The Parsing Program for Automatic Text-to-Speech Synthesis Developed at the Electrotechnical Laboratory in 1968

NORIKO UMEDA AND RYUNEN TERANISHI

Abstract—**The paper describes a parsing program developed at the Electrotechnical Laboratory, Tokyo, Japan, in 1968 for automatic speech synthesis from ordinary English spelling. The parser handles unique problems for a speech production system, especially of phrase-structure analysis in regard to stress and pause assignments. The parsing program consists of a dictionary of about 1500 most frequently used words, a simple syntactic analyzer and a breath-group delimiter. The syntactic analyzer, with the assistance of information stored in the dictionary, divides the sentence into phrases, and assigns pause markers at major syntactic boundaries; the breath-group delimiter decides actual pauses and sentence stress. The output of the parsing program consists of a sequence of phonemes with stress marks and of phrase termination marks. These letters and marks are transformed into vocal tract shapes, duration, and pitch signals in the subsequent part of the synthesis system. The parsing program, written in the PL/I language, consists of about 1900 statements.**

IN 1968 at the International Congress on Acoustics, the speech research group at the Electrotechnical Laboratory in Tokyo reported an entirely automatic system of speech synthesis from ordinary English spelling [1], [2]. The software part of the system was written in the PL/I language and ran on an IBM 360 computer. The system was divided mainly into two parts: a set of rules for assigning prosodic cues (stress and pause) using a dictionary and surface-structure analysis, and another set of rules for transforming these prosodic cues into acoustic properties and for assigning vocal tract area functions. The output was written on a magnetic tape, and controlled, off-line, the hardware vocal tract analog synthesizer. The intelligibility of the output voice was quite good for being the product of such a simple-minded automaton.

The parsing program (i.e., dictionary, surface-structure analysis and breath-group division) was originated and actualized mainly by Teranishi. Its basic characteristics were transplanted to Bell Laboratories by Umeda. Although the speech group then at the Electrotechnical Laboratory is now largely dispersed, the fundamental idea of this parsing program is still alive at Bell Laboratories as a part of the automatic text-to-speech synthesis system [3], and is undergoing further improvements. We therefore feel obligated to publish this useful piece of work that was done six years ago, and that, for various reasons, has never been published in complete form.

Manuscript received June 27, 1974.

N. Umeda is with the Acoustics Research Department, Bell Laboratories, Murray Hill, N. J. 07974.

R. Teranishi is with the Department of Acoustics Design, Kyushu Institute of Design, Fukuoka, Japan.

PROBLEMS FOR SPEECH SYNTHESIS—CHARACTERISTICS OF THIS PARSER

When a sentence is spoken, how much syntactic information do the soundwaves convey? They certainly convey clues to the grouping of word sequences into meaningful units, to the word(s) which deserve maximum attention, and to the locations of break points between units. But probably the clues are no more detailed than these in terms of the syntactic structure. For example, the identical pitch contour may be found on a noun phrase, consisting of an adjective and a noun, or on a verb-plus-object predicate. It seems that a complete parsing is the listener's task; he hypothesizes a most likely analysis to match with the input group of words.

Bearing this in mind, we see rather obviously that the problem of parsing for speech synthesis is very different from that of parsing for other purposes. Systems for syntactic structure analysis, language translation, information retrieval, etc., to varying degrees require a test for the grammaticalness or meaningfulness of the input sentence. At every stage of analysis, the system predicts a set of possibilities for the rest of the sentence such that it will meet a criterion of meaningfulness.

The parser for a speech production system need not consider a possible overall syntactic structure of the whole sentence in terms of either tree structure or immediate constituents; it is enough for the analyzer to process the phrase structure locally, assuming only one—the simplest and most probable—sentence structure as its grammar. In speech production systems, in order to help the listener carry out his parsing task, the analyzer may well provide him with extra processing time (a pause) at the point where the simplest grammar breaks down.

Our analyzer was developed quite independently of any other automatic parser of English at the time. Because we realized fully the special problems in parsing for speech synthesis, we avoided all complicated syntactic analysis by assuming only one simple phrase structure (never treed or nested). All unit sentences consist of seven syntactic constituents (initial sentence modifier, subject, etc.), occurring in a fixed order with the option of the omission

of some of them. A point where this fixed order is reversed is the indication of a clause boundary, and the analyzer resets its grammatical status regardless of the syntactic role of the clause. Thus, the analyzer assigns phrase boundaries, clause boundaries and syntactic roles of phrases (subject, object, etc.), depending on the constituents of the phrase and the successive order in which phrases occur.

The analyzer operates from left to right, with one-word lookahead and no backtracking. The grammar has three unique aspects each of which contributes to the reduction of complexity in the analyzer. The fixed order of the seven syntactic constituents requires no test of grammaticality of the input sentence. The presence or absence of any of these syntactic constituents is optional. The analysis restarts at any point where the analyzer can not find the expected syntactic order.

The dictionary has two characteristics which assist the syntactic analysis. First, word classes (if there are more than one for the word) are listed according to frequency of usage. The analyzer takes the most frequent usage first, and will not make any second consideration if the first one goes through.

Secondly, word classes are predetermined in a special way and stored in the dictionary. Function words usually have multiple usages. For example, "that" can be an adjective (a determinant), a pronoun, a relative pronoun or a conjunction. Instead of giving all of these usages in separate terms, a single word class "ADA" is given to this word, which is the sole member of its class. The analyzer, recognizing ADA, finds possible paths for the word "that". Many function words form one-word classes.

This feature-merging, fixed word-class definition does not allow a sophisticated and flexible analysis. But for speech synthesis the function of an analyzer is not to detect a precise relationship between words, phrases or clauses; but it is to determine whether a pause should be placed before or after the current word. And our analyzer serves this purpose quite well.

Many features of the above phrase-structure analysis have been carried over to Bell Laboratories. The form of the dictionary and the table-driven parsing program at Bell Laboratories are quite different from their original version at Electrotechnical Laboratory (decision tree). A moderate backtracking is implemented in the Bell Laboratories system. The idea of using the content–function distinction for stress assignment (see below) has been strongly expanded there. The Bell Laboratories analyzer uses pause-assignment procedure very different from the breath group assignment described later in this paper.

DICTIONARY

The dictionary contains the 1500 most frequently occurring English words. In the process of dictionary creation, which needs to be performed only once, a separate card is prepared for each entry. All derivational and inflectional forms are given separate entries. The following information is punched manually on the card: 1) ordinary spelling of the word, 2) phonetic transcription, 3) content–function distinction, and 4) grammatical usage of the word ("word class"). When the card deck is completed, the entire information is recorded on a magnetic tape. In the process of recording, the phonemic notation is converted by rule into allophonic notation. Fig. 1 shows a schematic diagram of the preparation of the dictionary.

The phonetic transcription of the word consists basically of a machine-readable translation from the Kenyon–Knott dictionary [4], i.e., the phonemic notation with stress marks and syllable division markers. As in Kenyon–Knott, the stress mark is omitted in monosyllabic words. For polysyllabic words, one of two levels of stress is assigned to each syllable in the word, namely stressed or unstressed. The primary stress in polysyllabic words is registered as "stressed". Secondary stress is omitted when it follows the primary stress. The secondary stress is promoted to "stressed" when it precedes the primary one.

All the words in the dictionary are classified as either content words or function words. The content–function distinction was originally proposed by K. L. Pike in 1945 [5], and was found useful by Fries for teaching English to foreign students [6]. We found this distinction far more useful for speech synthesis than any other recent theories of English prosody. Function words are

pronouns (personal, reflexive);
prepositions;
conjunctives (conjunctions, relative pronouns, etc.);
auxiliary verbs (including "be", "do" and "have");
articles; and
adverbs of degree.

All others are content words. In the rule for sentence stress assignment, content words are stressed, and function words are not.

Each word is assigned to one or more word classes. The definition of the word class appears in the dictionary as a code mark containing from one to three letters. The analyzer first interprets the leftmost letter, and ignores the middle or the rightmost one until further detailed classification is necessary in peripheral branches of the decision tree. All the nouns form only one class, "N". All verbs are given "V" as their major class, and subclassified according to the matrix of their function (transitive, auxiliary, etc.) and mode (present tense, gerund, etc.). Words having multiple grammatical usages often form one-word classes.

After the four kinds of information—ordinary spelling of the word, pronunciation, content–function distinction and word classes—have been punched on a card, then all the information on the card deck is recorded on a magnetic tape. In the course of this automatic procedure, number of syllables and of sound codes (i.e., phonemes, punctuation marks, stress marks and syllable markers) are added to each entry. At the same time, the phonemic notations of the words are converted by rule into allophonic notations specific for the particular synthesis procedure. All the allophones considered in the system were coarticulation

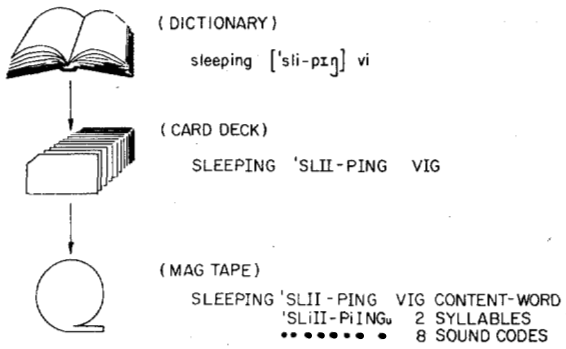

Fig. 1. A schematic diagram of dictionary preparation. Information is transferred manually from a standard dictionary to a deck of cards. The card deck is automatically processed on to magnetic tape, with conversion from phonemic to allophonic transcription.

allophones. For example, five allophones (i.e., five different vocal tract configurations), differing with the following vowels, were stored for the phoneme /k/. The vocal tract area function every ten ms is obtained by linear interpolation between two target configurations of the vocal tract with a fixed transition time (for most cases 50 ms). Therefore, allophones are required for the sake of intelligibility, and they are stored as target configurations. 35 phonemes were converted into 82 coarticulation allophones. Inside the computer, phonemes are represented by letter codes, and allophones by 8 bit codes.

SURFACE-STRUCTURE ANALYSIS

A simple surface-structure analysis is used in the speech synthesis system. Parsing starts at the head of the sentence. Subsequently, the grammatical status of the preceding words determines the status of the word under analysis. The words that follow, except for the one immediately following, do not affect the determination (see Fig. 3).

Every sentence is assumed to contain one or more "unit sentences" that correspond to "clauses" in the traditional grammar. A unit sentence consists of one or more of the following seven grammatical parts.

Part 1 transformational marker,
Part 2 sentence modifier,
Part 3 subject,
Part 4 verb group,
Part 5 object or complement,
Part 6 tail modifier, and
Part 7 terminal marker.

Interrogatives, interjections, etc. fall into Part 1. It is expected that some word order may be changed within the unit sentence when Part 1 appears in the sentence. Conjunctions, relative pronouns, adverbs and prepositional phrases form Part 2, the sentence modifier. Part 3, subject, is assigned to the first noun phrase in the unit sentence unless a verb group precedes the noun phrase. Verbs form Part 4. There is a strict priority order in the occurrence of verb classes in this part, so that the analyzer can process appropriately such a phrase as "what the fish foretold came true". Particles and adverbs immediately before or after the verb are also assigned to Part 4, the verb group. Part 5, object, consists of noun phrases or adjectives which follow the verb. Part 6, the tail modifier, accepts prepositional phrases and adverbs. Part 7, terminal marker, consists of one of the following: comma, colon, semicolon, dash, period, exclamation mark, quotation mark, or question mark. The most important assumption in the process of analysis is that these parts occur in the order of low number to high number in a unit sentence (some of the Parts may not appear). At a point where this order is reversed, a unit boundary is indicated. When this occurs, the analyzer resets its status to start testing for a new unit sentence.

When the input sentence is read, the first step in the analysis procedure is to fill the table called "SENTENCE" in core memory. An example of the table is shown in Fig. 2. The table contains a number of columns. The column called "LETTERS" is to be filled with the words from the input sentence in the order of their appearance. The column called "CONTENTS," that consists of several subcolumns, contains records of these words from the dictionary—phonemic and allophonic transcriptions with stress marks and syllable boundaries, numbers of syllables and phonemes, word class(es), and the content–function index. As the sentence is analyzed, the grammatical Part is assigned and written in the column called "REMARKS" (see Fig. 2). When Part 7 (punctuation) is assigned, the mark "M" is added to this column for later use in intonation assignment (see below).

A flow chart of the analysis procedure is shown in Fig. 3. The status of the analyzer is set initially for Part 1. The analyzer tests the possibility of the first word's falling into Part 1, looking, perhaps, for one of the interrogatives. If the test is negative, the status of the grammatical Part is advanced to "2". If the word is found to be a preposition, for example, the analyzer assigns Part 2 (sentence modifier) to the word and calls a subroutine "PCHAIN" in order to find a prepositional phrase. Inside the "PCHAIN", a match is attempted, with one word at a time, between the word classes of the input words and one of the possible types of prepositional phrase. At the same time, a phrase boundary mark and a Part assignment are given when they are necessary. If the word does not match any that may constitute Part 2 (sentence modifier), the analyzer advances the grammatical status to Part 3 (subject) and looks for a noun phrase as the subject. The search continues until the current word is given a status. Then the next word to the right is processed in a similar way; testing begins with the status of the word just finished (see Fig. 3).

The analyzer thus gives a syntactic role to each word in the sentence, groups words into phrases, and assigns a preliminary boundary marker at a clause boundary and at the boundary inside a clause.

BREATH GROUP DELIMITATION— PAUSE, STRESS, AND INTONATION

After a syntactic analysis has been given to the input sentence, the next step is to decide where actual pauses

WORD NO.	LETTERS FROM (INPUT CARD)	CONTENTS (20 Byte) (FROM DICTIONARY RECORD AREA) (PHONEME) (SOUND CODE) (USE)	REMARKS (PART#) (SPR)
1	SLEEPING	'SLI-PING 'SLiII-PiINGu VIG C	3
2	BEAUTY	'BYUU-TI 'BiYuUU-TiI N C	3
3	.		7 M.
4			
5			
.			
.			
.			
.			
.			
50			

Fig. 2. An example of the table "SENTENCE" created in the core memory in the process of analysis. The LETTERS column contains input words. The CONTENTS column contains the information stored in the dictionary for those words. The REMARKS column is filled during the analysis procedure with 1) the grammatical Part that is assigned to the word, 2) a breath group marker, and 3) an intonation marker.

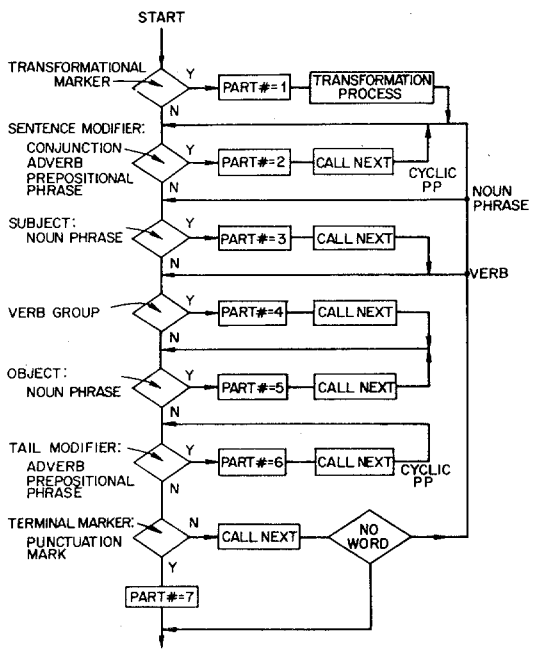

Fig. 3. Flow chart of the parsing program. Analysis starts from the head of the input sentence. The program handles one word at a time, assigning the grammatical Part to the word. At any place in the sentence where the analysis has to begin again from the top, a clause boundary is indicated.

should be placed in the sentence. The pause assignment process not only inserts silences, but also determines intonation contour and sentence stress. Sentence stress, intonation, and the syntactic value of the boundary where a pause is placed—all of these affect the duration and pitch of the syllable immediately preceding the pause. A unit (which consists of several words) between two pauses is defined as a "breath group". By studying an American actor's deliberately slow reading of several fairy tales, we found that almost no breath group exceeded 2400 ms in length. Accordingly, we set the breath-group delimiter at a ten-syllable maximum. This information also determined the size of the hardware buffer storage for producing a continuous time pattern of the vocal tract configurations and sound sources.

A pause code, which consists of a letter, is written at the end of a breath group in the "REMARKS" column of the table "SENTENCE". The letters "M" and "Z" are assigned at the boundary between unit sentences with and without a punctuation mark respectively. The breath-group delimiter acts when the constituents of a unit sentence exceed three words or ten syllables. "L" is given to the boundary between Parts 2 and 3, 3 and 4 or 2 and 4. "T" is assigned at the boundary between 4 or 5 and 6. "S" is assigned when a single Part consists of more than one phrase and the word or syllable count exceeds the limits. Rising intonation is given to a breath group that is terminated with "L" or "S" (indicated by $<$ in Fig. 4). Falling intonation is given to other breath groups (indicated by $>$ in Fig. 4).

In practice, as a consequence of the delimitation rule, a sentence was divided at almost every three or four words, and as it turned out, short breath groups compensated very well for flaws in phoneme definitions and helped the intelligibility of the output speech.

One of three stress levels—tonic, stressed, and unstressed—is assigned to each word in the breath group. Function words and unstressed syllables in polysyllabic content words are "unstressed". Monosyllabic content words and stressed syllables in polysyllabic content words are "stressed". The stressed syllable of the last content word in a breath group is given "tonic" stress. A preposition at the end of a breath group is always converted to a content word and given tonic stress.

The output of the parsing program is a sequence of strings which consists of phonetic symbols, stress values of syllables, and pause markers (see Fig. 5). These phonetic and prosodic notations are transformed into duration, pitch, and intensity in the second part of the synthesis program.

The program for derivation of prosodic cues consists of about 2800 statements in the PL/I language. The dictionary preparation occupies about 900 statements, which are mostly for the phoneme-allophone conversion. In the derivation part, the dictionary look-up takes about 400 statements, and the parsing program, which includes four major subroutines (adverb chain, prepositional phrase, noun phrase, and verb chain), consists of about 850 statements. The breath-group delimiter and stress assignment occupy about 650 statements. About 700 statements are used for the conversion from the internal results into output strings.

SUMMARY

There is a great difference between speculating about the possibility of an automatic simulation of human activities, and making a working system. The significance of the work described in this paper is that we made an automatic system of speech synthesis, using a simple surface-structure analysis. Most important, a slightly improved version of the parsing program with the dictionary is still working at Bell Laboratories, and no radical change is necessary.

In the past few years at Bell Laboratories, the acoustic realization of linguistic units has been extensively studied both from natural speech and from speech synthesis with articulatory control [3], [7], [8]. People are more aware

WORD	USED FNC	C/F	PART(I)	SEPARATOR	SEP.MARK
THE	TCE	F	3		
QUEEN	N	C	3		
HAD	VHR	F	4		
A	TCA	F	5		
LITTLE	A	C	5		
GIRL	N	C	5	Z	>
WHO	WHO	F	3		
WAS	VBP	F	4		
SO	ERS	C	4		
BEAUTIFUL	A	C	5	Z	>
THAT	ADA	F	2		
THE	TCE	F	3		
KING	N	C	3		
NEVER	BN	C	4		
TIRED	VAR	C	4	T	>
OF	P1F	F	6		
LOOKING	VAG	C	6	S	
AT	P1	F	6		
HER	M7S	F	6		
,	Z ,		7	M	,
AND	C1A	F	2		
IN	P1	F	2		
HIS	M2H	F	2		
GREAT	A	C	2		
JOY	N	C	2	L	<
HE	M1H	F	3		
ORDERED	VAR	C	4		
A	TCA	F	5		
MAGNIFICENT	A	C	5		
FEAST	N	C	5	Z	>
TO	P1T	F	3		
BE	VBO	F	4		
PREPARED	VAR	C	4		
.	Z .		7	M	.

Fig. 4. One of the computer output from syntactic analysis and breath group assignment. The first column consists of the words of the input sentence. The second column contains a word class that was chosen from the dictionary for each word by the analyzer. The third column indicates content–function index. The fourth column shows the grammatical Part the analyzer has assigned to each word. The fifth column indicates types of breath groups and the rightmost column intonation markers.

```
*** INPUT SENTENCE ***

THE QUEEN HAD A LITTLE GIRL WHO WAS SO BEAUTIFUL THAT THE
KING NEVER TIRED OF LOOKING AT HER, AND IN HIS GREAT JOY HE
ORDERED A MAGNIFICENT FEAST TO BE PREPARED.

*** STRING OF CODE ***
(EXPRESSED WITH FAMILIAR CHARACTERS)

DHUE'KWIIN-HAED-UE'LIT-L*GUERL>
HU-WUEZ'SOU'BYUUT-UE-FUEL>
DHUET-DHUE'KING'NEV-UER*TAIRD>
UEV'LUK-ING-AET-HUER,
AEND-IN-HIZ'GREIT*JHOI<
HI'ORD-UERD-UE-MAEG'NIF-UE-SNT*FIIST>
TUE-BII-PRI*PERD.
```

Fig. 5. An example of the final output of the analyzer. Though "STRING OF CODE" consists of phonemic notation of words and phrases as a computer print-out, the actual input to the succeeding part of the system is the corresponding allophonic notation.

now of the complex ways in which phoneme durations and allophonic variations are involved with stress, boundary, and phonemic sequences. This study of the detailed structure of speech has been, to a great extent, stimulated by the existence of a usable automatic system of speech synthesis.

The rules in the second part of our software system, the determination of acoustic and articulatory properties, were ad hoc and are now obsolete. But we need not apologize for them. Without this part, and without the test of actual output sounds, it would have been impossible to establish the workability of such a simple program of surface structure analysis for automatic speech synthesis by rule.

APPENDIX

1) Noun class:

N All nouns

2) Adjective class:

A All Adjectives, except for
- Q Quantitative adjectives
- # Numerals (cardinals, ordinals)
- D Demonstratives are further classified into:
  - A that
  - I this
  - O those
  - E these

Examples:

AQ: no, one, all, some, few, many, every, etc.
A#: nine, thirteenth, etc.

3) Interrogatives (as well as Relatives):

WHA what
WHO who
WHI which
WHN when
WHR where
WHY why
WHS whose
WHM whom
WHW how

4) Pronouns:

M Personal pronouns (combination of 1 thru 7 and I thru D)
- 1 Nominative
- 2 Possessive
- 3 Objective
- 4 Reflexive
- 5 Possessive pronoun
- 6 Nominative = objective
- 7 Possessive = objective
  - I First person singular
  - Y Second person
  - H he
  - S she
  - T it
  - W First person plural
  - D Third person plural

Examples:

M1D: they
M6I: it
M4Y: yourself, yourselves
M3H: his

5) Adverb class:

B All adverbs, except for
- I Degree adverb (quite, almost, very)
- ER Place adverb (there, here)
- N Negative (never, hardly, seldom)
  - N not
  - R Special adverbs are defined by the third-letters.
  - S soon, so
  - O only (not only . . . but)
  - R rather (rather . . . than)
  - N neither (neither . . . nor)
  - I either (either . . . or)
  - E even (even though)
  - B both (both . . . and)
  - C according (according to)

6) Articles:

TCA indefinite article (a, an)
TCE definite article (the)

7) Verb Class:

V Verb are all classified by the matrix of I thru H and O thru G.
- I Intransitive
- T Transitive
- A Intransitive and Transitive
- X Auxiliary
- B be
- D do
- H have
  - O Present and root
  - P Past
  - R Present = past
  - Q Past participle
  - C Present = past participle
  - S Present = past = past participle
  - G Gerund

Examples:

VHP: had
VXO: shall, will, can, may, etc.
VAG: living, playing

8) Prepositions:

P Prepositions are all classified as follows:
- 1 Preposition only (Some of then are given a third definition because of their more than one grammatical usages.)
  - F of
  - T to (also in the infinitive formation)
  - R for (also as conjunction. See C2)
  - B between (between . . . and)
- 2 Preposition as well as adverb
- 3 Preposition as well as conjunction

Examples:

P1: in, at, on, from, with, upon, beside, etc.
P2: out, over, down, off, under, above, etc.
P3: after, till, since, etc.

9) Conjunctions:

C Conjunctions are all classified as follows:
- 1 Coordinate conjunctions (are further subfied into)
  - A and
  - B but
  - R or
- 2 nor, yet, for
- 3 however, moreover, anyway, therefore, thus, hence
  - S unless, while, though, lest, because, whether
  - A as
  - I if

ACKNOWLEDGMENT

We appreciate our colleagues' (T. Suzuki and H. Omura) efforts of writing the program for the second part, building the synthesizer and taking care of rather unstable dc drift of the synthesizer. Especially, we would like to express our profound thanks to Dr. E. Matsui, who strongly supported the idea of automatic text-to-speech synthesis and supervised the work of the entire system and project.

REFERENCES

[1] R. Teranishi and N. Umeda, "Use of pronouncing dictionary in speech synthesis experiments," in *Rep. 6th Int. Congr. Acoustics*, Tokyo, Japan, 1968, B-155.
[2] N. Umeda, E. Matsui, T. Suzuki, and H. Omura, "Synthesis of fairy tales using vocal tract," in *Rep. 6th Int. Congr. Acoustics*, Tokyo, Japan, 1968, B-159.
[3] C. H. Coker, N. Umeda, and C. P. Browman, "Automatic synthesis for ordinary English text," *IEEE Trans. Audio Electroacoust. (Special Issue on 1972 Conference on Speech Communication and Processing)*, vol. AU-21, pp. 293–298, June 1973.
[4] J. S. Kenyon and T. A. Knott, *A Pronouncing Dictionary of American English*. Springfield, Mass.: Merriam, 1953.
[5] K. L. Pike, *The Intonation of American English*. Ann Arbor, Mich.: University of Michigan Press, 1945.
[6] C. C. Fries, *Teaching and Learning English as a Foreign Language*. Ann Arbor, Mich.: University of Michigan Press, 1945.
[7] C. H. Coker, "Speech synthesis by modelling the human articulatory system," Bell Laboratories, Murray Hill, N. J., unpublished Memo., 1969.
[8] N. Umeda and C. H. Coker, "Subphonemic details in American English; data and rule," *J. Phonetics*, to be published.