parameter on the basis of spectral comparisons between synthesis and a natural recording. Finally, generalizations are made concerning typical synthesis parameter values for breathy and glottalized onsets and offsets.

A. Copying reiterant utterances

Application of the procedures described in this section to two female utterances resulted in the control parameter values that are shown in Figs. 15 and 16. The first utterance is a [ʔɑ] reiterant imitation of the five-syllable sentence "Steve eats candy cane" spoken by female LK. The second is a [hɑ] reiterant imitation of the same sentence by LK. The new voicing source has seven control parameters that must be specified:

F0 fundamental frequency,

AV amplitude of voicing,

OQ open quotient (ratio of open period to total period),

TL extra spectral tilt of the source (dB down at 3 kHz),

AH amplitude of turbulent aspiration noise added to voicing,

FL flutter (slowly varying statistical fluctuations to the fundamental period),

DI double pulsing (temporal offset and reduced amplitude of alternate periods).

1. Step 1: Set f_0 contour

The fundamental frequency parameter **F0** is used to reset the fundamental period $T0$ at the beginning of each pitch period.[17] A harmonic sieve-pitch-tracking algorithm (Duifhuis *et al.*, 1982), employing a 25-ms Hamming window, was used to determine f_0 every 10 ms in the original recording. The analysis data were transferred to the **F0** synthesis parameter track as the very first step in synthesis specification. Some trial and error adjustment based on comparison of synthesis and natural waveforms was necessary to match the irregular periods of glottalized attacks and offsets. A good match to the f_0 contour facilitates spectral comparisons needed to optimize other synthesis parameters since natural and synthesis spectra then have the same harmonic locations.

2. Step 2: Set AV, the amplitude of voicing

The amplitude of voicing **AV**, in dB, controls the height of each glottal pulse. Initial values for **AV** were determined from a plot of overall rms energy in the waveform, as estimated every 10 ms[18] using a 25-ms Hamming window, combined with a knowledge of when voicing was present in the waveform. Incremental adjustments were then made to the **AV** parameter track on the basis of trial and error comparison of synthesis and natural spectral levels until there was a good match (within about 1 to 2 dB) in spectra sampled about every 30 ms. This adjustment was done several times, the last being after all other synthesis parameters had been optimized.

3. Step 3: Set formant frequencies and bandwidths

Formant-frequency locations were estimated from spectra sampled at regular 30-ms time intervals throughout the vocalic portions of the utterance. An attempt was made to find a single constant value for each formant that would be a satisfactory approximation to the (possibly time varying) formant position. It is likely that slightly better synthesis matches could have been achieved by varying formant frequencies over time, but we wished to concentrate efforts on the glottal source parameter behavior, and we wanted to be sure that we were not covering up possible deficiencies in source flexibility by substituting unrealistic formant-frequency changes over time. After specifying formant frequencies, formant bandwidth values were then adjusted so as to give the appropriate level and shape to each spectral peak. Formant bandwidths were generally set to constants for this synthesis so as to minimize possible bandwidth compensation for source deficiencies (bandwidths were increased when the glottis was open for [h]).

4. Step 4: Set OQ, the open quotient

The open quotient **OQ**, in percent of the total period, determines the time during which the waveform is nonzero.

TABLE XII. Control parameters that can be varied over time in the KLSYN88 synthesizer configuration. Each control parameter is assigned a name, an indication of whether it is a constant or can be made to vary over time, 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.

SYM	V/C	MIN	VAL	MAX	Description
F0	v	0	1000	5000	fundamental frequency, in tenths of an Hz
AV	v	0	60	80	amplitude of voicing, in dB
OQ	v	10	50	99	open quotient (voicing open-time/period), in %
SQ	v	100	200	500	speed quotient (rise/fall time of open period, LF model only), in %
TL	v	0	0	41	extra tilt of voicing spectrum, dB down @ 3 kHz
FL	v	0	0	100	flutter (random fluct in $f0$), in % of maximum
DI	v	0	0	100	diplophonia (pairs of periods migrate together), in % of max
AH	v	0	0	80	amplitude of aspiration, in dB
AF	v	0	0	80	amplitude of frication, in dB
F1	v	180	500	1300	frequency of the 1st formant, in Hz
B1	v	30	60	1000	bandwidth of the 1st formant, in Hz
DF1	v	0	0	100	change in $F1$ during open portion of a period, in Hz
DB1	v	0	0	400	change in B1 during open portion of a period, in Hz
F2	v	550	1500	3000	frequency of the 2nd formant, in Hz
B2	v	40	90	1000	bandwidth of the 2nd formant, in Hz
F3	v	1200	2500	4800	frequency of the 3rd formant, in Hz
B3	v	60	150	1000	bandwidth of the 3rd formant, in Hz
F4	v	2400	3250	4990	frequency of the 4th formant, in Hz
B4	v	100	200	1000	bandwidth of the 4th formant, in Hz
F5	v	3000	3700	4990	frequency of the 5th formant, in Hz
B5	v	100	200	1500	bandwidth of the 5th formant, in Hz
F6	v	3000	4990	4990	frequency of the 6th formant, in Hz (frication excited, or if NF = 6)
B6	v	100	500	4000	bandwidth of the 6th formant in Hz (only applies if NF = 6)
FNP	v	180	280	500	frequency of the nasal pole, in Hz
BNP	v	40	90	1000	bandwidth of the nasal pole, in Hz
FNZ	v	180	280	800	frequency of the nasal zero, in Hz
BNZ	v	40	90	1000	bandwidth of the nasal zero, in Hz
FTP	v	300	2150	3000	frequency of the tracheal pole, in Hz
BTP	v	40	180	1000	bandwidth of the tracheal pole, in Hz
FTZ	v	300	2150	3000	frequency of the tracheal zero, in Hz
BTZ	v	40	180	2000	bandwidth of the tracheal zero, in Hz
A2F	v	0	0	80	amplitude of frication-excited parallel 2nd formant, in dB
A3F	v	0	0	80	amplitude of frication-excited parallel 3rd formant, in dB
A4F	v	0	0	80	amplitude of frication-excited parallel 4th formant, in dB
A5F	v	0	0	80	amplitude of frication-excited parallel 5th formant, in dB
A6F	v	0	0	80	amplitude of frication-excited parallel 6th formant, in dB
AB	v	0	0	80	amplitude of frication-excited parallel bypass path, in dB
B2F	v	40	250	1000	bandwidth of frication-excited parallel 2nd formant, in Hz
B3F	v	60	320	1000	bandwidth of frication-excited parallel 3rd formant, in Hz
B4F	v	100	350	1000	bandwidth of frication-excited parallel 4th formant, in Hz
B5F	v	100	500	1500	bandwidth of frication-excited parallel 5th formant, in Hz
B6F	v	100	1500	4000	bandwidth of frication-excited parallel 6th formant, in Hz
ANV	v	0	0	80	amplitude of voicing-excited parallel nasal formant, in dB
A1V	v	0	60	80	amplitude of voicing-excited parallel 1st formant, in dB
A2V	v	0	60	80	amplitude of voicing-excited parallel 2nd formant, in dB
A3V	v	0	60	80	amplitude of voicing-excited parallel 3rd formant, in dB
A4V	v	0	60	80	amplitude of voicing-excited parallel 4th formant, in dB
ATV	v	0	0	80	amplitude of voicing-excited parallel tracheal formant, in dB

This waveform is later filtered by the "tilt" low-pass filter, which may increase the effective open time by rounding the waveform corner at closure. An appropriate value for **OQ** is very difficult to determine directly from spectral or waveform characteristics, so a typical default value of 50% for a male voice or 60% for a female voice is usually chosen as a departure point. The primary acoustic effect of changes in open time is to increase and decrease the amplitude of the first harmonic relative to adjacent harmonics. Thus trial and error adjustment of **OQ** to match first-harmonic amplitude was attempted. In some cases, locations of spectral zeros can be used to confirm the correct value for this parameter, but spectral zeros were not evident at low frequencies in the data of LK. Matching of first-harmonic amplitude by adjustments to **OQ** was generally successful and resulted in physiologically reasonable values for **OQ** over most of the utterance. As would be expected, **OQ** had to be decreased for glottalization, and increased for breathiness associated with adjacent voiceless consonants.There was one problem situation. Frequently, at the end of a voicing interval followed by silence, the first harmonic gained a large relative prominence that could not be matched only by changes to **OQ**; it was necessary to also increase spectral tilt **TL**, increase **AV**, and increase the bandwidths of the lower formants.

5. Step 5: Set TL, the spectral tilt

The spectral tilt **TL**, in dB, determines the amount of extra attenuation of higher harmonics of the voicing source spectrum. Initial values for the **TL** parameter were determined by observing whether the dft spectrum was perfectly periodic in the frequency region of the third and fourth formants. In cases where the spectrum is essentially periodic, **TL** is set to zero, while indications of random aspiration noise in place of harmonics implies a **TL** value of about 20 dB. Transitions between these two states can be gradual, but are sometimes quite abrupt. Simultaneous trial and error matching of both **TL** and **AH** may be necessary to get the right balance of harmonics and aspiration noise in each part of the spectrum.

6. Step 6: Set AH, the amplitude of aspiration noise

The amplitude of aspiration noise, **AH** in dB, that is used to generate [h] and the aspiration of [p,t,k] is also used to add breathiness noise to voiced portions of utterances. Initial values for the **AH** parameter were obtained on a trial and error basis by matching levels of $F3$ and $F4$ excitation in spectra that were clearly inharmonic (after the **TL** parameter had been optimized so as to attenuate higher harmonics). The process is complicated by the statistical variability of noise processes, but if one averages visually over several 25-ms windowed spectra, the results are more stable and reliable. An appropriate value for **AH** in an essentially perfectly harmonic vowel is either zero, or perhaps some value about 10 dB lower than in a clearly breathy vowel (i.e., some small degree of aspiration noise may be present even if no visible manifestation appears in the spectrum).

7. Step 7: Set DI, waveform of diplophonic double pulsing

The double-pulsing parameter is provided in order to be able to mimic a special form of laryngealization that is occasionally seen for some talkers—alternate periods are delayed and attenuated [recall Fig. 12(b)]. The **DI** parameter is calibrated in percent such that 100% means that the first of two pulses is delayed maximally and is attenuated to zero, while 50% would delay only half as much and attenuate to half its unperturbed amplitude. For speaker LK, several offsets were laryngealized in this way, as indicated by the **DI** parameter in Figs. 15 and 16. Appropriate initial values for **DI** were determined by studying the natural waveform for obvious examples of doubling pulsing. Refined estimates of appropriate synthesis values for **DI** were obtained by comparing waveforms of synthetic and natural speech.

In summary, the seven steps enumerated above provide a reasonably straightforward procedure for synthesizing a close imitation for reiterant [hV] or [ʔV] versions of any sentence. Several female and male voices have been successfully imitated by this process. It may even be possible to automate parts of the analysis (Ananthapadmanabha, 1984; Fujisaki and Ljungqvist, 1986).

B. Synthesis generalizations concerning breathiness and laryngealization

The two reiterant sentences synthesized using values shown in Figs. 15 and 16 are representative of the larger corpus of speakers in that, in general, breathiness noise increased toward the end of a vowel (see Fig. 15) and near voiceless consonants (see Fig. 16), and was greater in unstressed syllables and in the final syllable of a sentence. Similar results have recently been described by Chasaide and Gobl (1987). Typical voicing source parameter values for normal voicing and breathy voicing for this speaker were:

Par	Normal	Breathy
AV	60	60
OQ	60	80
TL	8	24
AH	0–40	52
DI	0	0
FL	25	25

It is clear from listening, and from the **TL** value for the normal voice, that LK has a somewhat breathy version of normal phonation, but it is also clear from the time variation in Figs. 15 and 16 that her voice becomes more breathy in specified predictable circumstances.

In order to match the reiterant spectra of the [ʔɑ] and [hɑ] utterances of LK, it was also necessary to make some modifications to the vocal-tract transfer function. The formant frequencies were left roughly constant (exactly constant in the [ʔɑ] sentence and constant except for a slight rise during the [h]'s of the [hɑ] sentence). Formant bandwidths **B1** and **B2** were increased whenever formant spectral peaks were flattened; these times corresponded very well with times at which source parameters indicated that the

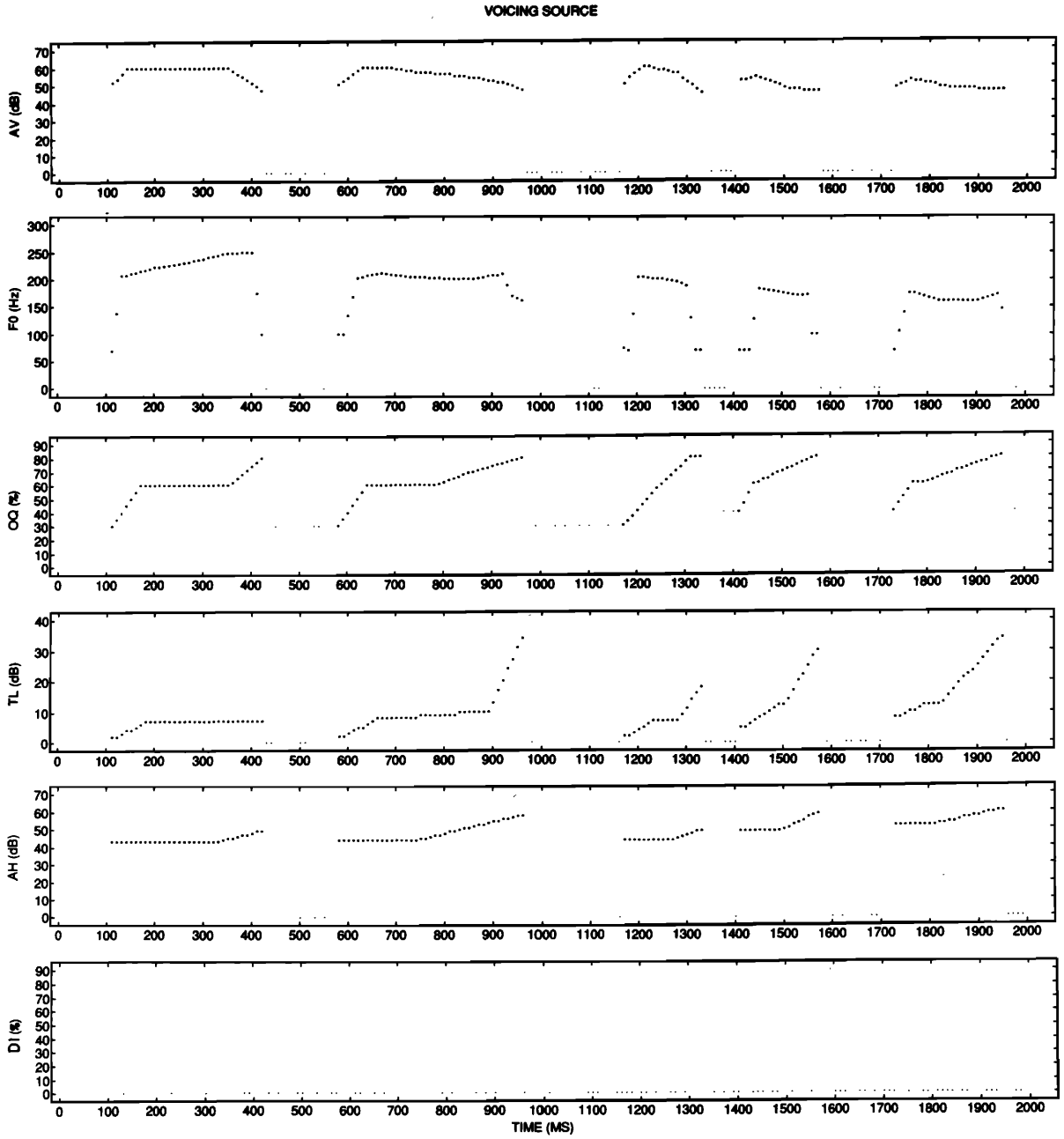


FIG. 15. Synthesis parameter values as a function of time for the [ʔɑ] reiterant imitation of "Steve eats candy cane" by female speaker LK. Bold data points indicate time intervals when voicing is on.

glottal opening was significantly increased. For example, there were many cases of vowel terminations where the first-formant peak essentially disappeared from the spectrum, and the first-formant bandwidth had to be increased substantially. For this speaker, acoustic coupling to the resonances of the trachea during breathy vowels and aspiration was minimal, so no use was made of the synthesis parameters available to introduce tracheal pole and zero pairs into the vocal-tract transfer function.

The reiterant speech data also reveal changes to a typical vowel induced by a glottal-stop onset or offset. There is a rapid (30 ms) fall in f_0 accompanying a glottalized onset or offset, and the amplitude of voicing **AV** is reduced by about 6 dB when f_0 is at its lowest. In a glottalized onset, the primary waveshape-controlling parameter to be affected is the open quotient, **OQ**, which is reduced to perhaps 30% in the vicinity of the glottal stop. It is likely that **AH** will be reduced and **TL** reduced in this interval if the vowel is otherwise somewhat breathy. Similar changes occur at a glottalized offset, but there is less of a reduction in breathiness correlates.

There were three examples of temporally offset glottal pulses (double pulsing) in the two sentences from LK. At these times, it was observed that the **TL** and **AH** parameters were not changed as much as for other vowel offsets, suggest-

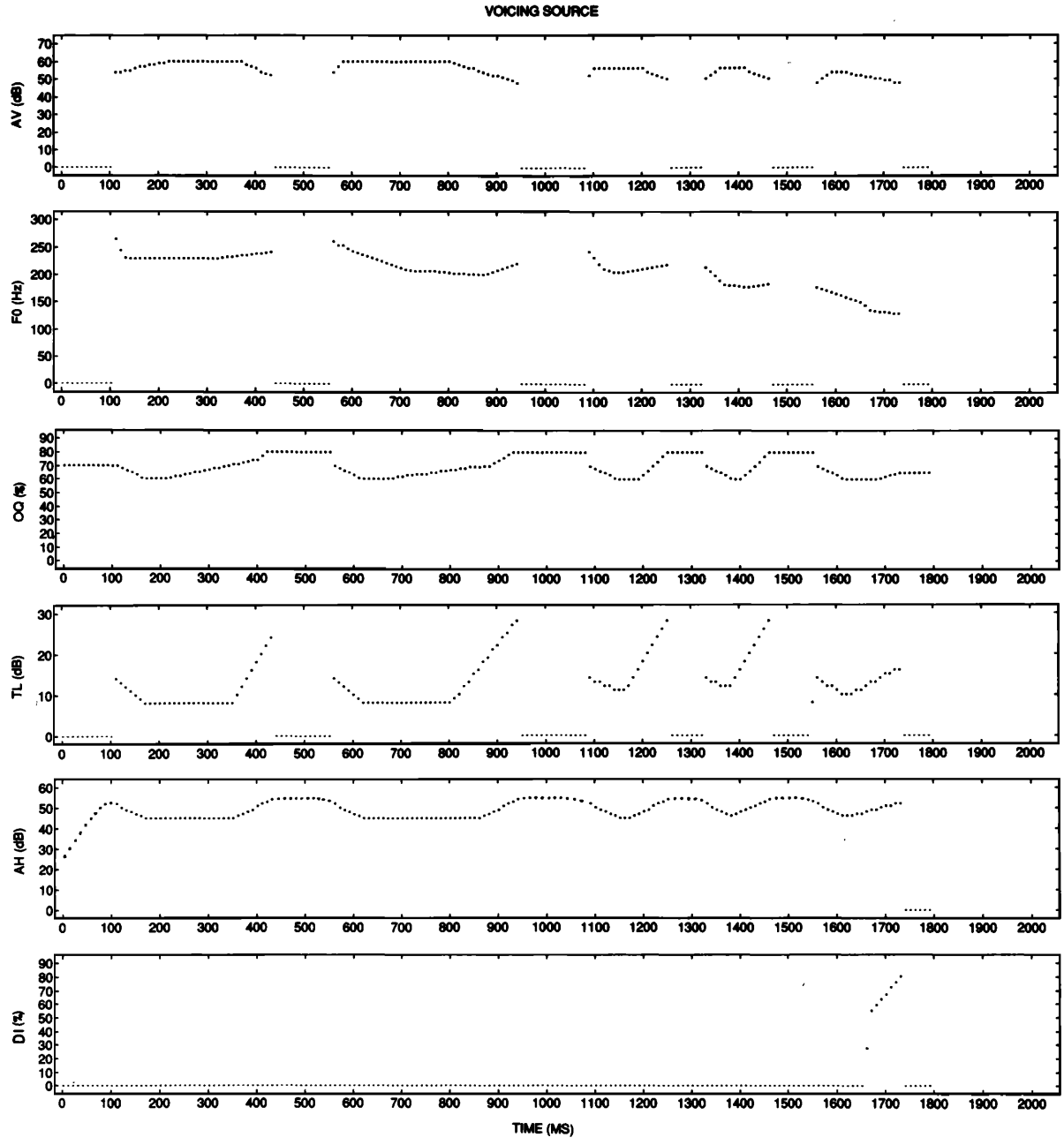


FIG. 16. Synthesis parameter values as a function of time for the [hɑ] reiterant imitation of "Steve eats candy cane" by female speaker LK.

ing a possible distinction between a laryngealized offset accompanied by double pulsing, and a more typical "breathy falling-pitch offset." It is not known whether the ability to mimic details of the sort exemplified by a few periods of double pulsing at voicing offset has much perceptual importance, but it is clear that there exist some voices that display a degree of double pulsing most of the time (Lieberman, 1963). For these voices, it is very likely that the parameter is of perceptual importance.

In summary, during breathy voicing, several of the voicing source parameters change together in such a way as to: (1) increase the relative strength of the first harmonic (by increasing the open quotient **OQ**), (2) reduce the strength of higher harmonics (by increasing spectral tilt **TL**), and (3) add in some aspiration noise at mid and high frequencies (by increasing the amplitude of turbulent aspiration noise **AH**). In addition, the partial abduction of the vocal folds associated with breathiness has effects on the vocal-tract transfer function; the bandwidths of lower formants, especially $F1$, are increased due to increased glottal losses, and additional formantlike spectral peaks and valleys may be introduced into the transfer function due to acoustic coupling to the trachea. These five acoustic effects can be observed not only in a breathy vowel, but also in transitions between normal vowels and voiceless consonants such as [h], or at the offset of a vowel into silence. There are regular

TABLE XIII. Values for synthesis constants that differ from default values for the reference stimulus used in the breathiness perception test.

Parameter	Value
NF	4
DU	300
OQ	65
TL	3
AH	40
F1	800
B1	200
F2	1300
B2	110
F3	2850
B3	180
F4	3700
B4	250

TABLE XIV. Synthesis parameters that have been changed for each of the stimuli contrasted with the reference vowel in the breathiness perception test. GO is the overall gain scale factor, in dB.

(1) Fundamental component boosted 6 dB	FTP and FTZ follow f_0, BTP = 50, BTZ = 100
(2) Fundamental component boosted 10 dB	FTP and FTZ follow f_0, BTP = 50, BTZ = 150
(3) Fundamental frequency lowered initially	f_0 down 10 Hz over first 100 ms
(4) Formant bandwidths increased	B1 = 500, B2 = 170
(5) Spectral tilt down 15 dB at 3 kHz	TL = 15, GO = 65
(6) Spectral tilt down 25 dB at 3 kHz	TL = 25, GO = 67
(7) Aspiration noise of 54 dB added	AH = 54
(8) Aspiration noise of 60 dB added	AH = 60
(9) Spectral tilt of 15 dB and aspiration of 55 dB	TL = 15, GO = 65, AH = 55
(10) Spectral tilt of 20 dB and aspiration of 50 dB	TL = 20, GO = 67, AH = 50
(11) Ditto, plus bandwidth widening and OQ increase	TL = 15, GO = 65, AH = 55, B1 = 400, B2 = 170, OQ = 75

acoustic manifestations of breathiness and laryngealization that occur in predictable locations over the course of typical English sentences. These manifestations have been parametrized and mimicked using a simple synthesis model. Utilization of these types of voice quality cues in speech synthesis by rule should enhance the naturalness of the speech so produced, especially for female voices.

IV. BREATHINESS PERCEPTION TEST USING SYNTHESIS

The stimuli used in the breathiness perception test included 12 synthetic vowels: a reference stimulus and a set of 11 modified stimuli. The reference stimulus was generated by using the default values in Tables XI and XII, except for the constant parameters in Table XIII, and the two time-varying parameters in Fig. 17. The modified stimuli are defined by indicating parameters that change from the reference values, Table XIV, and by showing spectral cross sections of each stimulus in Fig. 18. The reference stimulus was patterned after the LK female voice, but with no cues to breathiness included. Other stimuli add various individual breathiness cues (increased first harmonic amplitude, or bandwidth increases, or added aspiration noise, or spectral tilt at high frequencies, or tracheal pole-zero added) and several combinations of such cues (noise plus tilt, all cues together).

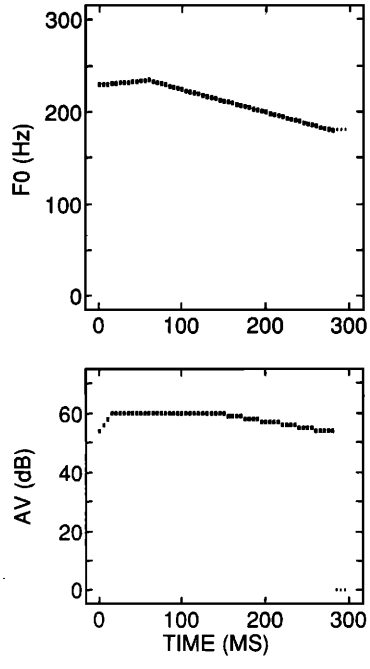


FIG. 17. Values for time-varying parameters for the reference synthetic stimulus used in the breathiness perception test.

The instructions to the subjects were as follows:

"You will hear pairs of synthetic vowels, in which the first vowel, the reference, is always the same. The second vowel of the pair has been modified in some way. Most of the modifications were intended to increase the perceived breathiness of the second member of the pair, but we were clearly not always successful. In fact, some of the changes result in no change, in a nasalized vowel, or in vowel sound qualities that would be difficult for a human to produce. We are interested in two things: (1) the degree to which the second vowel is perceived to be more breathy than the first, and (2) whether the change in sound quality is natural (could be produced by the same human talker who produced the reference, rather than, for example, by computer processing). Since these two judgments are more-or-less independent, we will play the 6-min tape to you twice, first asking for breathiness judgments, and then for naturalness judgments. Because some of the changes seem to have made some of the vowels sound nasalized, during the naturalness rating process, we would also like you to star any item that sounds nasalized.

Instructions before first playing: The breathiness scale will go from 0 (no change in breathiness between reference