XVI. SPEECH COMMUNICATION

Academic and Research Staff

Prof. Kenneth N. Stevens
Prof. Morris Halle
Dr. Sheila Blumstein*
Dr. Margaret Bullowa
Dr. René Carré†
Dr. William L. Henke
Dr. A. W. F. Huggins
Dr. Allan R. Kessler
Dr. Dennis H. Klatt
Dr. Paula Menyuk‡
Dr. Joseph S. Perkell
Dr. Raymond A. Stefanski**
Dr. Jacqueline Vaissière††
Mary M. Klatt

Graduate Students

Thomas Baer
Jared C. Bernstein‡‡
William E. Cooper
William F. Ganong II
Ursula G. Goldstein
Shinji Maeda
Bernard Mezrich
Stefanie R. Shattuck
Victor W. Zue

A. SIMILARITY STRUCTURES AMONG VOWELS

National Institutes of Health (Grant 2 RO1 NS04332-11)

Jared C. Bernstein

In recent publications on the auditory structure of vowels, a class of models has been proposed that is liable to direct test. These models propose that the auditory impression created by a vowel is representable by a simple combination of the effects of the formant frequencies. The research reported here indicates that under the test outlined below these models are probably valid within most of an acoustic vowel space and demonstrably invalid in parts of that vowel space. The models of vowel perception in question, which generalize the proposals of Hanson[1] or Pols et al.,[2] hypothesize that similarity judgments between pairs of vowels can be represented as a componentwise combination of effects where the components are the formant frequencies, or differences in log or linear formant frequencies within each vowel. A component or psychological dimension in these models can be a function of any formant frequency (F_n), any (F_n-F_m), or any (F_n/F_m). The models that were tested experimentally are all of the two-dimensional combinations of these components in which the two arguments contain information equivalent to two formant frequencies for each vowel, and all of the three-dimensional combinations of these components in which the three arguments contain information equivalent to three formant frequencies.

---

*Assistant Professor, Department of Linguistics, Brown University (on leave).

†On leave from Ecole Nationale Supérieure d'Electronique et de Radioélectricité, Grenoble, France.

‡Also Professor of Special Education, Boston University.

**N. I. H. Postdoctoral Fellow.

††Also Instructor of Phonetics, Department of French, Wellesley College.

‡‡Rackham Prize Fellow from University of Michigan, 1973-74.

The most usual rule in these models for combining the effects of the components or dimensions is the Euclidean distance function. Tversky and Krantz[3] have shown that the adequacy of models such as Euclidean distance can be tested directly. Certain classes of combination rules on a given dimensional representation of the similarity structure of a set of stimuli are equivalent to statements about the ordinal properties of similarity between pairs of stimuli.

$$\text{DECOMPOSABILITY: } d(V_1, V_2) = G[\psi_1(F_{11}, F_{12}), \psi_2(F_{21}, F_{22}), \psi_3(F_{31}, F_{32})]. \tag{1}$$

In Eq. 1 a distance function d on two vowels V_1 and V_2 is called decomposable relative to a representation of the vowels as their first three formant frequencies when distance can be expressed as a strictly monotonic increasing real-valued function G in three arguments, each argument being a real-valued function ψ on the appropriate formant frequencies of the two vowels.

$$\text{INTERDIMENSIONAL ADDITIVITY: } d(V_1, V_2) = G'\left[\sum_{i=1}^{3} \psi_i(F_{i1}, F_{i2})\right]. \tag{2}$$

A distance function is called interdimensionally additive when d is expressible as some G' on one argument which is the sum of the outputs of three ψ functions (see Eq. 2). As in the decomposable case, G' is a strictly monotonic increasing function and all functions are real-valued. Notice that an additive model is an example of a decomposable model and that Euclidean distance is an example of an additive structural model in which the ψ are squared differences and G' is the square root. A decomposable model of distance is violated if the order of distances between pairs of points in any one dimension is not constant regardless of the fixed levels in the other dimensions, and additivity between dimensions is violated by a similar inconsistency in the ordering of distances along pairs of dimensions.

Fig. XVI-1.
Bar representation of 8 vowels with an order of judged differences that violates decomposability.

Figure XVI-1 exemplifies a violation of decomposability in the second formant. Figure XVI-1 is a bar representation of the first three formant frequencies of 8 vowels (from left to right) A, B, C, D, E, F, G, H. The first four vowels have the same formant frequencies except for their second formants, and the second four vowels are also all the same except for their second formants. In their second formant frequencies, A, B, C, and D are equal to E, F, G, and H, respectively. A violation of perceptual

decomposability is implicit in the two judgments shown below the bar representations. A subject judged the difference between vowel A and vowel B to be less than the difference between vowel C and vowel D, but the same subject judged the difference between vowel E and vowel F to be greater than the difference between vowel G and vowel H. For any eight vowels like this, if the orders of distances are in the same direction, the judgments would be completely consistent with some decomposable model. There are, however, violations like the example shown in Fig. XVI-1 and all violations of decomposability that are found in the data are like this example in that they all involve some vowel in which two formants are close together in frequency.

Figure XVI-2 shows the same sort of violation in a plot of F_1 by F_2. The slanted edge in the lower right is where F_1 equals F_2. A violation of decomposability in this

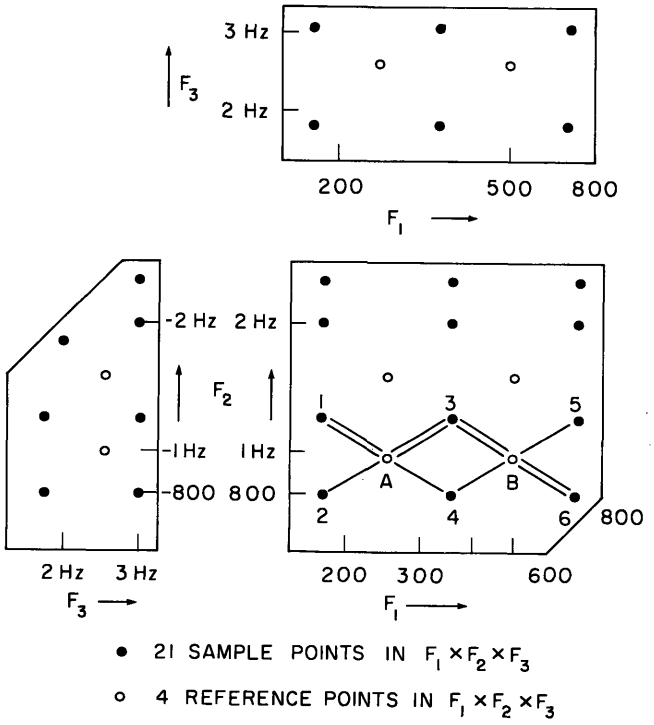

Fig. XVI-2. Four hypothetical orders of distance between pairs of vowels in vertical orientation. Distances drawn with double lines are greater than those drawn with single lines.

picture will occur if all distances drawn with double lines are judged to be greater than single-line differences in the same vertical orientation. For instance, if the distance from A to 1 were judged to be greater than the distance from A to 2 [$d(A, 1) > d(A, 2)$], and the distance from A to 3 were judged to be greater than the distance from A to 4

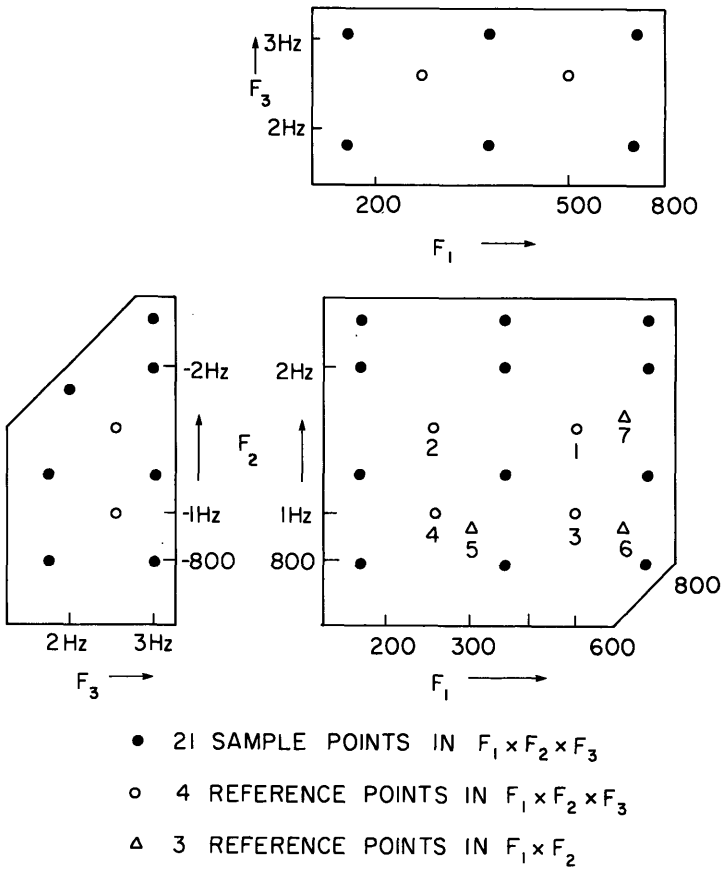

Fig. XVI-3. Judgments by a Persian and an American listener of differences between pairs of serially synthesized vowels.

Table XVI-1. Formant frequencies of the reference vowels and the sample vowels (Fig. XVI-3).

Formant frequencies of reference vowels

#1:	500,	1500,	2500 Hz
#2:	250,	1500,	2500 Hz
#3:	500,	1000,	2500 Hz
#4:	250,	1000,	2500 Hz
#5:	300,	925,	2500 Hz
#6:	625,	925,	2500 Hz
#7:	625,	1625,	2500 Hz

Sample vowels were all possible combinations of

F_1	F_2	F_3
175 350 700	800 1200 2000	1800 3000

and the three vowels

F_1	F_2	F_3
175 350 700	2500	3000

[d(A, 3) > d(A, 4)], and the distance from B to 3 were judged to be greater than the distance from B to 4 [d(B, 3) > d(B, 4)], but the distance from B to 5 were judged to be less than the distance from B to 6 [d(B, 5) < d(B, 6)], then there would be an inconsistency in this representation and hence a violation. The hypothetical perceptual data in Fig. XVI-2 would count as 4 pairs of distances relevant to testing the decomposability of F_2 relative to F_1. The judgments on 3 pairs of pairs of distances [(A1, A2; B5, B6), (A3, A4; B5, B6), (B3, B4; B5, B6)] violate decomposability, and 3 pairs of pairs of distances in Fig. XVI-2 [(A1, A2; A3, A4), (A1, A2; B3, B4), (A3, A4; B3, B4)] are consistent with some decomposable model.

One Persian and one American listener repeatedly judged differences between pairs of serially synthesized vowels. One member of each pair was a "reference" vowel positioned as indicated in Fig. XVI-3 by open circles or open triangles, and the other member of each pair was a 'sample' vowel positioned as indicated by closed circles. The formant frequencies of the reference vowels, and the sample vowels in Fig. XVI-3 are given in Table XVI-1. The stimuli were all 400 ms in duration with a fundamental frequency ramp that increased linearly from 85 Hz to 115 Hz. The amplitude of the impulse source to the serial synthesizer was initially 40 dB, it then increased linearly to 56 dB at 15 ms into each stimulus. The source remained at 56 dB until 350 ms into the stimuli, where it decreased linearly to 0 dB at 400 ms. When presented to the subjects, the stimuli were separated by a 100-ms silence, and each pair of stimuli was separated from the next pair by 9 seconds.

On the left in Fig. XVI-4 there is an F_2 by F_3 section in a vowel space, and on the right an F_1 by F_2 section. Both slanted edges are where the two formants are equal in frequency. On the basis of 24 magnitude judgments per subject per pair of vowels, confidence intervals were constructed around the means of the 24 responses per pair of stimuli. The confidence intervals were calculated as the mean (M) ±2 times the standard deviation (σ) of the responses around their mean divided by the square root of the number of responses ($\sqrt{n}$). If the confidence intervals for two pairs did not overlap, then the distances between the pairs were said to be unequal, otherwise they were considered equal. The American listener judged approximately 2200 sets of 8 vowels in a manner consistent with an interdimensionally additive model of vowel quality in the first three formants or formant differences. There are 6 very definite violations of decomposability in the American listener's data which involve either the open triangle or the closed circle in the lower right corner of the F_1 by F_2 section and there were several reliable violations of decomposability involving the three vowels closest to the upper left slanted edge of the F_2 by F_3 area. These were the only violations in that listener's data. The Persian listener had approximately 2000 eight-vowel sets consistent with an additive model, but he showed 12 extremely reliable violations of decomposability in the same areas. The violations all involve vowels that are near the slanted edges in Fig. XVI-4; that is where

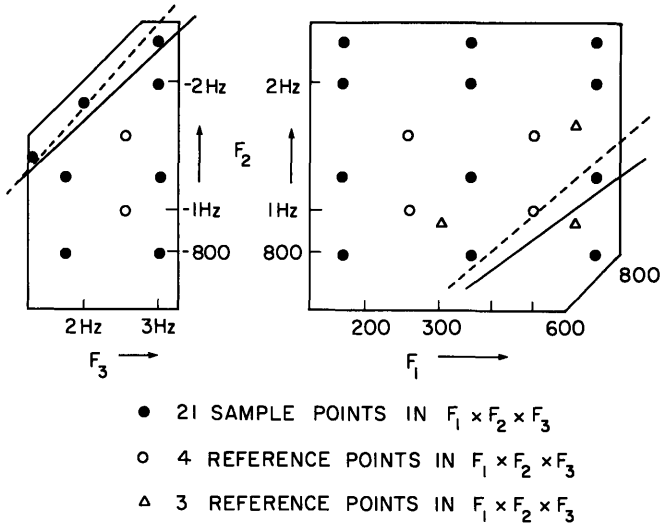

Fig. XVI-4. Vowel space partitioned into incompatible parts. Dashed lines: Persian listener. Solid lines: American listener. Confidence level, $M \pm 2(\sigma/\sqrt{n})$.

two adjacent formants are close in frequency. For the American subject, the areas between the slanted edges and the solid lines are not representable by the same decomposable distance model as the rest of the space, but there is strong evidence that the rest of the space is consistent with an interdimensionally additive model under the test outlined above. Preliminary analysis from a different test of decomposability indicates that no part of the vowel space that ranges over more than 200 Hz or 300 Hz in F_1 can be represented adequately as decomposable in three or fewer dimensions. The same conclusions hold for the Persian subject, except that the boundaries from his data are marked by the dashed lines. These violations are probably the result of changes in another perceptual dimension, such as the overall amplitude or the amplitude of the second formant, which is analytically dependent but not monotonically related to any other single dimension. Notice that two of the vowels that violate the general pattern (i.e., 625, 925, 2500 and 400, 1300, 1500) are easily made in an adult male vocal tract.

In conclusion, for these two listeners, no decomposable representation of auditory vowel similarity in any two or three dimensions that are formant frequencies or log or linear formant frequency differences is consistent with the perceived distances between vowels in all parts of this acoustic vowel space.

References

1. G. Hanson, "Dimensions in Speech Sound Perception: An Experimental Study in Vowel Perception," Ericsson Technics 23, 3-175 (1967).

2. L. Pols, L. van der Kamp, and R. Plomp, "Perceptual and Physical Space of Vowel Sounds," J. Acoust. Soc. Am. 46, 458-467 (1969).

3. A. Tversky and D. Krantz, "The Dimensional Representation of the Metric Structure of Similarity Data," J. Math. Psychol. 7, 572-596 (1970).

B. AN EFFECT OF SYNTAX ON SYLLABLE TIMING

National Institutes of Health (Grant 2 RO1 NS04332-11)

A. W. F. Huggins

1. Introduction

It is well known that (at least in citation form) a stressed syllable is longer when it occurs as a monosyllable than when it occurs in a trochee (i.e., followed by an unstressed syllable).[1,2] Thus the vocalic segment /ou/ gets progressively shorter in the series: ode, odor, odorous, odorously. This effect has only been studied within words – although the "words" used in Lindblom's study of the effect were strings of nonsense syllables such as "labalalal," which could probably be classified just as accurately as phrases. In view of the repeated claims for isochrony in English,[3] one might expect to find similar effects operating across word boundaries, but within the rhythmic foot. That is, one might expect a stressed monosyllable to be shortened by adding an unstressed syllable to the front of the following word.

In an earlier study just such an effect was found.[4] The sentence studied was "cheeses abounded about," which contains four unstressed syllables, one on each side of each of the two internal word boundaries. Sixteen almost-grammatical, meaningful versions of the sentence can be produced by including or excluding each of the four unstressed syllables independently. Measurements of spectrograms of each of the sixteen sentences, read five time each in irregular order by two speakers, showed that the vowel in "bound" was shortened almost as much by adding the unstressed syllable across the word boundary (i.e., "bound out" becomes "bound about") as by adding it within the word (i.e., "bound out" becomes "bounded out"). This was not true of the word "cheese," however. The expected shortening occurred when the unstressed syllable was added within the word (i.e., "cheese bound" becomes "cheeses bound"), but there was no shortening (and for one subject, there was even a slight lengthening) when the unstressed syllable was added to the start of the following word ("cheese bound" becomes "cheese abound"). Thus the shortening effect operated across the word boundary following "bound," but not across the word boundary following "cheese." One possible reason for this difference is that the MSB (Major Syntactic Break) of the sentence falls between "cheese" and "bound," whereas "bound" and "out" are not so divided. We might then hypothesize that the shortening effect operates across word boundaries, but not across a major syntactic boundary. The present experiment was designed to test the foregoing hypothesis.

1. A	The	slave	*	faced	the	wall				SENSE
B	The	slaver	*	faced	the	wall				
C	The	slave	*	defaced	the	wall				
2.	The	dive(r)	*	trade (betrayed)	the	soup				NONSENSE
3.	Dave (David)		*	(de)faced	the	wall				NAMES
4. A	The	brave	porter	*	disarmed	the	gunman			SENSE
B	The	bravest	porter	*	disarmed	the	gunman			
C	The	brave	supporter	*	disarmed	the	gunman			
5.	The	plain(er)	(at)tack	*	spoke	to	the	girl		NONSENSE
6.	Dave (David)		Cannon (Buchanan)	*	faced	the	wall			NAMES

* = Major Syntactic Break

Fig. XVI-5. Examples of three types of sentence triplets, with two positions of the indicator syllable relative to the major syntactic break.

2. Experiment

Figure XVI-5 shows representative examples of the material used. The complete set of sentences is listed in the appendix. Thirty-two base sentences were composed, each containing a sequence of two stressed syllables, of which the first was the "indicator" syllable that would show the effect (e.g., "slave" in sentence 1A). The indicator syllable was made as long as possible (long vowel, voiced final consonant) to make it maximally sensitive to the shortening effect.[5] In sixteen of the base sentences, the word boundary following the indicator syllable was also the MSB of the sentence, and in the other sixteen the MSB was later in the sentence (e.g., sentence 1A: "the slave/faced the wall," and sentence 4D: "the brave porter/disarmed the gunman"). Each base sentence had associated with it two derived forms, both identical with the base form, except that an extra unstressed syllable was added following the indicator syllable. In one derived form, the extra syllable was added to the end of the word containing the indicator syllable (e.g., "the slaver faced ..."; "the bravest porter ..."), and in the other, the extra syllable was added to the front of the word following the indicator syllable (e.g., "the slave defaced ..."; "the brave supporter ...").

Making up meaningful sentences with all of the foregoing properties proved to be extremely difficult, partly because pairs of words like slave-slaver, in which both members of the pair fall in the same narrow-form class, are very rare. It would have been even better had it been possible to choose indicator syllables starting and ending with voiced stops, because this would have simplified the task of measuring vowel

durations in the indicator syllables. Eight meaningful sentences were invented for each position of the MSB, and they are labeled "sense" in Figs. XVI-5, XVI-6, and XVI-7 and in the appendix. Four grammatically well-formed, but semantically anomalous, sentences were also included for each MSB position, labeled "nonsense." A final set of four sentences was included in which the subject noun-phrase was a name. These were the only sentences in which shortening was measured in the identical indicator syllable for both positions of the MSB.

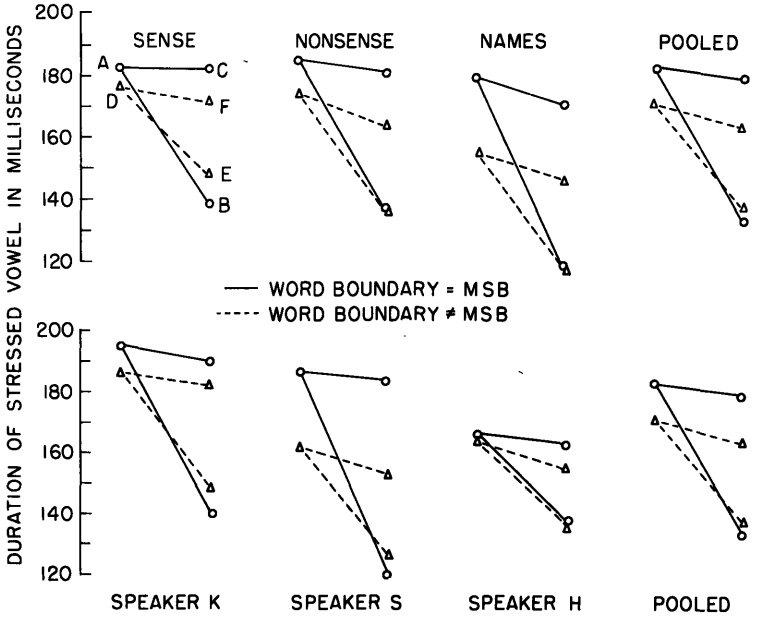

Fig. XVI-6. Pooled durations for the vocalic segment in the indicator syllable for the three types of material (upper), and for the three subjects (lower). Letters refer to the sentence types described in Fig. XVI-6, and in the text.

Three speakers read each of the 96 sentences once, in irregular order, at a normal-to-fast rate, and the duration of the vocalic part of the indicator syllables was measured from wideband spectrograms. Pooled results for the three types of material, sense, nonsense, and names, and for the three speakers, are presented in Fig. XVI-6. The data are broken down by subject in Fig. XVI-7.

The data from each set of three sentences, the base and two derived forms, are joined to form left-pointing arrows. The solid lines represent the results for indicator syllables that were in phrase-final position (i.e., followed by the MSB), and the dotted lines represent data from indicator syllables in nonfinal position.

The left point of each arrow represents the base form of the sentence (letters A, B, C, D, E, F apply to the same sentence types in Figs. XVI-6 and XVI-7 as in Fig. XVI-5).

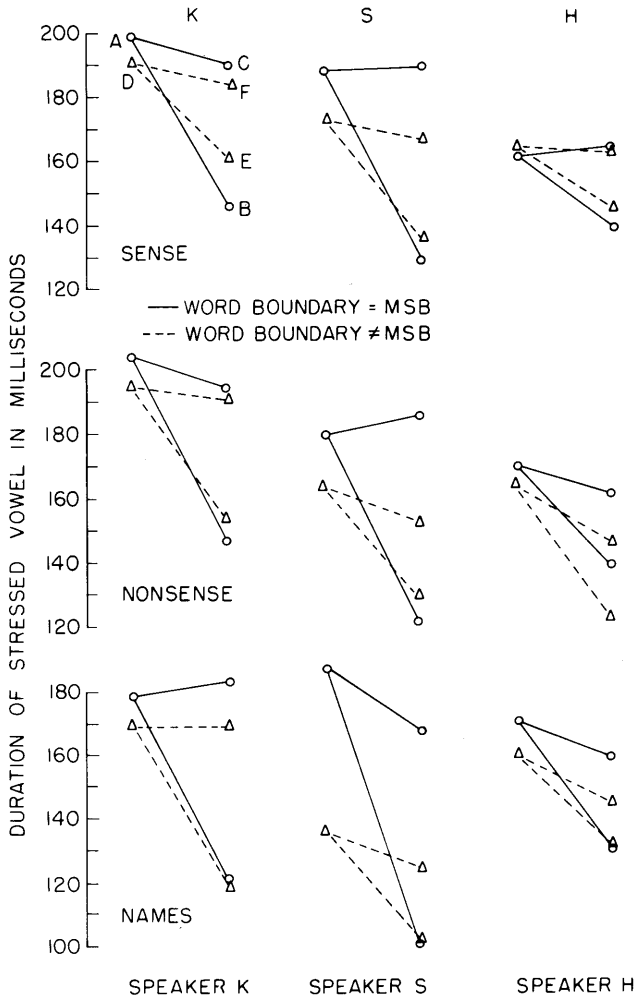

Fig. XVI-7. Pooled durations for the vocalic segment in the indicator syllable, broken down by type of material and subject.

The lower line of each angle illustrates the shortening of the indicator syllable caused by adding the extra syllable within the word, and the upper line represents the shortening when the extra syllable was added to the following word.

3. Results

The figures show that the shortening effect operating across a word boundary is a second-order effect for both positions of the MSB (the top lines are all nearly horizontal). The failure to find shortening across the word boundary within the noun-phrase was unexpected. The context in which it had been found previously[4] was between a verb and its participle ("cheese bound about" vs "cheese bound out"). Second, the results varied for the different sentences, some showing almost as much shortening effect across the word boundary as within the word. This was particularly true for subject S, who showed no shortening across the word boundary in any of the eight meaningful sentences,

when the word boundary was also the MSB, but strong shortening in three of the eight when the word boundary was not the MSB. The subjects did not agree on which sentences showed the shortening effect across a word boundary, so that other variables must be operating which were not controlled in the present study.

The results do show some clear inter-word timing effects. Consider the data for the "names" material, in which the same indicator syllable was always used. The only difference between the sentence triplets for each MSB position was the addition of a last name, for example, changing "Dave" to "Dave Cannon." As we have mentioned, the "names" sentences were much better controlled than the others for phonetic context effects. Consider the top lines of the arrows. The top dotted line is lower than the top solid line. Thus "Dave" was shortened by adding "Cannon" within the same phrase. Possible reasons for this inter-word effect include the following:

1. Changes in the stress on the word "Dave."
2. Changes in the number of syllables in the whole phrase (and indeed, in the whole sentence, although measurements showed that the timing of the rest of the sentence was not affected by adding the extra word).
3. Removal of the indicator syllable from final position within the phrase.

Inspection of Figs. XVI-6 and XVI-7 shows that the same effect occurs to a lesser extent in the other data (except subject H's "sense" data, in Fig. XVI-7). The main point of interest is that the leftmost point of each of the dotted angles is always (with one exception) lower than the leftmost point of the solid angles. Thus the indicator syllables are longer in phrase-final position. This observation has very interesting implications, and confirms a recent result found by Dennis Klatt.[6] Klatt compared stressed syllable durations with the median duration in a long paragraph. All 28 syllables containing vowels or consonants that were more than 40% longer than the median durations were the last stressed syllable before the MSB. (But note that the relationship was not symmetrical: some phrase-final stressed syllables were not lengthened.) The interesting implication, as pointed out by Klatt, is that the syntactic structure of a sentence may be marked far more clearly in the acoustic waveform than has previously been suspected. Pre-pausal lengthening is a well-known phenomenon: it now appears that phrase boundaries, as well as clause boundaries, may be marked by lengthening of the preceding syllable(s), even though no pause occurs.

The results may also solve a current controversy: Umeda and her colleagues at Bell Telephone Laboratories have reported that the timing rules developed for speech in citation form do not apply in conversational speech.[7] Inspection of the "names" data in Fig. XVI-7 shows, in agreement with Umeda's finding, that adding an unstressed syllable within the word has a smaller effect within the phrase than at the end of the phrase (points D-E$\langle$A-B in Figs. XVI-6 and XVI-7). The reason, however, seems to be that a singleton stressed syllable is considerably longer in phrase-final position, whereas

position within the phrase has little effect when the stressed syllable starts a bisyllablic word (point B = point E). The same argument also may apply to the "sense" and "nonsense" data, if the difference between points B and E were due to the less careful control of phonetic context across MSB position in these materials.

Appendix

Set of Sentences Used in the Experiment

SENSE

1. The (maid/maiden) * (moved/removed) her hat.
2. The (farm/farmer) * (missed/dismissed) the milkmaid.
3. The (slave/slaver) * (faced/defaced) the wall.
4. The (page/pages) * (liked/disliked) the Senator.
5. The (Board/boarder) * (acted/reacted) immediately.
6. The (squad/squadron) * (trusted/distrusted) the President.
7. The (ward/warden) * (tested/detested) the students.
8. The (Lodge/lodger) * (armed/disarmed) the watchman.

NONSENSE

13. The (dive/diver) * (trade/betrayed) the soup.
14. The (coal/cola) * (roused/aroused) the grape.
15. The (trees/treason) * (fined/defined) the house.
16. The (horn/hornet) * (posed/exposed) the dust.

NAMES

9. (Jane/Janey) * (moved/removed) her hat.
10. (Carl/Carlos) * (missed/dismissed) the milkmaid.
11. (Dave/David) * (faced/defaced) the wall.
12. (Hayes/Hazel) * (liked/disliked) the Senator.

* = MAJOR SYNTACTIC BREAK.

SENSE

17. The (wide/wider) (view/review) * pleased the architects.
18. The (broad/broader) (sign/design) * won the prize.
19. The (brave/bravest) (porter/supporter) * disarmed the gunman.
20. Our (firm/firmest) (trust/distrust) * was fully justified.
21. The (wise/wizened) (lector/collector) * read the notice.
22. A (warm/warmer) (light/delight) * enveloped us all.
23. The (gold/golden) (sample/example) * proved our argument.
24. The (strange/stranger) (quality/equality) * affected our judgement.

NONSENSE

25. The (gold/golden) (fusion/infusion) * spared the tires.
26. The (wide/wider) (seat/deceit) * used the factory.
27. The (rude/ruder) (vision/division) * drank the wine.
28. The (plain/plainer) (tack/attack) * spoke to the girl.

NAMES

9. (Jane/Janey) (Linski/Kalinski) * moved her hat.
10. (Carl/Carlos) (Duffy/McDuffy) * missed the milkmaid.
11. (Dave/David) (Cannon/Buchanan) * faced the wall.
12. (Hayes/Hazel) (Cord/McCord) * liked the Senator.

References

1. B. Lindblom, "A Note on Segment Duration in Swedish Polysyllables," QPSR 1-1964, Speech Transmission Laboratory, K. T. H., Stockholm, 1964.
2. T. P. Barnwell III, "An Algorithm for Segment Duration in a Reading Machine Context," Technical Report 479, Research Laboratory of Electronics, Massachusetts Institute of Technology, January 15, 1971.
3. D. Abercrombie, "Syllable Quantity and Enclitics in English," in D. Abercrombie et al. (Eds.), In Honour of Daniel Jones (Longmans, Green and Company, New York, 1964).
4. A. W. F. Huggins, "On Isochrony and Syntax," Proc. Symposium on Auditory Analysis and Perception of Speech, Leningrad, 1973 (to appear in the Journal of Phonetics).
5. D. H. Klatt, "Interaction between Two Factors That Influence Vowel Duration," J. Acoust. Soc. Am. 54, 1102-1104 (1973).
6. D. H. Klatt, "Vowel Lengthening Is Syntactically Determined in a Connected Discourse" (manuscript in preparation).
7. M. S. Harris and N. Umeda, "Parametric Analysis of Vowel Duration in Single- and Multi-Syllable Words," J. Acoust. Soc. Am. 55, 397(A) (1974).

C. MORE TEMPORALLY SEGMENTED SPEECH: IS DURATION OR SPEECH CONTENT THE CRITICAL VARIABLE IN ITS LOSS OF INTELLIGIBILITY?

National Institutes of Health (Grant 2 RO1 NS04332-11)

A. W. F. Huggins

1. Introduction

Cherry and Taylor[1] have reported that when a continuous speech message is switched back and forth between the left and right ears of a listener at a rate of approximately 3 Hz, its intelligibility is considerably reduced relative to that at higher or lower switching rates. This phenomenon is known as the Cherry effect. The critical alternation rate, where intelligibility is lowest, increases, however, when the playback speed of the speech is increased.[2] This finding rules out any explanation of the Cherry effect predicated on a temporal parameter of the perceptual apparatus such as the time taken to switch attention between the ears, as proposed by Cherry and Taylor. Instead, it suggests that the effect may be due to an interaction between the switching rate and a temporal parameter of the speech. Since then, several other authors[3] have relied on the last result to develop positions of their own, and it has become quite important to verify that the original result was correct. Some more recent work with "temporally segmented" speech has raised doubts about its validity.

Results very similar to those with alternated speech can be obtained simply by inserting silent intervals into continuous speech, thereby dividing it into speech intervals

separated by silent intervals.[4] In particular, when silent intervals of 200 msec were used to temporally segment a continuous message, its intelligibility declined from more than 90% to ~15% as the duration of the speech intervals was reduced from ~200 msec to ~30 msec.[5] Second, when the duration of the speech intervals was held constant at 63 msec, intelligibility rose again as the silent intervals were shortened from ~120 msec to ~60 msec. Clearly, these results have strong implications for a description of auditory processing, especially for short-term acoustic storage. Since the second of these results involves only the duration of a silent interval, it is hard to reconcile it with the earlier finding with alternated speech, which implicated a temporal parameter of the speech.

Therefore, an experiment was designed to confirm the alternated-speech result, using temporally segmented speech. The aim of the experiment was to discover whether the declining intelligibility of speech, temporally segmented by 200-msec silent intervals, is a function of the duration of the speech intervals or of their speech content.

2. Experiment

The experimental design is illustrated in Fig. XVI-8. Intelligibility was measured by having the subject shadow a 150-word passage at each speech-interval duration, of which only the middle 100 words were scored to avoid end effects. Using a sampling rate of 9 kHz, we were able to digitize and store a complete 150-word passage on the PDP-9 computer of the Speech Communication Group of the Research Laboratory of Electronics. The stored signal was then marked (in unused low-order data bits) to divide it into speech

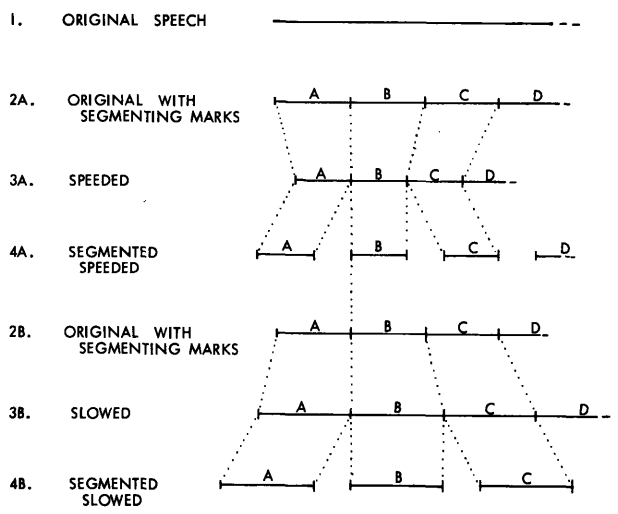

Fig. XVI-8. Diagram of the method of independently varying the duration and the speech content of the speech intervals in temporally segmented speech.

intervals of the required duration (lines 2A and 2B in Fig. XVI-8). Then we changed the sampling rate and set the 200-msec silent interval. The program then played back the stored speech, inserting a 200-msec silent interval wherever there was a mark on the waveform. Two recordings were made of each passage, one with the sampling rate increased (speeded) by $\sqrt[4]{2}$ (Fig. XVI-8, line 4A), and the other (line 4B) with the sampling rate slowed by the same factor. Thus the speech intervals in the two recordings contained exactly the same waveform, except that their time scales differed by a factor of $\sqrt{2}$. Two sets of passages were prepared, each of which had two practice passages followed by 7 experimental passages. Speech-interval duration increased from each experimental passage to the next by a factor of $\sqrt{2}$, covering a range of ~210-25 msec before speeding or slowing occurred.

Twelve subjects shadowed one set of slowed passages, and the other set speeded, with the order of presentation and the two sets of passages appropriately counterbalanced.

3. Results

Mean shadowing scores are plotted in Fig. XVI-9. The scores are plotted twice, on the left as a function of the duration of the speech intervals as heard by the subjects, and on the right as a function of the speech-interval duration before they were slowed or speeded. That is, on the right a single value on the abscissa corresponds to speech intervals with the same speech content (i.e., number of phonemes, syllables, etc.), but whose durations varied, depending on whether they were speeded or slowed in playback.

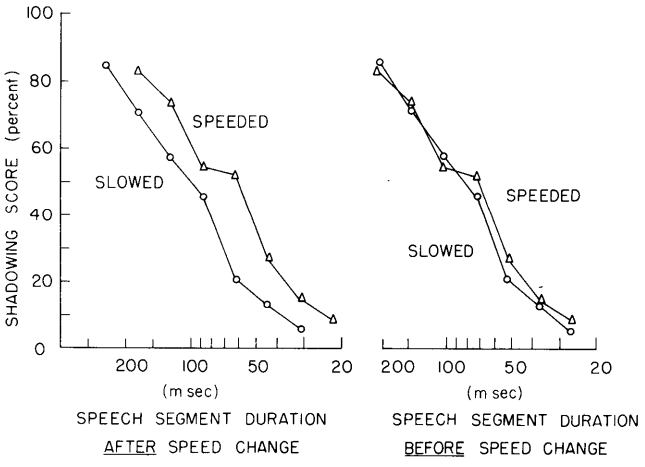

Fig. XVI-9. Pooled shadowing scores for 12 subjects are plotted on the left as a function of the absolute duration of the speech intervals, as presented to the subjects, and on the right as a function of their effective speech content. Each subject heard one passage set with the speech intervals slowed, and the other speeded by a factor of 1.189.

Clearly, the two sets of data agree very well on the right, and rather poorly on the left. Hence the critical variable in describing the declining intelligibility is the speech content of the speech intervals, not their duration, and this confirms the original finding with alternated speech. Incidentally, the effect of experimenter bias in scoring the results was accidentally controlled very well in the foregoing study. It was not until all data had been prepared for plotting that I realized I was wrong about which way the two sets of data should be displaced relative to each other to support my original finding with alternated speech. Thus any (unconscious) bias had an effect that worked against the reported result!

4. Earlier Failures

This experiment was not the first attempt at producing this result, so it is appropriate to discuss briefly the reasons for earlier failures, to rule out the possibility that the new finding is simply due to chance as a result of trying often enough. The designs of the experiments that failed were similar to those of the present experiment.[6] We attempted to vary independently the duration and the speech content of the speech intervals, in one case by increasing the playback speed and in two cases by time compression of the speech intervals. In the original result with alternated speech we used a speedup factor of 1.19, but the result would be much more convincing if a really large speedup factor could be used. This was impossible with alternated speech, since the maximum speech rate at which subjects can shadow undistorted speech (not temporally segmented) is quite limited. The silent intervals inserted into temporally segmented speech effectively slow it down, however, so that the speech rate in words per minute is the same as the original signal for doubled-speed speech temporally segmented into 200-msec speech intervals by 200-msec silent intervals. Therefore we tried speedup factors of 1.5 and even 2.0. The computer programs used to time-compress the speech, and to temporally segment it, were more primitive versions of the present programs, and were constrained to operate in real time. As a result, pulses had to be recorded on the second track of the tape to mark off the speech intervals so that the speech would be segmented in the same places for each experimental condition.

5. A Time-Compressor/Expander

A pitch-synchronous time-compression computer program was written for the PDP-9 computer, and it was used to produce a time-compressed master tape, which was then temporally segmented in the same way as the master tape. Since the earlier versions of the program were constrained to run in real time, it was actually a frequency dividing/multiplying program, whose output could be converted to time-compressed/expanded speech by restoring the signal frequencies by recording the signal and playing it back at a different speed.

The program had three functional parts: an input routine, which continuously sampled the input at the input-clock rate and stored the samples in a circular buffer; an output routine, which repeatedly played out a segment of the circular buffer at the output-clock rate; and a glottal-period detecting routine, which identified complete glottal periods (or equivalent intervals when voicing was not present) in the input signal, and passed them to the output routine. If the output clock was set to run at one-half the rate of the input clock, the period detector would pass arguments defining glottal periods to the output routine approximately twice as often as the output routine looked at them, so that alternate glottal periods (on the average) were omitted from the output waveform. The glottal period detector looked through a weighted time window at the rms energy integrated from one positive-going zero crossing of the waveform to the next, and selected as the start of the next glottal period the zero crossing that was closest to that predicted from the preceding period and associated energy integral. The routine was adaptive in both time and amplitude, in a manner similar to that described by Crystal.[7] It made surprisingly few mistakes, in view of the corners that had to be cut before the program would run in real time. Although no formal tests were run, the compressed speech that was produced seemed considerably more intelligible than that produced by a second program, which simulated the well-known Fairbanks method.[8] Since the output-clock rate was continuously variable, any desired time compression or expansion could be produced, but at the cost of some degradation of the signal because of the necessity of recording the signal at one tape speed and playing it back at another.

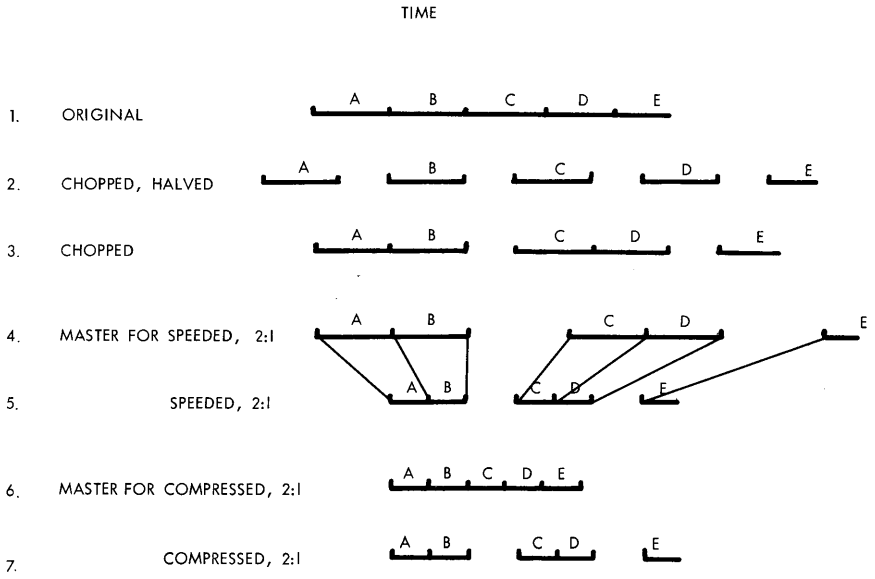

Fig. XVI-10. Diagram of the method of separating duration and speech content of the speech intervals, using speedup or time compression by a factor of 2.0.

The design for the experiments that failed is illustrated in Fig. XVI-10, which enables us to compare the intelligibility of the 2:1 "speeded" speech (Fig. XVI-10, line 5) both with that of the "chopped" speech (line 3), with which it shares the speech content of the speech intervals but not their duration, and with that of the "chopped-halved" speech (line 2), with which it shares the duration but not the speech content of the speech intervals. Similar comparisons can be made for the 2:1 "compressed" speech (line 7).

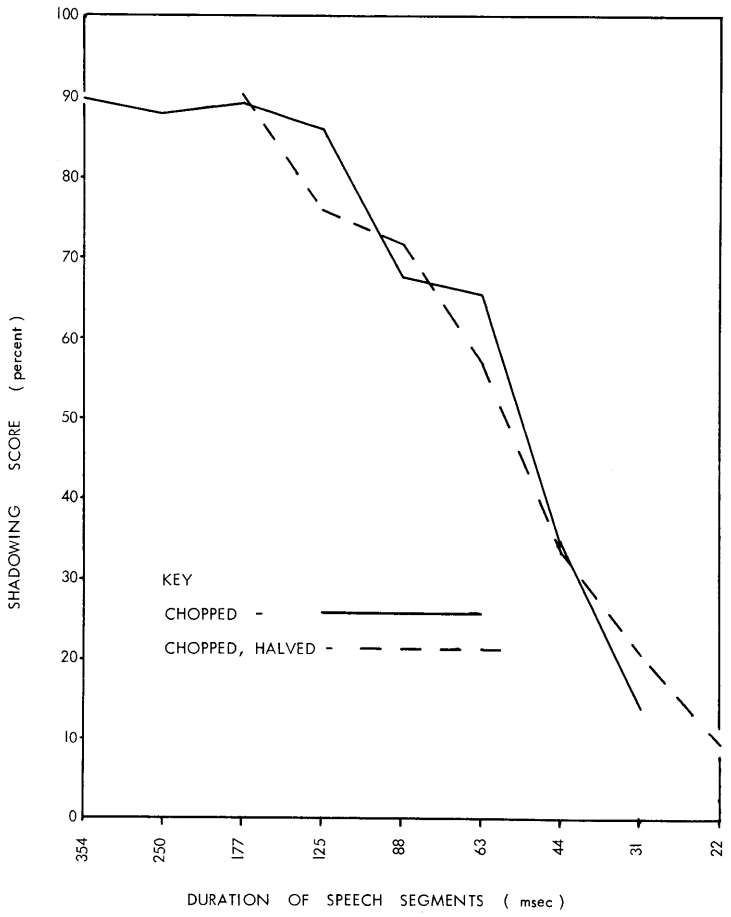

Fig. XVI-11. Pooled scores of 8 subjects, each of whom shadowed two sets of passages covering an overlapping range of speech intervals, as a function of speech-interval duration.

The "chopped" and "chopped-halved" conditions, of course, represent replications of a single experiment, with different speech passages used. Both were included so that the comparisons that we have just outlined could all be made on the same speech passages. To make sure that, in fact, they do represent replications of the same experiment, 8 subjects shadowed two sets of passages, one "chopped" and the other "chopped-halved."

The $\underline{n}^{th}$ passage in the "chopped" set had speech intervals of the same duration as those in the $\underline{n+2}^{th}$ passage in the "chopped-halved" set, since passages were presented in ascending order of speech-interval duration. The pooled shadowing scores are presented in Fig. XVI-11 and it can be seen that there are only minor differences because of the use of different passages.

Each subject in the experiments that failed (Fig. XVI-10) shadowed two sets of passages, one "chopped" and the other either "speeded" or "compressed" by a factor of 1.5 or 2.0. Subjects who heard the "speeded" speech were first given 15 minutes practice of shadowing speech that had not been temporally segmented, but had been frequency-multiplied by the same factor as the speeded speech they were to hear.

The results of the experiments that failed are shown in Fig. XVI-12 where mean shadowing scores are plotted against the duration of the speech intervals. The heavy

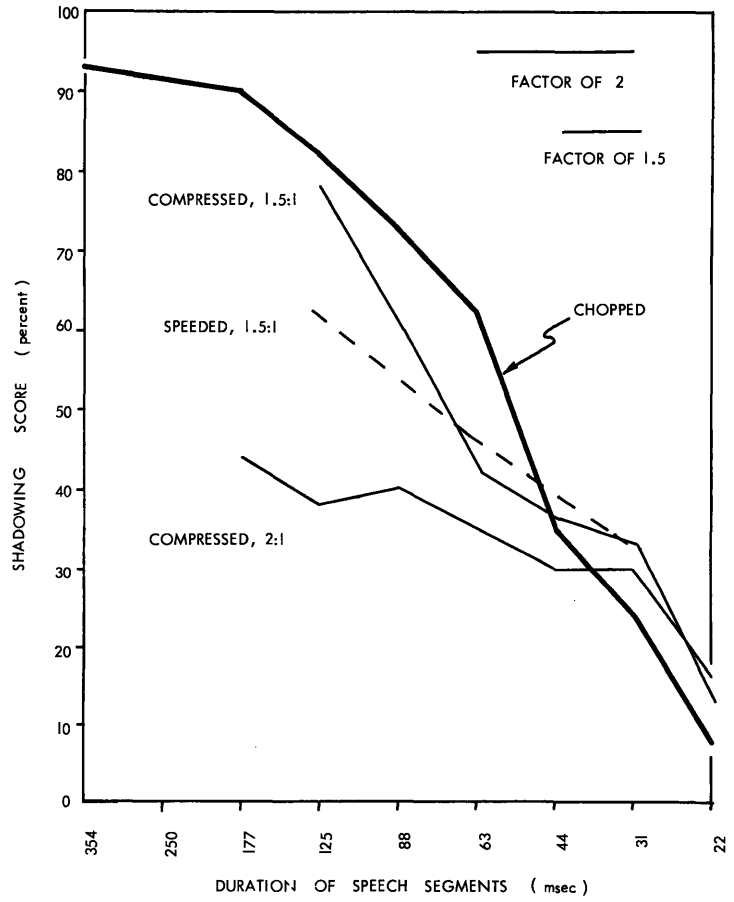

Fig. XVI-12. Shadowing scores for temporally segmented speech with 3 types of time compression compared with those with normal speech (the control condition, labeled "chopped") as a function of speech-interval duration.

curve, labeled "chopped," represents all data collected until then under comparable conditions (that is, included are the data from the "chopped" vs "chopped-halved" experiment described above, and the data reported in an earlier experiment,[5] as well as the control data collected in the present experiment). The six middle data points represent data from 23 subjects; those for the two longest and the shortest speech intervals represent 16, 19, and 5 subjects.

If the data from the speeded or compressed speech is to provide an adequate answer to the experimental question, the data should either coincide with the heavy curve in Fig. XVI-12 if the duration is the critical attribute, or lie to the right of it, by an amount indicated at the top right of Fig. XVI-12 for each speedup factor, if the speech content of the speech intervals is the critical attribute. Clearly, the data do not agree with either of these options. The intelligibility of the speeded or compressed speech rises far more slowly as speech intervals are lengthened than does the intelligibility of the normal speech. We must conclude that speech speeded or compressed by a factor of 1.5 or 2.0 is significantly less intelligible than the normal speech used as a control, and this fact undermines the experimental design. It was partly for this reason that in the successful experiment described at the beginning of this report we used symmetrical speedup and slowdown. (Fletcher had shown that speedup and slowdown by small factors produce roughly equal decrements in intelligibility.[9])

One final point should be made. At the right-hand side of Fig. XVI-12 where speech intervals are very short, the data from the speeded and compressed conditions do lie to the right of the heavy curve, which tends to support the importance of the speech content of the speech intervals rather than their duration. The break occurs when the speech intervals are made longer than ~30-40 msec, which corresponds to ~60 msec of the prespeeded or precompressed speech. When the speech intervals are shorter than this, subjects apparently are unable to distinguish between the normal speech and the compressed speech. Therefore the cues to speech rate must be extracted from speech intervals longer than 60 msec. The figure of 60 msec has appeared in several recent experiments,[5, 10] and further experiments are under way to discover its meaning and significance.

References

1. E. C. Cherry and W. K. Taylor, "Some Further Experiments upon the Recognition of Speech, with One and with Two Ears," J. Acoust. Soc. Am. 26, 554 (1954).
2. A. W. F. Huggins, "Distortion of the Temporal Pattern of Speech: Interruption and Alternation," J. Acoust. Soc. Am. 36, 1055-1064 (1964).
3. U. Neisser, Cognitive Psychology (Appleton-Century-Crofts, Inc., New York, 1967).
4. A. W. F. Huggins, "The Perception of Temporally Segmented Speech," Proc. VII International Congress of Phonetic Sciences, Montreal (Mouton, The Hague, 1972).

5. A. W. F. Huggins, "Second Experiment on Temporally Segmented Speech," Quarterly Progress Report No. 106, Research Laboratory of Electronics, M.I.T., July 15, 1972, pp. 137-141.

6. A. W. F. Huggins, "Temporally-Segmented Speech III: Does Intelligibility Depend on Duration or Speech Content of the Speech Intervals?" J. Acoust. Soc. Am. 54, 300(A) (1973).

7. T. H. Crystal, "Methodology and Results on Laryngeal Disorder Detection through Speech Analysis," Final Report, Contract No. PH-86-68-192, Signatron, Inc., Lexington, Mass., see especially p. 115.

8. G. Fairbanks, W. L. Everitt, and R. P. Jaeger, "Method for Time or Frequency Compression-Expansion of Speech," IRE Trans., Vol. AU-2, No. 1, pp. 7-12, January-February 1954.

9. H. Fletcher, Speech and Hearing (D. Van Nostrand Company, Inc., New York, 1929).

10. N. Guttman and B. Julesz, "Lower Limits of Auditory Periodicity Analysis," J. Acoust. Soc. Am. 35, 610(L) (1963).

D. A CHARACTERIZATION OF FUNDAMENTAL FREQUENCY CONTOURS OF SPEECH

National Institutes of Health (Grant 2 RO1 NS04332-11)

U. S. Navy Office of Naval Research (Contract N00014-67-A-0204-0069)

Shinji Maeda

1. Introduction

This is a study of American English intonation based on the analysis of the fundamental frequency (F_0) contours of speech signals. It is well known that voice pitch, which is correlated with the F_0 values, fluctuates greatly during speech. The fluctuation is not random, but highly organized in some way, depending on the particular language. In each language the organization of pitch fluctuation is somewhat standard, so that all speakers of the language speak in a similar manner, but this may differ markedly from that of other languages. This organized voice pitch fluctuation is called "intonation."

The most noteworthy studies of intonation have been made by linguists. Before Pike's study,[1] linguists had focused their attention upon the relation between meanings of sentences and the variation of intonation contours.[2,3] After Pike's study, more attention was given to the manner of representation of intonation, which then allowed linguists to develop an elaborate theory of intonation. Using a discrete representation of intonation (such as the four-stress level representation of Trager and Smith,[4] the relationship between intonation pattern and syntax was investigated intensively by a number of researchers, for example, Stockwell,[5] Chomsky and Halle,[6] Halle and Keyser,[7] and, quite recently, Vanderslice and Ladefoged.[8] These works provide various insights into intonational phenomena. Perhaps the most basic, as Stockwell clearly states, is the

fact that there exist "neutral" or "colorless" intonation contours for every sentence, serving as a basic line against which all other possible contours are contrastable.[9] Pike expressed this as the "intonational minimum" of speech.

Most of these studies were based on perceptual impressions of the sound. For this reason, intonation was sometimes called "the abstracted characteristic of sentence melodies."[1] We must admit the importance of the abstraction procedure, since speech signals carry various kinds of information besides intonation. However, abstraction using auditory perception was sharply criticized by Lieberman,[10] in the sense that the perceptual impression is not always consistently related to the physical reality of speech, particularly to the F_0 contours.

Recently, researchers have tended to study intonational phenomena from the physical aspects,[11,12] such as F_0 values (which are the primary physical correlates of the voice pitch), intensity, and segmental duration, and from the physiological aspect,[13-15] such as measurement of the electrical activity – or electromyography – of the muscles associated with the phonatory apparatus, and measurement of the subglottal pressure. In these studies, most attention has been focused upon intonational differences between word pairs such as OBject vs obJECT[11,13] and between ambiguous phrase pairs such as American#history teacher vs American history#teacher.[13] At a higher level (the level of the sentence) only the emphatic effects seem to have been systematically investigated[13-15] and the published data concern very short sentences (3 or 4 lexical words). The only study that we found based on the analysis of F_0 variations during longer sentences[16] was directed toward automatic segmentation of continuous speech into sentences and phrases. As far as we know, more basic studies such as how the fluctuations in F_0 are organized at the level of the sentence have not yet been undertaken.

In our study we investigated first how the F_0 contours can be described in terms of certain attributes that specify the intonation. Our study is based strictly on the analysis of F_0 in sentences. We are not interested in a quantitative (or statistical) analysis, which is not suitable for our purpose, since F_0 values are influenced not only by intonational phenomena but also by nonlinguistic, physiological and acoustic factors of speech production. Rather, we try to abstract a regular pattern in F_0 contours from visual inspection of the contours. If the regular patterns are closely related to the semantic organization of sentences and/or the syntactic structure of sentences, it may be safe to state that the regular patterns that are abstracted are manifestations of intonation.

Second, in order to obtain deeper insight into the generation of intonation patterns, we try to look for the physiological correlates of the intonational attributes specifying the regular pattern of F_0 contours. Any speaker of a language controls his phonatory apparatus in a specific manner to generate regularity in his speech. Thus some physiological activities must be consistently related to intonational attributes. We hope that this study will provide sufficient knowledge to postulate a generative model of intonation,

in which the intonational attributes (features) associated with a sentence are mapped into the physiological features, and then into the F_o contours.

In this report, we shall describe our schematic analysis of F_o contours leading to a specification of the pattern in terms of a limited number of intonational attributes. Also, we shall discuss briefly a measurement of the larynx height during speech, based on cineradiographic data.

2. Procedure

The corpus to be analyzed has two parts. One is a set of 60 isolated sentences in which the length of noun phrases is systematically manipulated, in terms of the number of syllables in the words and the number of words in the noun phrases. The second part is a text composed of 14 sentences (entitled "Chickens"). All sentences are declarative. The first part was used for studying how F_o patterns depending on the length of the sentence and its constituents are organized. The second part was used to investigate how far the results obtained from the first part could be generalized to sentences with various grammatical structures.

In the experiment, three speakers were asked to read the corpus: 60 sentences were written on separate cards, and each card was presented to the speaker after he had completely read the preceding one; 14 sentences of the text were written on a single page. The speech signals, and the glottal signals (detected by using an accelerometer) were recorded on two-channel tapes. The F_o contours were calculated from the glottal waves by using the absolute difference sum method.[17] Hard copies of F_o contours and the amplitude envelopes computed from the corresponding speech waves (which are displayed on the computer interactive display oscilloscope), have been used for this investigation.

3. Schematic Analysis of F_o Contours

Before going into the details of the data we shall show how regular patterns are abstracted from F_o contours. In Fig. XVI-13, we show the F_o contours of the beginning of the text: "In the jungles of Asia, there is a large bird with brilliant colors, ..." read by the three speakers. Even though these contours differ quantitatively from each other, we can see some similarities in the curves. These similarities may be abstracted visually and described qualitatively or approximated by schematic drawings. For instance, any contour in Fig. XVI-13 may be described as follows: The contour is raised during the first stressed syllable in the word "jungle," and this rise is associated with a peak, indicated by the letter "P". Then the contour is lowered during "of", raised in the first syllable of "Asia," and lowered in its second syllable. During the first phrase "In the jungles of Asia" the whole contour can be regarded as falling gradually (indicated by dashed lines), and this gradual declination line or baseline is raised at the beginning of the next phrase. The contour is raised again in the word "large" and this rise is

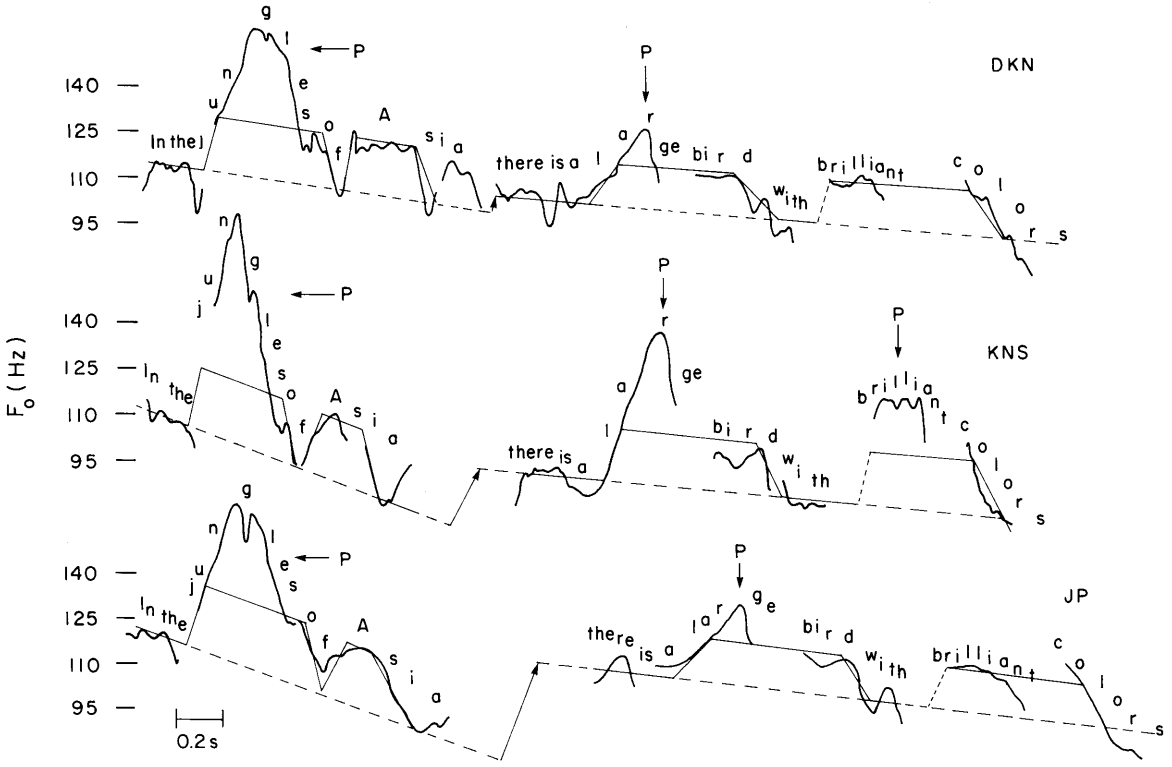

Fig. XVI-13. F_o contours of the sentence read by three speakers and the corresponding schematic patterns.

associated with a smaller peak than that of the first phrase (on the word "jungle"). Then the contour is lowered at the word "bird," and so on. This description may be represented by the schematic drawing shown in Fig. XVI-13 in which the piecewise-linear approximation is superimposed on each original contour.

The basic elements of the schematized pattern are the baseline, represented by the dashed line, the piecewise-linear trapezoidal pattern with a rise, a fall, a plateau that is parallel to the baseline, and a peak P that in these examples occurs immediately after the initial rise. Other attributes of the pattern will be introduced later. In this analysis, the small F_o valley observed at /g/ in "jungles" for every speaker is completely ignored, since this is considered as an acoustic effect of the manner of production of the voiced stop consonant /g/[12,13] upon the F_o values. Other localized segmental influences are also ignored in the schematic representation.

It is observed that the sentences are divided into smaller units, and some of the units (noun phrases) are demarcated by the combination of the rising contour associated with the peak and the lowering contour at the end of the unit. Thus these schematic patterns indicate some grammatical function. It is not unreasonable, therefore, to state that the schematic description represents the intonation pattern and is characterized by the

intonational attributes such as the rising, the lowering, and the peak (prominence).

We shall now describe separately the results obtained from the analysis of noun phrases in isolated sentences as units. All noun phrases were embedded in sentences with the following grammatical structure: S (sentence) = NP (noun phrase) + V (verb) + NP. The noun phrases include the following structures, all of which we describe briefly.

NP = Det (determiner) + Adj (adjective) + N (noun)

NP = Det + Adj + Adj + N

NP = Det + Adj + N + N

NP including prepositions

The sentences have the following structure:

S = NP + V + NP

Since a similar manner of intonational organization was found for all three speakers, we shall use examples from only one speaker.

Det + Adj + N in Predicate

The F_o contours of the noun phrases (NP=Det+Adj+N) are well characterized by the location of the lexically stressed syllable in each content word (adjective and noun), as shown in Fig. XVI-14. Besides the phonological differences, those phrases differ also in the number of syllables and in the location of the lexical stress in each content word. It is clearly observed that the lexical stress (in this case, the primary stress shown by the superfix "1" in Chomsky and Halle's notation in Fig. XVI-14) is consistently related to the F_o patterns. In every phrase, the F_o contour is raised in the stressed syllable of the adjective and lowered from the end of the stressed syllable in the noun. This rising (R) and lowering (L) constitutes the "hat-pattern" introduced by J. 't Hart and A. Cohen[18] from their perceptual study of Dutch intonation. In Fig. XVI-14, the hat patterns corresponding to the noun phrases "the enormous kangaroo" (Fig. XVI-14a) and "the paralyzed kangaroo" (Fig. XVI-14b) are superimposed on each F_o contour by using the connected straight lines (which represents a sort of "hat" shape). It will be recognized later that these risings (R's) and lowerings (L's) that compose the hat pattern are the most basic intonational attributes.

It can be seen that the rising, R, in each stressed syllable in the adjective is not only a simple rise but is also associated with the peak, indicated by "P" in Fig. XVI-14. This peak often occurs with the rising in the first stressed syllable in the unit. The rising, R, and the peak, P, should be recognized as independent intonational attributes, as we shall explain.

Two other remarks should be made. First, the F_o contour for the nonstressed syllables between the two stressed syllables falls gradually, representing the plateau of the hat pattern. This gradual falling (baseline) can be observed not only during the

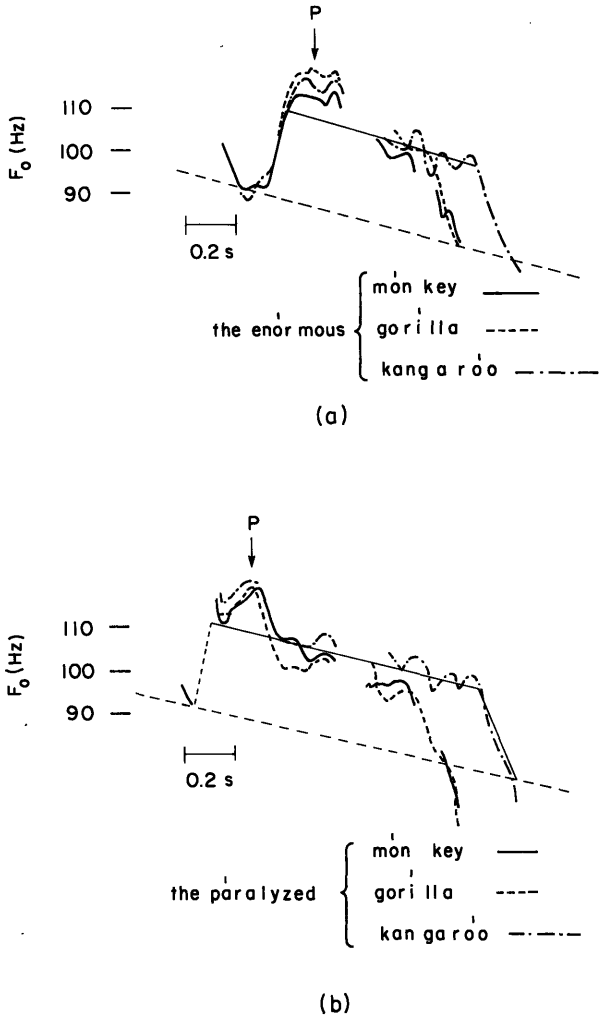

Fig. XVI-14. F_0 contours for NP=Det+Adj+N, and the corresponding hat patterns and the peaks (P).

noun-phrase portions but also through the whole sentences, as shown by dashed lines in Fig. XVI-14.

Second, the F_0 contour in the word "paralyzed" (Fig. XVI-14b) begins with a high F_0 value and there is no rising contour. This is considered as a phonetic influence of the unvoiced stop consonant /p/. Several authors have discussed why this happens.[19,20] No definite conclusion concerning this influence has been reached, however, because of lack of physiological evidence.

The examples that we have shown represented NP composed of multisyllabic words. The patterns found for NP=Det+Adj+N with monosyllabic words are also well characterized by the hat pattern; the contour is raised during the adjective and lowered during the noun.

Det + Adj + Adj + N and Det + Adj + N + N in Predicate

We shall now describe how the local structure underlying NP is manifested in the F_o patterns. There are three possible structures for a NP composed of three content words (W): the structures may be described as groupings of the words as a parallel structure (WWW), a left-branched structure ((WW)W) and a right-branched structure (W(WW)). Two typical examples of F_o contours in NP = Det + Adj + Adj + N are shown in Fig. XVI-15a: "the enormous yellow monkey" and "the magnificent yellow alligator." (The two corresponding contours are superimposed by adjusting the beginning of the adjective "yellow" in the two phrases.) Again, the hat pattern associated with the peak, P, can be observed. Each contour is raised during the stressed syllable in the first content word and lowered from the end of the stressed syllable in the final content word in NP. Each F_o contour

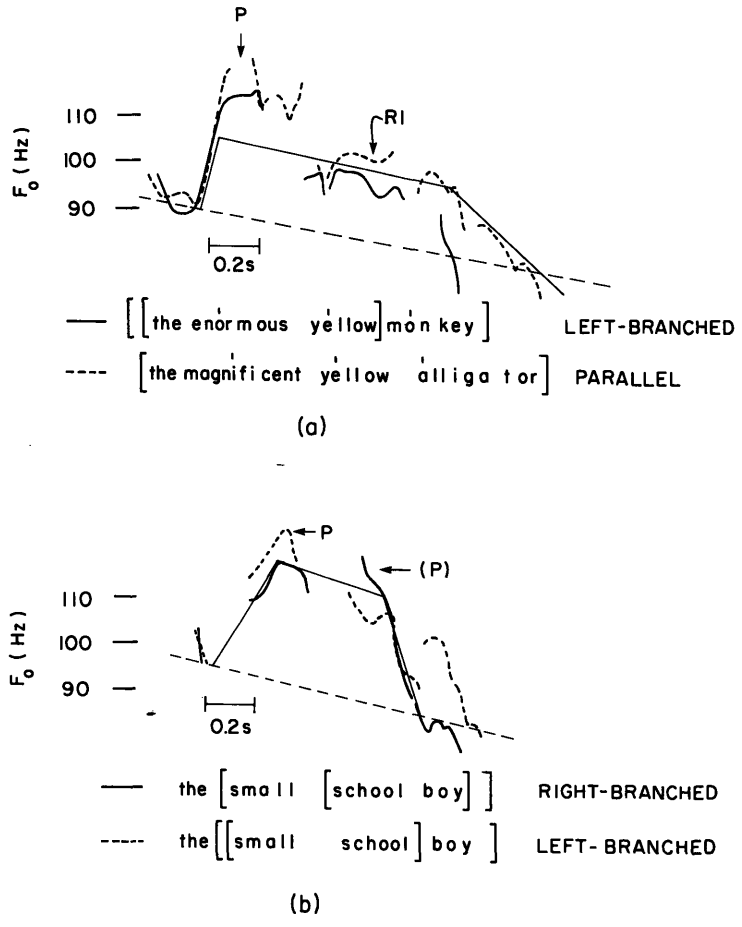

Fig. XVI-15. (a) F_o contours for NP=Det+Adj+Adj+N (the hat pattern corresponds to NP, "the magnificent yellow alligator").
(b) F_o contours for NP=Det+Adj+N+N (the hat pattern corresponds to the right-branched NP).

for the middle content word "yellow" corresponds to the plateau of the hat pattern. In detail, the contour for "yellow" in the phrase "the enormous yellow monkey," shown by the solid line, seems to be gradually falling. The two adjectives in this case may be interpreted as a single compound adjective, since the contour is similar to that of a single adjective. On the contrary, in the second phrase "the magnificent yellow alligator," the adjective "yellow" has a slowly rising contour, compared with the same adjective in the first phrase. This difference may be the manifestation of different structure: the falling contour of the word in the intermediate position (the adjective "yellow") could correspond to the left-branched structure, and the slowly rising (we call this rising R1) to the parallel structure. The speaker was not asked to specify the structure; therefore, the grouping of the words may be arbitrary, depending on his semantic organization for each sentence.

In order to obtain deeper insight into the relationships between the intonational organization and the grouping of the words, the same speaker was asked to read the noun phrase "the small school boy" embedded in a sentence, and to specify the following structures:

(1) right-branched: (small (school boy))
(2) left-branched: ((small school) boy)

A remarkable difference can be seen between the two corresponding F_o contours in Fig. XVI-15b. The solid line indicates the contour of the right-branched NP=(Adj(N N)). The pattern is well characterized by a hat pattern that is quite similar to the pattern found for NP=A+N. (Note that the attribute P is assigned on "school.") Therefore, it may be stated that the two nouns are compounded into a single word. On the other hand, the F_o contour for the left-branched NP=((Adj N) N), shown by the broken line in Fig. XVI-15b, is lowered in the second content word, but raised again and lowered in the final content word. The noun phrase is divided into two groups: (Adj N) and N, by composing one hat pattern for (Adj N) and the other for N.

It will be shown that the grouping (or chunking) of the words in a sentence by forming hat patterns seems to be the primary manifestation of the intonational organization of the sentence. In previous studies of Stockwell[5] and Lieberman[13] it was emphasized that the acoustic cue for differentiating the two structures, such as (small (school boy)) and ((small school) boy) is mainly a matter of disjuncture as A#N N and A N#N. The words are grouped properly, depending on the meaning of the phrase, by inserting a pause. Our example has shown, however, that the grouping of the words can also be made by the intonation contour, by forming the hat patterns.

We now try to see how the structural differences are reflected in the F_o patterns of the noun phrases NP_1=A+A+N and NP_2=A+N+N. For NP_1, the parallel and left-branched structures seem to be related to the slowly rising (R1) and to the gradually falling

contour of the middle adjective, but the difference is not always clear. For example, we have observed that the direction of the contour corresponding to the adjective in the intermediate position could change continuously from falling to rising as shown in Fig. XVI-16. The three contours for NP, "the small black cat," were sampled from various contextual environments. Observe the directions of the F_o contours for the adjective "black" indicated by arrows. It may be suggested, then, that the difference

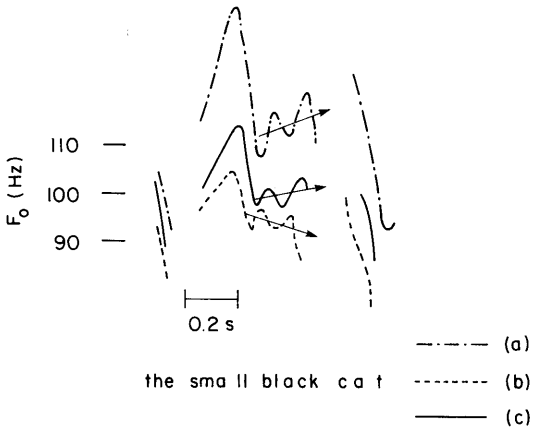

Fig. XVI-16.

F_o contours for NP, "the small black cat" from various contextual environments: (a) "(NP) likes the dog," (b) "The big white dog likes (NP)," (c) "The dog likes (NP)."

between the parallel and the left-branched structures in NP=A+A+N may be manifested in the F_o pattern, but in a continuous form, and there is no clear boundary for distinguishing one structure from another. In NP_2=A+N+N, however, the manifestation is marked and is discrete, as already described. The radical difference in the manner for specifying the structures in NP_1 and NP_2 is well related to the degree of the semantic difference, depending on the different grouping of the words in each phrase. In NP_1, the grouping of the words, for example, ((small black) cat) or (small black cat), has no great effect on the meaning of the phrase. On the contrary, the grouping in NP_2 such as (small (school boy))—the boy is small—, and ((small school) boy)—the school is small—, changes the meaning of the phrase. The speaker seems to spend more physiological effort in contrasting the structures in NP_2 than in NP_1. Here, we postulate a sort of economical principle working under the organization of the intonation patterns, taking account of the balance between the semantic importance and the physiological effort.

Noun Phrase with Prepositions

The analysis of the previous examples has suggested that the words in the noun phrases are grouped into larger units by forming hat patterns. This grouping can be observed more clearly in noun phrases containing prepositions. An example is shown in Fig. XVI-17a which displays the F_o contour of the noun phrase, "the yellow alligator in the mud." This phrase is clearly separted into two groups:

R L RL
(3) the (yellow alligator) in the (mud)

where the superfixes R and L indicate the rising R and the lowering L, respectively. In this description the attribute "peak" (P) is not specified, since in most cases P occurs with the rise. Only when P is not associated with R is it necessary to indicate its location. For convenience, we use the term "phonetic group" (PG) to designate a group of words chunked by the combination of the attributes R and L (i.e., a hat pattern). PG's may be represented in parenthesis as shown in NP (3).

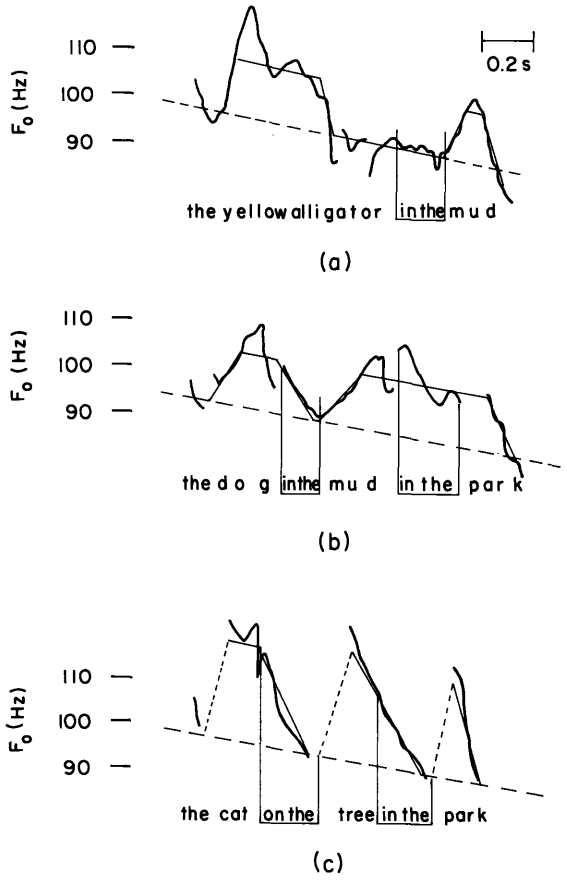

Fig. XVI-17. F_0 contours for NP's containing prepositions, and the corresponding schematic patterns.

On the basis of Fig. XVI-17a, it would seem that a phonetic group has only content words (such as adjectives and nouns), and not function words. This is not always true, however; counterexamples are shown in Fig. XVI-17b and 17c for NP's "the dog in the mud in the park" and "the cat on the tree in the park." The pattern corresponding to each phrase represents the following groupings of the words:

(4) R L R L
the (dog in the) (mud in the park)

(5) RL RL RL
the (cat on the) (tree in the) (park)

As we have mentioned, when the initial consonant at the first stressed syllable in a phonetic group is a voiceless stop, the F_0 contour begins with a high value and often the rising contour cannot be seen. Perhaps the laryngeal gesture for the rising is incorporated with the gesture for the initial stop consonant. It may be reasonable, therefore, to locate the attribute R on the stop consonant, as shown in the phrase (5).

These examples show that function words (prepositions and determiners) can be located in any portion of the hat pattern, except the rising portion. This may be explained as follows: Even though the grammatical structures of the noun phrases (4) and (5) are similar, the words in each phrase are divided into particular groups, depending on the semantic organization of the phrase. In NP (4), the speaker emphasizes the words "dog" and "mud" compared with the word "park." On the other hand, in NP (5), the three content words "cat" "tree" and "park" are assigned equal importance in specifying the meaning of the phrase. It is not unreasonable to assume, therefore, that there is an original grouping of the words in a phrase which is determined by its semantic organization. For instance, the original (semantic) grouping of NP (5) is probably as follows:

(6) the (cat) on the (tree) in the (park)

The grouping is then modified in such a way that less physiological effort (in some sense that is still to be quantified) is required for the generation of the intonation pattern.

Sentences

We have indicated how the F_0 contours of the predicate noun phrases can be well characterized by a hat pattern. Nevertheless, a contour for a noun phrase composed of only one noun often appears as a part of the hat pattern which includes not only this noun phrase but also the preceding phrase or the following phrase.

In Fig. XVI-18 the F_0 contours of four sentences (with identical structure, S=NP+$(V+NP)_{VP}$), and the corresponding schematic patterns are illustrated. It is observed that the contour of the shortest sentence, "The dog likes the elephant" (Fig. XVI-18a), is characterized by only one hat pattern. (The equivalence between a short sentence and a single phrase has also been established for French.[21]) The whole sentence corresponds to one phonetic group:

(7) R L
The (dog likes the elephant)

When one of the two noun phrases in a sentence forms its own phonetic group (hat

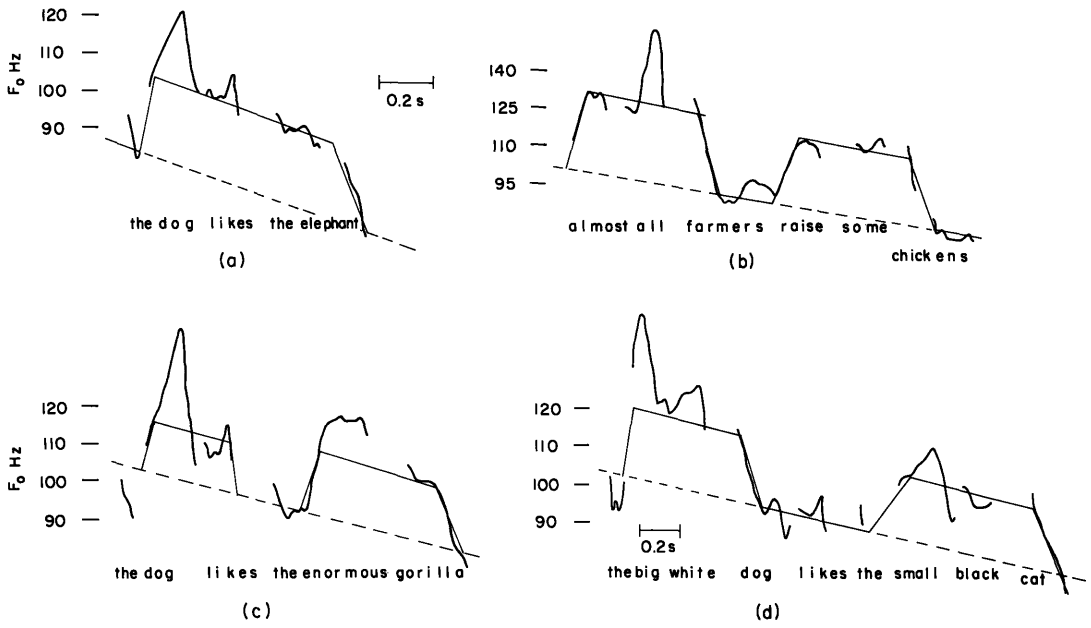

Fig. XVI-18. F_o contours and schematic patterns for the sentences in which the structure is described as $S=NP+(V+NP)_{VP}$.

pattern), and the other noun phrase contains only one content word (a noun), an interesting intonational organization of the sentence can be observed, as shown in Fig. XIV-18b and 18c. In Fig. XVI-18b the sentence is divided into the two groups:

R P L R L
(8) (Almost all farmers) (raise some chickens)

This grouping of the words is well related to the syntactic structure of the sentence, but this does not mean that the structure of a sentence is always related to the grouping by its intonation pattern. In the sentence shown in Fig. XVI-18c the F_o contour suggests that the words in the sentence are grouped as follows:

R L R L
(9) The (dog likes) (the enormous gorilla)

This grouping clearly opposes the grouping derived from the grammar in which the sentence is analyzed as S=NP+VP. Another grouping, which is between S (8) and S (9), can be seen in the sentence shown in Fig. XVI-18d. The contour in this case indicates the following grouping of the words:

R R1 L R L
(10) The (big white dog) likes the (small black cat)

where "R1" indicates the attribute "slowly rising" which we have described. In this sentence, the contour of the verb "likes" takes the lower portion between

the two phonetic groups.

These four sentences indicate that a verb can take any portion of the intonational pattern, depending on the context, an observation that has also been made for the function words. In other words, different grouping of words can be specified by the intonation pattern for sentences in which the grammatical structure is identical. This phenomenon may provide some insight into the generation of the intonation patterns. Perhaps the generation of intonation involves two levels of processing: the first is the linguistic (particularly the semantic) level, and the second is the physiological level. At the semantic level a sentence may be divided uniquely into semantic groups, depending on the semantic organization of the sentence. At the physiological level the semantic grouping is modified so that the intonation pattern is generated in a physiologically efficient manner. This modification is constrained as follows: each lexical stressed syllable in the important words has to receive at least one of the intonational attributes, rising R (perhaps associated with peak P), lowering L, and rising R1. Such a model would predict that the lexically stressed syllables in nouns and adjectives always take the rising or/and the lowering portion of the hat pattern, while the contours for the verbs or for function words are determined by physiologically efficient organization. This principle of efficient organization could generate different grouping of words for sentences with the same grammatical structure.

Three more remarks should be made concerning the attributes that are superimposed on the hat pattern: the peak P, a continuation rising, and baseline BL.

The F_o contour in Fig. XVI-18b provides insight into the nature of the attribute P. In the noun phrase, "Almost all farmers," P is assigned on "all," as shown in S (8). This suggests that R at the beginning of the phonetic group and P must be recognized as independent attributes.

The amplitude of the peak P varies greatly, depending on the location of the phonetic group in the sentence. For the nonemphatic sentences investigated in this study, the amplitude is largest at the beginning of the sentence, and smallest (or often missing) at the end of the sentence (see Fig. XVI-13). Therefore the attribute P should be considered as a continuous feature, and the amplitude may be a manifestation of the degree of emphasis that reflects the attitude of the speaker toward the sentence.

A continuation rising can appear at the end of a phonetic group (PG) which is not located at the end of a sentence. In sentence-final position, the direction of the lowering contour does not change, as can be seen in the examples displaying the noun phrases in the predicates. When PG is not final, however, the lowering contour is connected to the next rising contour in the following PG by a curve located on the nonstressed syllables between the two stressed syllables. When there is no nonstressed syllable(s) between the two PG's a continuation rising can occur at the very end of the lowering contour. An example is shown in Fig. XVI-16a (word "cat" in the subject phrase).

The baseline (BL) seems to be another intonational attribute. We cannot describe this in detail, since the corpus was not designed for the study of the baseline. Therefore we shall describe briefly its characteristics which were found from the analyzed material. The baseline for one of the three speakers gradually falls through the whole sentences, and its degree of falling is a function of the length of the sentence; shorter sentences show a steeper falling. For the other two speakers, the baseline falls only at the beginning of sentences (roughly until the third to the fourth initial content words). When the sentence is divided into phrases, all speakers produce a rising of the baseline at each major syntactic boundary, as shown in Fig. XVI-13 (this boundary may or may not be associated with a pause). Thus we consider the baseline to be one of the linguistically significant intonational attributes.

4. Physiological Correlates of the Intonational Attributes

We have shown that the F_o patterns are well characterized by certain intonational attributes, such as rising R, lowering L, peak P, and so on. These attributes are abstracted visually from F_o contours of speech. When we discuss the physiological efficiency, the number of the attributes specified within a phrase or a sentence might be used as a gross measure of the effort required for its realization. In this discussion, we implicitly assumed that each attribute is related to certain physiological activities. The next straightforward step to be investigated, then, is to determine how the attributes are related to the physiological reality of speech. If the attributes truly represent the intonational organization, we should find a consistent relationship between the attributes and some physiological activities.

The primary physiological factor that determines the F_o values is the state of the vocal cords, which includes the vocal-cord tension and the thickness of the vibrating portion of the vocal ligaments. The vocal-cord tension is controlled primarily by the intrinsic laryngeal muscles, such as the cricothyroid muscles, the lateral cricoarytenoid, and so on.[14, 22] The extrinsic laryngeal muscles which suspend the larynx are also active in the control of F_o. The activity of the extrinsic muscles is observed as the change of the larynx height, which is often correlated with the F_o values.[23] It is a common observation that the vertical movement of the larynx seems to correspond to large variations of the voice fundamental frequency, and we suggest that these variations are related to the hat pattern.

The reason why the larynx height is correlated with the F_o values is not so well known. An interpretation is given by A. Sonninen,[24] who says: The coordinated activities of the sternothyroid muscle and the thyrohyomandibular muscle chain (this activity is termed the "external frame function") produce a resultant force that causes not only a vertical movement of the thyroid cartilage (a change in the larynx height), but also a horizontal movement (posterior to anterior) of the thyroid. This horizontal

movement of the thyroid cartilage (under the assumption that horizontal movement of the cricoid cartilage is prevented) may change the length of the vocal folds, and consequently the F_0 value.[25] Therefore, it is the horizontal movement that must be the essential factor in determining the F_0 value, although there is some correlation between the horizontal and vertical movements of the larynx.

In order to examine the relation between a hat pattern and the larynx movements, we have tried to measure the movements of the larynx, using available x-ray motion pictures for one of the three speakers (KNS).[26]

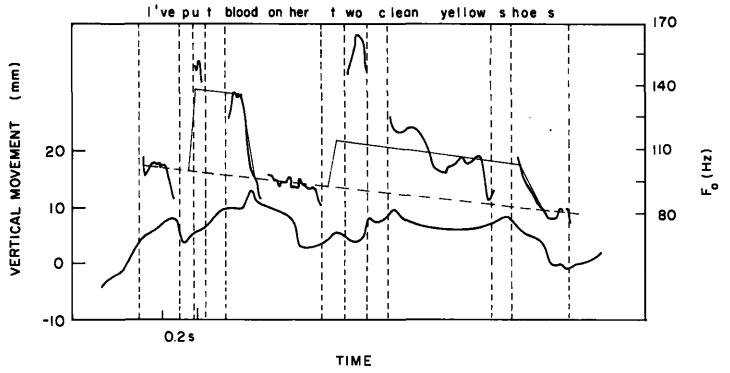

Fig. XVI-19. F_0 contour (upper curve) and the vertical movement of the larynx (lower curve) for the sentence.

Unfortunately, we could obtain meaningful data only for vertical movement. In Fig. XVI-19 the vertical movement of the anterior glottis edge and the F_0 contour calculated from the speech signals (recorded simultaneously with the x-ray films) are shown. Correlation between some aspects of the two contours was found. The sentence "I've put blood on her two clean yellow shoes." is divided into two PG's, "put blood" and "two clean yellow shoes." Correspondence is observed between the hat patterns and the vertical movement, especially for the second PG "two clean yellow shoes." It should be noted that the rising of the hat pattern is much faster than the corresponding larynx movement. In fact, there is no peak in the larynx movement that could correspond to the peak in the F_0 contour. It might be suggested, therefore, that the hat pattern (the attributes R and L) and the peak P are generated by different laryngeal gestures: probably by gestures of the extrinsic and the intrinsic muscles, respectively.

This hypothesis seems to be well supported by the electromyographic (EMG) data reported by Atkinson.[15] Averaged EMG signals obtained from the cricothyroid muscles (CT) and from the sternothyroid muscles (ST), and F_0 contours are sampled from

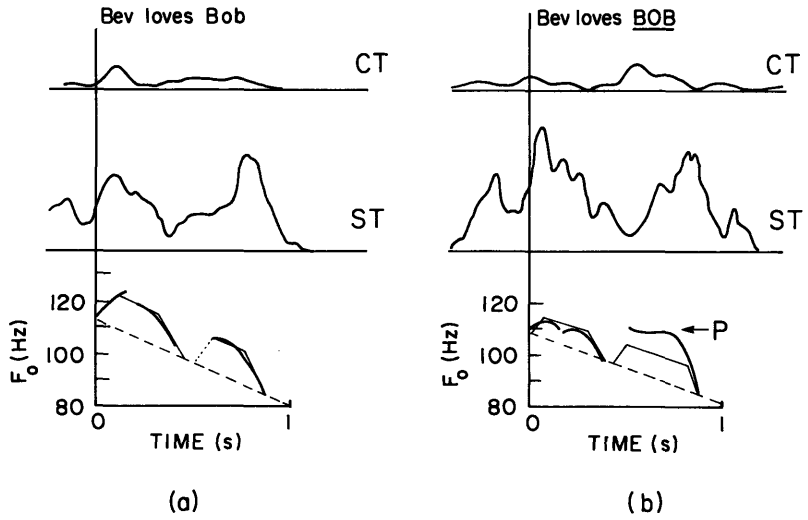

Fig. XVI-20. The averaged EMG activity of the cricothyroid muscle (CT) and of the sternothyroid muscle (ST), and the F_o contour, for (a) the sentence without emphasis, and (b) with emphasis (after Atkinson[15]). The schematic pattern superimposed on each F_o contour was added by the author.

Atkinson's thesis, and are shown for "Bev loves Bob" in Fig. XVI-20a, and "Bev loves BOB," with emphasis on "BOB," in Fig. XVI-20b. Each F_o contour for the two sentences indicates the following description of the intonation patterns.

(11) R L RL
(Bev loves) (Bob)

P
(12) R L RL
(Bev loves) (BOB)

It is observed for each sentence that the two broad peaks in ST activity correspond roughly to the two consecutive hat patterns. Surprisingly, the ST activities are spread across both rising R and lowering L portions of the hat pattern. This may be explained as follows. Recalling Sonninen's interpretation[24] of the external frame function of the larynx, we postulate that the state of the vocal cords is determined by both sternothyroid muscle activities and the thyrohyomandibular muscle chain (THM) activities. Perhaps, for rising R, both ST and THM are active, thereby causing a lengthening of the vocal folds, and hence a raising of the F_o value. On the other hand, for lowering L, ST activities are dominant and cause a lowering of the F_o value. This interpretation must, of course, be tested through an experiment in which EMG data are taken simultaneously from ST and THM.

The cricothyroid (CT) activities show a good correspondence with the attribute P.

In Fig. XVI-20a (S (11)), a large peak in CT activity is observed, which corresponds to the attribute P associated with rising R in the first phonetic group (PG). In Fig. XVI-20b (S (12)), the large peak in CT can be seen at the emphasized word "BOB" in the second PG, while there is a smaller peak at the beginning of the first PG. We observe again that the attributes P and R may be controlled independently, but the sharp rising of the F_o contour at the beginning of PG must be generated by the coordination of both.

Another correlate between the intonational attributes and the vertical movement of the larynx is found in the baseline (BL). Observe that the larynx height, as well as the F_o contour, gradually falls along the whole sentence, as shown in Fig. XVI-19. This observation suggests that BL is also related to the external frame function. A possible explanation for the gradual falling of the larynx height is the following. During speech the lung volume decreases monotonically, thereby causing a continuous lowering of the sternum[27] and hence of the larynx height, since there are muscular and ligamental connections between the sternum and the thyroid cartilage and hyoid bone. As a consequence of the falling larynx position, there is a falling in F_o. (This interpretation was derived during a discussion with K. N. Stevens.)

The physiological investigation has shown some insight into the generation of intonation patterns, in the sense that the attributes specifying the grouping of the words in a sentence such as rising R, lowering L, and baseline BL are correlated with the activities of the extrinsic laryngeal muscles. The other attributes, such as peak P, reflecting the speaker's attitude, seem to be related to the intrinsic laryngeal muscle activities.

We strongly feel that more extensive organized experiments should be performed in order to obtain deeper understanding of the physiological aspects of intonational phenomena.

5. Summary

The fundamental role of intonation is to divide a sentence into smaller groups of words. The words in the sentence are arranged into phonetic groups (PG's), where the corresponding F_o contour is schematized as a "hat pattern" that is described by the intonational attributes, rising R, and lowering L. An upward shift of baseline (BL), which is also an intonational attribute, signals the division of the sentence into large phrases, which are composed of one or more PG's. The other attributes are peak P which often occurs with rising R, continuation rising, and rising R1 that occurs on a intermediate content word in PG. These last three attributes are somewhat continuous in degree, and seem to depend on the attitude of the speaker.

The organization of the intonation patterns seems to be constrained by two primary factors: the semantic organization of a sentence and a principle of physiological economy. It is tempting to postulate that the attributes specifying the grouping of the words in a sentence, R, L, and BL, are related to the extrinsic laryngeal activities, and the

remaining attributes, P, continuation rising, and R1, to the intrinsic muscle activities. Our data are not sufficient to verify these hypotheses, and we feel that more work has to be done in the physiological area.

References

1. K. L. Pike, The Intonation of American English (University of Michigan Press, Ann Arbor, Michigan, 1945).
2. L. E. Armstrong and I. C. Ward, Handbook of English Intonation (B. G. Teubner, Leipzig and Berlin, 1926).
3. D. Jones, An Outline of English Phonetics (W. Heffer and Sons Ltd., Cambridge, England, 9th ed., 1960).
4. G. L. Trager and H. L. Smith, An Outline of English Structure, Studies in Linguistics No. 3 (Battenburg Press, Norman, Oklahoma, 1959).
5. P. R. Stockwell, "The Place of Intonation in a Generative Grammar of English," Language 36, 360 (1960).
6. N. A. Chomsky and M. Halle, The Sound Pattern of English (Harper and Row Publishers, Inc., New York, 1968).
7. M. Halle and S. J. Keyser, English Intonation (Harper and Row Publishers, Inc., New York, 1971).
8. R. Vanderslice and P. Ladefoged, "Binary Supersegmental Features," Working Paper in Phonetics, University of California, Los Angeles, 17, 6-24, 1971.
9. R. P. Stockwell, "The Role of Intonation: Reconsideration and Other Considerations," in D. Bolinger (Ed.), Intonation (Penguin Books, Middlesex, England, 1972).
10. P. Lieberman, "On the Acoustic Basis of the Perception of Intonation by Linguists," Word 21, 40-54 (1965).
11. D. B. Fry, "Experiment in the Perception of Stress," Language and Speech 1, 126-152 (1958).
12. P. Lieberman, "Some Acoustic Correlates of Word Stress in American English," J. Acoust. Soc. Am. 32, 451-454 (1964).
13. P. Lieberman, Intonation, Perception, and Language (The M.I.T. Press, Cambridge, Mass., 1967).
14. J. Ohala, "Aspects of the Control and Production of Speech," Working Papers in Phonetics, University of California, Los Angeles, 15, 1-192, 1970.
15. J. E. Atkinson, "Aspects of Intonation in Speech: Implications from an Experimental Study of Fundamental Frequency," Ph. D. Thesis, The University of Connecticut, 1973.
16. W. A. Lee, "Syntactic Boundaries and Stress Patterns in Spoken English Texts," Univac Report PX1046, Univac Park, Minnesota, 1973.
17. A. R. Meo and G. Gignini, "A New Technique for Analyzing Speech by Computer," Acoustica 25, 261-268 (1971).
18. J. 't Hart and A. Cohen, "Intonation by Rule: A Perceptual Quest," Manuskript 242/II, Institute voor Perceptie Onderzoek, Eindhoven, Netherlands, 1973.
19. M. Halle and K. N. Stevens, "A Note on Laryngeal Features," Quarterly Progress Report, No. 101, Research Laboratory of Electronics, M.I.T., April 15, 1971, pp. 198-213.

20. J. Ohala, "The Physiology of Tone," in L. M. Hyman (Ed.), Consonant Types and Tone, Southern California Occasional Papers in Linguistics No. 1, Los Angeles, California, 1973.

21. J. Vaissière, "Generative Rules for French Prosody," (BB 13), J. Acoust. Soc. Am., Vol. 55, Supplement, S56, Spring 1974.

22. M. Sawashima, "Laryngeal Research in Experimental Phonetics," Status Report on Speech Research, Haskins Laboratories, New Haven, Connecticut, SR-23, 69-115, 1970.

23. J. Ohala and W. Ewan, "Speed of Pitch Change," (DD 6), J. Acoust. Soc. Am. 53, 345 (1973).

24. A. Sonninen, "The External Frame Function in the Control of Pitch in Human Voice," Ann. N.Y. Acad. Sci. 115, 68-90 (1968).

25. P. H. Damste, H. Hollien, P. Moore, and T. Hurry, "An X-ray Study of Vocal Fold Length," Folia Phonat. 20, 349-359 (1968).

26. K. N. Stevens and S. E. G. Ohman, "Cineradiographic Studies of Speech," Quarterly Progress and Status Report, Royal Institute of Technology, Stockholm 2, 9-11 (1963).

27. F. D. Minife, T. J. Hixson, and F. Williams (Eds.), Normal Aspects of Speech, Hearing and Language (Prentice Hall, Englewood Cliffs, N.J., 1973), Chap. 3 and Chap. 4.

E. ON FRENCH PROSODY

National Institutes of Health (Grant 2 RO1 NS04332-11)

Jacqueline Vaissière

This report describes the organization of the fundamental frequency (F_o) variations in French sentences. Interpretation of F_o contours in French is quite different from the interpretation of F_o variations in English. Stress plays an important role in English, whereas in French each syllable of a multisyllabic word pronounced without emphasis receives almost the same amount of stress.[1] But phoneticians also feel that there is a somewhat stronger stress on the last sounded syllable of words pronounced in isolation and on the last sounded syllable of the sense groups embedded in a sentence.[1,2] As Delattre[2] noticed, neither intensity nor F_o are consistent correlates of this final stress ("accent final"): "... les variations d'intensité ne sont pas proéminentes dans l'accent, mais partout plutôt effacées"[3] and furthermore: "Il faut donc admettre que le rôle de la hauteur, si important qu'il soit comme facteur de l'accent, reste accessoire."[3] Only duration is generally recognized as a consistent correlate of the final stress: "La durée est le seul des trois éléments acoustiques qui soit toujours, par sa proéminence, un facteur de l'accent."[4]

This study is based on the analysis of the F_o patterns of declarative sentences read by six native speakers of French (three males and three females). It does not concern spontaneous speech. Signals from an accelerometer attached to the throat and from a conventional microphone were recorded using a two-channel tape recorder. F_o contours were detected from the accelerometer waveform using a computer program (written by Shinji Maeda) and were displayed with envelopes of the corresponding speech signal on an oscilloscope. The reading material includes a paragraph and isolated sentences (listed in the appendix). The selection of the material was influenced by the previous results[5] of a systematic analysis of the F_o patterns for one professional speaker.

1. Use of F_o Contours to Demarcate Constituents within a Sentence

What is the role of F_o variations within sentences read by native speakers of French in a nonemphatic manner? In French, as in English, intonation is related to the grammatical organization of the sentences. (The relationship between intonation and syntax was observed quite early for English,[6] and some aspects of this relationship have been studied for French.[5,7-9]) The hypothesis within which our data are interpreted is that F_o variations have essentially a demarcative function: along with pauses and the longer duration of the final syllable F_o patterns mark the boundaries of the constituents of the sentence, and they generate the acoustic image of a simplified prosodic-syntactic tree structure. The prosodic-syntactic structure has, essentially, three levels: the first level is the sentence, the second is the sense group (or phrase) and the third is the word.

a. Sentence Level

Figure XVI-21 is the schema of the common code used by the seven speakers (1+6) for demarcating the boundaries between sentences: a sharp fall in F_0 before a pause indicates the end of a sense group in final position in a sentence; a rise followed by a

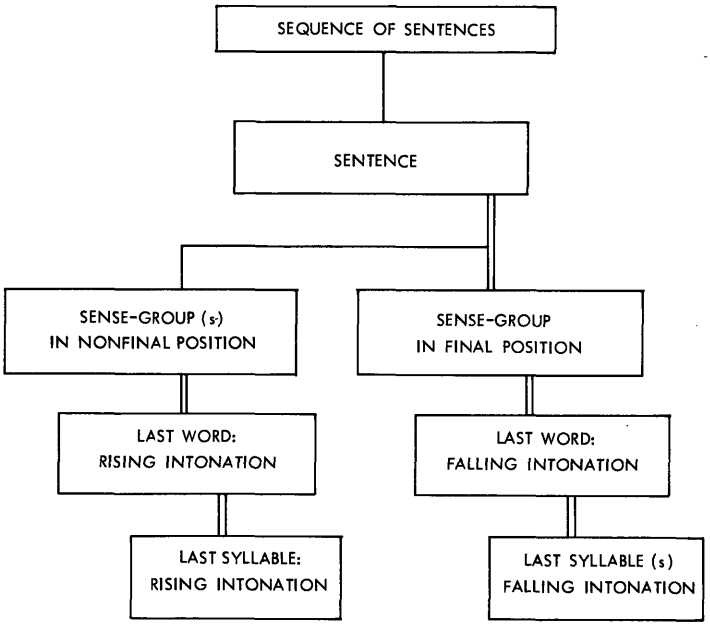

Fig. XVI-21. Schematic representation of strategy used for producing F_0 contours before pauses in nonfinal position and in sentence-final position.

pause indicates the end of a sense group in a nonfinal position in the sentence. In our analysis we have found only one exception: F_0 falls at the end of the major boundary in the fourth sentence of the paragraph (see the appendix) in the reading of one speaker.

b. Sense Group Level

The position of the sense group (final or nonfinal position) in a sentence not only determines the intonation of the last syllable of the group (falling or rising) but also influences the overall F_0 pattern for the whole group. Figure XVI-22 illustrates typical F_0 patterns for nonfinal phrases (NFP) and for final phrase (FP). Schematized contours for each phrase are indicated by dashed lines sketched above the actual contours. Figure XVI-22a and 22b represents contours for phrases composed of only one lexical word, in nonfinal and final position in a sentence, respectively (sentences 19a and 19b in the appendix). Figure XVI-22c and 22d displays the contours of the last two phrases

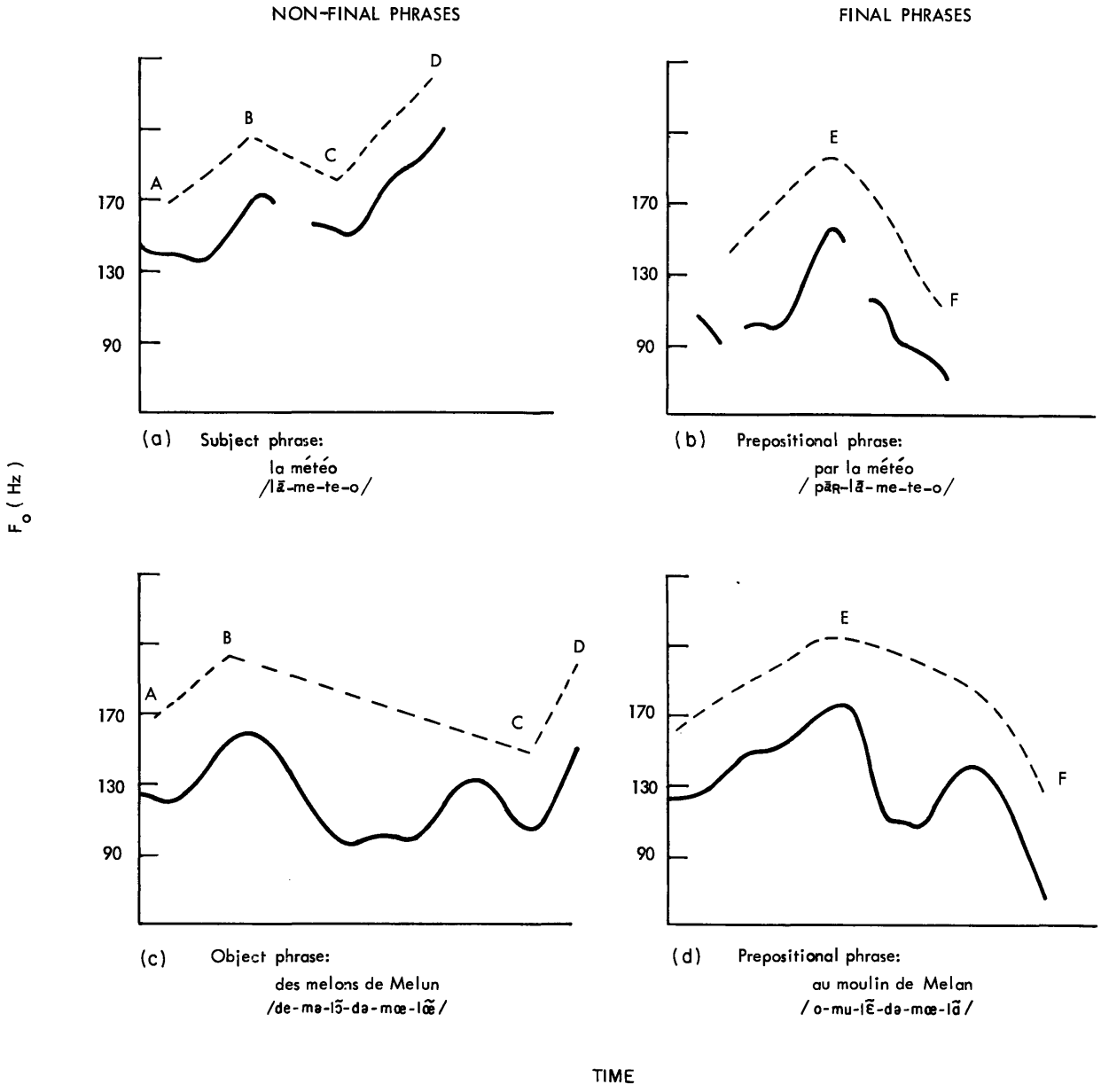

Fig. XVI-22. Typical F_o patterns for nonfinal phrases and for final phrases. Dashed lines are schematized representations of the contours in which local segmental influences have been ignored.

in the same sentence (sentence 5a in the appendix): the object phrase "des melons de Melun," and the prepositional phrase "au moulin de Melun." (Notice that the two content words of each sense group have the same semantic relationship.)

The F_0 pattern for a NFP is characterized primarily by a rise on its last syllable (segment CD on the schematized contours in Fig. XVI-22). An F_0 rise (segment AB) occurs at the beginning of the group, generally from the beginning (first syllable) of the first lexical word. (A smaller rise can be observed sometimes during the preceding function word, particularly at the beginning of a sentence.) F_0 reaches its higher value (point B) during the first lexical word of the group. The only exception that we found in our analysis was for the NFP "Grâce à ces associations avec les universités voisines ..." (Thanks to its associations ...): the maximum F_0 value was found on the word "grâce" (4 speakers) and also on the word "associations" (2 speakers). Then the overall F_0 contours for the whole group (segment BC) gradually fall until the final rise on the last syllable. Rising intonation (segment CD) creates an impression of stress on that syllable. For example, in Fig. XVI-22c the syllable "lun" in the word "Melun" is perceived as stressed. This final stress at the end of the group generally masked partially or completely the potential stress on the last syllable of every lexical word inside the group.

The F_0 pattern for an FP is quite different. This pattern is characterized primarily by a fall of F_0 from the beginning or near the beginning of the last lexical word in the sentence, if the last sense group is only one lexical word (preceded or not by function words). The fall is indicated by the segment EF on the schematized contour in Fig. XVI-22b. On the other hand, this fall (segment E'F' in Fig. XVI-22d) starts from the last syllable of the penultimate lexical word in the sentence if the last sense group is composed of more than one lexical word. (When FP is a monosyllabic content word preceded by function words, the maximum F_0 value generally occurs during the function words.) If the last lexical word of the group is a long word (more than 3 syllables, for example), the maximum F_0 value can be found either on the last or on the penultimate word: the F_0 maximum occurs on the syllable "tut" of the FP "... de l'institut de technologie" (sixth sentence of the paragraph) for three speakers, on the syllable "tech" for two speakers, and on the syllable "no" for another speaker. The lowest F_0 of the whole sentence is reached during the last syllable: F_0 not only falls close to the extreme lower limit of the speaking voice, but also the intensity level becomes so low that, as Coustenoble and Armstrong noticed,[1] it is difficult for a foreigner to hear it. This last syllable in the group is not perceived as stressed and the rule of stressing the last syllable of a sense group is not applicable at the end of a sentence. The perception of stress is shifted to the syllable in the group which has higher pitch. Using synthetic speech, Rigault[10] has shown that pitch is more effective than duration and intensity in producing the impression of stress for a French listener.

c. Word Level

Figure XVI-23 illustrates the next division, the division of the sense group into successive words. This figure displays the F_o pattern of a subject noun phrase (sentence 18a in the appendix), composed of three lexical words: "Un retour offensif de l'hiver ...". It can be observed in Fig. XVI-23 that the beginning of each lexical word is characterized

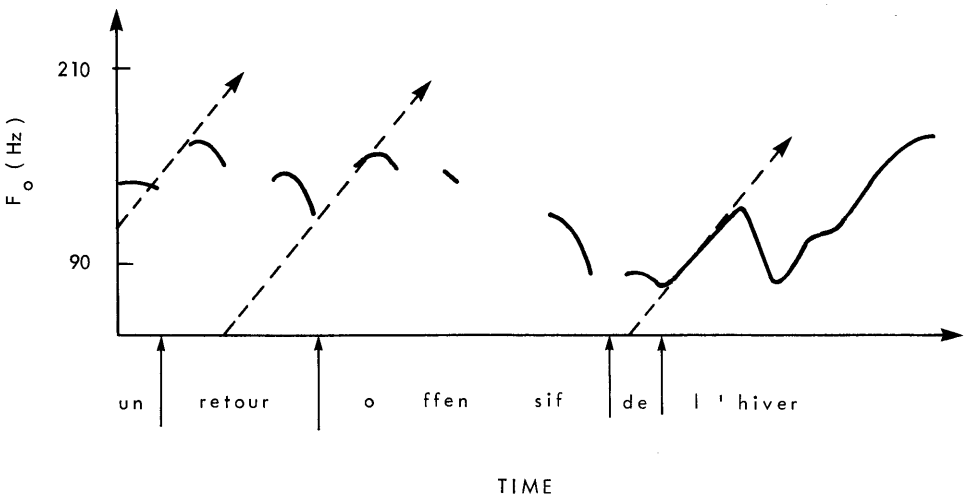

Fig. XVI-23. Use of the F_o contour to divide a sense group into words. Dashed lines indicate initial rise associated with each lexical word.

by an F_o rise. (The initial rises are indicated by dashed lines.) An initial rise may not occur for some words in the sentences, especially for those words that the speaker does not consider important. For example, when a lexical word is repeated twice in the sentence (such as the word "sciences" in the first sentence of the paragraph, and the word "consonnes" in sentence 21), the rise can be omitted in the second repetition of the word (three speakers omit the rise in the second word "consonnes"). Figure XVI-24 illustrates one of the cases in the NFP "... des consonnes initiales, des consonnes finales, ...". It may also happen that two successive lexical words are pronounced with the F_o pattern of one single word; the regrouping of the words has a tendency to shorten the total duration of the words. Regroupings are most likely to happen in rapid speech. Figure XVI-25 illustrates the F_o patterns found for the same NFP as in Fig. XVI-24. The speaker was asked to repeat the sentence more rapidly. In fact, the durations of the segments are almost unchanged, but the change in the F_o pattern creates the impression of more rapid speech. Initial rises are more or less important, depending on the speaker and the speed of elocution, but we noticed the same tendencies for all speakers:

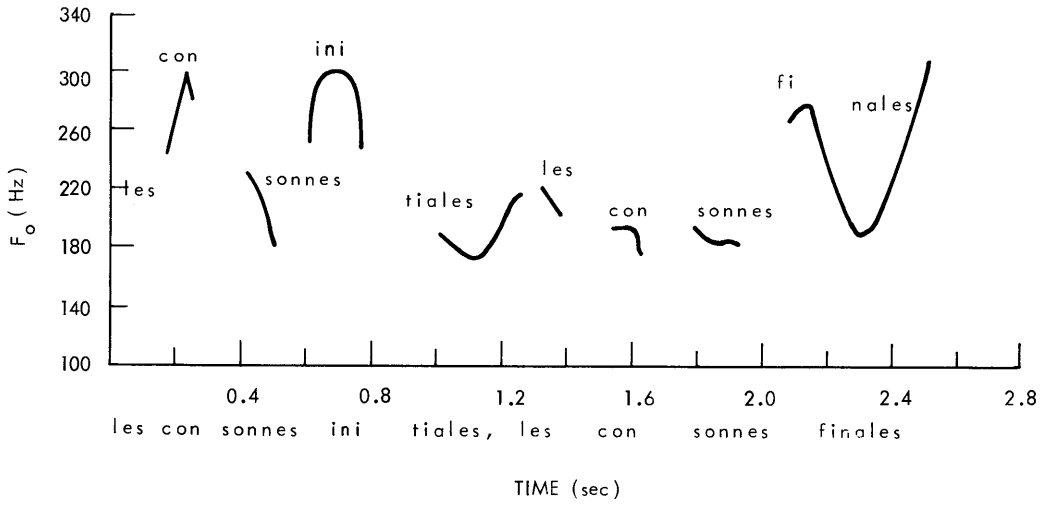

Fig. XVI-24. Lack of an initial F_o rise when a word ("consonnes") is repeated in a sentence.

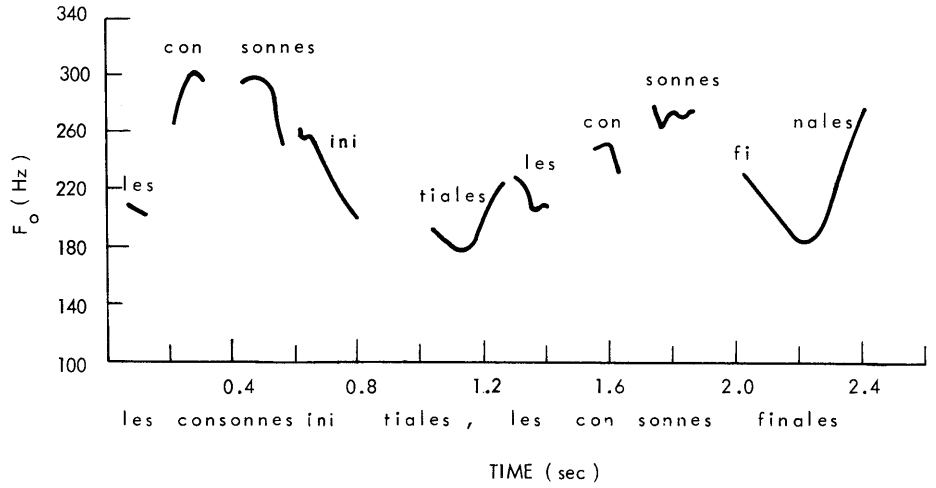

Fig. XVI-25. Pronunciation of a group of more than one word with the F_o pattern of a single word, giving the impression of more rapid speech.

(i) The initial rise on the first lexical word of the group is larger than initial rises on the following words.

(ii) Longer words (more than 2 syllables) have an initial rise more consistently than shorter words. (Three speakers suppress the initial rise in two-syllable words with rising intonation.)

(iii) The initial rise appears more clearly after function words than after another lexical word, and more clearly after a lexical word with falling intonation than after a

lexical word with rising intonation.

What is remarkable in the organization of the F_0 variations is not only the demarcative function of the F_0 patterns but also the subordination of each constituent to the one immediately above it. This fact is probably due to physiological constraints, as suggested by S. Maeda in Section XVI-D. The F_0 pattern for a shorter constituent is often shifted to the pattern of the immediately higher level constituent. For example, a short sentence (composed, say, of three or four lexical words and less than eight syllables) is equivalent to a sense group in final position in a longer sentence. The pattern for a sense group formed by two or three lexical words containing less than approximately five syllables is acoustically equivalent to the pattern of a single word. The boundaries between the successive words in the sense group cannot be seen from the F_0 pattern (such as in the NFP "les sciences de l'ingénieur" pronounced by all six speakers). Rapid speech produces the same effect: the constituents of the lower level (words) are less clearly demarcated, and the pause between two sense groups may be suppressed.

2. Interspeaker Differences in Actualization of F_0 Contours

All seven speakers in our experiment use F_0 patterns for demarcating the constituents of the sentences, but the actual F_0 contours differ from one speaker to another, and the differences are more marked for some speakers than for others. These differences are not only ascribed to anatonomical differences of individual speakers but also to an individual manner of the actualization of the demarcative features. We have found two kinds of major difference for F_0 patterns given by different speakers for the same sentence. The first kind of difference may be acoustically perceived and, to a certain extent, interpreted as semantically relevant; the second kind concerns only individual variants in the production of the prototype F_0 pattern, corresponding probably to differences in laryngeal gestures to accomplish a contour.

a. Speaker's Interpretation of the Relation between Words

The final word in a sense group has a rising or falling intonation, depending on the position of the group in the sentence. The intonation of the words within a sense group depends more on the speaker's own judgment of the closeness of the relation between the successive words. A falling intonation indicates the close semantic relation of a word to the next word. For example, an adjective followed by a noun has a falling intonation in the patterns for each of the seven speakers. (The two words, as we have mentioned, can also be regrouped into a single prosodic word.) A rising intonation, on the other hand, indicates that the word is relatively independent of the next word. Figure XVI-26 illustrates the actual F_0 contours of the subject phrase: "L'institut de technologie du Massachusetts ..." (first sentence of the paragraph), spoken by four speakers. Informal listening to the four phrases indicates that in the first case (Fig. XVI-26a), all of the

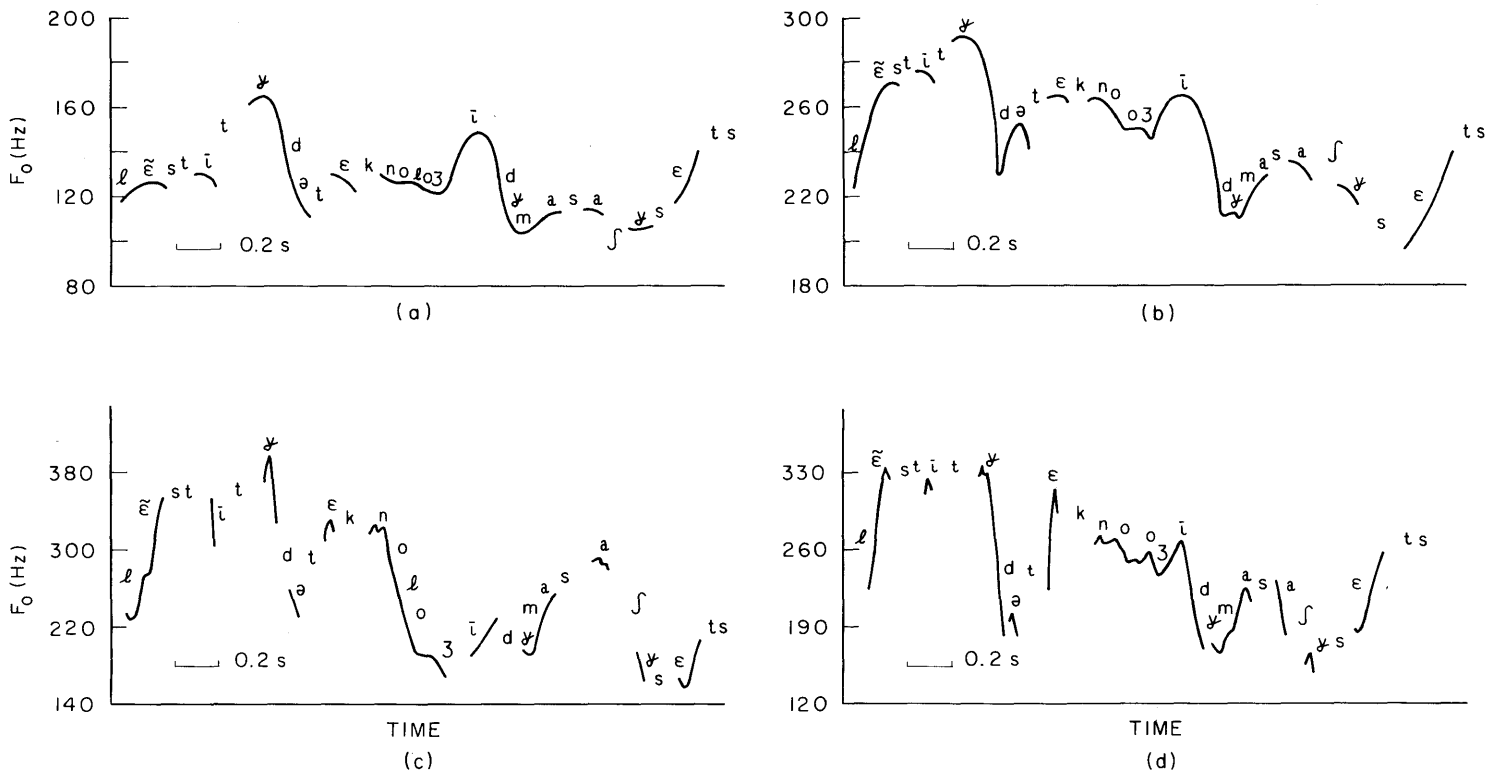

Fig. XVI-26. F_0 contours for the same phrase (l'institut de technologie du Massachusetts ...) spoken by four talkers, illustrating different strategies for grouping the words.

words seem to be equally important. In the second case, a slightly stronger stress is perceived at the end of the group. In the third case (Fig. XVI-26c), the relation between the first two words ("L'institut de technologie") seems to be much more close than the relation between the last two words of the group ("... de technologie du Massachusetts). In the fourth case (Fig. XVI-26d), the four words seem to form only one unit, and only the last syllable of the group is heard as prominent. It is possible, by a change in the F_o pattern, to give more information about the semantic relation between the words in sequences in a sense group. One of our speakers uses this approach in almost all sentences; the five other speakers use it occasionally. (The professional speaker used it only to disambiguate some sentences, and he gives a falling intonation to every word inside a nonfinal sense group in nonambigous sentences.) A greater degree of interspeaker variation in this respect has been found in the isolated sentences than in the text, where the context often provided enough information to specify the semantic relation between the words.

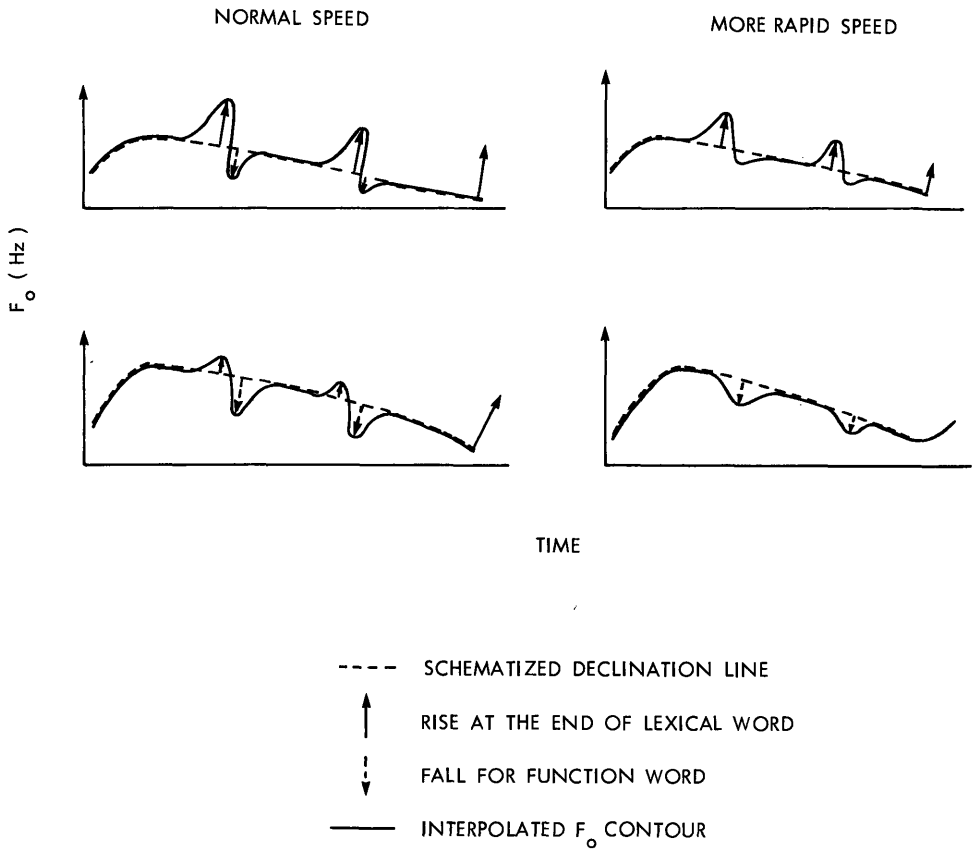

Fig. XVI-27. Schematized F_o contours for two speakers (upper and lower) at two rates of talking, illustrating individual differences in modifying the contours in rapid speech.

b. Individual Differences in Producing F_0 Patterns

It can be seen in Fig. XVI-26 that F_0 patterns for individual words differ greatly from one speaker to another. The differences among speakers may be due to different laryngeal gestures. Each speaker has his own characteristic pattern which he repeats in the various sentences: the similarities from one pattern to another are probably due to the same laryngeal gesture, with adjustments attributable to different phonological conditions.

Some of the components of the gesture (to increase or to decrease F_0 values) to accomplish a contour are more time-dependent for some speakers than for others. For example, when speaker 1 (Fig. XVI-26a) speaks more rapidly, the fall during function words tends to disappear, but the rise corresponding to the last syllable of the lexical word can still be easily detected visually from the F_0 contours. On the contrary, for the second speaker (Fig. XVI-26b) only the fall during the function words can be detected in rapid speech. Figure XVI-27 illustrates the two different tendencies.

Another speaker-dependent characteristic concerns the exact location of the final rise in the last syllable of a nonfinal sense group. For some speakers, the rise begins at the onset of the syllable (see one example in Fig. XVI-26a), while for others the rise starts only with the vowel (see one example in Fig. XVI-26b). Physiological data are needed before an adequate interpretation of these differences can be developed.

Appendix

Reading Material

1. Paragraph:

L'institut de technologie du Massachusetts est une institution mixte et privée, dont les centres d'intérêts principaux sont les sciences de l'ingénieur, les sciences pures et l'architecture. Il a contribué à la majorité des progrès technologiques des vingt dernières années et il continue à développer sa participation dans les techniques de pointe. La gamme des programmes de recherche est très vaste, et elle s'étend de la biologie à l'économétrie, en passant par la linguistique, l'électronique et les sciences nucléaires. Grâce à ses associations avec les universités voisines, les étudiants ont accès aux cours et aux recherches les plus variés qui soient. Le laboratoire de recherche en électronique a été construit à la fin de la seconde guerre mondiale. C'est le premier laboratoire de recherche interdépartemental de l'institut de technologie. Trois cent cinquante étudiants y conduisent des recherches, encadrés par une centaine de professeurs.

2. Isolated sentences:

1a – Il parle de la compatibilité entre deux êtres.
b – Il parle de l'incompatibilité entre deux êtres.
2a – Il dessine un rat d'eau au milieu de la marre.
b – Il dessine un radeau au milieu de la marre.
3a – Il dessine un chat dans un sac.
b – Il dessine un chateau sur un lac.
4a – La vie de ton ami est très intéressante.
b – L'avis de ton ami est très intéressant.
5a – J'ai acheté des melons de Melun au moulin de Melan.
b – J'ai acheté des melons de Melun au moulin.
c – J'ai acheté des melons de Melun.
d – J'ai acheté des melons.
e – Au moulin de Melan j'ai acheté des melons de Melun.
6 – La nouvelle bonne nous a annoncé une bonne nouvelle.
7 – La bonne, nouvelle victime, est bien vite repartie.
8 – La nouvelle, bonne à entendre, la réconforta.
9a – Qui va à Paris avec vous?
b – Ta cousine, Sophie, Roger, Bertrand et Raphael.
c – Ta cousine Sophie, Roger Bertrand et Raphael.
d – La cousine de Sophie Roger, Bertrand et Raphael.
10 – C'est en Espagne que j'ai vendu ma maison: j'ai vendu ma maison en Espagne.
11 – J'ai vendu la maison que j'avais en Espagne: J'ai vendu ma maison en Espagne.
12 – Le pilote ferme la porte.
13 – Le pilote, ferme, la porte.
14a – Il y a un fer à repasser dans le tiroir de la commode.
b – Il y a une robe à repasser dans le tiroir de la commode.
15 – La confédération générale du travail a organisé des manifestations et le conflit s'aggrave de jour en jour.
16 – Elle a organisé des manifestations importantes et le conflit s'aggrave.
17 – L'administration, les routes, les constructions avaient donné à cette contrée un certain essor.
18a – Un retour offensif de l'hiver est annoncé par la météo.
b – Un retour de l'hiver est annoncé par la météo.
c – L'hiver est annoncé par la météo.
19a – La météo annonce un retour offensif de l'hiver.
b – La météo annonce un retour de l'hiver.
c – La météo annonce l'hiver.

20a – Je pars en vacances.

b – Je pars en vacances cet après-midi.

c – Je pars en vacances cet après-midi à Trégastel.

21a – Nous allons parler successivement des consonnes initiales et des consonnes finales, des voyelles initiales et des voyelles finales.

b – Nous allons parler successivement des consonnes initiales et des voyelles finales, des consonnes finales et des voyelles initiales.

References

1. H. N. Coustenoble and L. E. Armstrong, Studies in French Intonation (W. Heffer and Sons Ltd., Cambridge, England, 1934).
2. P. Delattre, "L'accent final en français: accent d'intensité, accent de hauteur, accent de durée," in Studies in French and Comparative Phonetics (Mouton and Co., London, 1966).
3. Ibid., p. 66.
4. Ibid., p. 68.
5. J. Vaissière, "Contribution à la synthèse par règles du français," Thèse de troisième cycle, Université des Langues et Lettres de Grenoble, France, 1971.
6. D. Jones, An Outline of English Phonetics (W. Heffer and Sons Ltd., Cambridge, England, 1909).
7. G. Faure, "Contribution à l'étude du statut phonologique des structures prosodématiques," in Analyse des faits prosodiques. Studia Phonetica 3 (Didier, Montréal, 1971).
8. B. Malmberg, Les domaines de la Phonétique (PUF, Paris, 1971).
9. M. Rossi, "L'intonation prédicative dans les phrases transformées par permutation," Linguistics 103, 64-94 (1973).
10. A. Rigault, "Rôle de la Fréquence, de l'Intensité et de la Durée Vocalique dans la Perception de l'Accent en Français," Proc. Fourth International Congress of Phonetic Sciences, 1962, pp. 735-748.

F. SIGNALS FROM EXTERNAL ACCELEROMETERS DURING PHONATION: ATTRIBUTES AND THEIR INTERNAL PHYSICAL CORRELATES

U. S. Navy Office of Naval Research (Contract N00014-67-A-0204-0069)

William L. Henke

1. Introduction

Small accelerometers (less than 2 grams) which are easily affixed directly to external body surfaces in the region of the larynx transduce signals that may be interpreted to yield data on several aspects of laryngeal activity during phonation. Comparative evaluation of such signals across varying conditions of articulation, transducer locations, gas density (sound velocity), and speaker has suggested the association of specific features of the signals with mechanical and acoustical correlates, such as open and closed regions of the glottal cycle, distinct subglottal and supraglottal resonances, and an aspect of the characteristic wave motion of the vocal folds.

2. Measurement Technique

Small accelerometers are simply attached to external surfaces of the subject with double-sided adhesive tape. Accelerometers are available with a mass low enough (less than 2 grams) that the presence of the accelerometer does not significantly perturb the system that is being measured. This was verified by adding mass to the transducer and observing no change in the signals until the total mass became several times that of the unloaded transducer. Small electret microphones which were held onto the skin by tape on the distal side of the transducer yielded somewhat similar signals, although the accelerometer signal showed some of the features discussed here somewhat more clearly.

The transducer position for most measurements was on the midline in the suprasternal notch, with the sense axis in an anterior-posterior orientation. In such a position the transducer is typically 2-3 cm below the glottis. This position is to be assumed unless otherwise specified.

3. Typical Signals

Figure XVI-28 displays a simultaneous microphone signal and subglottal accelerometer signal. The time grid values are in seconds, and Fig. XVI-28 encompasses 51 ms of signal. In making visual comparison of features in the two waveforms it should be noted that distance from the glottis to the microphone was approximately 27 cm, which gives a propagation time delay of slightly less than 1 ms. We shall interpret various features of such waveforms.

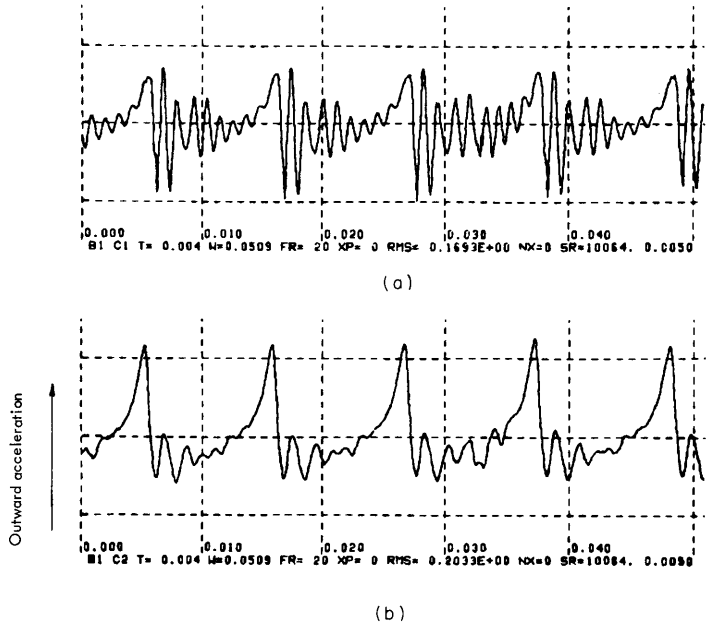

Fig. XVI-28. (a) Signal from audio microphone 10 cm from lips, vowel /a/.
(b) Simultaneous signal from an external accelerometer located on the midline in the suprasternal notch, with an AP sense axis.

4. Determination of Open and Closed Intervals

Figure XVI-29 shows a "synchronous sequential" plot of consecutive periods of the subglottal accelerometer signal for a changing supraglottal articulation, which can be interpreted to show the open and closed intervals of the glottis. Varying the supraglottal articulation varies the waveform of the acoustic pressure above the glottis, and when the glottis is open this varying pressure waveform propagates down through the glottis and manifests itself as a varying component of a signal transduced by a subglottal accelerometer. During the closed interval the accelerometer signal would be more independent of the supraglottal articulation. Thus, for the case shown in Fig. XVI-29, the left half of each waveform corresponds to the closed interval of the glottis, and the right half corresponds to the open interval. It can be seen that during the closed interval the accelerometer signal is very independent of the supraglottal articulation, and this also demonstrates that the accelerometer is quite free of coupling to the external or ambient acoustic field (that is, the normal speech signal).

For this particular display, the signal was segmented automatically at a prominent feature that occurs closely following the instant of closure, and is observed consistently over almost all conditions of phonation. This particular feature is called the "flyback stroke." (The segmentation algorithm includes linear and nonlinear signal processing, followed by event detection, which is followed by a discrete event selection logic that incorporates memory of neighboring periods.) These displays are plotted with inverted

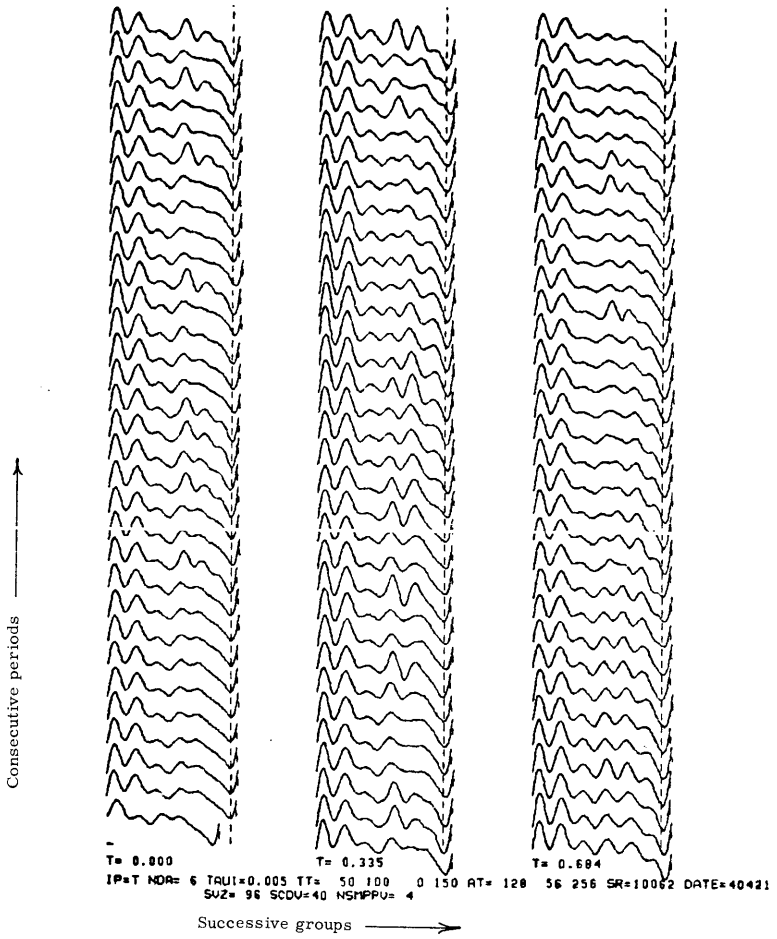

Fig. XVI-29. Consecutive periods of signal from an accelerometer in the suprasternal notch for changing suproglottal articulation — the vowel glide /i-æ/.

polarity, that is, outward acceleration is plotted downward.

Figure XVI-30 is a display of simultaneous consecutive periods of the signals transduced by a subglottal accelerometer and an audio microphone, for a vowel glide selected to show prominently a rise in F1 in the microphone signal. It can be seen that there is negligible effect of the supraglottal F1 in the subglottal signal during the closed interval. Also, an increase in damping in the microphone signal during the open interval is clearly observable. In some periods a glitch in the microphone signal is seen which corresponds to the instant of glottal opening.

5. Subglottal Resonances

Some of the stable features of the subglottal accelerometer signal observed during the closed interval of the glottis seem to be manifestations of characteristic resonances. Figure XVI-31 shows smoothed spectra of a 51-ms sample of simultaneous microphone

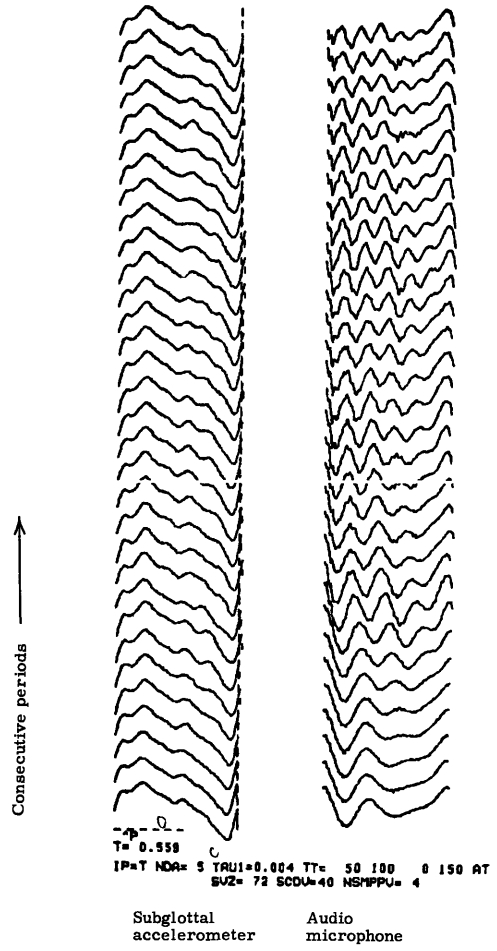

Fig. XVI-30.

Simultaneous subglottal accelerometer and normal audio signals for the vowel glide /i-æ/.

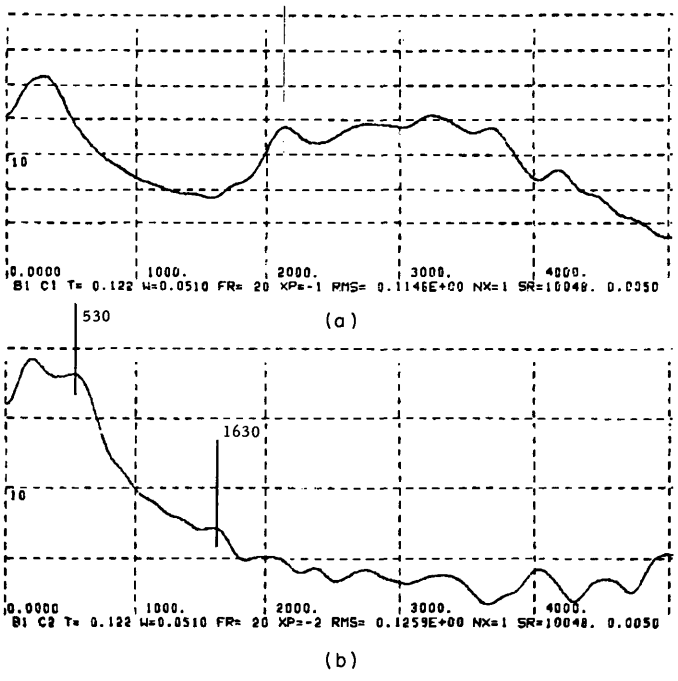

Fig. XVI-31.

Smoothed spectra of a simultaneous microphone signal (a) and subglottal accelerometer signal (b) for a male speaker saying the vowel /i/ in normal air. Horizontal grid lines are spaces 10 dB apart.

and subglottal accelerometer signals. (Spectrum smoothing lowpass time was 5 ms.) Two prominences in the subglottal spectrum, at 530 Hz and 1630 Hz, are unrelated to the supraglottal formants. The frequencies of these "resonances" are unaffected by additional mass loading of the accelerometer, an observation that suggests that the resonance is not an artifact of the measurement technique. When the subject breathed helium rather than air, the resonant frequencies significantly increased. Similar results were obtained for several different subjects, and for different supraglottal articulations. These data suggest that the resonances are due to acoustic rather than mechanical factors, and can be attributed to being "subglottal" or "lung" resonances.

Other reports of subglottal resonance frequencies have been given by van den Berg,[1] who extrapolated from measurements on the cadaver of a large dog to suggest the first

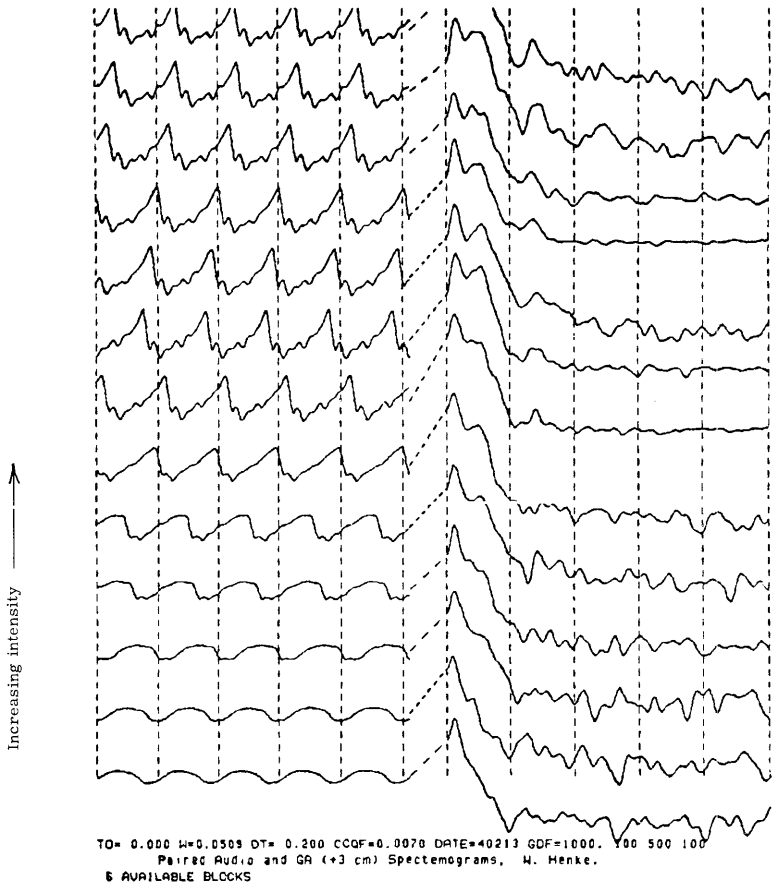

Fig. XVI-32. Paired time signals and smoothed spectra from a subglottal accelerometer, for different intensities of phonation (vowel /i/). As the intensity increases, the closed quotient increases and features attributable to subglottal resonances become more apparent in the associated spectra.

two human resonances at 314 Hz and 890 Hz; and by Fant et al.,[2] who reported measurements of the tracheal input impedance of 5 laryngectomized subjects and arrived at mean resonant frequencies of 640 Hz and 1400 Hz. The relative simplicity and noninvasive nature of the present measurement technique would recommend it for studies of cross-subject variations of subglottal resonances.

The effect of lung volume upon subglottal resonances has been studied informally with only a few subjects, but the following observations seem to be quite stable. There is relatively little change of resonance frequency with lung volume, but there appears to be a more significant change in the "prominence," or perhaps in the "bandwidth" or "damping," in a way that is consistently repeatable for some individuals, but in different ways for different individuals.

6. Relationship of Closed Quotient to Intensity

Figure XVI-32 shows successive samples of a single utterance during which the speaker slowly increased the intensity. The samples are 51 ms long, and spaced 200 ms in time. It would appear that the closed quotient (fraction of a period during which the glottis is closed) at low intensities was zero, that is, complete closure was never obtained, and that an increase in intensity caused an increase in the closed quotient. Associated with each time sample is the smoothed spectrum of that sample. The increasing prominence of the first and second subglottal resonances is clearly observable as the intensity increases. This is to be expected, since the subglottal resonances are clearly manifested in the subglottal accelerometer signal only during the closed portion of the glottal period.

7. The "Flyback Stroke"

A signal feature characterized by a rapid change from outward to inward acceleration immediately following the maximum outward acceleration that occurs at or shortly after the instant of closure (whenever complete closure obtains) is the most prominent and stable signal characteristic. We refer to this feature of the signal as the "flyback stroke" (so named because of its signal characteristics rather than any production mechanism correlate). The time of the negative-going zero crossing of this stroke thus provides a stable segmentation point for delimiting individual periods. Such points are useful for synchronous displays and analyses, and for pitch period and period jitter measurements.

Our suggestion of the primary physical correlate of the flyback stroke is as follows. Simultaneous observations with different transducer locations suggest that this feature is attributable to tracheal sound pressure rather than to mechanical coupling of the accelerometer to the vibrating structures. The precursor of outward acceleration is caused by an increase in subglottal pressure as the glottis is narrowing. Several authors[3] have

reported on the characteristic wave motion of the vocal folds, wherein closure occurs first at the lower extremity, and the lowest point of closure propagates up as the lower edges begin to separate again. Thus there is an effective expansion of the volume of the immediate subglottal region during the closed interval.

Dynamic measurements by Baer[4] on dog larynxes of approximately the same size and shape as human larynxes show a frontal section area expansion (below the lowest point of closure) during the first quarter of the closed interval of approximately 5 mm^2 for a moderately loud level of phonation. During the remainder of the closed interval expansion continues, but at a much reduced rate. For a vocal cord length of 15 mm this results in a volume expansion of .075 cc. The corresponding time interval is approximately 1 ms and so the effective volume velocity is 75 cc/s. This volume flow from the characteristic impedance of a trachea of 3 cm^2 area would generate a pressure of 1050 dyn/cm^2. Such a pressure drop is comparable to the magnitude of the total pressure variation. Hence the flyback stroke might reasonably be attributed to subglottal volume expansion immediately following closure.

8. Summary of Microstructure Interpretations

Figure XVI-33, which is an augmented version of Fig. XVI-28, summarizes our interpretation of the physical correlates of a signal from a subglottal accelerometer. The

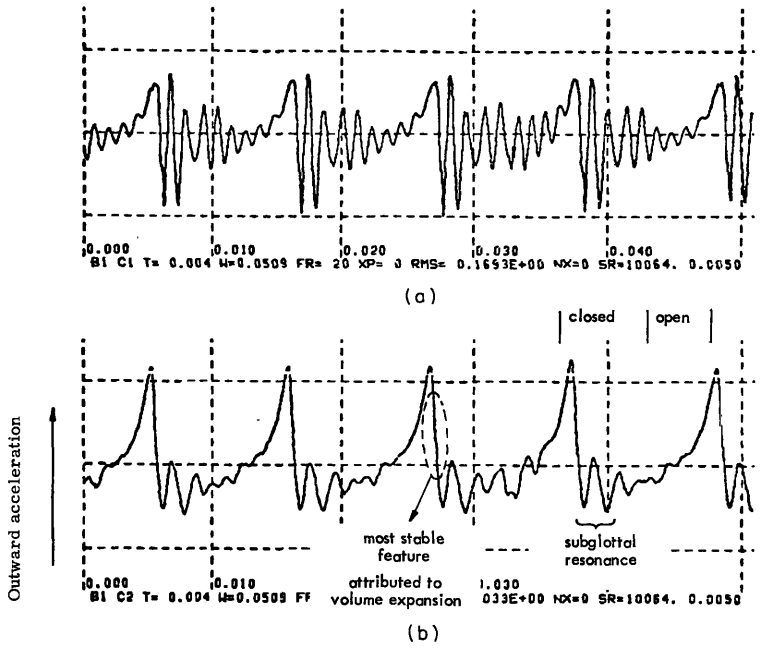

Fig. XVI-33. (a) Signal from audio microphone 10 cm from lips, vowel /a/.
(b) Simultaneous signal from an external accelerometer located on the midline in the suprasternal notch, with an AP sense axis.

open and closed regions are marked, as is the manifestation of the lowest subglottal resonance. In the first quarter of the closed interval the flyback stroke attributed to subglottal volume expansion can be seen.

The intraperiod features are also useful in studying interperiod variations. The synchronous sequential display format gives visual prominence to the characteristics of voicing onset and offset, and shows aperiodic anomalies which may be of interest in pathological cases. A display format showing tracks of fundamental frequency with a simultaneous measure of amplitudes of individual pulses indicates clearly the prosodic aspects of phonation. In a subsequent report we shall be concerned with these issues.

References

1. Jw. van den Berg, "An Electrical Analogue of the Trachea, Lungs, and Tissues," Acta Physiol. Pharmacol. Neerl. 9, 361-385 (1960).
2. C. G. M. Fant, K. Ishizaka, J. Lindqvist, and J. Sundberg, "Subglottal Formants," Speech Transmission Laboratory Quarterly Progress Status Report 1/1972, Royal Institute of Technology, Stockholm, pp. 1-12.
3. For example, S. Smith, "Remarks on the Physiology of the Vibrations of the Vocal Cords," Folia Phoniatr. 6, 166-178 (1954).
4. Thomas Baer, "Investigation of Phonation Using Excised Larynxes," Ph. D. Thesis, Department of Electrical Engineering, M. I. T. (to be submitted in September 1974).