have a shorter open quotient than females, except for male TW who is more breathy than two females according to this measure.[10] The range of H1 amplitude across subjects is considerable—from a maximum of 17.1 dB for CB to a minimum of 4.6 dB for male JG—i.e., the range is 12.5 dB. Within-subject variability is generally low, but there are clear examples of particular sentences that are uttered with a different average first-harmonic amplitude. A value in Table III is followed by an asterisk if it differs from the average for that subject by more than 2 dB. There are only seven such examples in the database, suggesting that subjects are free to select differing modes of vibration and degrees of breathiness at will but tend to stay at one mode during an experiment of this sort.

1. Stress and syllable amplitude

Data on rms amplitude measured at syllable midpoint with a 25-ms Hamming window are presented in Table IV. Stressed syllables, on average, are 3.0 dB more intense than unstressed syllables in this corpus for females, and 4.3 dB more intense for males. Some of the difference between stressed and unstressed syllables is simply due to differences in fundamental frequency (a vowel uttered at a higher fundamental has more pulses per second within the analysis window, all else being equal), while part of the difference is presumably due to vocal effort. Other confounding factors are the possibility of vowel reduction and a lowered $F1$ for unstressed vowels, and the possible increased breathiness of unstressed vowels that might, among other effects, increase the first-formant bandwidth and thereby reduce the amplitude of the vowel. Thus, while we can say that there are moderately large differences in intensity between stressed and unstressed vowels, it is not easy to quantify the contribution of the various factors that may be involved.

In addition to a stress effect, there appears to be a general tendency for syllable amplitude to be less at the end of a sentence, even when terminated by a nominally stressed syllable. Across all 16 subjects, the average rms amplitude of syllables 1, 3, and 5 for sentence S1 (ὸὸόσσ) is 58.1, 56.7, and 49.9 dB, and for sentence S2 (σσόὸό) is 55.8, 58.2, and 53.4 dB. If we first subtract 3 dB from each stressed syllable to normalize out the previously described stress effect on syllable amplitudes, the utterance position effect is found to be a 1.0-dB fall from syllable position 1 to position 3, and a more pronounced 4.3-dB fall from syllable position 3 to final position 5. Presumably, this amplitude reduction in utterance-final position is associated with reduced lung volume and is a natural consequence of lowered subglottal pressure at the end of a breath group (Lieberman, 1967). In addition, there appears to be a general relaxation of muscular activity and a preparation for breathing in the larynx musculature that could also contribute to the reduction in voicing source amplitude.

TABLE IV. Rms amplitude in dB at the midpoint of each syllable in several five-syllable reiterant sentences, as averaged across speakers of a given gender.

Female averages Sentence	Syll1	Syll2	Syll3	Syll4	Syll5	Av
S1 [ʔa] (ὸὸόσσ)	58.6	57.0	56.7	52.2	50.8	55.0
S1 [ha](ὸὸόσσ)	58.8	58.4	57.8	54.0	51.6	56.1
S2 [ʔa](σσόὸό)	57.7	54.8	59.1	58.0	55.7	57.1
S2 [ha] (σσόὸό)	57.2	53.3	58.2	56.2	52.6	55.5
Male averages Sentence	Syll1	Syll2	Syll3	Syll4	Syll5	Av
S1 [ʔa] (ὸὸόσσ)	56.8	55.7	55.3	50.7	48.0	53.3
S1 [ha](ὸὸόσσ)	57.5	55.2	55.8	50.7	47.7	53.4
S2 [ʔa] (σσόὸό)	53.0	51.6	58.2	56.2	54.4	54.7
S2 [ha] (σσόὸό)	52.8	51.7	56.3	55.0	50.0	53.2

The fundamental frequency was also measured at vowel midpoint. The average f_0 of females, relative to the male data, was found to be remarkably systematic—i.e., about 1.7 times that of males for each vowel position in each sentence. Within each gender, there was considerable variation in average f_0. It is possible that f_0 could be a secondary perceptual cue to judged breathiness. For example, a high f_0 might indicate a desire to be "feminine" and thus breathy, or a very low f_0 might be indicative of laryngealization. However, correlation data to be presented below indicate that, within each sex, average f_0 of a talker is not at all correlated with perceived breathiness rating.

In summary, the most striking aspect of first-harmonic amplitude is the large variation across speakers of a particular gender. On average, the relative amplitude of the first harmonic for females is about 6 dB greater than that for males, but, within the class of female speakers, the range is over 10 dB. As an indirect measure of open quotient, first-harmonic amplitude indicates that the final vowel of a reiterant utterance is likely to be slightly laryngealized (reduced open quotient), and lower in overall amplitude.

C. Results II: Aspiration noise in the *F3* region of the spectrum

A second potential acoustic correlate of the degree of breathiness of a vowel is the amount of noise present at higher frequencies in the spectrum. One way to estimate the relative strength of noise components is to isolate the third formant, using a bandpass filter.[11] The filtered waveform can be displayed, as in Fig. 4, and examined visually to determine whether the waveform is periodic (repeating itself identically) or has indications of random variation due to the introduction of noise. Note that it is difficult to determine whether noise is present by examining the unfiltered speech waveform shown in parts (2a) and (3a) in Fig. 4 because the periodic components at low frequencies dominate the visual impression due to their greater energy.

Reiterant sentences involving [ha] were processed in three steps: (1) a broadband spectrogram was produced, and the frequency of the third formant, $F3$, was estimated visually (see column 2 of Table V); (2) a four-pole Butterworth bandpass filter, having a center frequency set equal to $F3$ and a bandwidth of 600 Hz, was used to create a filtered version of the original digitized waveform; and (3) a plot of the waveform was examined subjectively to determine the degree of random noise present. A four-step scale, described in Table V, was used to quantify the presence or absence of

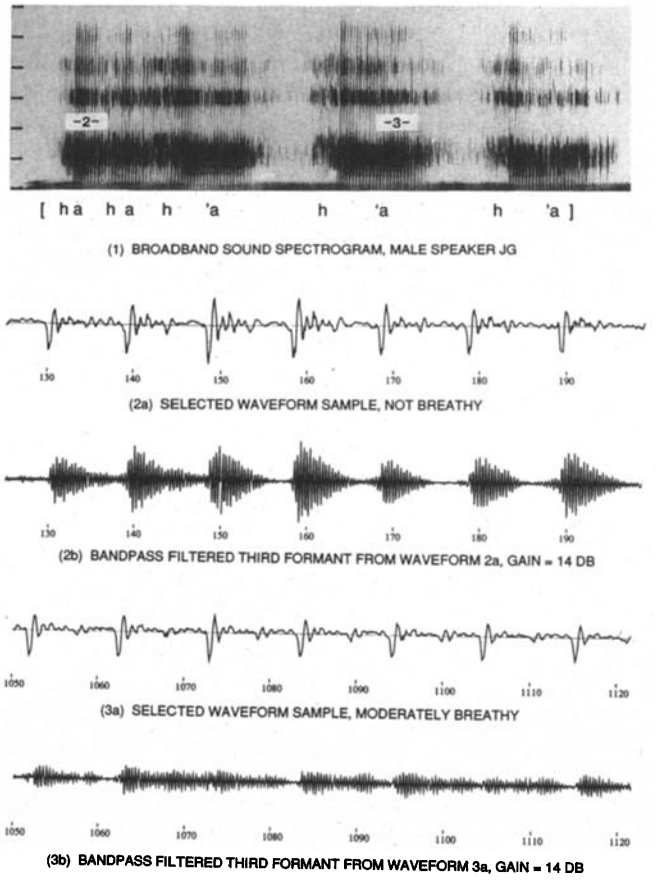

FIG. 4. (1) Broadband spectrogram of male speaker JG indicating locations where waveform samples have been extracted to show (2) a vowel with little or no aspiration noise, and (3) a vowel with appreciable aspiration noise, as evidenced by the presence/absence of noise in the third formant region of the spectrum.

noise over the course of each vowel. If the filtered $F3$ waveform consisted of a periodic damped sinusoid, in synchrony with the unfiltered waveform, the vowel was judged to be periodic and free from aspiration noise. If there was no visible periodicity in synchrony with the original waveform, the vowel was judged to have strong aspiration noise. While it has been assumed that the noise source is at the glottis (aspiration), it is possible that some speakers employ a pharyngeal constriction to augment [h] noise with the frication noise of a pharyngeal fricative before a vowel such as [ɑ].

The degree of aspiration noise in each of the five syllables of the reiterant utterance S1, "Steve eats candy cane," is summarized in the middle five columns of Table V. This rating of noise presence in $F3$ for each syllable, as estimated by the first author, is tabulated separately for female and male subjects.

Comparing the noisiness rating across syllable position, we observe that noise increases toward the end of the utterance. This tendency may be a characteristic of all utterances, or it may be due to the fact that this particular utterance ends with two unstressed syllables. We will return to the question of whether a syllable tends to have more aspiration noise if unstressed and/or if in utterance-final position after examining data from the utterance S2, "The debate hurt Bob."

TABLE V. Degree of periodicity versus noise excitation of $F3$ for [hɑ] reiterant imitations of sentence S1, "Steve eats candy cane." The 4-point subjective scale ranges from: (1) periodic, no visible noise, (2) periodic but occasional noise intrusion, (3) weakly periodic, clear evidence of noise excitation, and (4) little or no periodicity, noise is prominent.

Females	$F3$	Syll1	Syll2	Syll3	Syll4	Syll5	Av
KK	2300	1	1	1	2	2	1.4
LK	2900	3	3	3	4	4	3.4
CB	2950	2	3	3	3	4	3.0
LG	2650	3	2	3	3	3	2.8
SS	2800	2	3	3	3	4	3.0
LL	2550	1	2	1	2	2	1.6
ND	2400	2	3	3	4	3	3.0
SH	2700	2	2	2	2	2	2.0
CE	2950	3	2	3	4	4	3.2
JW	2900	3	3	4	4	3	3.4
Av		2.2	2.4	2.6	3.1	3.1	2.7
Males							
KS	2700	1	1	1	1	2	1.2
MR	2200	1	1	2	2	3	1.8
JG	2700	1	2	2	3	3	2.2
JP	2350	1	1	1	1	2	1.2
MP	2600	1	1	1	1	1	1.0
TW	2200	1	1	2	4	4	2.4
Av		1.0	1.2	1.5	2.0	2.5	1.7

Comparing the noisiness rating between female and male subjects, we see that females generate more noise, on average, than males. The average noisiness rating for females is 2.7, while for males it is 1.7. However, there is a wide range of degrees of noise presence within each gender. The noise measurements for the female speakers KK, LL, and SH are approximately the same as those for the average male. The remaining female speakers exhibit even more noise than the average difference between the sexes would imply.

A reiterant sentence with a different stress pattern, "The debate hurt Bob," was analyzed in the same way as the sentence in Table V. The results are summarized in Table VI, which indicated the noisiness rating for each syllable as tabulated separately for female and male subjects.

Comparing the average noisiness rating across syllable position, we see that noise presence increases for unstressed syllables, and that noise presence increases slightly toward the end of the utterance. Factoring out these two effects from the S1 and S2 data sets, we find a stress difference (unstressed more noisy) of 0.6, and a final position effect (more noisy relative to utterance-initial position) of 0.55 subjective units. While these tendencies are not large compared with individual differences, they reflect effects that may be important for the synthesis of natural variation in voicing source characteristics over a sentence.

There seems to be a paradox when comparing the "noise-in-$F3$" measure at utterance offset with the "first-harmonic-amplitude" measure described in the previous section. In this section, we find that there is more noise, indicating a greater glottal airflow in an utterance-final syl-

lable, but in the previous section we found a weaker first-harmonic amplitude in an utterance-final syllable, indicative of a pressed voice with a slightly shorter duration open quotient. We conjecture that the natural tendency to open the larynx in preparation for breathing at the end of an utterance indeed occurs in virtually all cases, resulting in an increase in posterior glottal chink size and an increase in aspiration noise. However, most speakers simultaneously rotate the anterior tips of the arytenoid cartilages inward, presumably to maintain voicing, but actually partially laryngealize, leading to a somewhat novel breathy-laryngealized mode of vibration.

The average difference between males and females is not as great in Table VI (0.4 subjective units) as it was in Table V (1.0 subjective units). However, those female speakers who show less $F3$ noise in Table V also have less noise for the second sentence. Again, individual variation within a gender is large compared with average differences between genders.

In summary, aspiration noise is very commonly present in the waveforms extracted from the third-formant region throughout the vowel portion of [hɑ] reiterant utterances from both sexes. The aspiration noise can strongly dominate harmonic excitation, implying a changed voicing vibration pattern in which the harmonic spectrum is tilted down at high frequencies with respect to normal voicing. On average, there is more noise infusion in female than male utterances, but three of the females are not very breathy by this measure. Variation in amount of aspiration noise in $F3$ across an utterance appears to be systematic in that there is more noise; i.e., the glottis can be inferred to be slightly more spread, in unstressed syllables and toward the end of a breath group.

TABLE VI. Degree of periodicity versus noise excitation of $F3$ for [hɑ] reiterant imitations of sentence S2, "The debate hurt Bob." The four-point subjective scale ranges from: (1) periodic, no visible noise, (2) periodic but occasional noise intrusion, (3) weakly periodic, clear evidence of noise excitation, and (4) little or no periodicity, noise is prominent.

Females	$F3$	Syll1	Syll2	Syll3	Syll4	Syll5	Av
KK	2500	2	1	2	1	1	1.4
LK	2900	4	3	2	4	3	3.2
CB	3000	2	4	2	2	3	2.6
LG	2600	3	3	3	3	2	2.8
SS	2750	3	4	2	3	3	3.0
LL	2550	2	3	2	2	2	2.2
ND	2500	1	4	2	2	2	2.2
SH	2450	1	2	1	1	1	1.2
CE	2700	4	4	2	3	3	3.2
JW	2800	3	3	3	3	2	2.8
Av		2.5	3.1	2.1	2.4	2.2	2.5
Males							
KS	2700	1	1	1	1	1	1.0
MR	2100	4	3	2	3	2	2.8
JG	2750	1	3	2	3	4	2.6
JP	2400	2	2	2	1	2	1.8
MP	2500	2	1	1	1	1	1.2
TW	2250	2	3	3	3	4	3.0
Av		2.0	2.2	1.8	2.0	2.3	2.1

D. Results III: Tracheal coupling

The acoustic effects of tracheal coupling on the normal transfer function of the vocal tract for a vowel include (1) possible addition of poles and zeros associated with the tracheal and lung system below the glottis and (2) increased losses at the glottal termination, which primarily affect the first-formant bandwidth. It is difficult to objectively quantify the extent to which these potential perturbations are present in the data from our 16 speakers. Therefore, we will present as much of the primary data as possible when describing our interpretations of the data in the following analysis.

1. Extra poles and zeros

The best way to get an idea of possible locations of tracheal poles and zeros is to choose a speech sound in which the glottis is as open as possible, subject to the constraint that there is still sound generation at the larynx. Therefore, aspiration spectra during the production of the [h] portions of the [hɑ] reiterant sentence "Steve eats candy cane" were obtained and analyzed. An example of the analysis process is shown in Fig. 5. Aspiration spectra were produced using a 51.2-ms rectangular window in order to obtain a stable estimate of the spectrum of a random process (Shadle, 1987). Note the extra pole at about 2100 Hz in Fig. 5, which is prominent in the aspiration noise spectrum and is visible as a local spectral maximum between $F2$ and $F3$ in the harmonic spectrum of the following vowel.

Samples of individual [h] noise data are presented in Fig. 6, which summarize spectra of the [h] aspiration before the second vowel of the [hɑ] reiterant imitation of "Steve eats candy cane." Peaks associated with $F2$ and $F3$ of the following vowel are identified in the figure. A clear $F1$ peak is usually not visible due to increased losses and the falloff in

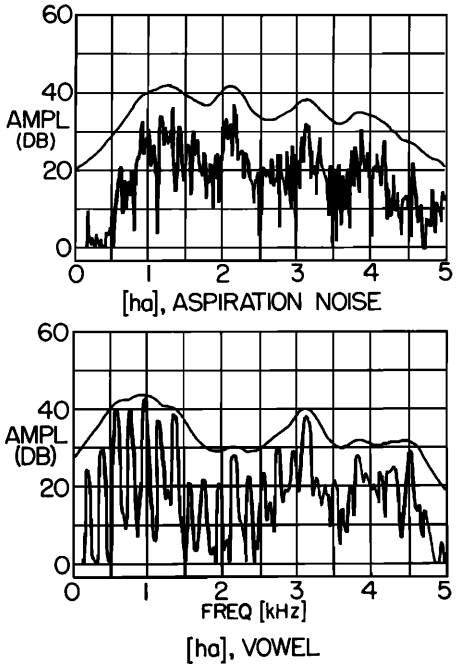

FIG. 5. A tracheal resonance at 2100 Hz is identified in an aspiration spectrum (top) and in the initial portion of the following vowel (bottom).

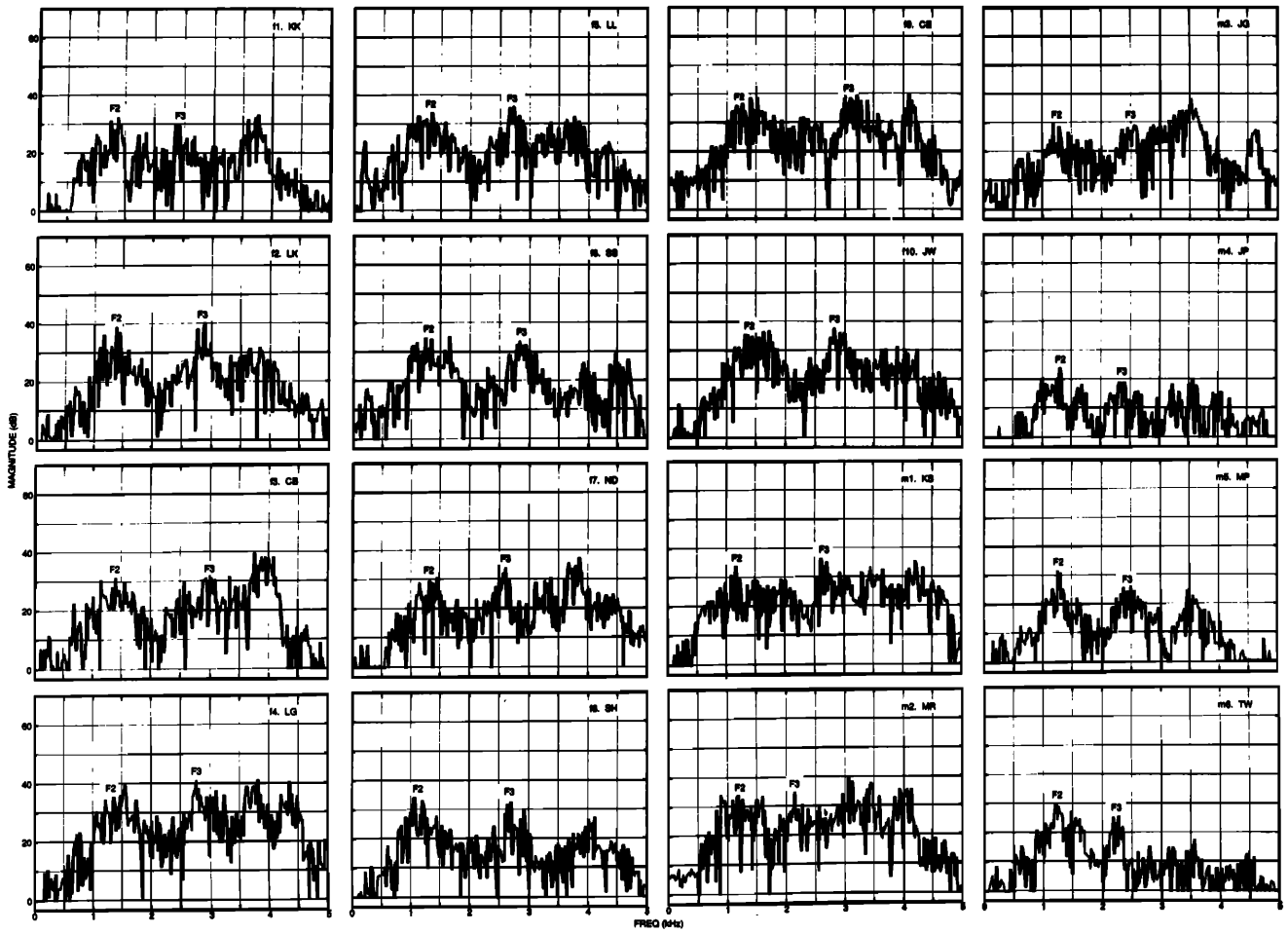

FIG. 6. Fifty-ms dft magnitude spectra of [h] aspiration noise from 16 speakers, as obtained just prior to the second vowel in [hɑ] reiterant imitations of the sentence "Steve eats candy cane."

aspiration source spectral energy at low frequencies. An appreciation for the importance of cross-speaker variability in developing an understanding of speech perception can be obtained by study of the figure. For example, a single spectral template representative of [h] before [ɑ] is unlikely to be able to account for the presumed perceptual similarity between these spectra even if powerful normalization procedures are applied to the spectral data prior to comparison with the idealized template.

Examination of aspiration spectra has revealed extra poles and zeros that often creep into adjacent vowel spectra whenever the glottis is partially spread and tracheal pole-zero coupling is possible. Therefore, as a second step in the analysis, each vowel following [h] was examined for evidence of extra poles and zeros. A 25-ms windowed dft spectrum obtained at [ɑ] vowel midpoint following the [h] is plotted for each subject in Fig. 7. The figure indicates that it is sometimes possible to see an extra formantlike peak at about 2000 Hz. When present, this peak could easily be confused with a normal formant. Other expected peaks at lower frequencies were not as easy to detect, perhaps due to the presence of close-by formants, as well as the fact that the high f_0 of females provides relatively few harmonics from which to determine spectral shape of the vocal-tract transfer function. At vowel midpoint, the effects of tracheal coupling are not great for most speakers, but during the initial and final parts of the vowel, greater departures from normal all-pole spectra were more frequently seen.

Peaks in the aspiration spectrum were compared with peaks in the spectrum of the following vowel; any peak that could not be associated with a formant of a normal [ɑ] vowel was assumed to be a tracheal resonance.[12] These extra resonances are listed in Table VII for each speaker. Also listed are the locations of prominent dips in the aspiration spectrum, which are presumably caused by tracheal zeros. Extra peaks and dips of uncertain status (either weak or present in only one or two of the five [h] tokens analyzed in the reiterant sentence) are indicated in the table by a parentheses notation. In general, the results are fairly consistent across speakers. If a pole or zero is visible, it tends to be at about the same frequency location for each speaker. An average (median) calculation across speakers of each sex yields clear poles at 1650 and 2350 Hz, as well as weak indications of poles at 750 and 3150 Hz for the females. These values are consistent with results published previously on the frequency locations of tracheal poles (Ishizaka *et al.*, 1976; Cranen and Boves, 1987). Values for our male speakers are slightly lower in frequency, as would be expected given the larger body size of males. The frequency locations of zeros associated with tracheal coupling are usually close

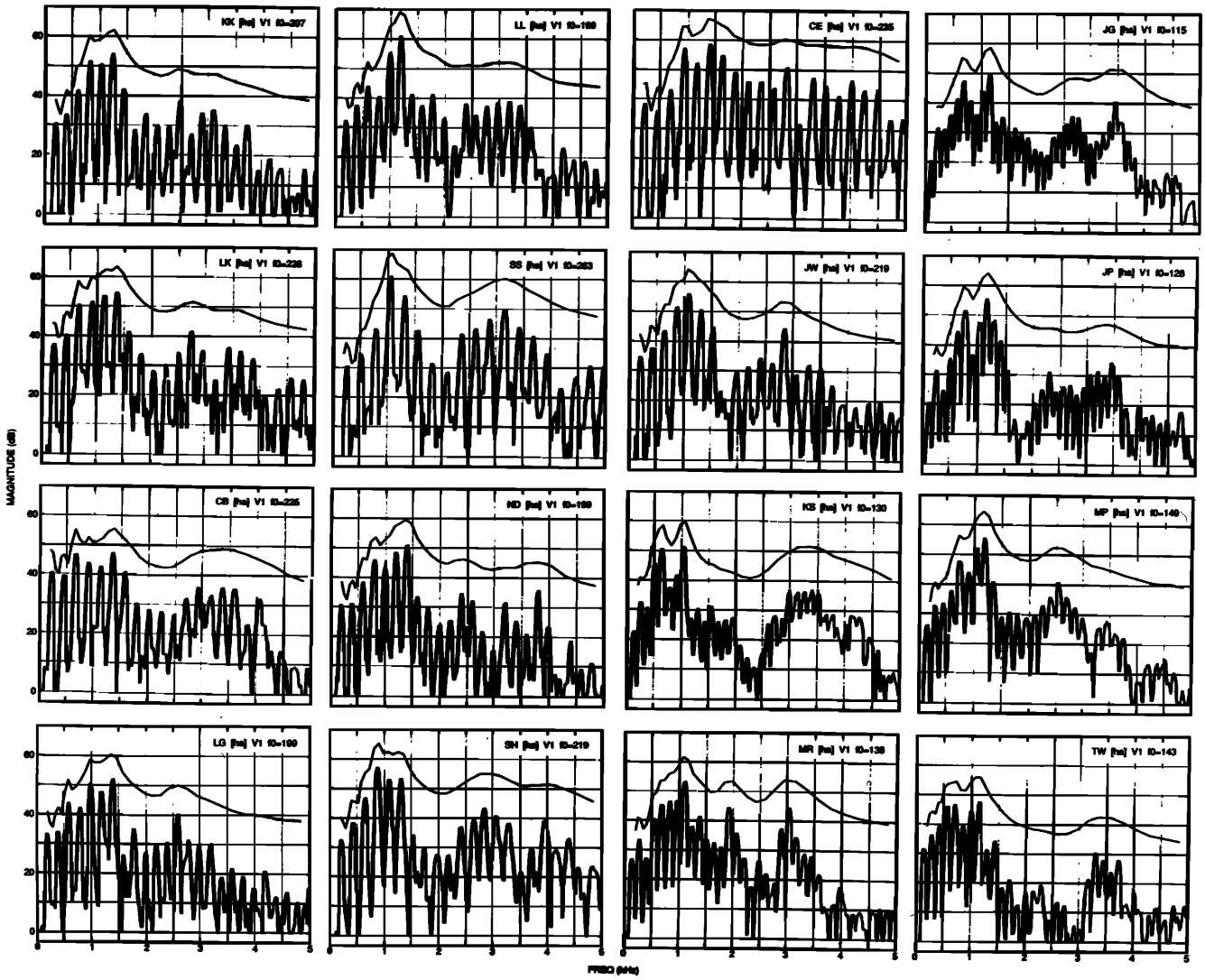

FIG. 7. Twenty-five-ms dft magnitude spectra of [ɑ] obtained from 16 speakers. Spectra were sampled at the midpoint of the vowel in the first syllable in [hɑ] reiterant imitations of the sentence "Steve eats candy cane." The smooth curve was produced by averaging dft energy over a critical band.

TABLE VII. Frequency locations of extra (tracheal) poles and zeros in aspiration spectra for ten female and six male talkers. Values in parentheses are of uncertain status. See text.

Females	Poles P1	P2	P3	P4	Zeros Z1	Z2	Z3	Z4
KK	(750)	1800	2650	(3150)	(900)	1550	2100	2900
LK	700	(1650)	(2500)		850	(1800)	2100	3200
CB		1700	(2600)		900	(1850)	2200	
LG	750	(1600)		3100	850	1800	(2300)	
SS		1650	2350		900	1950	2400	
LL	700	1650	2400	3250	850	(1800)	2200	3050
ND	750				(900)			
SH		1800	2600			1550	2500	3200
CE	800	(1650)	2300	2600	950	1750	2700	
JW		1700	2400		900		2200	3100
Median:	(750)	1650	2350	(3150)	900	1800	2200	(3100)
Males								
KS		1500	2000			1300	1750	2400
MR		1500		3400		1800		3200
JG		1400	1700	3250			2100	3050
JP		1650	2550	(3300)		1800	(2050)	(3000)
MP		(1600)	(2200)			(1800)	(2050)	
TW		1600	2800	(3200)		(1400)	1900	2400
Median:		1550	2200	3275		1800	2050	3000

to an observed extra pole; the average frequency locations for females are 900, 1800, 2200, and 3100 Hz.[13] Data for males are similar.

In summary, extra tracheal poles often distort vowel spectra to varying degrees. A pole at about 2100 Hz is frequently seen in the [ɑ] vowel for our ten female talkers. When present, tracheal resonances tend to be located at frequencies consistent with previous measurements of pole locations.

2. Bandwidth of F1

The bandwidth of the first formant of the vocal-tract transfer function determines several aspects of the acoustic output. It determines the width of the resonance peak, and it determines the relative strength or prominence of the first-formant peak. An extreme example of the effect of increased B1 on the spectrum of a vowel in our corpus is shown in Fig. 8. There is virtually no indication of the presence of $F1$ in the spectrum for the example of [ɑ] in the right panel. In an attempt to quantify the variation in prominence of the $F1$ spectral peak across speakers of our database, we have defined two measures. The first is the amplitude of the $F1$ peak relative to some reference amplitude. The second is a subjective estimate of how easy it is to see the location of $F1$ in the spectrum. Both measures were applied to the spectrum sampled at the midpoint of the first vowel in [hɑ] reiterant imitations of "Steve eats candy cane" (sentence S1). $A1_{re2}$ is an indirect estimate of first-formant bandwidth, as indicated by the amplitude in dB of the first formant, relative to amplitude of second formant, as measured at the beginning, middle, and end of the first vowel in [hɑ] reiterant imitations of sentence S1. $F1_{vis}$ is an indirect estimate of the first-formant bandwidth, as indicated by the visibility of a local $F1$ maximum in spectrum: 2 = obvious local spectral maximum, 1 = inflection point in smoothed spectrum, 0 = no evidence of $F1$ peak. $F1_{vis}$ is the sum of this measure at the beginning, middle, and end of the first vowel in [hɑ] reiterant imitations of S1.

Results of this analysis are presented in Table VIII. Due to natural variability in the levels of formants, it appears that $A1_{re2}$ is not a particularly useful measure of the distinctiveness of the $F1$ peak. In any case, there is no difference between the genders in this measure. It also appears that we might have been better off looking at a short unstressed vowel, because it is more likely to be influenced by glottal opening for adjacent consonants and thus have a less distinct first-formant peak.

TABLE VIII. Two measures of the prominence of the $F1$ peak in vowel spectra adjacent to [h]: the amplitude of the first formant relative to the amplitude of the second formant, and the subjective visibility of $F1$ as a distinct local spectral maximum.

Speaker	$A1_{re2}$	$F1_{vis}$
KK	− 1.0	5
LK	− 2.0	6
CB	− 1.3	5
LG	− 3.3	5
LL	− 8.0	5
SS	− 6.0	2
ND	− 4.0	6
SH	3.7	6
CE	− 2.3	6
JW	− 0.3	3
Av	− 2.5	5
KS	0.7	6
MR	− 4.0	5
JG	− 0.3	6
JP	− 3.0	6
MP	− 5.7	5
TW	− 3.3	5
Av	− 2.7	5.5

Hawkins and Stevens (1985) have shown that vowel nasalization can have a similar flattening effect on the spectrum of $F1$, due to increased losses in the nasal tract and due to the splitting up of $F1$ into a pole–zero–pole complex. Thus it is not clear what perceptual attributes to assign to the change in the spectrum associated with an increase in B1 (breathiness or nasality). We will return to this issue in the design of a perceptual experiment.

In summary, the partially open glottis of a breathy vowel can cause the first-formant bandwidth to increase, sometimes obliterating the spectral peak at $F1$ entirely. This effect and the appearance of extra tracheal pole-zero pairs can cause serious problems for formant trackers and for models of perception that presuppose a formantlike representation of speech sounds for men, women, and children.

E. Correlation analysis of acoustic and perceptual data

A listening test was prepared in which reiterant imitations of "Steve eats candy cane" for the 16 speakers were randomized and played to a panel of eight listeners who rated the breathiness of the speakers' vowels on a seven-point scale. The scale is defined at the top in Table IX. Judgments were obtained for reiterant imitations using both [ʔɑ] and

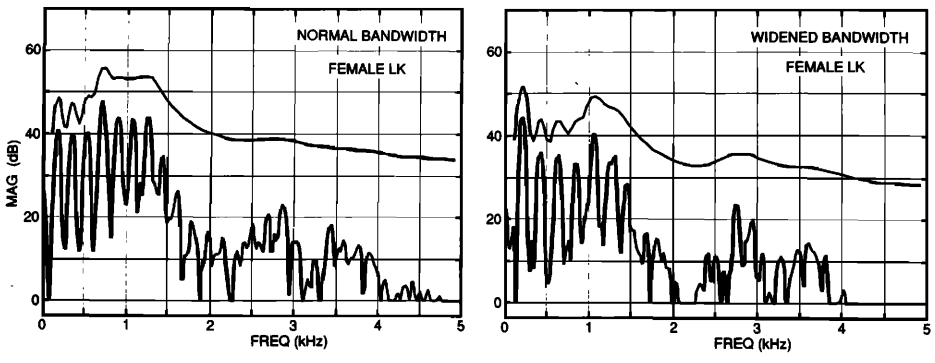

FIG. 8. Spectra of normal and breathy versions of the vowel [ɑ] are compared. The breathy token (right panel) provides an example in which the first formant bandwidth is increased to a point where it is difficult to see any evidence of a spectral peak in the expected location of $F1$ (about 750 Hz).

TABLE IX. Average ratings of perceived breathiness in vowel portions of the [ʔɑ] and [hɑ] reiterant imitation of "Steve eats candy cane," using the 7-point scale defined below. The rightmost column indicates average breathiness ratings for a vowel excised from the [hɑ] imitation.

Rating	Description		
1	not breathy		
2			
3	slight breathiness for some or all syllables		
4			
5	moderate breathiness for some or all syllables		
6			
7	strongly breathy		
Speaker	[ʔɑ] sentence	[hɑ] sentence	Excised vowel
KK	2.6	2.7	3.6
LK	5.2	5.0	4.6
CB	5.5	5.8	6.0
LG	3.8	5.1	5.0
LL	3.1	4.3	3.5
SS	2.1	3.7	2.8
ND	3.6	4.0	3.8
SH	1.8	2.7	2.9
CE	4.6	6.0	3.7
JW	4.6	5.4	3.2
Av	3.7	4.4	3.9
KS	2.7	3.5	2.7
MR	2.0	5.0	2.3
JG	3.0	4.7	3.0
JP	3.1	3.4	2.8
MP	2.2	2.0	2.3
TW	4.0	4.9	5.3
Av	2.8	3.9	3.1

[hɑ] syllables; average results based on four separate randomizations of each block of 16 trials are presented in the Table.

As general trends, it can be seen that females are judged, on average, to be slightly more breathy than males, and that sentences involving the [hɑ] syllable are perceived to be more breathy than sentences involving [ʔɑ] (even though instructions to the subjects were to rate the breathiness of the vowel portions of the utterances). Subjects judged to be most breathy include females CE, CB, JW, LG, and LK, and males TW, MR, and JG.

The perceptual judgment data may be contaminated by several factors not directly related to the breathiness of individual vowels. For example, the level of the aspiration noise could influence vowel judgments, as could the details of transitions between voicing and voicelessness. For this reason, we performed a second listening test in which the first vowel was excised from the [hɑ] reiterant imitation of "Steve eats candy cane." The first three glottal pulses and the last three pulses were deleted from the vowel and the remaining vowel had its onset and offset modified by a 10-ms half-Hanning window. A panel of three listeners produced four judgments of breathiness for each speaker. The averages are shown in the last column of Table IX.

The pattern of breathiness rankings for subjects changes a bit in this new test; CE, CW, and MR are not perceived to be as breathy as before. The correlation between the two sets of perceptual data in columns 2 and 3 of Table IX is only 0.55. Informal inspection of the acoustic data suggests that the level of the aspiration noise may well account for the perceptual difference.

Correlations between subjective breathiness ratings and a number of acoustic measures are presented in Table X. It is impossible to determine causation from such an analysis, but the few correlations reaching significance are easily interpreted in familiar terms. The acoustic measures are defined below:

PRC_1: group results of breathiness judgments for the first vowel excised from [hɑ] reiterant imitation of S1;

PRC_5: group results of breathiness judgments for [hɑ] reiterant imitation of S1;

$F0_{mid}$: fundamental frequency in Hz, as measured at the midpoint of the first vowel in [hɑ] reiterant imitation of S1;

$H1_{re2}$: amplitude of first harmonic, in dB, at first vowel midpoint in [hɑ] reiterant imitation of S1, relative to second-harmonic amplitude, offset by 10 dB to make positive;

$NOIS_1$: degree of breathiness noise visually present in $F3$ waveform, judged by DK in first vowel of [hɑ] reiterant imitation of S1;

$NOIS_5$: degree of breathiness noise visually present in $F3$ waveform, judged by DK, average of all five vowels of [hɑ] reiterant imitation of S1;

ASP_1: rms level, in dB, of aspiration noise during the [h] preceding the first vowel, relative to overall rms level of the first vowel, as measured at vowel midpoint, in the [hɑ] reiterant imitation of S1;

ASP_2: rms level, in dB, of aspiration noise during the [h] preceding the second vowel, relative to overall rms level of the second vowel, as measured at vowel midpoint, in the [hɑ] reiterant imitation of S1;

$A1_{re2}$: indirect estimate of first-formant bandwidth, as indicated by the amplitude in dB of the first formant, relative to amplitude of second formant, as measured at the beginning, middle, and end of the first vowel in [hɑ] reiterant imitation of S1, average, in dB;

$F1_{vis}$: another indirect estimate of the first-formant bandwidth, as indicated by the visibility of a local $F1$ maximum in spectrum: 2 = obvious local spectral maximum, 1 = inflection point in smoothed spectrum, 0 = no evidence of $F1$ peak; sum of measurements at beginning, middle, and end of first vowel in [hɑ] reiterant imitation of S1;

$A3_{re2}$: estimate of general spectral tilt above 1 kHz, as indicated by the amplitude of the third formant, in dB, relative to the amplitude of the second formant, average of measurements at beginning, middle, and end of the first vowel in [hɑ] reiterant imitation of S1;

$A5_{re2}$: another estimate of general spectral tilt above 1 kHz, as indicated by the amplitude of the third, fourth, or fifth formant (whichever is the greatest), in dB, relative to the amplitude of the second formant, average of measurements at beginning, middle, and end of the first vowel in [hɑ] reiterant imitation of S1.

Only two correlations with the perceptual judgments reach statistical significance: The amplitude of the first harmonic relative to H2 ($H1_{re2}$) is closely tied to subjective breathiness of the isolated first vowel (PRC_1), and the

TABLE X. Selected acoustic correlates of breathiness for ten female and six male talkers (top), and correlation coefficients between two subjective measures of perceived breathiness and these acoustic measurements (bottom).

SPKR	PRC_1	PRC_5	$F0_{mid}$	$H1_{re2}$	$NOIS_1$	$NOIS_5$	ASP_1	ASP_2	$10A1_{re2}$	$F1_{vis}$	$A3_{re2}$	$A5_{re2}$
KK	3.6	2.7	207	13.0	1	1.4	−14	−19	−10	5	−13	−13
LK	4.6	5.0	228	12.6	3	3.4	−23	−14	−20	6	−11	−11
CB	5.9	5.8	225	17.1	2	3.0	−15	−7	−13	5	−6	−6
LG	4.9	5.1	199	12.6	3	2.8	−8	−5	−33	5	−13	−13
LL	3.5	4.3	199	10.3	2	3.0	−23	−17	−80	5	−15	−15
SS	2.8	3.7	263	9.4	1	1.6	−14	−18	−60	2	−13	−13
ND	3.8	4.0	199	11.9	2	3.0	−8	−15	−40	6	−15	−14
SH	2.9	3.7	219	8.4	2	2.0	−19	−19	37	6	−8	−8
CE	3.7	6.0	235	12.5	3	3.2	−14	−12	−23	6	−9	−9
JW	3.2	5.4	219	11.8	3	3.4	−12	−10	−3	3	−14	−14
Av:	3.9						−15	−14	−25	5	−12	−12
KS	2.7	3.5	130	5.8	1	1.2	−7	−9	7	6	−14	−8
MR	2.3	5.0	138	5.3	1	1.8	−7	−8	−40	5	−8	−7
JG	3.0	4.7	115	4.6	1	2.2	−12	−9	−3	6	−9	−4
JP	2.8	3.3	128	4.9	1	1.2	−26	−23	−30	6	−21	−21
MP	2.3	2.0	149	6.0	1	1.0	−24	−20	−57	5	−17	−17
TW	5.3	4.9	143	9.7	1	2.4	−9	−13	−33	5	−20	−15
Av:	3.1						−14	−14	−27	5.5	−15	−12

	PRC_1	PRC_5	$F0_{mid}$	$H1_{re2}$	$NOIS_1$	$NOIS_5$	ASP_1	ASP_2	$A1_{re2}$	$F1_{vis}$	$A3_{re2}$	$A5_{re2}$
PRC_1	···	0.55	0.30	0.83	0.41	0.59	0.14	0.39	0.01	0.08	0.09	0.07
PRC_5		···	0.28	0.57	0.63	0.81	0.35	0.72	0.11	−0.01	0.47	0.45
$F0_{mid}$			···	0.50	0.60	0.50	−0.10	−0.06	−0.03	−0.43	0.36	0.03
$H1_{re2}$				···	0.54	0.66	0.24	0.45	0.08	−0.03	0.35	0.23
$NOIS_1$					···	0.84	−0.03	0.34	0.12	0.03	0.30	0.06
$NOIS_5$						···	0.11	0.45	−0.01	0.00	0.31	0.18
ASP_1							···	0.70	0.19	−0.12	0.22	0.44
ASP_2								···	0.19	0.00	0.49	0.63
$A1_{re2}$									···	0.33	0.40	0.50
$F1_{vis}$										···	0.07	0.20
$A3_{re2}$											···	0.88
$A5_{re2}$												···

amount of noise replacing harmonics in the $F3$ region of the spectrum for the five vowels of the [hɑ] reiterant imitation of S1 ($NOIS_5$) is closely tied with subjective breathiness of the entire [hɑ] reiterant imitation of S1 (PRC_5).

Low correlations with f_0, in spite of a known tendency for females to be heard as more breathy than males, must mean that, within each gender, f_0 is a poor predictor. Low correlations with $NOIS_1$, the estimate of noise in $F3$ for the first vowel of the sentence, may be due, in part, to the highly quantized nature of the data; for example, all males were assigned a value of 1 (periodic with no evidence of noise). Stronger correlations with $NOIS_5$, the average over five vowels of the same measure, support this conjecture. Low correlations with ASP_1, the level of aspiration noise in the [h] preceding the first vowel, appear to indicate that this level is highly variable and not at all predictive of the noise level at the middle of the first vowel ($NOIS_1$). The correlation of $NOIS_1$ with ASP_2, the level of the aspiration noise in the [h] preceding the second vowel, is somewhat higher, reinforcing a general observation that the first consonant of an utterance varies more in level than do utterance-internal consonants.

The correlations suggest that both (1) the relative amplitude of the first harmonic and (2) the presence of noise during voicing affect judgments of breathiness, but we cannot conclude that these are the only factors involved. Instead, we will use a speech synthesis experiment to establish the perceptual importance of manipulations to a large set of variables.

We turn next to a description of a new speech synthesizer that was developed on the basis of recent research on source mechanisms in speech production. Perceptual experiments using this synthesizer will then be described.

II. SOURCE MODELS FOR SPEECH SYNTHESIS

As indicated in Fig. 9, the effective sound source during vowel production is the glottal volume velocity waveform, $U_g(t)$. This waveform acts as the input to the vocal-tract transfer function, which introduces resonant structure to the output lip volume velocity. Sound pressure measured some distance from the lips is then proportional to the temporal derivative of lip volume velocity (Fant, 1960).

Recent efforts to characterize the essential features of the voicing source waveform $U_g(t)$ for different male and

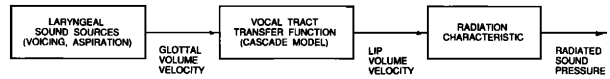

FIG. 9. Block diagram illustrating the acoustic theory of speech production.

female voices have led to several new parametric models of glottal output (Ananthapadmanabha, 1984; Fant, 1979, 1982b; Titze, 1984; Fant *et al.*, 1985; Allen and Strong, 1985; Fujisaki and Ljungqvist, 1986; Klatt, 1987b; Rosenberg, 1971, 1975).

In an updated version of a laboratory formant synthesizer (which we have called KLSYN88, to distinguish it from the older version KLSYN), two different new models of the glottal source have been incorporated. One of these is a slightly modified version of the Liljencrants–Fant (LF) model (Fant *et al.*, 1985). The other model, which has some characteristics in common with the LF model but incorporates some additional features, is called KLGLOTT88.

In the models, the characteristics of the waveform are 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 and (2) **TL**, spectral tilt—or the additional spectral change associated with "corner rounding" in which closure is nonsimultaneous along the length of the vocal folds.

A. The modified LF model

The LF model was chosen over other candidates because both Fant *et al.* (1985) and Fujisaki and Ljungqvist (1986) have shown it to be superior to other models of the same complexity when the objective is to model natural speech with minimum rms error. 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).[14]

The LF model does not consider the possible importance of turbulence noise generation at the glottis due to a constant flow leakage between partially spread arytenoid cartilages. Data from Holmberg *et al.* (1988) indicate that dc flows are very common for male and female speakers of English in the environment of aspirated stops. If the dc flow component causes the generation of simultaneous aspiration noise at the glottis, as is very likely, then the modeling of $U_g(t)$ should include provisions for introduction of aspiration noise (Pandit, 1957; Fujimura, 1968; Dolansky and Tjernlund, 1968; Rothenberg, 1974; Rothenberg *et al.*, 1975; Ladefoged and Antoñanzas-Barroso, 1985; Klatt, 1986b, 1987b; Hunt, 1987). Voicing source models have been devised for a formant synthesizer that are intended to increase the naturalness of the output speech by permitting a mixture of an impulse train and noise as the source waveform (Kato *et al.*, 1967; Holmes, 1973). The strategy is to specify a cutoff frequency below which the source consists of harmonics, and above which the source is flat-spectrum noise. Similar strategies for mixed-excitation synthesis have been described by Rothenberg *et al.* (1975) and Makhoul *et al.* (1978). The KLGLOTT88 voicing source model, to be described next, includes a parameter, **AH**, which controls the amplitude of aspiration noise that can be added to the $U_g(t)$ waveform. The aspiration noise has a spectrum that falls off at about 6 dB/oct of frequency increase, but, when the radiation characteristic is folded into the source models, the result is a relatively flat spectrum. Thus, as noise is added to the voicing source, it is most noticeable at high frequencies where the harmonic spectrum of voicing is weaker. If the glottis is partially spread, as in a breathy vowel, **TL** and **AH** will be increased, and higher harmonics of the source spectrum will be replaced by aspiration noise.

B. The KLGLOTT88 voicing source model for KLSYN88

A block diagram of the new voicing source model for the synthesizer originally described in Klatt (1980) is presented in Fig. 10. In order to distinguish it from its predecessor, the new model will be called the KLGLOTT88 model. Source control parameters identified in the figure 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 (quasirandom fluctuations) in f_0, in percent of maximum; **DI**, degree of diplophonic double-pulsing irregularity in f_0, in percent of maximum; and **AH**, amplitude of aspiration (breathiness) noise, in dB.

During the open phase of a glottal cycle, the volume velocity waveform has been parametrized to obey a relationship first proposed by Rosenberg (1971); i.e., $U_g(t) = at^2 - bt^3$, where a and b are constants whose values depend on the amplitude of voicing and the duration of the open period. The spectral consequences of varying open quotient, spectral tilt, and aspiration level are shown in Fig. 11. The aspiration source provides noise energy with a relatively flat spectrum, as indicated in the last panel of the figure. The behavior of the remaining control parameters of the KLGLOTT88 voicing source will be explicated in the next few sections.

Current acoustic models of the $U_g(t)$ waveform are rather simple, capturing only the first-order shapes and spectral aspects of observed natural glottal waveforms. It is hoped that, for speech synthesis purposes, the models will turn out to be useful. However, the following qualifications suggest that several additional modeling complications may be necessary to achieve high-quality synthesis of male and female voices.

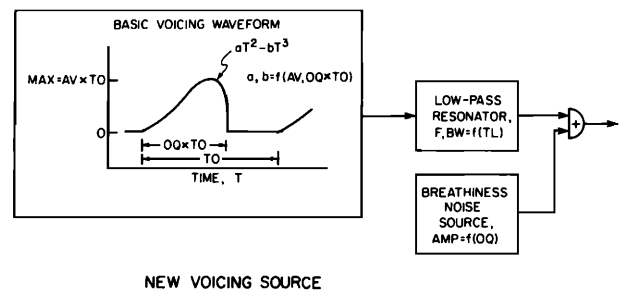

FIG. 10. Block diagram of the voicing source for the KLSYN88 formant synthesizer. The effects of the radiation characteristic have also been folded into the source models, resulting in a voicing source spectral output (Fig. 11) that falls off at about 6 dB/oct [corresponding to $U_g'(t)$] and an aspiration source spectrum that is essentially flat over the frequency range of interest.

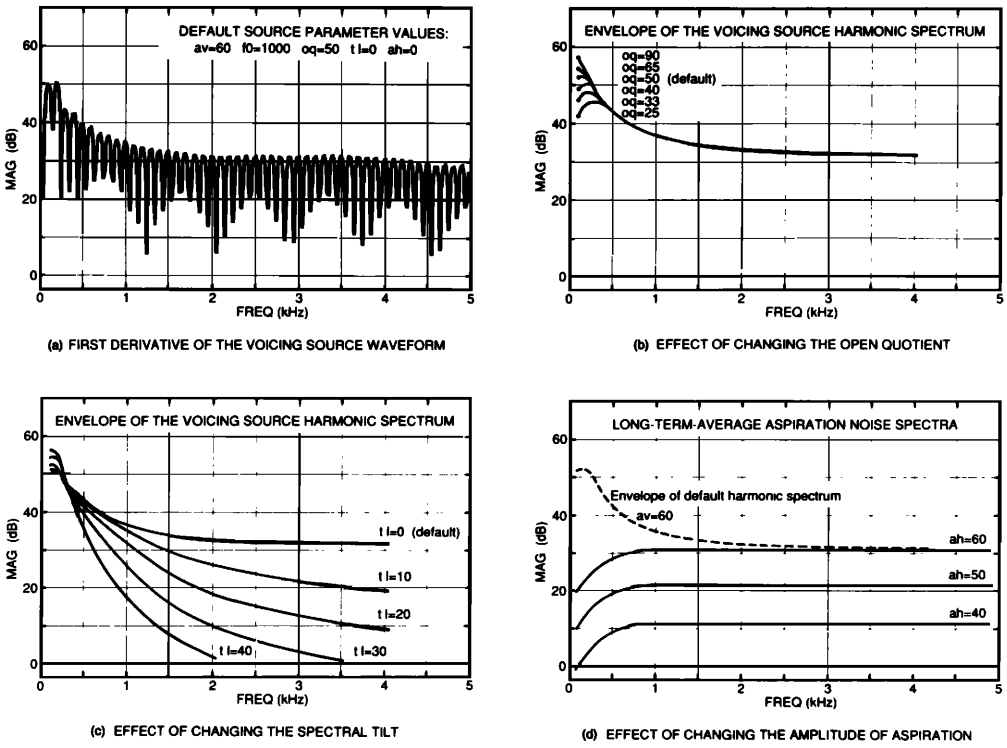

FIG. 11. Dft magnitude spectra are shown of $U_g'(t)(U_g(t)$ modified by a first difference approximation to the radiation characteristic), as synthesized by the KLGLOTT88 voicing source model at several values for each of three control parameters. (a) Spectrum of a train of pulses with $f_0 = 100$ Hz; (b) and (c) spectral envelope of a harmonic spectrum that would result from synthesizing a train of such pulses; (d) spectra of aspiration noise source.

1. Complications I: Glottal pulse timing irregularities

The waveshape 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. The discussion then shifts to acoustic interactions between the source and vocal tract that may contribute to naturalness.

a. 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 Fig. 12. While the wavering nature of the f_0 trace may in small part be due to analysis artifacts,[15] 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.

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 Fig. 12 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, 1981, 1986). If the appropriate parameter to characterize jitter and shimmer is the standard deviations 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 KLGLOTT88 voicing source model includes a mechanism for introducing a slow quasirandom drift to the f_0 contour through the **FL** flutter control parameter (shimmer is not modeled). The term "flutter" has been adopted since jitter has a well-defined meaning that differs from our synthesis strategy. Instead of using a random process to simulate jitter, we add to the nominal f_0 a quasirandom component that is, in fact, the sum of three slowly varying sine waves:

$$\Delta f_0 = (\mathrm{FL}/50)(F0/100)[\sin(2\pi 12.7t) + \sin(2\pi 7.1t) + \sin(2\pi 4.7t)]\ \mathrm{Hz}. \quad (1)$$