FROM TEXT TO SPEECH
THE MITALK SYSTEM

Presented by
Najeeb Khan
2013-5-31

LEXICAL STRESS PLACEMENT

LEXICAL STRESS PLACEMENT

- Application of the stress assignment rules proceeds in two phases
  - The first phase consists of three ordered rules which are applied cyclically, first to root, then to root and leftmost suffix combined and so on. Cyclic phase only concerns primary stress
  - The second noncyclic phase includes the application of rules to the entire word and reduces all but one of the primary stress marks to secondary or zero stress

LEXICAL STRESS PLACEMENT

LEXICAL STRESS PLACEMENT

- In the context of stress rules
  - Syllable means a vowel followed by any number of consonants (including 0)
  - Weak Syllable means a short vowel followed by at most one consonant before the next vowel

LEXICAL STRESS PLACEMENT

LEXICAL STRESS PLACEMENT

- Stress placement rules are given in terms of formulas. Each formula is a phonetic segment string pattern matching expression

LEXICAL STRESS PLACEMENT

- Stress placement rules are given in terms of formulas. Each formula is a phonetic segment string pattern matching expression
- The symbols used in the formulas are
  - C: matches a single consonant
  - V: matches a single vowel
  - X and Y: match segment strings of any length
  - []: denote the features with vowels
  - (): denote an optional term
  - {}: list of alternative terms

LEXICAL STRESS PLACEMENT

LEXICAL STRESS PLACEMENT

- The overall structure of the rule is

LEXICAL STRESS PLACEMENT

- The overall structure of the rule is

  V → feature / pattern

LEXICAL STRESS PLACEMENT

- The overall structure of the rule is

  V → feature / pattern

- A vowel receives the feature in the context of pattern

LEXICAL STRESS PLACEMENT

- The overall structure of the rule is

  V → feature / pattern

- A vowel receives the feature in the context of pattern
- Where the pattern contains the symbol __ in the position where the vowel is to appear

LEXICAL STRESS PLACEMENT

LEXICAL STRESS PLACEMENT

- Main Stress Rules
  - V → [1-stress] / X__C_0 {[short V]C_0^1/V} {[short V]C_0/V}
  - V → [1-stress] / X__C_0 {[short V]C_0/V}
  - V → [1-stress] / X__C_0
  - Some exceptions are discussed

LEXICAL STRESS PLACEMENT

- Main Stress Rules
  - V → [1-stress] / X__C_0 {[short V]C_0^1/V} {[short V]C_0/V}
  - V → [1-stress] / X__C_0 {[short V]C_0/V}
  - V → [1-stress] / X__C_0
  - Some exceptions are discussed
- Stressed Syllable Rules
  - In Below rules Y contains no primary stress
  - V → [1-stress] / X__C_0 {[short V]C_0^1/V}VC_0[1-stress V]Y
  - V → [1-stress] / X__C_0VC_0[1-stress V]Y

Lexical Stress Placement

LEXICAL STRESS PLACEMENT

- Alternating Stress Rules
  - $V \rightarrow [1\text{-stress}] \ / \ X\_\_C_0VVC_0[1\text{-stress } V]C_0$
  - $V \rightarrow [1\text{-stress}] \ / \ X\_\_C_0VC_0[1\text{-stress } V]C_0$

LEXICAL STRESS PLACEMENT

- Alternating Stress Rules
  - V → [1-stress] / X__C_0VVC_0[1-stress V]C_0
  - V → [1-stress] / X__C_0VC_0[1-stress V]C_0
- Destressing Rule(noncyclic)
  - V → [-stress] / C_0VC_0X__C[1-stress V]Y
  - V → [-stress] / C_0[short __]C[stress V]Y

LEXICAL STRESS PLACEMENT

- Alternating Stress Rules
  - V → [1-stress] / X__C_0VVC_0[1-stress V]C_0
  - V → [1-stress] / X__C_0VC_0[1-stress V]C_0
- Destressing Rule(noncyclic)
  - V → [-stress] / C_0VC_0X__C[1-stress V]Y
  - V → [-stress] / C_0[short __]C[stress V]Y
- Compound Stress Rule(noncyclic)
  - V → retain / X[1-stress__]YVC_0 IY
  - V → retain / X[1-stress__]YVC_0
  - V → retain / X[1-stress__]Y

LEXICAL STRESS PLACEMENT

- Alternating Stress Rules
  - V → [1-stress] / X__C_0VVC_0[1-stress V]C_0
  - V → [1-stress] / X__C_0VC_0[1-stress V]C_0
- Destressing Rule(noncyclic)
  - V → [-stress] / C_0VC_0X__C[1-stress V]Y
  - V → [-stress] / C_0[short __]C[stress V]Y
- Compound Stress Rule(noncyclic)
  - V → retain / X[1-stress__]YVC_0 IY
  - V → retain / X[1-stress__]YVC_0
  - V → retain / X[1-stress__]Y
- Vowel Reduction Rule
  - V → reduce / X[-stress , short __]Y

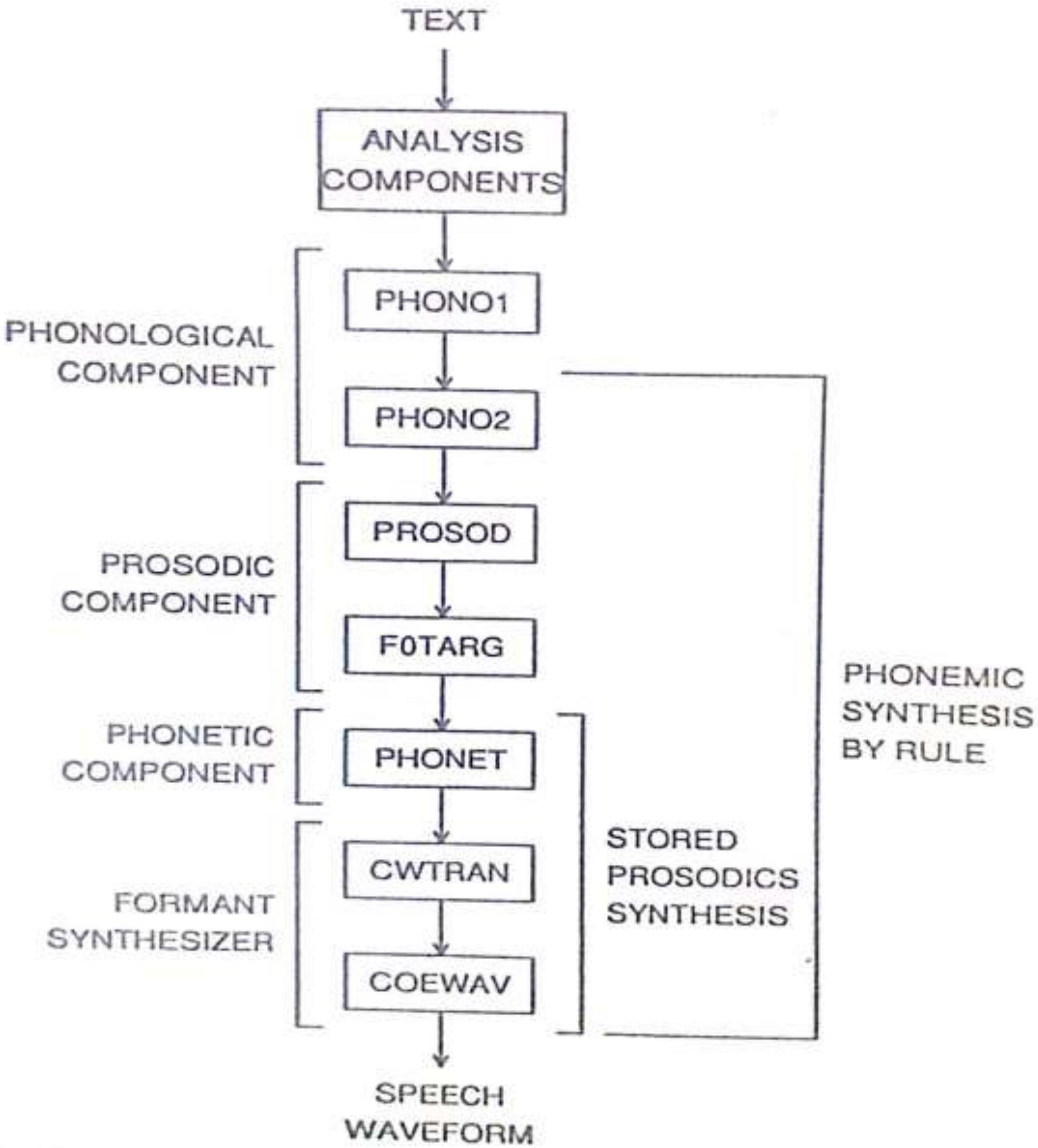
SURVEY OF SPEECH SYNTHESIS TECHNOLOGY

SURVEY OF SPEECH SYNTHESIS TECHNOLOGY

- The MITalk modules can be used in three ways

SURVEY OF SPEECH SYNTHESIS

- The M ree ways



SYNTHESIS TECHNIQUES

SYNTHESIS TECHNIQUES

- Word Assembly
  - Use prerecorded words concatenated into sentences
  - To reduce memory use LP representation
  - Or use formant trajectories extracted from prerecorded words. This allows for smoothing at boundaries and duration and f0 adjustments to match the accent of a speaker
  - Advantage: simplicity
  - Disadvantage: general timing and f0 rules that adjust the prosodic characteristics of a word as a function of sentence structure are more easily defined at a segmental level

SYNTHESIS TECHNIQUES

SYNTHESIS TECHNIQUES

- Syllables Assembly
  - Any English word can be broken into syllables consisting of vowel nucleus and adjacent consonants
  - Advantage: context conditioned acoustic changes to consonants are automatically preserved to a great extent
  - Disadvantage: Coarticulation across syllables is not treated well, if syllables are stored as prerecorded waveforms there is no way to mimic the prosodic contour of the intended message and the syllable inventory is very large

SYNTHESIS TECHNIQUES

SYNTHESIS TECHNIQUES

- DemiSyllable
  - The demisyllable is defined as half of a syllable (construct→ co-, -on, stru-,-uct)
  - There are less than 1000 demisyllables needed to synthesize any English utterance
  - Each demisyllable can be represented in terms of a set of LP frames
  - Disadvantages: how to adjust the durations to match the desired pattern for a sentence. The lengthening and shortening of speech tends to take place during the steady state, whereas the demisyllable is a mixture of SS and transitions

SYNTHESIS TECHNIQUES

SYNTHESIS TECHNIQUES

- Diphones
  - The diphone is defined as half of one phone followed by half of the other
  - Coarticulatory influence of one phoneme does not extend much farther than halfway into the next phoneme, thus minimal smoothing at the boundaries of diphones will be required

SYNTHESIS TECHNIQUES

SYNTHESIS TECHNIQUES

- Phoneme synthesis
  - Phonemes are considered the basic speech units because there are only bout 40 of them in English
  - There is no possibility of extracting phonemic sized chunks from natural speech in such a way that they can be reassembled into new utterances because of the large acoustic changes to a phoneme that occur in different phonetic environments

SYNTHESIS TECHNIQUES

SYNTHESIS TECHNIQUES

- Formant based Synthesis
  - The formant synthesizer accepts input time functions that determine formant frequencies, voicing, frication and aspiration amplitudes and fundamental frequency.
  - The synthesizer produces an output waveform that is intended to approximate the perceptually most relevant acoustic characteristics of speech

THE PHONOLOGICAL COMPONENT

THE PHONOLOGICAL COMPONENT

- The phonological component accepts input from the text analysis routines and produces an output that is sent to the prosodic component PROSOD

THE PHONOLOGICAL COMPONENT

- The phonological component accepts input from the text analysis routines and produces an output that is sent to the prosodic component PROSOD
- The phonological component consists of two modules PHONO1 and PHONO2

THE PHONOLOGICAL COMPONENT

- The phonological component accepts input from the text analysis routines and produces an output that is sent to the prosodic component PROSOD
- The phonological component consists of two modules PHONO1 and PHONO2
- PHONO1 uses information from the parser to specify the syntactic markers that influence the spoken output

THE PHONOLOGICAL COMPONENT

- The phonological component accepts input from the text analysis routines and produces an output that is sent to the prosodic component PROSOD
- The phonological component consists of two modules PHONO1 and PHONO2
- PHONO1 uses information from the parser to specify the syntactic markers that influence the spoken output
- PHONO2 contains a set of segmental recoding rules that are activated to select an appropriate allophone for each phoneme

THE PHONOLOGICAL COMPONENT

THE PHONOLOGICAL COMPONENT

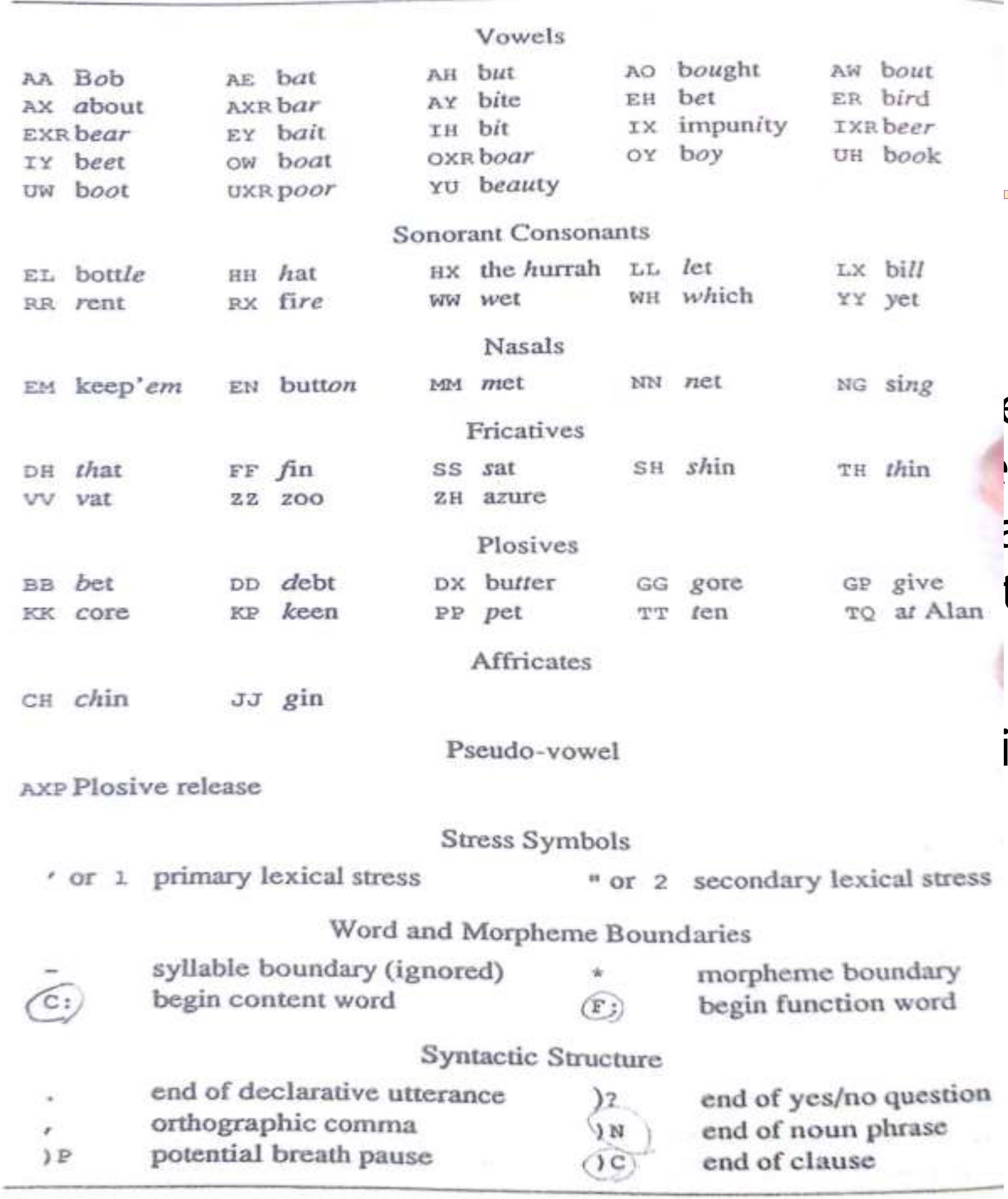
- Input:
  - The input to PHONO1 consists of a phonemic representation for each word(i.e. as spoken in isolation), lexical stress pattern, and syntactic information concerning pos and phrasal structure

THE PHONOLOGICAL COMPONENT

- Input:
  - The input to PHONO1 consists of a phonemic representation for each word(i.e. as spoken in isolation), lexical stress pattern, and syntactic information concerning pos and phrasal structure
- Output
  - The output from PHONO1 consists of a single string of symbols for each sentence
  - The symbol inventory used in the PHONO1 and PHONO2 is shown

T T

- In
  - T emic r en in i actic i ture
- Ou
  - T ingle s
  - T and P



Vowels

AA	Bob	AE	bat	AH	but	AO	bought	AW	bout
AX	about	AXR	bar	AY	bite	EH	bet	ER	bird
EXR	bear	EY	bait	IH	bit	IX	impunity	IXR	beer
IY	beet	OW	boat	OXR	boar	OY	boy	UH	book
UW	boot	UXR	poor	YU	beauty				

Sonorant Consonants

EL	bottle	HH	hat	HX	the hurrah	LL	let	LX	bill
RR	rent	RX	fire	WW	wet	WH	which	YY	yet

Nasals

EM	keep'em	EN	button	MM	met	NN	net	NG	sing

Fricatives

DH	that	FF	fin	SS	sat	SH	shin	TH	thin
VV	vat	ZZ	zoo	ZH	azure				

Plosives

BB	bet	DD	debt	DX	butter	GG	gore	GP	give
KK	core	KP	keen	PP	pet	TT	ten	TQ	at Alan

Affricates

CH	chin	JJ	gin

Pseudo-vowel

AXP Plosive release

Stress Symbols

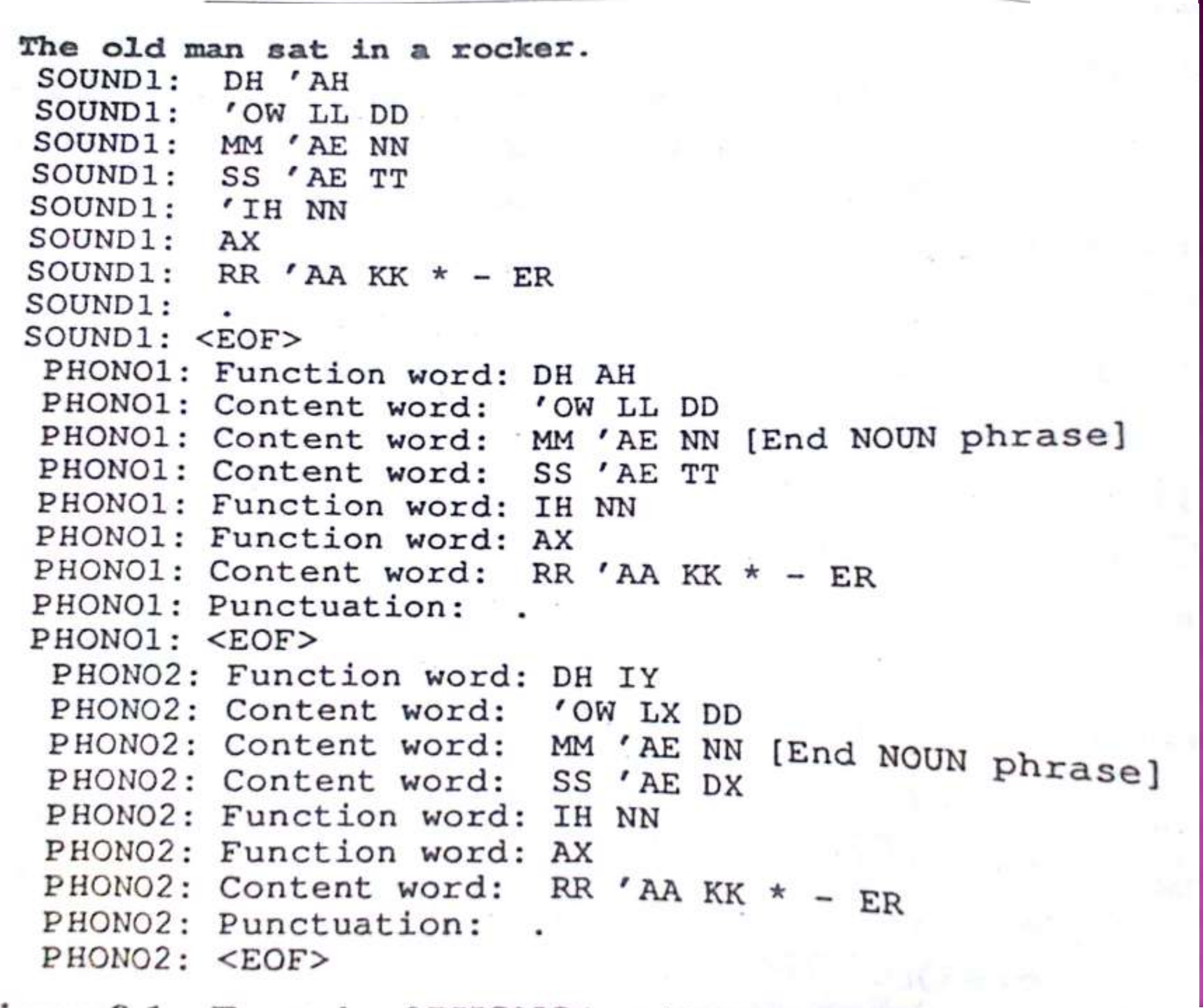
′ or 1	primary lexical stress	" or 2	secondary lexical stress

Word and Morpheme Boundaries

-	syllable boundary (ignored)	*	morpheme boundary
C:	begin content word	F:	begin function word

Syntactic Structure

.	end of declarative utterance	)?	end of yes/no question
,	orthographic comma	)N	end of noun phrase
)P	potential breath pause	)C	end of clause



```
The old man sat in a rocker.
 SOUND1:  DH 'AH
 SOUND1:  'OW LL DD
 SOUND1:  MM 'AE NN
 SOUND1:  SS 'AE TT
 SOUND1:  'IH NN
 SOUND1:  AX
 SOUND1:  RR 'AA KK * - ER
 SOUND1:  .
 SOUND1: <EOF>
  PHONO1: Function word: DH AH
  PHONO1: Content word:  'OW LL DD
  PHONO1: Content word:  MM 'AE NN [End NOUN phrase]
  PHONO1: Content word:  SS 'AE TT
  PHONO1: Function word: IH NN
  PHONO1: Function word: AX
  PHONO1: Content word:  RR 'AA KK * - ER
  PHONO1: Punctuation:  .
  PHONO1: <EOF>
   PHONO2: Function word: DH IY
   PHONO2: Content word:  'OW LX DD
   PHONO2: Content word:  MM 'AE NN [End NOUN phrase]
   PHONO2: Content word:  SS 'AE DX
   PHONO2: Function word: IH NN
   PHONO2: Function word: AX
   PHONO2: Content word:  RR 'AA KK * - ER
   PHONO2: Punctuation:  .
   PHONO2: <EOF>
```

The Phonological Component

THE PHONOLOGICAL COMPONENT

- Lexical Stress
  - Each stressed vowel in the input is preceded by a stress symbol (‘ or “)

THE PHONOLOGICAL COMPONENT

- Lexical Stress
  - Each stressed vowel in the input is preceded by a stress symbol (‘ or “)
- Stress Reduction
  - Many closed class function words are reduced in stress in PHONO1 so that they do not receive a pitch gesture associated with primary stress

THE PHONOLOGICAL COMPONENT

- Lexical Stress
  - Each stressed vowel in the input is preceded by a stress symbol (‘ or “)
- Stress Reduction
  - Many closed class function words are reduced in stress in PHONO1 so that they do not receive a pitch gesture associated with primary stress
- Syntactic Structure
  - Syntactic structure symbols appear just before the word boundary symbols
  - They are important determiners of sentence stress , rhythm and intonation

THE PHONOLOGICAL COMPONENT

THE PHONOLOGICAL COMPONENT

- Errors in the Analysis routines:
  - An error made by the analysis routine need not to be an error in some abstract linguistic sense, but only an error in the sense that the symbol is not the one that is desired by the synthesis routines

THE PHONOLOGICAL COMPONENT

- Errors in the Analysis routines:
  - An error made by the analysis routine need not to be an error in some abstract linguistic sense, but only an error in the sense that the symbol is not the one that is desired by the synthesis routines
- Phonetic Transcription Errors
  - There are 25 phonetic transcription errors (in the test paragraphs analyzed by MITalk) most of which concern the difference between “I” and schwa

THE PHONOLOGICAL COMPONENT

- Errors in the Analysis routines:
  - An error made by the analysis routine need not to be an error in some abstract linguistic sense, but only an error in the sense that the symbol is not the one that is desired by the synthesis routines
- Phonetic Transcription Errors
  - There are 25 phonetic transcription errors (in the test paragraphs analyzed by MITalk) most of which concern the difference between "I" and schwa
- Stress Errors
  - Certain common words such as 'might' and 'each' should be marked with primary stress in the lexicon because they almost always attract a certain amount of semantic focus

THE PHONOLOGICAL COMPONENT

THE PHONOLOGICAL COMPONENT

- Morpheme Boundary Problems
  - The morpheme boundary symbol is used to prevent word such as back*ache from having a strongly aspirated kk. However in words such as applic*ation a strongly aspirated kk is desired
  - Perhaps the morpheme boundary symbol should be removed between roots and bound suffix but not between two root morphemes

THE PHONOLOGICAL COMPONENT

- Morpheme Boundary Problems
  - The morpheme boundary symbol is used to prevent word such as back*ache from having a strongly aspirated kk. However in words such as applic*ation a strongly aspirated kk is desired
  - Perhaps the morpheme boundary symbol should be removed between roots and bound suffix but not between two root morphemes
- Syntactic Errors
  - There are a large number of syntactic errors
  - The trade off between adding breath pauses to break the speech up into fewer processing chunks versus insertion of breaks at a syntactically unacceptable place has yet to be optimized

THE PHONOLOGICAL COMPONENT

THE PHONOLOGICAL COMPONENT

- Stress Rules
  - The phonological component assigns a feature stress value to each phonetic segment in the output string
  - Stressed consonants are defined to be affiliated with the following vowel while unstressed consonants are affiliated with a preceding vowel

THE PHONOLOGICAL COMPONENT

- Stress Rules
  - The phonological component assigns a feature stress value to each phonetic segment in the output string
  - Stressed consonants are defined to be affiliated with the following vowel while unstressed consonants are affiliated with a preceding vowel
- Pauses
  - Pauses of 800ms sufficient for a real speaker to take a breath are introduced after any sentence of more than 5 words
  - 1200ms pause appears at the end of a paragraph
  - Brief sentence internal pauses (400ms) are triggered by punctuation marks contained in the text

PROSODIC COMPONENT

PROSODIC COMPONENT

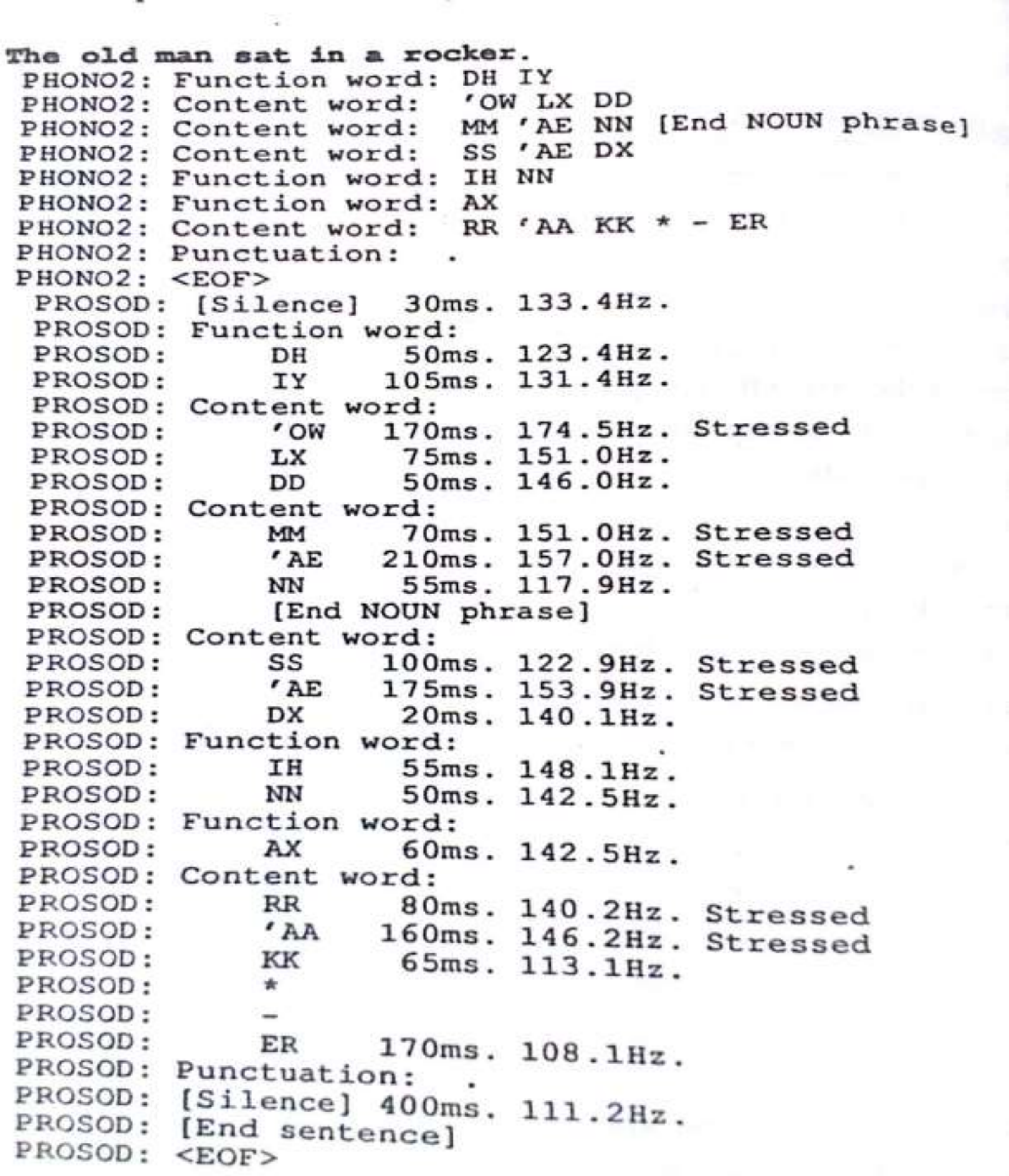
- The sentence representation produced by the phonological component serves as input to the prosodic component PROSOD

PROSODIC COMPONENT

- The sentence representation produced by the phonological component serves as input to the prosodic component PROSOD
- The output consists of a string of phonetic segments, with each segment assigned a stress feature and a duration in msec

PROSODIC COMPONENT

- The sentence representation produced by the phonological component serves as input to the prosodic component PROSOD
- The output consists of a string of phonetic segments, with each segment assigned a stress feature and a duration in msec
- The f0 targets are computed by an obsolete algorithm and are replaced latter



```
The old man sat in a rocker.
 PHONO2: Function word: DH IY
 PHONO2: Content word:  'OW LX DD
 PHONO2: Content word:  MM 'AE NN [End NOUN phrase]
 PHONO2: Content word:  SS 'AE DX
 PHONO2: Function word: IH NN
 PHONO2: Function word: AX
 PHONO2: Content word:  RR 'AA KK * - ER
 PHONO2: Punctuation:   .
 PHONO2: <EOF>
  PROSOD: [Silence]  30ms. 133.4Hz.
  PROSOD: Function word:
  PROSOD:      DH     50ms. 123.4Hz.
  PROSOD:      IY    105ms. 131.4Hz.
  PROSOD: Content word:
  PROSOD:      'OW   170ms. 174.5Hz. Stressed
  PROSOD:      LX     75ms. 151.0Hz.
  PROSOD:      DD     50ms. 146.0Hz.
  PROSOD: Content word:
  PROSOD:      MM     70ms. 151.0Hz. Stressed
  PROSOD:      'AE   210ms. 157.0Hz. Stressed
  PROSOD:      NN     55ms. 117.9Hz.
  PROSOD:      [End NOUN phrase]
  PROSOD: Content word:
  PROSOD:      SS    100ms. 122.9Hz. Stressed
  PROSOD:      'AE   175ms. 153.9Hz. Stressed
  PROSOD:      DX     20ms. 140.1Hz.
  PROSOD: Function word:
  PROSOD:      IH     55ms. 148.1Hz.
  PROSOD:      NN     50ms. 142.5Hz.
  PROSOD: Function word:
  PROSOD:      AX     60ms. 142.5Hz.
  PROSOD: Content word:
  PROSOD:      RR     80ms. 140.2Hz. Stressed
  PROSOD:      'AA   160ms. 146.2Hz. Stressed
  PROSOD:      KK     65ms. 113.1Hz.
  PROSOD:      *
  PROSOD:      -
  PROSOD:      ER    170ms. 108.1Hz.
  PROSOD: Punctuation:  .
  PROSOD: [Silence] 400ms. 111.2Hz.
  PROSOD: [End sentence]
  PROSOD: <EOF>
```

- Th the ph t to th
- Th etic se d a st
- Th lete al

PROSODIC COMPONENT

PROSODIC COMPONENT

- Segmental Durations
  - Only a few of the rule governed durational changes are perceptually discriminable
  - The goal is to characterize these perceptually important first order effects

PROSODIC COMPONENT

- Segmental Durations
  - Only a few of the rule governed durational changes are perceptually discriminable
  - The goal is to characterize these perceptually important first order effects
- Durational definitions
  - Closure for a stop
  - Interval of visible frication noise for fricatives
  - For sonorants, the segmental boundary is defined to be the half-way point in the formant transition for that formant having the greatest extent of transition

PROSODIC COMPONENT

PROSODIC COMPONENT

- Segmental Durations
  - Each segment is assigned a duration by a set of rules
  - The rules operate within the framework of a model of durational behavior which states that
    - Each rule tries to effect a percentage change in the duration of the segment
    - Segments cannot be compressed shorter than a certain minimum duration
  - Dur = ((INHDUR-MINDUR)*PRCNT)/100 +MINDUR

PROSODIC COMPONENT

PROSODIC COMPONENT

- Ten rules are applied , where each rule modifies the PRCNT value obtained from the previous applicable rules by an amount PRCNT1

PROSODIC COMPONENT

- Ten rules are applied , where each rule modifies the PRCNT value obtained from the previous applicable rules by an amount PRCNT1
- The duration of the segment is then computed by inserting the final value of PRCNT into the model equation and finally rule 11 is applied

PROSODIC COMPONENT

PROSODIC COMPONENT

- Rules
  - Pause insertion Rule: Insert a 200msec pause before each sentence internal main clause
  - Clause final Lengthening rule: The vowel or syllabic consonant in the syllable just before a pause is lengthened by 140%
  - Non-Phrase final Shortening: Syllabic segments are shortened by 60% if not in the phrase final position
  - Non Word-final shortening: Syllabic segments are shortened by 60% if not in the word final position
  - Polysyllabic Shortening: syllabic segments in a polysyllabic word are shortened by 80%
  - Non initial consonant shortening: Consonants in non-word-initial position are shortened by 85%
  - Unstressed Shortening: unstressed segments are half-again more compressible than stressed segments (MINDUR=MINDUR/2)

PROSODIC COMPONENT

PROSODIC COMPONENT

- Rules
  - An emphasized vowel is lengthened by 140%
  - Postvocalic Context of words: The influence of a postvocalic consonant on the duration of a vowel depends on the type of consonant

Thank You