COMMUNICATION SCIENCES
AND
ENGINEERING

IX. SPEECH COMMUNICATION*

Academic and Research Staff

Prof. K. N. Stevens	Dr. Mary C. Bateson	Dr. D. H. Klatt
Prof. M. Halle	Dr. Margaret Bullowa	Dr. Paula Menyuk
Prof. W. L. Henke	Dr. A. W. F. Huggins	Dr. J. S. Perkell
Prof. A. V. Oppenheim	Dr. R. D. Kent	A. R. Kessler

Graduate Students

D. E. Dudgeon	B. Mezrich	H. A. Sunkenberg
R. W. Hankins	M. R. Sambur	R. N. Weinreb
Emily F. Kirstein	J. S. Siegel	M. L. Wood, Jr.
R. M. Mersereau		V. W. Zue

A. ON PREDICTING THE DURATION OF THE PHONETIC SEGMENT [s] IN ENGLISH

1. Introduction

There have been several studies of segmental duration in English,[1-3] but most of this work has been restricted to very limited phonetic contexts. Very little is known about the overall framework within which the known rule-governed aspects of duration operate, and the completeness of the phenomena hitherto described is open to question.

Our pilot study investigates several issues that arise in connection with a generative theory of consonantal duration in English. We chose a single phonetic segment [s], having a rather well-defined and measurable duration, in order to restrict the data base to manageable dimensions.

We wish to determine which word-level phenomena have an influence on the duration of [s]. Specific questions concern the influence of the number of syllables in the word, the [s] position within the word, the stress level assigned to the various syllabic nuclei, and the morphemic and phonetic structure of the word. Words to be studied are spoken in a single-frame sentence at a moderate speaking rate. Thus any systematic effects of speaking rate,[4] syntactic position,[3] semantic importance,[5] or rhythmic temporal compensation[6] will not be considered.

2. Experiment

A list of words was constructed in such a way that the sound segment [s] appears in several phonetic environments. The word list was randomized and recorded in an

*This work was supported in part by the U. S. Air Force Cambridge Research Laboratories under Contract F19628-69-C-0044; and in part by the National Institutes of Health (Grant 5 RO1 NS04332-09) and M.I.T. Lincoln Laboratory Purchase Order CC-570.

Table IX-1. The duration of the single consonant [s] in ms for a list of words recorded by three speakers.

INITIAL PRESTRESSED (PRIMARY)	RK	KNS	DHK
1. sat	175	160	115
2. my seat	165	160	120
3. sister	150	130	105
4. sissy	145	135	115
5. new supper	135	130	120
6. Samson	135	145	100
7. soapsuds	140	130	100
8. singsong	160	145	110
9. seaside	160	150	105
10. sesame	130	130	90
11. somersault	125	135	105
12. sensible	130	130	100
13. saxophone	135	135	90
14. satisfy	120	135	100
15. seersucker	130	150	115
INITIAL PRESTRESSED (SECONDARY)			
16. sextet	145	125	95
17. semipro	120	150	110
18. situation	110	105	85
19. satisfaction	115	115	100
20. sensibility	120	130	75
INITIAL UNSTRESSED			
21. severe	130	130	95
22. seduce	125	130	105
23. sedation	130	120	100
24. secession	100	100	100
25. September	115	105	85
26. sensational	120	120	85
27. sophisticated	105	125	100
MEDIAL PRESTRESSED (PRIMARY)			
28. ascend	165	140	105
29. aside	145	135	105
30. beside	130	145	105
31. missout	125	115	100
32. secession	115	110	115
33. Mississippi	135	140	120

Table IX-1. (continued)

MEDIAL PRESTRESSED (SECONDARY)	RK	KNS	DHK
34. seaside	150	135	100
35. bisect	120	115	90
36. somersault	130	100	100
37. seersucker	135	115	90
MEDIAL UNSTRESSED			
38. crisis	130	105	85
39. rhesus	115	135	100
40. person	130	115	110
41. arson	120	125	100
42. crescent	90	130	90
43. blossom	110	120	100
44. sissy	130	125	105
45. bossy	110	110	100
46. crossing	115	90	85
47. blessing	115	115	85
48. missing	125	115	90
49. cursive	125	115	110
50. purser	145	135	100
51. placid	105	110	90
52. rustle	100	135	100
53. missile	100	130	110
54. personal	100	130	90
55. ascertain	115	100	95
56. asinine	105	115	90
57. sesame	115	100	95
58. Mississippi	105	105	95
WORD-FINAL			
59. geese	145	140	115
60. goose	150	110	90
61. mice eat	140	110	100
62. noose upper	130	105	95
63. seduce	145	145	95
64. misplace	135	125	100
65. crisis	90	115	85
66. rhesus	105	115	100
67. priceless	130	120	—

Table IX-2. The duration of [s] in ms for a list of words containing [s]-embedded consonant clusters is tabulated for three speakers.

MEDIAL PRESTRESSED (PRIMARY)	RK	KNS	DHK
1. misplace	90	70	55
2. mistake	115	95	85
3. misspell	105	155	130
4. misstep	115	160	95
5. sextet	90	85	55
6. crustacean	65	75	75
7. sensational	120	110	80
8. satisfaction	90	90	65
MEDIAL PRESTRESSED (SECONDARY)			
9. curbstone	70	85	60
10. soapsuds	130	95	95
11. singsong	120	115	80
12. boxcar	75	75	65
MEDIAL UNSTRESSED			
13. crusty	55	90	75
14. sister	70	100	70
15. crystal	70	90	55
16. Boston	45	85	55
17. Samson	100	120	70
18. boxer	115	100	85
19. priceless	115	120	85
20. crystalline	60	55	60
21. currency	130	105	95
22. saxophone	85	75	60
23. sensible	90	90	70
24. sensibility	95	95	60
25. sophisticated	65	50	45
26. satisfy	105	75	60
WORD-FINAL			
27. beets	130	135	85
28. beaks	130	120	85
29. beeps	145	115	85
30. crest	80	120	85
31. crisp	75	115	65
SYLLABLE-FINAL (ILLEGAL CLUSTER)			
32. crossroad	105	100	65
33. misread	85	105	70
34. baseball	60	85	45
35. misdeed	90	115	70

anechoic chamber by 3 adult male speakers. The words were spoken in the frame sentence "Say ____ instead" in order to control speaking rate and to avoid the durational variability that can occur in either the initial or final position in an utterance.

The word list is given in Tables IX-1 and IX-2. In Table IX-1, the segment [s] appears as a single consonant in initial, intervocalic, and final position in words containing different numbers of syllables and different stress patterns. A limited number of consonant clusters containing [s] in medial and word-final position are listed in Table IX-2. Word-initial clusters containing [s] will be examined as a part of a more extensive future study of cluster duration in English.

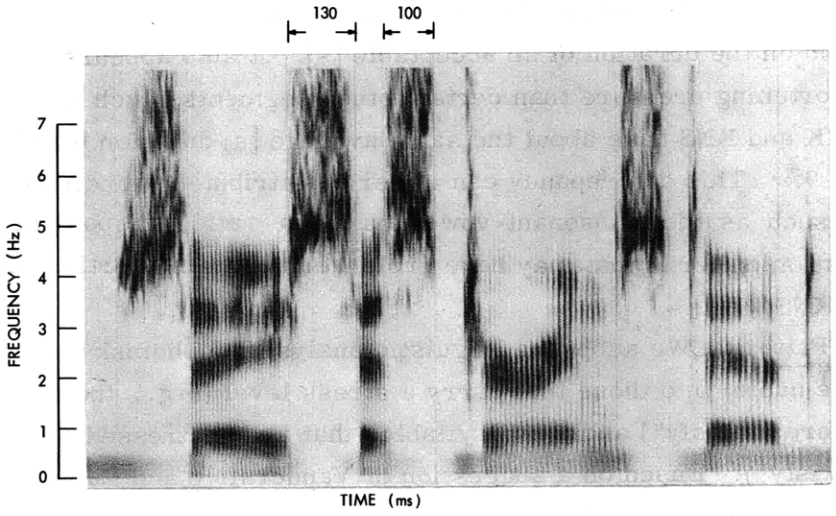

Fig. IX-1. Broadband spectrogram of the utterance "Say sister instead" spoken by KNS. Vertical lines have been drawn to aid in the measurement of the [s] duration.

A broadband spectrogram was made of each word. The duration of an [s] segment was determined by a technique illustrated in Fig. IX-1. Vertical lines were drawn to delimit the onset and offset of [s]. In general, these times were clearly descernible as a sudden onset or offset of turbulence noise in the frequency region from ~4 to 7 kHz. The length between the vertical lines was then determined in cm and converted to time in ms. We estimate that this procedure has a measurement error of ~±5-10 ms, which is considerably less than the measured variability of the speakers.

3. Results

Measured durations for each [s] in our word list are given in Tables IX-1 and IX-2 for the three speakers. A number of uniform tendencies can be seen in these data; some of them are consistent for all speakers, while others are probably due to individual habits.

1. Speaking Rate. The average speaking rates of the three subjects RK, KNS, and DHK, can be compared by computing the average duration of the utterance "Say _____ instead" for the entire 79-item word list. The average duration from the onset of visible frication in the s of "Say" to the instant of closure for the d of "instead" was 1650 ms for RK, 1800 ms for KNS, and 1540 ms for DHK.

Speaker DHK produced phonetic segments faster than the other speakers, in part by shortening the inherently long prestressed [s] more than they did. Averaged across all of the data in Table IX-1, the duration of the segment [s] was 127 ms for RK, 125 ms for KNS, and 100 ms for DHK. In word-initial prestressed position, the average [s] duration in ms of the three speakers was 142 (RK), 140 (KNS), and 106 (DHK). In other phonetic environments, the speakers differed much less, which might suggest that there is a lower bound on the duration of an acceptable [s]. It also appears that [s] is more resistant to shortening pressure than certain other segments, such as vowels.

Note that RK and KNS have about the same average [s] duration but that their speaking rates differ by 9%. This discrepancy can either be attributed to individual differences in a variable such as the consonant-vowel duration ratio, or possibly to different utterance-duration patterns that may have prolonged the frame portion of the sentence in the case of KNS.

2. Stress Pattern. We adopt the linguistic analysis of Chomsky and Halle[7] which divides syllable nuclei into those that carry a stress level (e.g., the first and third syllables of the word "satisfy") and those syllables that are unstressed (e.g., the second syllable of "satisfy"). Based on a suggestion of Vanderslice and Ladefoged,[8] we have divided stressed syllables into those that carry primary stress (e.g., the first syllable of "satisfy") and those that have stress levels lower than primary (e.g., the third syllable in "satisfy"). In the following analysis, the latter set will be called secondary stress. Speakers recited the frame sentence in such a way that primary stress always fell on a syllable of the word that was inserted in the frame.

The duration of the single consonant [s] appearing in word-initial and in medial intervocalic position has been plotted in Fig. IX-2 as a function of the stress level of the following vowel. This figure indicates that there is a clear tendency for the duration of prestressed [s] to be longer than the duration of pre-unstressed [s]. If we ignore single-syllable words (which have the longest [s]'s), then the prestressed [s] is approximately 15% longer in duration than pre-unstressed [s] in a word with an equivalent number of syllables.

It appears that secondary stress and primary stress have approximately the same effect on the duration of [s] when the effect of the number of syllables in each word is considered. This result, however, is not conclusive because there are only a limited number of secondary-stress cases. There may be a slight tendency for [s] to be somewhat shorter preceding vowels with secondary stress. Nevertheless, in most of the

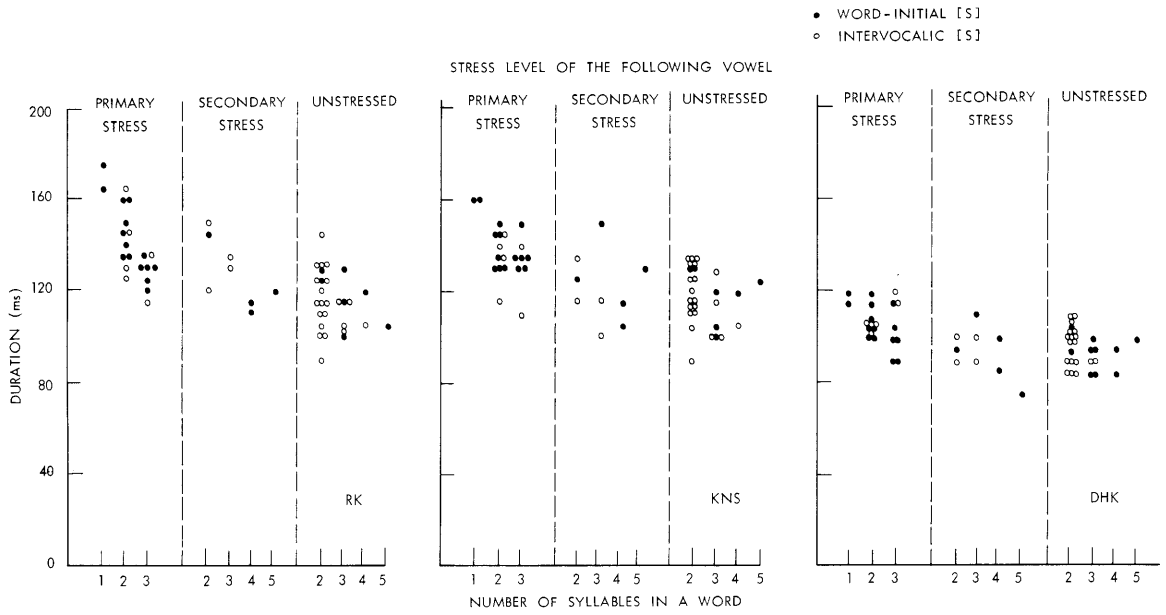

Fig. IX-2. Duration in ms of [s] plotted for three speakers as a function of the number of syllables in the word, the stress on the following vowel, and the position of the [s] in the word.

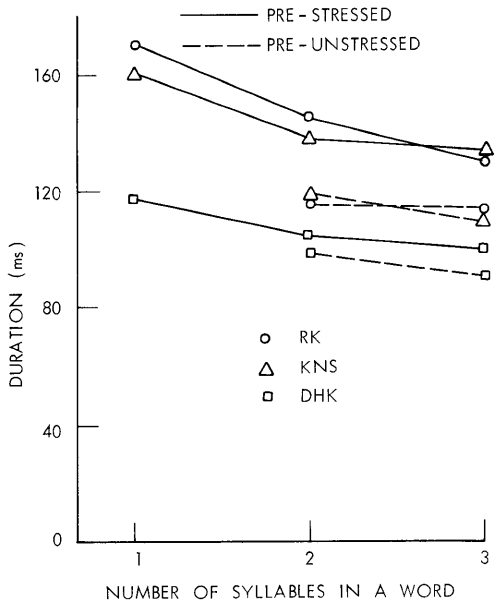

Fig. IX-3.

Average duration in ms of [s] plotted for three speakers as a function of the number of syllables in the word for prestressed and pre-unstressed [s].

following discussion, all stressed vowels will be grouped together.

3. Number of Syllables. Some authors have found that the duration of a word with many syllables is shorter than one might expect from the observed durations of single-syllable words. For example, Barnwell[3] found that a vowel nucleus in a two-syllable word is approximately 70% of its duration in a one-syllable word, and this figure approaches 50% in a multisyllable word.

The data of Fig. IX-2 suggest that a similar phenomenon holds for the duration of [s]. The data have been replotted in Fig. IX-3 to indicate the average [s] duration for one-, two-, and three-syllable words. The [s] duration in prestressed position is 15% shorter in two-syllable words and 20% shorter in three-syllable words than the [s] duration of one-syllable words. In pre-unstressed position, there is also an approximate 5% difference between the [s] duration of two- and three-syllable words.

The observed multisyllable shortening effect appears to be less pronounced on [s] than it is on vowels. This is perhaps not surprising when we consider that consonants are typically shorter than vowels, and that in English consonants carry a greater information load.

4. Syllable Boundaries and Morpheme Boundaries. The words "person" and "missing" both contain a medial pre-unstressed [s]. It could be argued, however, that for words of this type (words 38-58 in Table IX-1) the [s] is sometimes more properly called post-stressed and sometimes called pre-unstressed either on the basis of morpheme boundary placement or according to some kind of syllable-boundary definition. Thus it can be hypothesized that the rather large variance within the pre-unstressed classification could be accounted for by regrouping the words according to some independently motivated criterion. The following test was performed in order to check this alternative.

The 21 words with a pre-unstressed [s] were rank ordered according to [s] duration for each of the three speakers. The rank orders did not seem to be consistent from speaker to speaker, so a pairwise Spearman rank-order correlation was computed. The correlations were −0.17, +0.07, and +0.40. The last value is significant at the 0.1 level although we might expect one of the correlations to be that high 30% of the time in random normally distributed data.

Even though the speaker uniformity was not pronounced, we decided to continue the test by reordering the list according to the size of the sum of the rank orders of the three individual speakers. This ordering is presented in Column 1 of Table IX-3. If the list contained words from two different classes having inherently different [s] durations, members of one class would tend to cluster at the top of the list and members of the other class cluster at the bottom. The placement of a morpheme boundary before or after the [s] is indicated in the second column of Table IX-3. A syllable-boundary definition, based on the concept that the first vowel in the word captures the [s]

Table IX-3. Combined rank order (from short to long fricative duration) of words containing intervocalic pre-unstressed [s]. The presence of a morpheme boundary, phonetic syllable boundary (see text), or dictionary syllable boundary before (B) or after (A) the [s] is indicated, whenever possible, in columns 2, 3, and 4 respectively.

	WORD	Morpheme Boundary	Phonetic Syllable Boundary	Dictionary Syllable Boundary
1.	crossing	A		A
2.	placid		A	
3.	Mississippi		A	
4.	asinine		A	A
5.	blessing	A	A	A
6.	crescent		A	
7.	ascertain		A	
8.	sesame		A	A
9.	crisis			B
10.	personal		A	B
11.	bossy	A		A
12.	missing	A	A	A
13.	blossom			
14.	rustle		A	A
15.	missile		A	
16.	cursive		A	B
17.	arson			B
18.	rhesus			B
19.	person		A	B
20.	sissy		A	
21.	purser	A	A	A

if the vowel is lax, is given in column 3. Column 4 contains a syllable-boundary definition based on the orthography given in Webster's New World Dictionary.[9] None of these processes produced the expected clustering, although use of the dictionary definition came closest to showing such an effect. We conclude that morpheme structure and syllable-boundary placement do not influence the determination of the duration of [s] in intervocalic pre-unstressed position in English.

Within the pre-unstressed category, word-initial and medial [s] have the same durational characteristics. Therefore the two classes can be grouped together in computation of consonantal duration.

5. Word Final [s]. The duration of word-final [s] is generally short. Taking into account the previously noted effect of the number of syllables on the duration of [s], the average word-final [s] duration is equal to the average pre-unstressed [s] duration. As spoken in the particular frame sentence, all word-final [s]'s are also pre-unstressed. Therefore it is unclear whether a word-final [s] is short because it is word final or because the vowel in the next word happens to be unstressed. Some additional minimal-pair data (described below) support the notion that the key factor is that the [s] is in word-final position.

Throughout the data reported thus far, the critical parameters determining the durational characteristics of [s] are (i) whether the following vowel is stressed or unstressed, and (ii) the number of syllables in the word. This simple pattern is maintained if we group all word-final [s]'s together with the pre-unstressed class.

6. Word Boundary. Table IX-1 contains two minimal pairs which differ in the placement of a word boundary: "my seat" vs "mice eat" and "new supper" vs "noose upper." They were included because Lehiste[10] had observed previously that in her set of minimal pairs involving the placement of an internal open juncture, [s] was usually longer in word-initial position than in word-final position.

In the data reported here, the [s]'s were longer in word-initial positions for both pairs for each speaker. The average difference in [s] duration is 25 ms which is in substantial agreement with the average difference between prestressed and pre-unstressed [s].

7. Clusters Containing [s]. Since the durational data on clusters containing [s] are incomplete, only the more dramatic cluster effects will be described. The most striking example is the situation in which [s] is followed by a stop. The [s] duration is shorter in this case than for any other phonetic environment. The average pre-stressed [s] duration is 76 ms, and in pre-unstressed position, when followed by a stop it is only 67 ms.

The word list contains a few examples of medial clusters involving a nasal or a plosive followed by [s]. The [s] in these two-element clusters is somewhat shorter than a single [s] would be in the same environment. The average durational reduction is

approximately 15% with respect to a single [s] in prestressed position. A pre-unstressed cluster [s] is also shortened approximately 15% relative to a single [s] in pre-unstressed position. The data are in substantial agreement with data reported by Haggard[11] and by Schwartz.[12]

8. Underlying Double [s]. As previously mentioned, the morphemic decomposition of a word does not influence the duration of a medial pre-unstressed [s]. Another possible influence of morpheme structure is illustrated by the word pairs "misspell" vs "misplace" and "misstep" vs "mistake." The first member of each pair has an underlying representation with a double [s]. For example, the final [s] of the morpheme "mis-" is concatenated with the initial [s] of the morpheme "spell" to form the word "misspell."

Durational measurements indicate a consistent difference between members of each word pair. The underlying double [s] is longer in both cases for all speakers. A double [s] followed by a stop averages approximately 25% longer than would be expected for a single [s] followed by a stop.

9. Variability. Most of the rule-governed durational changes that we have observed are small. It is therefore relevant to examine more precisely the variability in the data. The standard deviation of the measured mean [s] duration was computed for two cases for which there are sufficient data to obtain a reliable result.

The standard deviation for prestressed [s] in two-syllable words for RK, KNS, and DHK is 12.9, 10.4, and 7.3 ms, respectively. For medial pre-unstressed [s] in two-syllable words, the somewhat larger figures are 13.7, 12.4, and 8.1 ms, respectively. Thus a prestressed [s] duration is more carefully timed than a pre-unstressed [s], even though a prestressed [s] has a longer basic duration.

Speaker DHK, who had the fastest average speaking rate, exhibited the smallest variability in his durational data and the smallest relative variance.[13] This may indidate that timing generally becomes more precise as the speaking rate of a talker increases.

4. Summary and Discussion

While our numerical results obviously apply directly only to the phonetic segment [s], the factors that influence [s] duration are probably involved in the determination of duration for many consonants in English. Some of these factors may be physiologically conditioned, and thus may operate as universal tendencies in language. In order to stimulate study of these questions, we present a tentative model to be used in predicting [s] duration in English words. We have not considered the possible influence of syntactic and semantic variables on consonant duration.

1. Durational Model for [s]. Our model is based on average data from three subjects speaking at a moderate rate. The expected duration of the phonetic segment [s]

Table IX-4. The combined average duration of [s] is tabulated as a function of the number of syllables in the word for prestressed and non-prestressed [s]. The numbers in parentheses are the number of sample durations from which the averages were computed.

Number of Syllables	Duration (ms)	
	Prestressed	Pre-unstressed and word-final
1	145 (6)	119 (12)
2	127 (42)	112 (68)
3	119 (30)	104 (15)
4 or more	113 (12)	106 (6)

depends on certain characteristics of the word in which it is embedded. The duration decreases as the number of syllables in a word increases. The duration is greater if the [s] appears in prestressed position and less if it appears in pre-unstressed position or in word-final position. The duration does not depend on the degree of stress on a stressed syllable nor on the position of the syllable within the word. These effects have been quantified, and the results are summarized in Table IX-4 for the case in which the [s] is not a part of a consonant cluster.

If [s] is followed by a plosive in a two-element cluster, the [s] duration is shortened to 60% of the value given in the table. If [s] is preceded by a nasal or plosive, the [s] duration is shortened to 85% of the given value. Other clusters involving [s] have yet to be studied.

The [s] duration is not influenced by the position of any morpheme boundaries within a word. An exception to this general rule has been discovered for words containing an underlying double [s]. If a morpheme final [s] is concatenated with a morpheme initial [s] as in the word "misspell," the double [s] is approximately 25% longer than one would expect for a single [s] in the same environment.

2. Implications for Speech Perception. The (partial) theory of segmental duration in English that has been outlined in this report, if substantially correct in this form, has some interesting implications for the way in which sentences are perceived by the listener. Durational cues appear to be organized according to a generative rule system, thus adding segmental duration as another aspect of language structure that is most easily set in a generative framework.[7] Assuming that speech-decoding rules cannot be organized in a way essentially similar to the generative rules, this report gives support at an acoustic level to the theory of Halle and Stevens[14] that speech perception follows an analysis-by-synthesis strategy. In this strategy, decoding rules of thumb of general validity are used to generate hypotheses about the contents

of an unknown utterance. The generative rules are then used to compute a synthetic utterance having these characteristics. The synthetic utterance is compared with the input representation, and discrepancies are used to generate new hypotheses until a satisfactory match is obtained.

The importance of the decoding rules of thumb in this process cannot be overemphasized. While they may not be correct in every case, these rules provide the only means of processing acoustic data, aside from strict memorization. At least two useful rules are suggested by this research.

i. Duration serves as a cue to word-boundary location. In our data, word boundaries may occur before long [s], but never after long [s]. A word boundary may occur before or after a short [s].

ii. A series of short syllables with very short [s] durations suggests the presence of a multi-syllable word. A very long [s] suggests the beginning of a single-syllable word.

We have not observed many large rule-governed durational changes in [s]. Typical average changes in duration are 30 ms or less, and random or unexplained variations in the measured durations are comparable. One standard deviation ranges from ~5-12 ms. It is therefore questionable whether we can distinguish systematic changes in duration to aid us in understanding speech perception, or whether such changes are buried in the random component of segmental duration.

This question is best answered by observing [s] duration in sentence material rather than with single words because although normally a speaker is not very careful in adhering to the durational rules, he will accentuate the tendencies described in the rules when he attempts to speak a phrase clearly or to remove ambiguities from a phrase or sentence.

The possibility of determining word-boundary locations directly from acoustic cues in English has been of interest for some time. Certain consonant clusters cannot occur within a word but can occur in a sequence of two words, in which case word boundaries are easily established. Lehiste[10] has catalogued several acoustic cues that are employed by listeners to clarify carefully spoken word pairs differing only in word-boundary location. Interpreting her data, we postulate that these cues include (a) laryngealization of a word-initial stressed vowel if preceded by a sonorant, (b) de-aspiration of the plosive in an [s]-plosive cluster only if no word boundary appears between the [s] and the plosive, and (c) lengthening of word-initial consonants with respect to word-final consonants. Our data indicate that the durational cue is not restricted to minimal word-pair situations, but is a general rule characterizing English word structure and is available to the listener at all times.

3. Implications for Linguistic Descriptions of English. Chomsky and Halle[7] have attempted to formalize a system of phonological rules that will generate a phonetic

transcription of an English sentence from its underlying representation. A phonological rule system must, among other things, assign a duration to each phonetic segment in an utterance. The rules that we have summarized should form a part of this phonological system.

Bimorphemic words containing an underlying double [s] present a problem for the current formulation of the Chomsky-Halle phonological rules. An existing word-level phonological rule always deletes the second member of a double consonant in the underlying representation of a word. A prior rule erases all morpheme-boundary symbols. Therefore, after applying the Chomsky-Halle rules of word-level phonology, it would be impossible to derive a correct prolonged duration for the [s] in a word like "misspell" from its phonological representation. If our data are substantially correct, some modifications of the structure or content of the phonological rules of English are required.

4. Why Are There Rules Involving the Duration of [s]? Languages, according to Lehiste,[15] display a wide range of organizational principles with regard to segmental duration. Nevertheless, some low-level rules of the type that we have described may have their origin in deep-seated physiological constraints on the speaker. Let us speculate a little on this topic.

(a) The shortening of multi-syllabic words may have evolved because of the obvious advantage of sending more information to the listener per unit time. It may be true that multi-syllabic English words do not carry significantly more semantic information than short words. If so, the greater phonetic segmental redundancy in long words permits the speaker to speak them more rapidly.

We do not know whether shortening of long words occurs in many languages; perhaps it does not occur in languages in which duration serves another function within a word. In Dutch, for example, word boundaries are marked by a lengthening of word-final syllables.[16] Perhaps another example of a situation for which long words would not be shortened would be the case in which information content increases with word length.

(b) The shortening of [s] when followed by a stop can be explained as a natural consequence of the sequence of articulatory requirements involved in producing this cluster. The production of a fricative such as [s] or a vowel requires a controlled articulatory gesture toward a target configuration. In a stop, the articulatory gesture is more ballistic in nature, the rapid closing motion ceasing abruptly when closure has occurred. In an intervocalic [s], the tongue tip must make two controlled movements in opposite directions, and synchronization of laryngeal activity is also required. When [s] is followed by a stop, the second movement (i) is ballistic, (ii) is in the same direction or with an independent articulator, and (iii) does not involve laryngeal coordination. Any of these factors permits an earlier onset of closure motion and a more rapid cessation of frication if [s] is followed by a stop. A stop preceding [s] should not produce a symmetrical shortening effect because the [s]-vowel sequence that follows stop release

involves a sequence of two controlled articulatory gestures. Indeed, this cluster situation resulted in only a slight shortening of the [s].

(c) Prestressed consonants and stressed vowels generally appear to be of longer duration in English than pre-unstressed consonants and unstressed vowels. That the consonant and vowel in a consonant-vowel sequence behave similarly in the durational rules provides additional support to the contention of Khozhevnikov and Chistovich[4] that the syllable is an articulatory programming unit.

(d) Prestressed [s] is both longer and of less variable duration than pre-unstressed [s]. It seems likely that unstressed syllables are not only spoken more rapidly, but also with a reduced muscle tonus and/or relaxed criteria for the attainment of target configurations.

5. Future Research. We are now making durational studies of the same speakers reading sentence material. Future work will include measurements of consonant duration in clusters and measurements to determine whether durational adjustments are employed by a speaker to make his utterances seem more rhythmical.

A long-range goal is to formulate a complete set of durational rules for English within a generative phonological framework. Some durational rules have already been written for a speech-synthesis-by-rule program that has been recently compiled by the author. The perceptual importance of individual rules is being systematically investigated with the aid of the speech-synthesis program.

D. H. Klatt

References

1. G. E. Peterson and I. Lehiste, "Duration of Syllabic Nuclei in English," J. Acoust. Soc. Am. 32, 693-703 (1960).
2. A. House, "On Vowel Duration in English," J. Acoust. Soc. Am. 33, 1174-1178 (1961).
3. T. P. Barnwell III, "An Algorithm for Segment Durations in a Reading-Machine Context," Technical Report 479, Research Laboratory of Electronics, M.I.T., January 15, 1971.
4. V. A. Khozhevnikov and L. A. Chistovich, Speech: Articulation and Perception (Joint Publications Research Service 30, Washington, D.C., 1965), p. 543.
5. N. Umeda and C. H. Coker, "Some Prosodic Details of American English," J. Acoust. Soc. Am. 49, 123(A) (1971).
6. G. D. Allen, "The Place of Rhythm in a Theory of Language," Working Papers in Phonetics No. 10, University of California, Los Angeles, 1968, pp. 60-84.
7. N. Chomsky and M. Halle, The Sound Pattern of English (Harper and Row Publishers, Inc., New York, 1968).
8. R. Vanderslice and P. Ladefoged, "Binary Suprasegmental Features," Working Features in Phonetics No. 17, University of California, Los Angeles, 1971, pp. 6-24.
9. D. B. Guralnik (Ed.) (World Publishing Co., New York, 1958).

10. I. Lehiste, "An Acoustic-Phonetic Study of Internal Open Juncture," Phonetica 5 (Suppl.), 1-54 (1960).

11. M. P. Haggard, "Effects of Clusters on Segment Durations," Speech Synthesis and Perception 5 (Psychological Laboratory, Cambridge, England, 1971), pp. 1-50.

12. M. F. Schwartz, "Intraoral Air Pressures for /p/ in Oral and Whispered Vowel Environments," J. Acoust. Soc. Am. 46, 480(L) (1969); "Duration of /s/ in /s/-plosive Blends," 47, 1143(L) (1970).

13. G. D. Allen, "Temporal Structure in Speech Production," J. Acoust. Soc. Am. 47, 58(A) (1970).

14. M. Halle and K. N. Stevens, "Analysis by Synthesis," in W. Wathen-Dunn and L. F. Woods (Eds.), Proc. of the Seminar on Speech Compression and Processing, AFCRL-TR-59-198, Paper D7, 1959.

15. I. Lehiste, Suprasegmentals (The M.I.T. Press, Cambridge, Massachusetts, 1970).

16. S. G. Nooteboom, Personal communication, 1971.

B. THE PERCEPTION OF TEMPORALLY SEGMENTED SPEECH

The usual way of studying speech perception is to find transformations of the speech wave that drastically interfere with its intelligibility, and then try to discover how the distortion has its effect. In this approach it is presumed that the perceptual apparatus relies heavily on the information that has been destroyed by the distortion.

In 1954, Cherry[1] discovered that running speech can be made virtually unintelligible by switching it alternately to the left and right ears of listeners at ~3 Hz. He found that higher or lower rates of switching had little effect. In this transformation the main parameter was duration, and the most dramatic effect occurred when the speech intervals reaching each ear of the listener lasted ~200 ms, that is, approximately the duration of a syllable. It was hoped that study of this effect would lead to a better understanding of how temporal properties of speech are used during perception.

Cherry argued that this effect was the result of a "dead time" in the "switch" that transfers the listener's attention from one ear to the other. At the critical rate, the signal and the listener's attention are out of phase, so that perception is frustrated. This explanation is not satisfactory, however, because intelligibility is similarly frustrated by interrupting the speech at the same rate with no "switching of attention" needed. Other work[2] suggested that intelligibility was destroyed because the speech reached each of the listener's ears in bursts, separated by silence. The present experiments tested this idea in greater detail.

Passages of continuous speech were "temporally segmented" with the aid of a computer. The operation is equivalent to cutting into pieces a tape carrying the message and splicing in a silent interval at each cut. Two sets of nine 100-word experimental passages of speech were cut into "intervals" whose duration increased in

9 log steps from 31 ms in the first passage in each set to 500 ms in the last. Three experimental tapes were then made from each set of passages. In the three tapes, labeled "short," "equal," and "long," the silent intervals were 41% shorter than the adjacent speech intervals, equal to them, and 83% longer than the speech intervals, respectively. The only difference between the tapes was the duration of the silent intervals that were spliced in.

This material had four advantages: (i) the speech reached the listener in bursts, as required, (ii) no switching of attention was required (unlike experiments using alternation), (iii) all of the speech reached the listener (unlike experiments using interruption), and (iv) silent intervals and speech intervals could be independently varied.

One group of 16 subjects shadowed the "short" and "equal" tapes, and a second group shadowed the "equal" and "long" tapes, with appropriate counterbalancing. Listeners whose first exposure to the material was the "long" tape showed a learning effect over the first four passages. Therefore the data from the first tape they encountered were discarded.

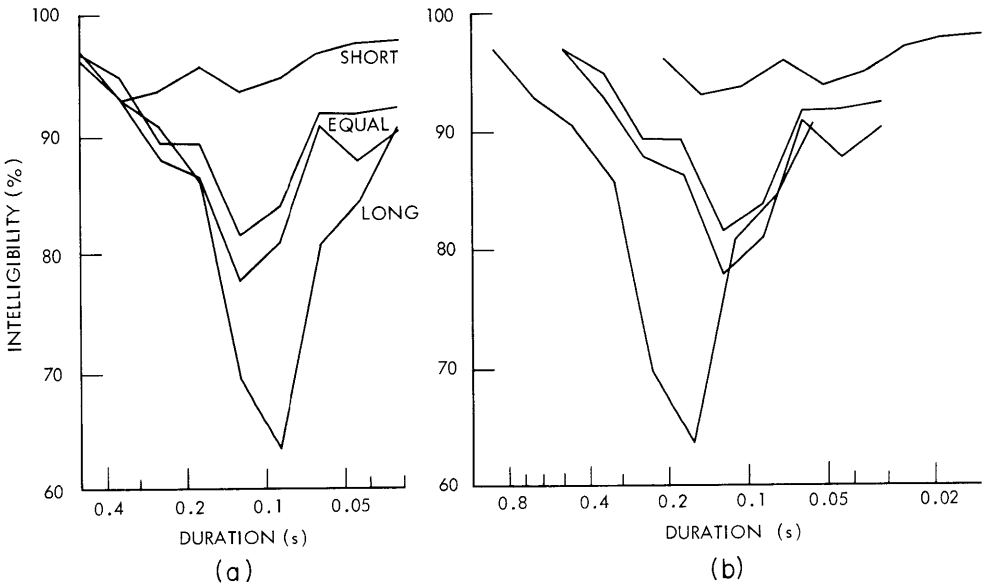

Fig. IX-4. Intelligibility (%) vs duration of (a) speech intervals, (b) silent intervals.

In Fig. IX-4a, intelligibility is plotted as a function of the duration of the speech intervals. The left-hand side of the three sets of data seems to lie on a single function, as if intelligibility progressively decreases as the duration of the speech intervals decreases. The minima occur, however, at different speech interval durations for the

"short," "equal," and "long" functions. The recovery from the minima occurs at progressively shorter speech-interval durations as the silent intervals are lengthened from one tape to another. This suggests that perhaps the recovery could also be described by a single function depending only on the durations of the silent intervals. To test this possibility the data were replotted in Fig. IX-4b as a function of the duration of the silent interval. This replotting brings the right-hand side of the curves into agreement, as if the intelligibility progressively increases as the duration of the silent interval decreases (the reverse of the speech case). A similar conclusion was recently reported by Powers and Speaks.[3]

This analysis suggests that the minimum of the Cherry effect is simply the crossing point of two functions, whose relationship is determined by the speech-silence ratio (which is unity in most alternation or interruption experiments).

What might the two functions represent? Consider first the decline of intelligibility, at the left side of Fig. IX-4a. Here, long speech intervals are separated by long silences. As the duration of the speech intervals becomes shorter, the speech becomes less intelligible. But this is just a restatement of a finding by Pickett and Pollack[4] that brief excerpts of fluent speech become increasingly intelligible as they become longer. Their data lie somewhat to the left of the data in Fig. IX-4a, that is, longer excerpts were required for a fixed intelligibility, but their experiment was different, too. Their excerpts were presented in isolation, and were always a small number of whole words, and their responses were also limited to whole words.

Extrapolating from the left-hand side of the curve in Fig. IX-4a points to a critical minimum sample duration of 60-70 ms for speech. More likely, the critical minimum sample is one acoustic segment of the speech, since other work has suggested that it is the speech content of the interval rather than its duration that is critical.[2]

What about the recovery shown in Fig. IX-4b? Experiments with trains of pulses[5] have suggested that silent intervals shorter than approximately 100 ms are integrated as part of an acoustic event (or sequence), whereas longer intervals act to break up the sequence into separate events, separated by pauses. If a sequence of speech samples, each too short to be recognized in isolation, can be integrated into a single ongoing acoustic event, then the samples may again become recognizable, whereas when intervening silences are sufficiently long that successive samples cannot be integrated into a single event, they may remain unrecognizable. Extrapolating from the right-hand side of the curves in Fig. IX-4b suggests a critical maximum silent interval of approximately 200-250 ms (which corresponds approximately to the duration of a syllable).

This analysis is supported by subjective impressions derived from listening to the tapes. When the silent intervals are short (at rates above the critical), the speech sounds as if it is being played at reduced speed. At the critical rate the speech sounds

very broken up. At slower rates words and phrases are heard, separated by pauses.

The difficulty with this explanation is that it should presumably apply to all sounds, not only speech, and thus represent a temporal parameter of the ear. As mentioned above, however, when speech is concerned the critical parameter seems to be the contents of the speech sample (that is, how many syllables, glottal cycles, and so forth it contains) rather than its duration.

This conflict will have to be resolved by further work.

A. W. F. Huggins

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 Patterns of Speech: Interruption and Alternation," J. Acoust. Soc. Am. 36, 1055-1064 (1964).
3. G. Powers and C. Speaks, "Intelligibility of Temporally Interrupted Speech," J. Acoust. Soc. Am. 50, 130(A) (1971).
4. J. M. Pickett and I. Pollack, "Intelligibility of Excerpts from Fluent Speech: Effects of Rate of Utterance and Duration of Excerpt," Language and Speech 6, 151-154 (1964).
5. A. W. F. Huggins, "Perceived Rate of Dichotically Alternated Clicks," J. Acoust. Soc. Am. 46, 88(A) (1969).

C. DIGITAL LADDER FILTER STRUCTURES AND COEFFICIENT SENSITIVITY

1. Introduction

Recent studies (Fettweis[1, 2]) show that digital filter structures can be molded after classical LC ladder structures. Since LC ladder structures are noted for the relative insensitivity of their frequency response to the element values, it has been conjectured that the digital structures should also have this desirable insensitivity.

In digital signal processing, recursive filters are generally implemented as cascade or parallel structures of first- and second-order filters. One reason for this choice is to achieve relative coefficient insensitivity. In order to compare the coefficient-sensitivity properties of digital ladder structures with those of the conventional cascade or parallel structures, we realized a seventh-order Tchebyshev lowpass filter as both ladder and cascade structures. The degeneration in the frequency responses of the two realizations is compared as the coefficients are quantized.

2. Digital Ladder Filters

We shall not duplicate the detailed description of synthesis procedures for digital ladder filters given by Fettweis.[1] We shall, however, describe briefly the steps taken

to synthesize the Tchebyshev lowpass filter, using Fettweis' notation.[1]

Figure IX-5 shows the three elements needed to realize the lowpass ladder structure: the distributed (Richards plane) inductor, the series adapter, and the unit element.

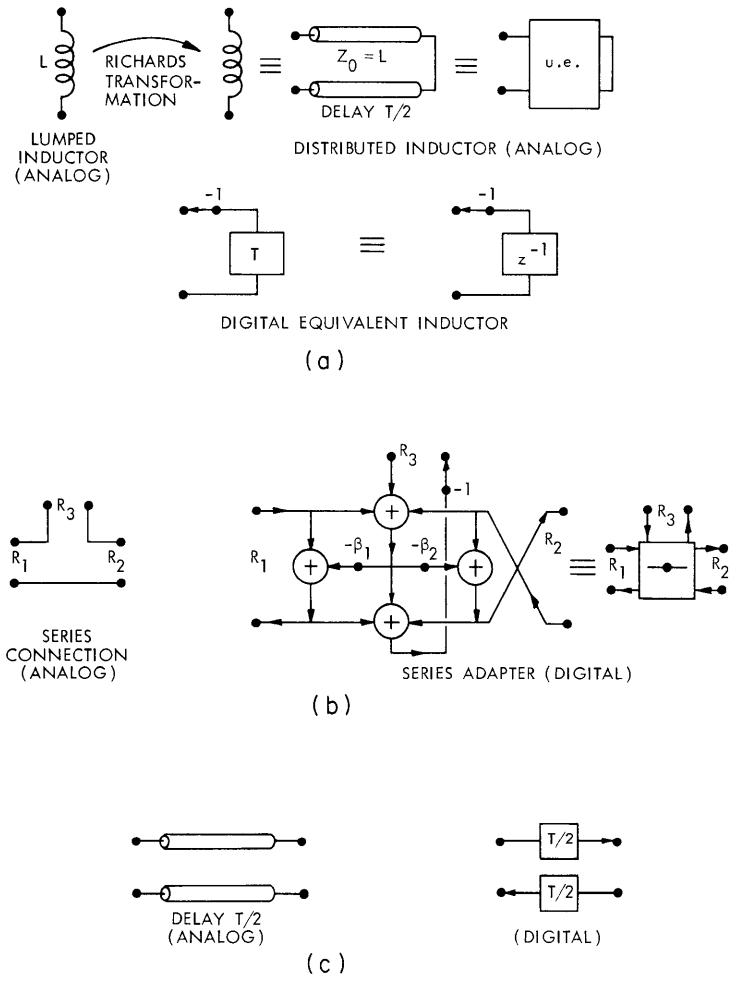

Fig. IX-5. Digital realization of (a) inductor; (b) series connection; (c) unit element.

A similar lowpass filter could be realized with distributed capacitors, parallel adapters, and unit elements. In the series adapter (or the parallel adapter) the three ports can be interchanged; that is, the dependent port need not be the one connected to the series element. The dependent port chosen for this example was the one with the largest value of β_k (see Fettweis[1]).

Figure IX-6 shows the synthesis of the digital ladder structure. The first step is to realize the analog filter, generally the easiest step, since tables or formulas are available for the most piecewise-constant filter forms. Next (see Fig. IX-6b) we replace

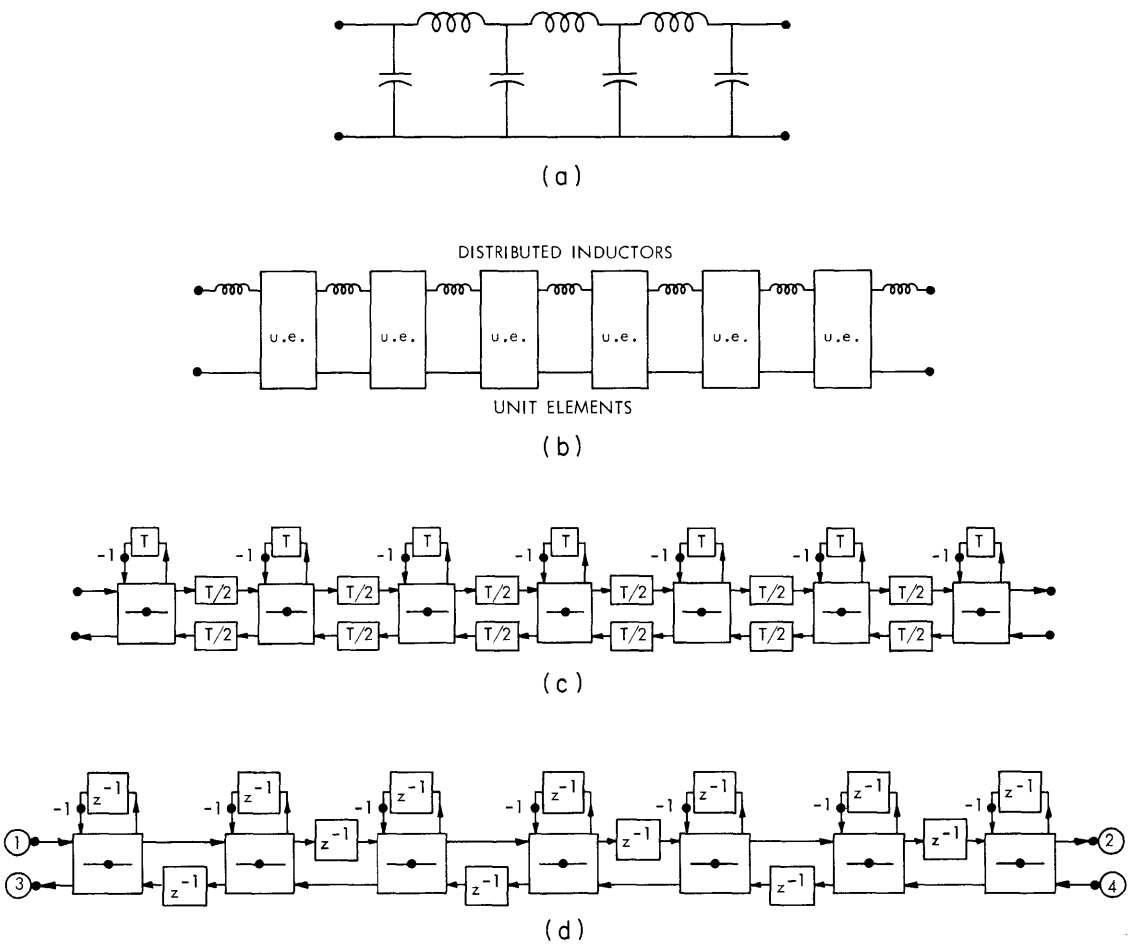

Fig. IX-6. (a) Analog circuit (lumped element).
(b) Analog circuit (distributed elements).
(c) Digital equivalent circuit of distributed analog circuit.
(d) Digital equivalent circuit modified to have only unit delays.

lumped elements with distributed elements (Richards transformation) and apply Kuroda's identities.[3] The application of Kuroda's identities is necessary to transform the network into a form with unit elements (delay) between the series and parallel connections. This ensures that the corresponding digital circuit will be realizable; that is, the digital circuit will contain no inner loops without delay.

The third step is to "digitize" the circuit according to the techniques described by Fettweis.[1] In this example the digital equivalent inductors (Fig. IX-5a) are connected to the circuit through the series adapters (Fig. IX-5b). The resulting digital circuit is shown in Fig. IX-6c. The upper paths between series adapters represent forward traveling waves and the lower paths represent reverse traveling waves. It is assumed that the data are sampled at rate 1/T.

The form of the circuit of Fig. IX-6c contains delays of T/2 and may not be

convenient for digital construction or programming. This ladder structure can be modified to a form containing only unit delays (delays of T). Such a modification is shown in Fig. IX-6d.

The filter can be used in either of two ways. Data entering the input at (1) are lowpass-filtered and emerge at (2). The stop-band frequencies appear as data at (3). Alternatively, we can insert the data at (4). In this case the passband frequencies appear at (3) and the stop-band frequencies at (2). In essence the filter can be used as a lowpass and a highpass filter simultaneously. Note that the filter in Fig. IX-6d requires 14 multiplications and 42 additions per cycle. Since the filter is symmetrical, only 7 coefficients need be stored. In comparison, the cascade representation of the same filter requires 7 multiplications and 14 additions. Seven coefficients need to be stored.

3. Degeneration of Frequency Response with Coefficient Quantization

We define the error criterion to measure the degeneration of the filter response as follows.

$$\text{Relative error} = \begin{cases} \dfrac{H_{max} - H_{min} - AM}{AM} & \text{for } H_{max} - H_{min} > AM \\ 0 & \text{for } H_{max} - H_{min} \leq AM \end{cases}$$

where

H_{max} = max $|H(e^{j\theta})|$ in the passband (dB)

H_{min} = min $|H(e^{j\theta})|$ in the passband (dB)

AM = originally specified ripple (dB) in passband

FC = normalized cutoff frequency (normalized to sampling frequency).

Nine examples were computed for the Tchebyshev lowpass filter with ripples of 1 dB, 0.2 dB, and 0.05 dB and cutoff frequencies FC = 0.250, 0.125, 0.0625. The average error for all of the examples is presented in Fig. IX-7a, followed by examples of individual filters. The dashed lines in Fig. IX-7a represent a best approximation through the set of average points for the ladder examples and the cascade examples, respectively. The same lines are included in Fig. IX-7b through IX-7f as a basis of comparison for individual examples.

In all of the results the relative error for a given level of coefficient quantization was greater for the cascade structure. When an average was taken over all of the

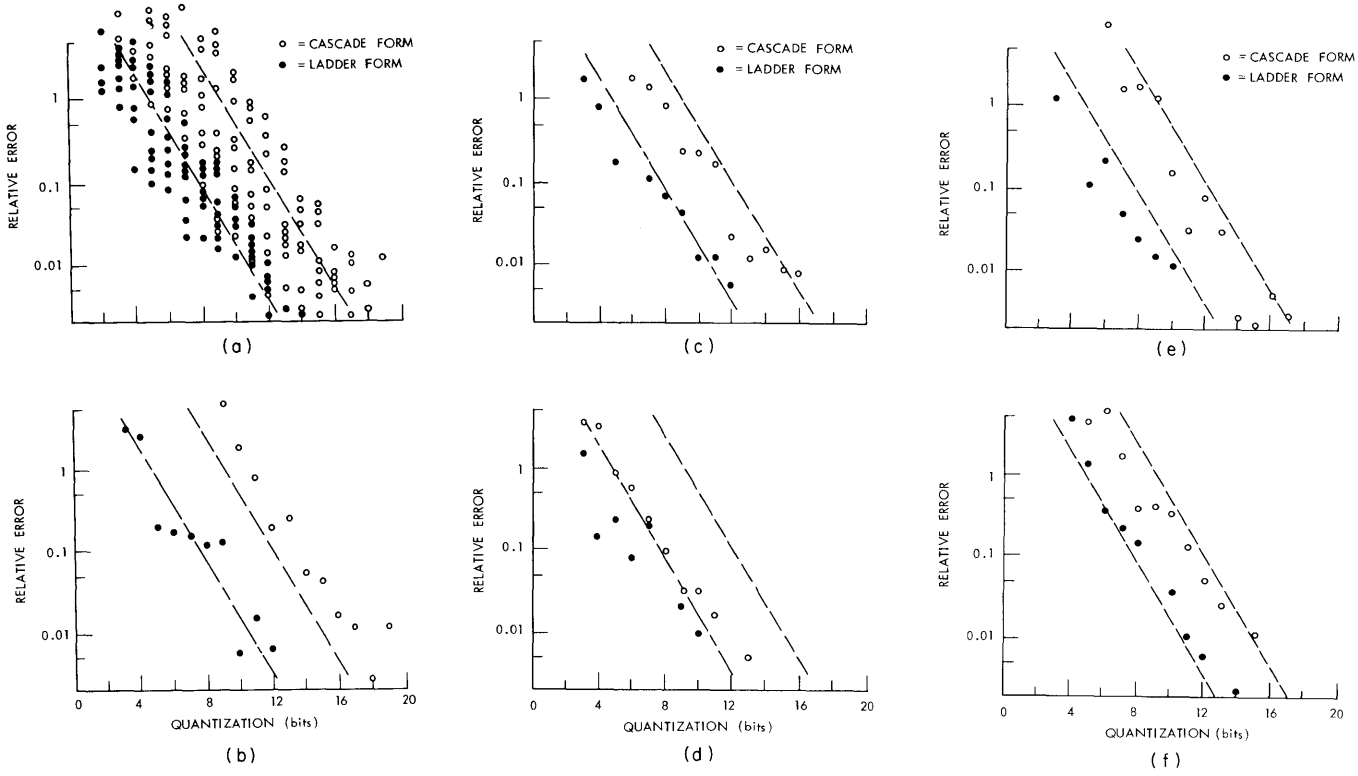

Fig. IX-7. Experimental results.
(a) Composite of nine examples.
(b) AM = 0.05 dB, FC = 0.0625.
(c) AM = 0.2 dB, FC = 0.125.
(d) AM = 1 dB, FC = 0.250.
(e) AM = 1 dB, FC = 0.0625.
(f) AM = 0.05 dB, FC = 0.250.

examples the difference in coefficient sensitivities was of the order of 4 bits. When individual filters were compared, however, the improvement in sensitivity was found to be dependent upon the cutoff frequency and the specified ripple. For example, with AM = 1 dB and FC = 0.250 (Fig. IX-7d) there was little improvement. For the example AM = 0.05 dB and FC = 0.0625 (Fig. IX-7f), however, an improvement of 4 or 5 bits was obtained by using the ladder structure. The ladder structure appears to be less sensitive than the cascade structure for situations of low cutoff frequency (or high sampling rate) and tight ripple specifications.

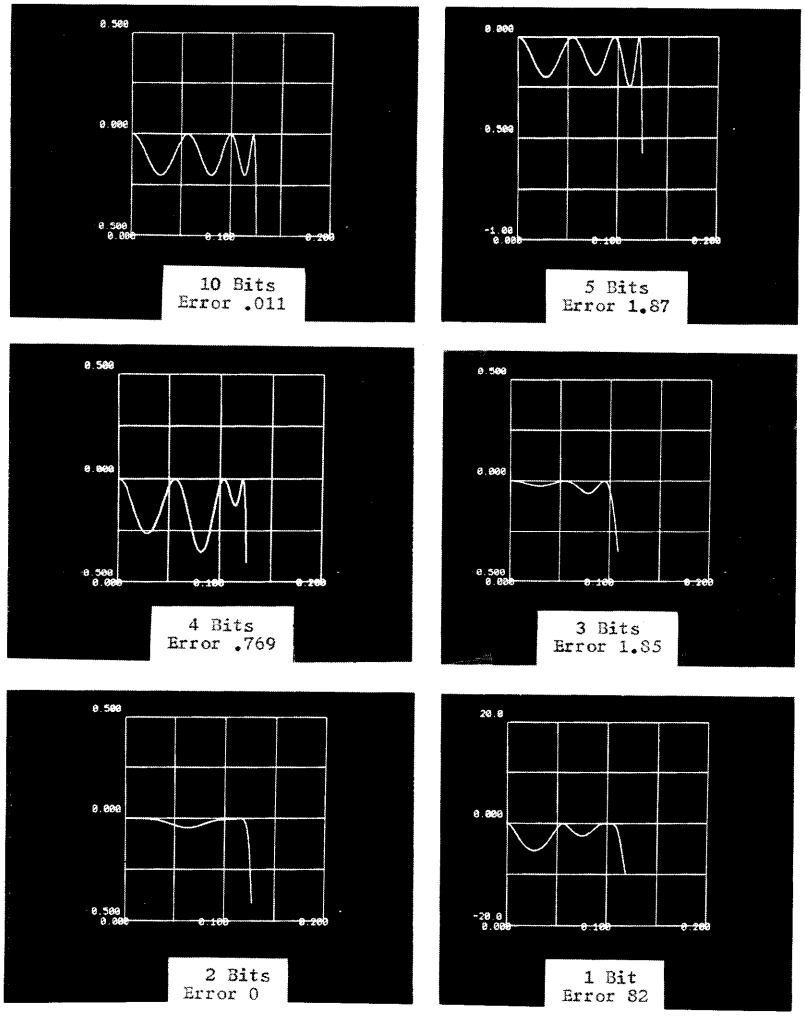

Fig. IX-8. Frequency response (passband) of ladder structure under coefficient quantization: AM = 0.2 dB, FC = 0.125.
Vertical axis: $|H(e^{j\theta})|$ in dB.
Horizontal axis: normalized frequency.

To observe the effect of coefficient quantization on the filter characteristics, we made some plots of frequency response at various quantization levels. The results are given in Fig. IX-8 for the ladder structure and in Fig. IX-9 for the cascade structure.

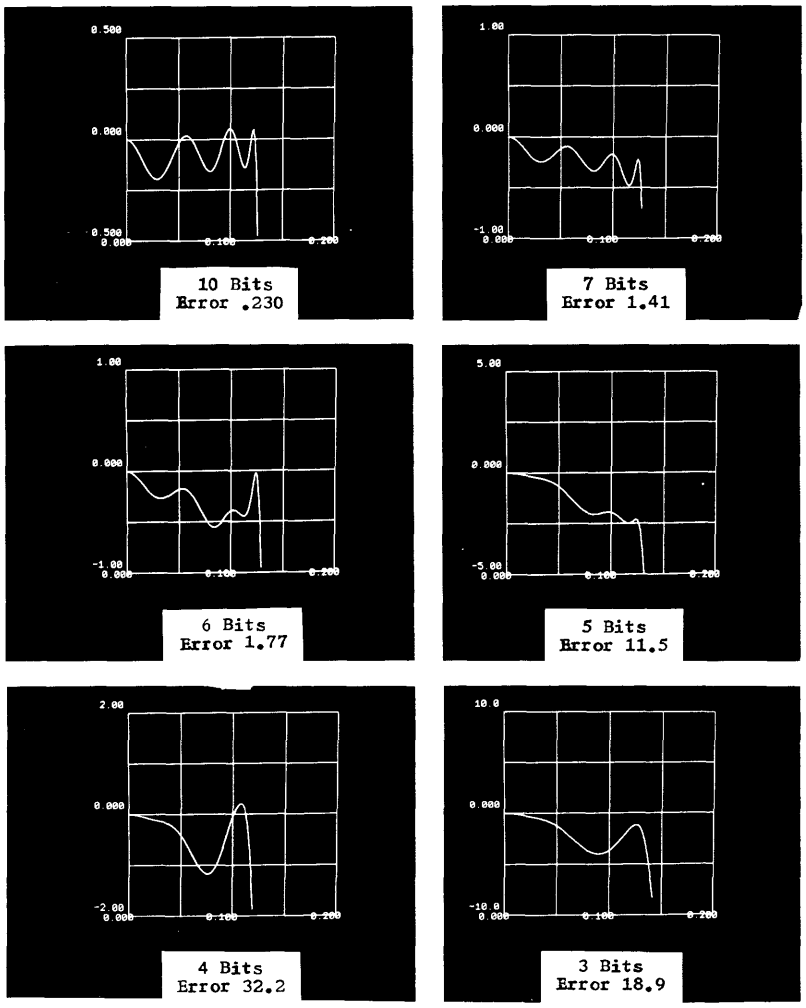

Fig. IX-9. Frequency response (passband) of cascade structure under coefficient quantization: AM = 0.2 dB, FC = 0.125.
Vertical axis: $|H(e^{j\theta})|$ in dB.
Horizontal axis: normalized frequency.

The example chosen for both cases was for AM = 0.2 dB and FC = 0.125. Note in Fig. IX-8 that the passband ripples for the ladder structure always reach their peak at zero dB.

A tenth example was computed with a normalized cutoff frequency of FC = 0.01 and ripple AM = 0.2 dB. The results are given in Fig. IX-10. In this example an

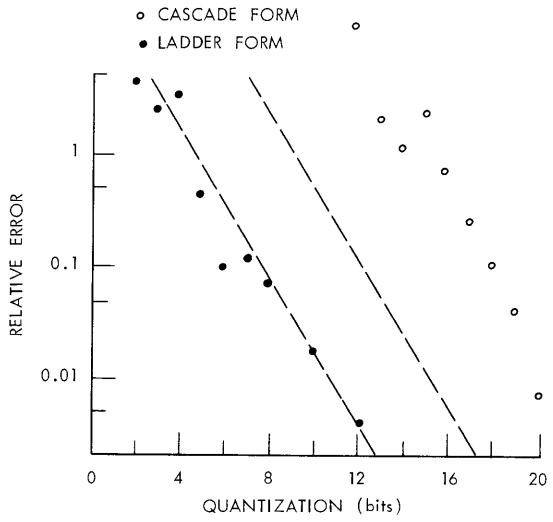

Fig. IX-10. Experimental result. AM = 0.2 dB, FC = 0.01.

improvement of 8-10 bits in coefficient insensitivity was gained by using a ladder structure instead of a cascade structure. (The dashed lines in Fig. IX-10 represent composite results of the first nine filter examples.)

4. Discussion

The digital ladder structure appears to have better overall insensitivity properties than the cascade structure, especially in situations of low cutoff frequency (or high sampling rates). In some examples we observed an improvement of 4-5 bits or more, while in other examples in which the filter requirements were less stringent little improvement (1 or 2 bits) was observed. The ladder structure requires nearly twice as many multiplications and three times as many additions per cycle as the cascade structure. The question, then, is whether the use of fewer bits per coefficient merits the additional number of operations that would be required.

As well as coefficient sensitivity, another important factor in comparing the ladder structure with conventional structures is that of noise performance. Some work has been done in this area[4] and there is reason to believe that the ladder structure may have some merit.

When series and parallel adapters with dependent ports are used and the digital filter is realized from a lossless LC analog, filter, the digital filter will be lossless and stable even under coefficient quantization. The passband ripples should reach their peak at zero dB (see Fig. IX-8).

If the techniques of Fettweis[1] are used to realize a digital filter from a lumped LC analog filter, the frequency response of the digital filter will be a bilinear transformation of the frequency response of the lumped analog filter.

For analog filter configurations with single inductors or capacitors appearing in series or shunt with the line, Kuroda identities can be applied in the digitization process.[3] For elliptic types of filters with zeros in the frequency response, the analog filter will contain series LC branches in shunt with the line or parallel LC branches in series with the line. For these forms it may be necessary to use the generalized Kuroda transformations cited by Levy.[5] Series and parallel resonant circuits can be realized by the methods described by Fettweis.[6]

R. E. Crochiere

References

1. A. Fettweis, "Digital Filter Structures Related to Classical Filter Networks," Arch. Elek. Übertrag., Vol. 25, p. 79 ff, February 1971.
2. A. Fettweis, "Some Principles of Designing Digital Filters Imitating Classical Filter Structures," IEEE Trans., Vol. CT-18, pp. 314-316, March 1971.
3. R. J. Wenzel, "Exact Design of TEM Microwave Networks Using Quarter-Wave Lines," IEEE Trans., Vol. MTT-12, pp. 94-111, January 1964.
4. J. Bingham, "A New Type of Digital Filter with a Reciprocal Ladder Configuration," Proc. 1970 IEEE International Symposium on Circuit Theory, Digest of Technical Papers, pp. 129-130.
5. R. Levy, "A General Equivalent Circuit Transformation for Distributed Networks," IEEE Trans., Vol. CT-12, pp. 457-458, September 1965.
6. Op. cit., see Sec. 5.2.