SYNTHESIS BY RULE OF SEGMENTAL DURATIONS IN ENGLISH SENTENCES

Dennis H. Klatt, Mass. Inst. of Tech., Cambridge, MA 02139.

In this paper, we are concerned with prediction of the (acoustically defined) durations of phonetic segments in spoken sentences. The durational definitions that have been adopted correspond to the closure for a stop (any burst and aspiration at release are assumed to be a part of the following segment). For fricatives, the duration corresponds to the interval of visible frication noise (or to changes in the voicing source if no frication is visible). For sonorant sequences, the segmental boundary is defined to be the half-way point in the formant transition for that formant having the greatest extent of transition. The definitions represent a convenient largely reproducible measurement procedure, but the physiological and perceptual validity of these boundaries have not been established.

In a review of the factors that influence segmental durations in spoken English sentences (Klatt, 1976a and references cited therein), it was concluded that only some of the systematic durational changes were large enough to be perceptually discriminable. The goal of this paper is to describe these first-order effects by rules.

Input Representation for a Sentence

The durational rule system to be presented is a part of a speech synthesis by rule program (Klatt, 1976b). The phonological component of this program accepts as input an abstract linguistic description of the utterance to be synthesized. The output of the phonological component is a detailed phonetic and prosodic representation of the utterance, including an acoustic duration for each segment. The symbol inventory is shown in Table 1; it includes 52 phonemes, 3 stress markers, 3 types of boundary indicators, and 6 syntactic structure indicators. An example of the use of some of these symbols is provided in Figure 1.

Phonemic Inventory. A traditional phonemic analysis of English is assumed, except that:

(a) Vowel+/R/ syllables are transcribed with the special vowel nuclei /IR/ ("beer"), /ER/ ("bear"), /AR/ ("bar"), /OR/ ("boar"), and /UR/ ("pure"). Words like "player" and "buyer" should be transcribed with two syllables, i.e. /EY/+/RR/ and /AY/+/RR/.


```
(M #F DH AX    #C 1 OW L D    #C M 1 AE N )N    #C S 1 AE T
          #F IH N    #F AX    #C R 1 AA K RR ).
```

Phone	Stress	Dur	Phone	Stress	Dur
SI	0	200	AE	1	165
DH	0	40	DX	0	20
IY	0	85	IH	0	65
OW	1	145	N	0	50
LX	0	65	AX	0	65
D	0	35	R	1	80
M	1	70	AA	1	140
AE	1	225	K	0	50
N	0	60	RR	0	175
S	1	105	SI	0	200

Figure 1. Input representation for "The old man sat in a rocker" and a listing of the output of the phonological component, i.e. the phonetic string, stress feature, and duration predictions in msec.

(b) The glottal stop [Q], dental flap [DX], glottalized alveolar stop [TQ], and velarized lateral [LX] listed in Table 1 are not really phonemes, but are allophones that are inserted in lexical forms before segmental durations are computed.

Lexical Stress. Each stressed vowel of an utterance must be preceeded by a stress symbol (1, 2, or !), where 1 is primary lexical stress (reserved for vowels in open-class content words, only one 1-stress per word). The secondary lexical stress "2" is used in some content words (e.g. the first syllable of "demonstration"), in compounds (e.g. the second syllable of "baseball"), in the strongest syllable of polysyllabic function words (e.g. "until"), and for pronouns (excluding personal pronouns like "his"). Emphatic stress "!" can be assigned to a semantically prominent syllable in a phrase.

Morpheme and Word Boundaries. There is no input symbol to indicate a syllable boundary. The symbol "*" can be used to mark morpheme boundaries. Each word of an utterance to be synthesized must be immediately preceded by a word boundary symbol. The distinction between content and function words is indicated by using "#C" and "#F". Open-class words (nouns, verbs, adjectives and adverbs) are content words. The program will check to see that no function word carries primary stress. A compound such as "apple cart" is indicated in the input representation by replacing the word boundary between "apple" and "cart" by a morpheme boundary and by reducing the lexical stress on the second word "cart" by one.

Table 1. The legal input symbols for synthesis of an utterance. Also given are a basic or inherent duration for each phonetic segment type and a minimum stressed duration in msec.

Vowels		INHDUR	MINDUR			INHDUR	MINDUR
IY	beet	160	50	IH	bit	130	40
EY	bait	190	70	EH	bet	150	60
OW	boat	220	70	AH	but	140	50
UW	boot	210	60	UH	book	160	50
AE	bat	230	60	AA	Bob	240	80
AO	bought	240	80	RR	bird	180	60
AY	bite	250	90	AW	bout	260	100
OY	boy	280	110	YU	beauty	230	100
AX	about	120	40	IR	beer	230	100
ER	bear	270	100	AR	bar	260	100
OR	boar	240	100	UR	poor	230	100
Sonorant Consonants							
W	wet	80	60	Y	yet	80	40
R	rent	80	30	L	let	80	40
WH	which	70	60	H	hat	80	20
EL	bottle	160	110	LX	bill	90	70
Nasals							
M	met	70	60	N	net	65	35
NG	sing	80	50	EM	keep'em	170	110
EN	butten	170	100				
Fricatives							
F	fin	120	60	V	vat	60	40
TH	thin	110	40	DH	that	50	30
S	sat	125	50	Z	zoo	75	40
SH	shin	125	50	ZH	azure	70	40
Plosives							
P	pet	85	50	B	bet	80	50
T	ten	65	40	D	debt	65	40
K	core	65	50	G	gore	65	50
DX	butter	20	20	TQ	at Alan	65	50
Q	Ma_opted	20	20				
Affricates (closure, frication)							
CH	chin	70	50	J	gin	70	50
		60	40			30	20

Stress Symbols

1	primary lexical stress
2	secondary lexical stress
!	emphatic stress

Word and Morpheme Boundaries

*	morpheme boundary
#C	begin content word
#F	begin function word

Syntactic Structure

.	end of declarative utterance
)?	end of yes/no question
(M	begin main clause
,	orthographic comma
)N	end of noun phrase
(R	begin relative clause

Syntactic structure. Syntactic structure symbols are important determiners of sentence stress, rhythm, and intonation. Syntactic structure symbols appear just before the word boundary symbol. Only one syntactic marker can appear at a given sentence position. The strongest syntactic boundary symbol is always used (the stronger symbols appear higher in the list in Table 1).

An utterance must end with either a period "." signalling a final fall in intonation, or a question mark ")?" signalling the intonation pattern appropriate for yes-no questions. Each clause must be preceeded by either "(M" to indicate the beginning of a main clause, or "(R" to indicate the beginning of a relative clause. If clauses are conjoined, a syntactic symbol is placed just before the conjunction. If a comma could be placed in the orthographic rendition of the desired utterance, then the syntactic comma symbol "," should be inserted. Syntactic commas are treated as full clause boundaries in the rules; they are used to break up larger units into chunks in order to facilitate perceptual processing. The end of a noun phrase is indicated by ")N". Segments in the syllable prior to a syntactic boundary are lengthened. Based on the results of Carlson, Granstrom, and Klatt (1979), an exception is suggested in that any)N following a noun phrase that contains only one primary-stressed content word should be erased. The NP + VP is then spoken as a single phonological phrase with no internal phrase-final lengthening.

Rules

The representation for a sentence discussed above serves as input to the phonological component of the synthesis-by-rule program. The form of the output from the phonological rules is shown at the bottom in Figure 1. The abstract string of symbols has been converted to a string of phonetic segments, with each segment being assigned a stress feature and duration in msec. Before presenting details of the duration algorithm, we summarize some of the rules that must be executed prior to duration prediction.

Stress Rules. The phonological component assigns a feature Stress (value = 0 or 1) to each phonetic segment in the output string. The default value is 0 (unstressed). Vowels preceded by a 1 or 2-stress in the input are assigned a value of 1. Consonants

preceding a stressed vowel are also assigned a value of 1 if they are in the same morpheme and if they form an acceptable word-initial consonant cluster. Segmental stress is used in rules that determine segmental duration, fundamental frequency, plosive aspiration duration, and formant target undershoot.

Rules of Segmental Phonology. There are presently very few phonological rules of a segmental nature in the program. A number of rules that are sometimes attributed by linguists to the phonological component (e.g. palatalization) are realized in the phonetic component because they involve graded phenomena (e.g. the [S] of "fish soup" is partially palatalized, but not identical to [SH:]. The segmental (within-word and across-word-boundary) phonological rules that are described below are extremely important. They are not "sloppy speech" rules, but rather rules that aid the listener in hypothesizing the locations of word and phrase boundaries. For example, the second rule ensures that a word-final /T/ is not perceived as a part of the next word by inserting simultaneous glottalization to attenuate any release burst. Rules are expressed in a feature-based notation that is compiled into Fortran code for computer simulation of the phonological component (Klatt, 1976b). Rules 1 and 2 below are stated in this way, while the others are expressed in ordinary English.

1. [L] --> [LX]/(+VOWEL)...(-STRESS)
   Substitute a postvocalic velarized allophone [LX] for [L] if the [L] is preceded by a vowel and followed by anything except a stressed vowel in the same word.
2. ([T] or [D]) --> [DX]/(+SONOR -NASAL)...(-STRESS +VOWEL)
   Replace [T] or [D] by the alveolar flap [DX] within words and across words boundaries (but not across phrase and clause boundaries) if the plosive is followed by a non-primary-stressed vowel and preceded by a nonnasal sonorant. Examples: "butter", "ladder", "sat about".
3. A word-final [T] preceded by a sonorant is replaced by the glottalized dental stop TQ (i.e. has a glottal release rather than a t-burst) if the next word starts with a stressed sonorant (unless there is a clause boundary between the words, in which case the [T] is released into a pause). Examples: "that one", "Mat ran".
4. A Voiceless plosive is not released if the next phonetic segment is another voiceless plosive within the same clause.
5. A glottal stop [Q] is inserted before a word-initial stressed vowel if the preceding segment is syllabic (and not a determiner), or if the preceding segment is a voiced nonplosive and there is an intervening phrase boundary. Example: "Liz eats".

6. Unstressed [OR] is replaced by syllabic [RR], as in "for him" or "forget". (There are many rules of this type.)

Duration Rules. Each segment is assigned a duration by a set of rules presented in detail below. The rules are intended to match observed durations for a single speaker (DHK) reading paragraph-length materials. The rules operate within the framework of a model of durational behavior which states that (1) each rule tries to effect a percentage increase or decrease in the duration of the segment, but (2) segments cannot be compressed shorter than a certain minimum duration (Klatt, 1976a). The model is summarized by the formula:

$$DUR = ((INHDUR-MINDUR)*PRCNT)/100 + MINDUR \quad (1)$$

where INHDUR is the inherent duration of a segment in msec, MINDUR is the minimum duration of a segment if stressed, and PRCNT is the percentage shortening determined by applying rules 1 to 10 below. The program begins by obtaining values for INHDUR and MINDUR for the current segment from Table 1, and by setting PRCNT to 100. The inherent duration has no special status other than as a starting point for rule application; it is roughly the duration to be expected in nonsense CVCs spoken in the carrier phrase "Say bVb again" or "Say CaC again". The following ten rules are then applied, where each rule modifies the PRCNT value obtained from the previous applicable rules according to the equation:

$$PRCNT=(PRCNT*PRCNT1)/100 \quad (2)$$

The duration of the segment is then computed by inserting the final value for PRCNT into Equation 1 and, finally, Rule 11 is applied.

1. PAUSE INSERTION RULE: Insert a 200 msec pause before each sentence-internal main clause and at boundaries delimited by a syntactic comma, but not before relative clauses. The "(R" symbol functions like a ")N" in the duration rules.
2. CLAUSE-FINAL LENGTHENING: The vowel or syllabic consonant in the syllable just before a pause is lengthened by PRCNT1=140. Any consonants between this vowel and the pause are also lengthened by PRCNT1=140.
3. NON-PHRASE-FINAL SHORTENING: Syllabic segments (vowels and syllabic consonants) are shortened by PRCNT1=60 if not in a phrase-final syllable. A phrase-final postvocalic liquid or nasal is lengthened by PRCNT1=140.
4. NON-WORD-FINAL SHORTENING: Syllabic segments are shortened by PRCNT1=85 if not in a word-final syllable.
5. POLYSYLLABIC SHORTENING: Syllabic segments in a polysyllabic word are shortened by PRCNT1=80.

6. NON-INITIAL-CONSONANT SHORTENING: Consonants in non-word-initial position are shortened by PRCNT1=85.

7. UNSTRESSED SHORTENING: Unstressed segments are half again more compressable than stressed segments (i.e. set MINDUR=MINDUR/2). Then both unstressed and 2-stressed segments are shortened by a factor PRCNT1 that is tabulated below for each type of segment. The result is that segments assigned secondary stress are shortened relative to 1-stress, but not as much as unstressed segments.

Context	PRCNT1 for Unstr. and 2-stress
syllabic (word-medial syll)	50
syllabic (others)	70
prevocalic liquid or glide	10
all others	70

8. LENGTHENING FOR EMPHASIS: An emphasized vowel is lengthened by PRCNT1=140 percent.

9. POSTVOCALIC CONTEXT OF VOWELS: The influence of a postvocalic consonant or sonorant-stop cluster on the duration of a vowel is given below. (Cs must be in the same morpheme as the V and must have the feature unstressed.) In a postvocalic sonorant-obstruent cluster, the obstruent determines the effect on the vowel and on the sonorant.

Context	PRCNT1
open syllable, word-final	120
before a voiced fricative	160
before a voiced plosive	120
before an unstr. nasal	85
before a voiceless plosive	70
all others	100

The effects are greatest at phrase and clause boundaries: if non-phrase-final, change PRCNT1 to be 70 + 0.3*PRCNT1

10. SHORTENING IN CLUSTERS: Segments are shortened in consonant-consonant sequences (disregarding word boundaries, but not across phrase boundaries), and segments

are also modifed in duration in vowel-vowel sequences.

Context	PRCNT1
vowel followed by a vowel	120
vowel preceded by a vowel	70
consonant surrounded by consonants	50
consonant preceded by a consonant	70
consonant followed by a consonant	70

11. LENGTHENING DUE TO PLOSIVE ASPIRATION: A 1-stressed or 2-stressed vowel or sonorant preceded by a aspirated plosive is lengthened by 25 msec.

When the rules are applied to the /RR/ of "rocker" in Figure 1, the second rule sets PRCNT to 140, the fifth rule reduces PRCNT to 112, the seventh rule reduces MINDUR to 30 msec and PRCNT to 78.4, and the ninth rule increases PRCNT to 94. Then INHDUR, MINDUR, and PRCNT are inserted in Equation 1 and the resulting duration is rounded up to the nearest 5 msec to obtain the value of 175 msec.

The resulting durations are determined in part by a variable that controls the nominal speaking rate SPRATE which can be set to any number between 60 and 300 words per minute. The default value is 180 words per minute. At rates slower than 150 wpm, a short pause is inserted between a content word and a following function word. (At a normal speaking rate, brief pauses are inserted only at the ends of clauses.) Individual segments are lengthened or shortened slightly depending on speaking rate, but most of the rate change is realized by manipulating pause durations.

Evaluation

The rules constitute only a first-order approximation to many of the durational phenomena seen in sentences (e.g. consonant interactions in clusters) and the rules completely ignore other factors. Nevertheless, as a first approximation, the rules capture a good deal of the systematic variation in segmental durations for speaker DHK. When compared with spectrograms of new paragraphs read by this speaker, the rule system produces segmental durations that differ from measured durations by a standard deviation of 17 msec (excluding the prediction of pause durations), and the rules account for 84 percent of the observed total variance in segmental durations. Seventeen msec is generally less than the just-noticeable difference for a single change to segmental duration in sentence materials (Klatt, 1976a).

A perceptual evaluation of the performance of the rule system is discussed by Carlson, Granstrom and Klatt (1979). The perceptual results are encouraging in that both naturalness and intelligibility ratings of sentences synthesized by these rules are very similar to ratings of the same sentences synthesized using durations obtained from a natural recording.

References

Carlson, R., Granstrom, B., and Klatt, D.H. (1979), "Some Notes on the Perception of Temporal Patterns in Speech", 9th International Congress of Phonetic Sciences, Copenhagen.

Klatt, D.H. (1976a), "Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence", J. Acoust. Soc. Am. 59, 1208-1221.

Klatt, D.H. (1976b), "Structure of a Phonological Rule Component for a Speech Synthesis by Rule Program", IEEE Trans. Acoustics, Speech, and Signal Processing ASSP-24, 391-398.