Speech Synthesis

Text to Speech

"Text-to-Speech software is used to convert words from a computer document (e.g. word processor document, web page) into <u>audible speech</u> spoken through the computer speaker"

Text to Speech Synthesis System

Text Normalization

Conversion of Text that include non standard word

- Abbreviation (like)
- Numbers (like)

Text processing

- Grapheme to Phoneme (Pronunciation)
- Prosody
  - Syllabification
  - Phrase/clause marking
  - Prosody and Intonation

Synthesis

Numbers

- Deciding how to convert numbers is another problem TTS systems have to address.
- It is a fairly simple programming challenge to convert a number into words, like 1325 becoming "one thousand three hundred twenty-five".
- However, numbers occur in many different contexts in texts, and 1325 should probably be read as "**thirteen twenty-five**" when part of an address (1325 Main St.) and as "one three two five" if it is the last four digits of a social security number.
- Often a TTS system can infer how to expand a number based on surrounding words, numbers, and punctuation, and sometimes the systems provide a way to specify the type of context if it is ambiguous.

Abbreviations

- Similarly, abbreviations like "**etc.**" are easily rendered as "et cetera", but often abbreviations can be ambiguous.
- For example, the abbreviation "**in.**" in the following example: "Yesterday it rained 3 in. Take 1 out, then put 3 in."
- "**St.**" can also be ambiguous: "St. John St."
- TTS systems with intelligent front ends can make educated guesses about how to deal with ambiguous abbreviations, while others do the same thing in all cases, resulting in nonsensical but sometimes comical outputs: "Yesterday it rained three in." or "Take one out, then put three inches."

Text-to-phoneme challenges

- Speech synthesis systems use two basic approaches to determine the pronunciation of a word based on its spelling, a process which is often called text-to-phoneme or grapheme-to-phoneme conversion, as phoneme is the term used by linguists to describe distinctive sounds in a language.

Dictionary Based approach

- The simplest approach to text-to-phoneme conversion is the **dictionary-based** approach, where a large dictionary containing all the words of a language and their correct pronunciation is stored by the program. Determining the correct pronunciation of each word is a matter of looking up each word in the dictionary and replacing the spelling with the pronunciation specified in the dictionary.

Pronunciations lexicon format of W3C (PLS)

```xml
<?xml version="1.0" encoding="UTF-8"?>
<!-- This pronunciation lexicon is licensed under the GPL. -->
<lexiconversion="1.0"
xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.w3.org/2005/01/pronunciation-lexicon
     http://www.w3.org/TR/2007/CR-pronunciation-lexicon-20071212/pls.xsd"
alphabet="ipa" xml:lang="bn">
<lexeme>
<grapheme>আমি</grapheme>
<phoneme>ɐmi</phoneme>
</lexeme>
<lexeme>
<grapheme>ছাড়া</grapheme>
<phoneme>ʧʰɐɽɐ</phoneme>
</lexeme>
<lexeme>
<grapheme>দ্বিতীয়</grapheme>
<phoneme>ditijo</phoneme>
</lexeme>
<lexeme>
<grapheme>কেউ</grapheme>
<phoneme>keu</phoneme>
</lexeme>
</lexicon>
```

Rule based approach

- The other approach used for text-to-phoneme conversion is the **rule-based** approach, where rules for the pronunciations of words are applied to words to work out their pronunciations based on their spellings. This is similar to the "sounding out" approach to learning reading.

Hybrid approach

- Articulatory
- Parametric
  - Formant Synthesis
  - HMM based TTS
- Concatenative
  - Di-Phone Synthesis
  - Element Based (ESNOLA)
  - Unit selection
  - Unit Selection with Prosodic Modification
  - HMM based speech synthesis

Articulatory synthesis

- In an articulatory synthesis, models of the human articulators (tong, lips, teethes, jaw) and vocal ligament are used to simulate how an airflow passes through, to calculate what the resulting sound will be like. It is a great challenge to find good mathematical models and therefore the development of articulatory synthesis is still in research. The technique is very computation-intensive but memory requirements is almost nothing.

Formant Synthesis

This synthesis is a sort of source-filter-method that is based on mathematic models of the human speech organ. The approach pipe is modelled from a number of resonances with resemblance to the formants (frequency bands with high energy in voices) in natural speech. The first electronic voices Voder, and later on OVE and PAT, were speaking with totally synthetic and electronic produced sounds using formant synthesis. As with articulatory synthesis, the memory consumption is small but CPU usage is large.

Formant Synthesis

- Formant synthesis does not use any human speech samples at runtime. Instead, the output synthesized speech is created using an acoustic model.
- Parameters such as frequency amplitude etc are varied over time to create a waveform of artificial speech.

Concatenating synthesis

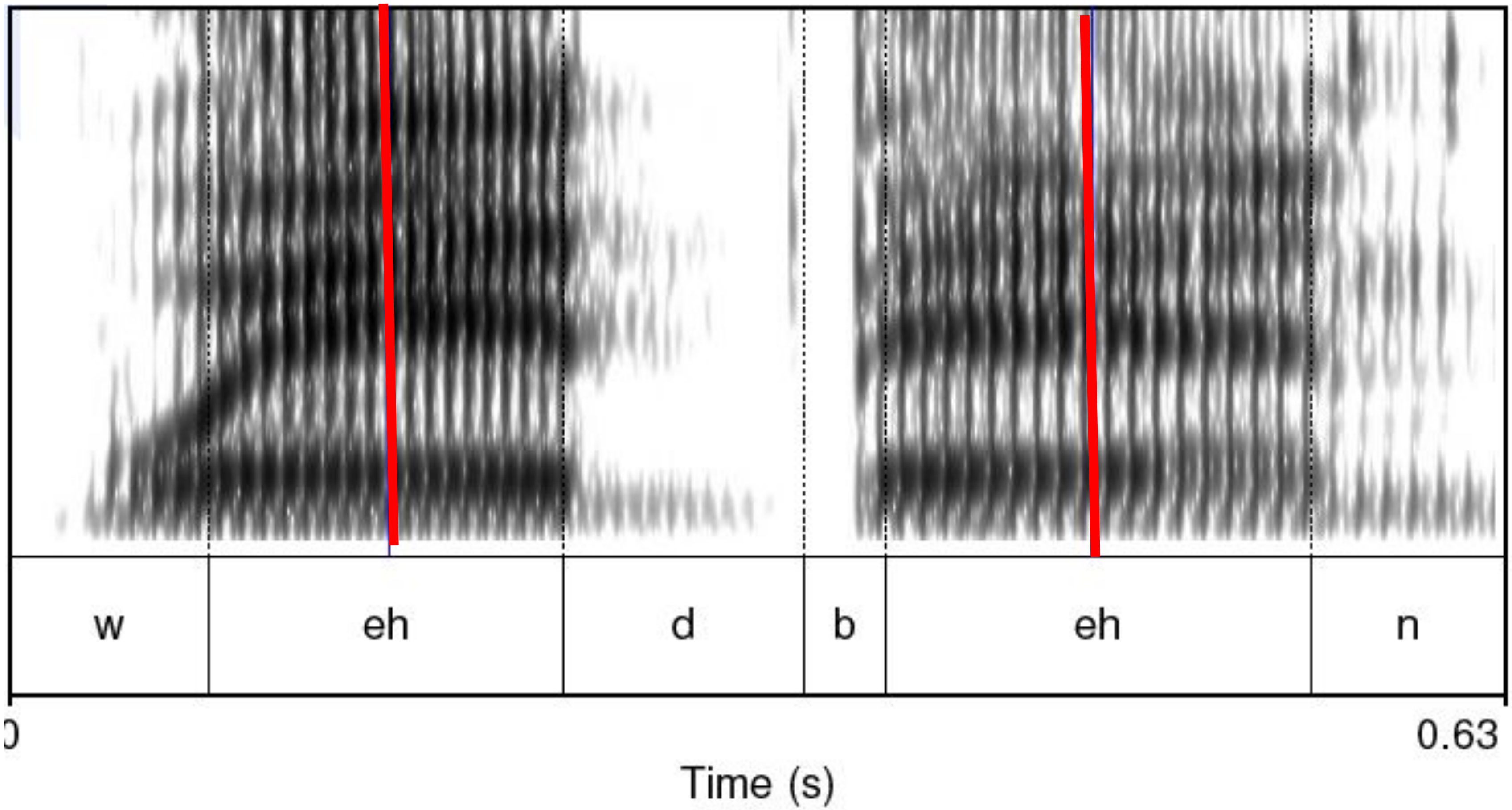
- A concatenating synthesis is made of recorded pieces of speech (sound-clips) that is then unitized and formed to speech.
- Depending on the length of sound-clips that are used it become a diphone or a polyphonic synthesis.
- The latter in a more developed version is also called a Unit Selection synthesis, where the synthesizer has access to both long and short segments of speech and the best segments for the actual context is chosen

Diphone TTS architecture

- **Training:**
  - Choose units (kinds of diphones)
  - Record 1 speaker saying 1 example of each diphone
  - Mark the boundaries of each diphones,
    - cut each diphone out and create a diphone database
- **Synthesizing an utterance,**
  - grab relevant sequence of diphones from database
  - Concatenate the diphones, doing slight signal processing at boundaries
  - use signal processing to change the prosody (F0, energy, duration) of selected sequence of diphones

Diphones

Mid-phone is more stable than edge:



Designing a diphone inventory: Nonsense words

- **Build set of carrier words:**
  - pau t aa b aa b aa pau
  - pau t aa m aa m aa pau
  - pau t aa m iy m aa pau
  - pau t aa m iy m aa pau
  - pau t aa m ih m aa pau
- **Advantages:**
  - Easy to get all diphones
  - Likely to be pronounced consistently
    - No lexical interference
- **Disadvantages:**
  - (possibly) bigger database
  - Speaker becomes bored

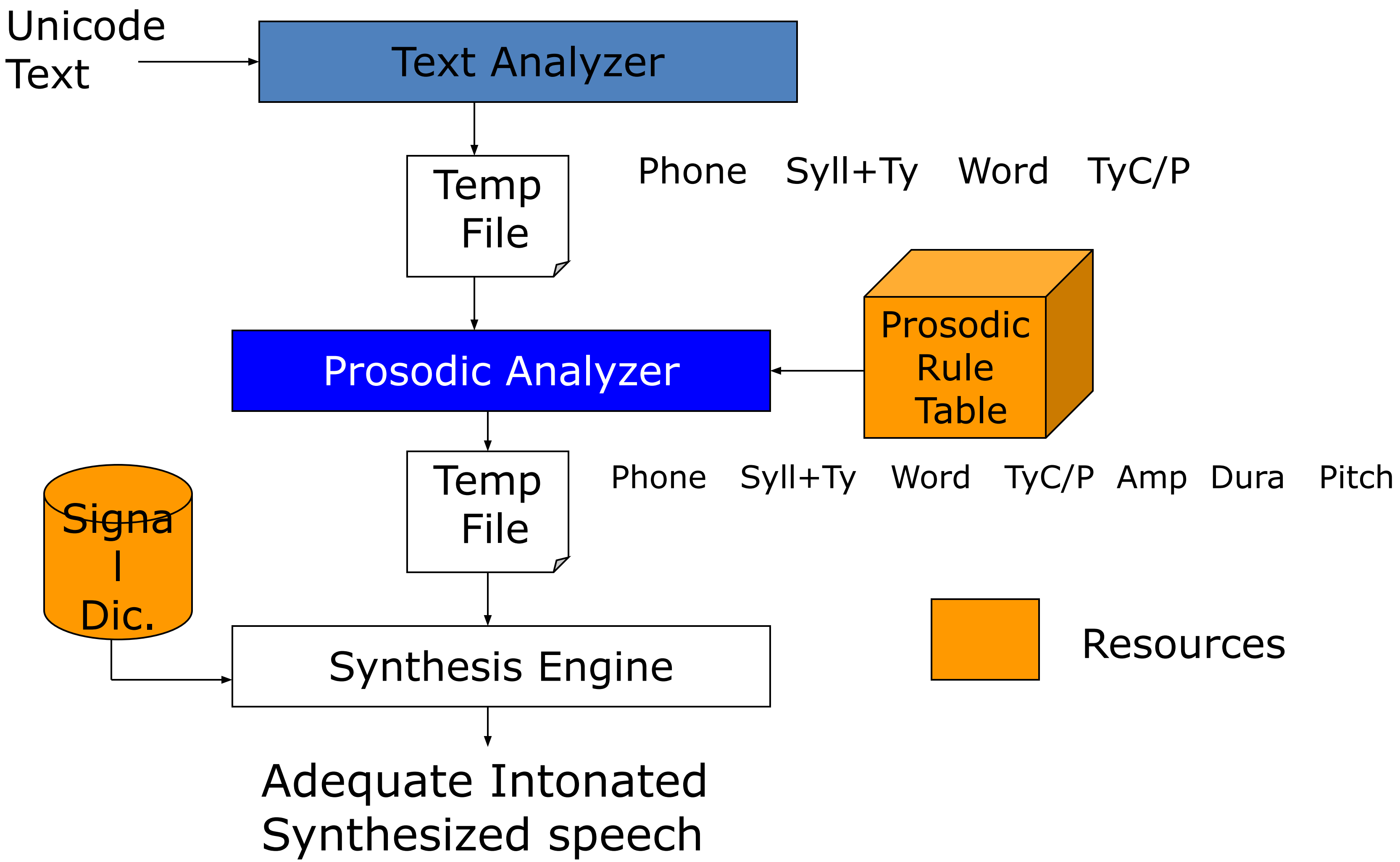
Designing a diphone inventory: Natural words

- Greedily select sentences/words:
  - Quebecois arguments
  - Brouhaha abstractions
  - Arkansas arranging
- Advantages:
  - Will be pronounced naturally
  - Easier for speaker to pronounce
  - Smaller database? (505 pairs vs. 1345 words)
- Disadvantages:
  - May not be pronounced correctly

Making recordings consistent:

- Diiphone should come from mid-word
  - Help ensure full articulation
- Performed consistently
  - Constant pitch (monotone), power, duration
- Use (synthesized) prompts:
  - Helps avoid pronunciation problems
  - Keeps speaker consistent
  - Used for alignment in labeling

Architecture of the ESNOLA System



*

1. CVCV. $\longrightarrow$ $C + CV + V + VC + C + V + V_o$

बाजे */baaje/* $\longrightarrow$ $b + baa + aa + aaj + j + je + e + e_o$

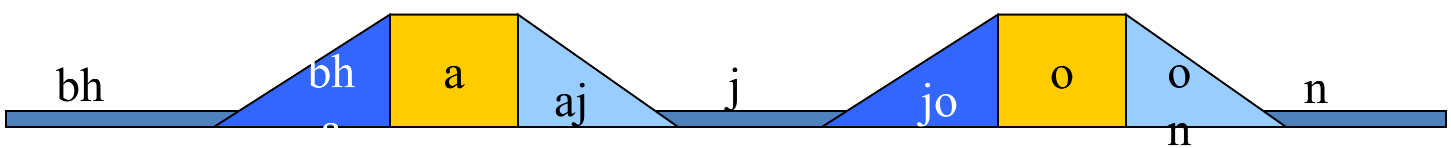
2. VCV. $V_i + V + VC + C + CV + V + V_o$

आगे */aage/* $\longrightarrow$ $aa_i + aa + aag + g + ge + e + e_o$

3. CVYV. $C + CV + V + VY + YV + V_o$

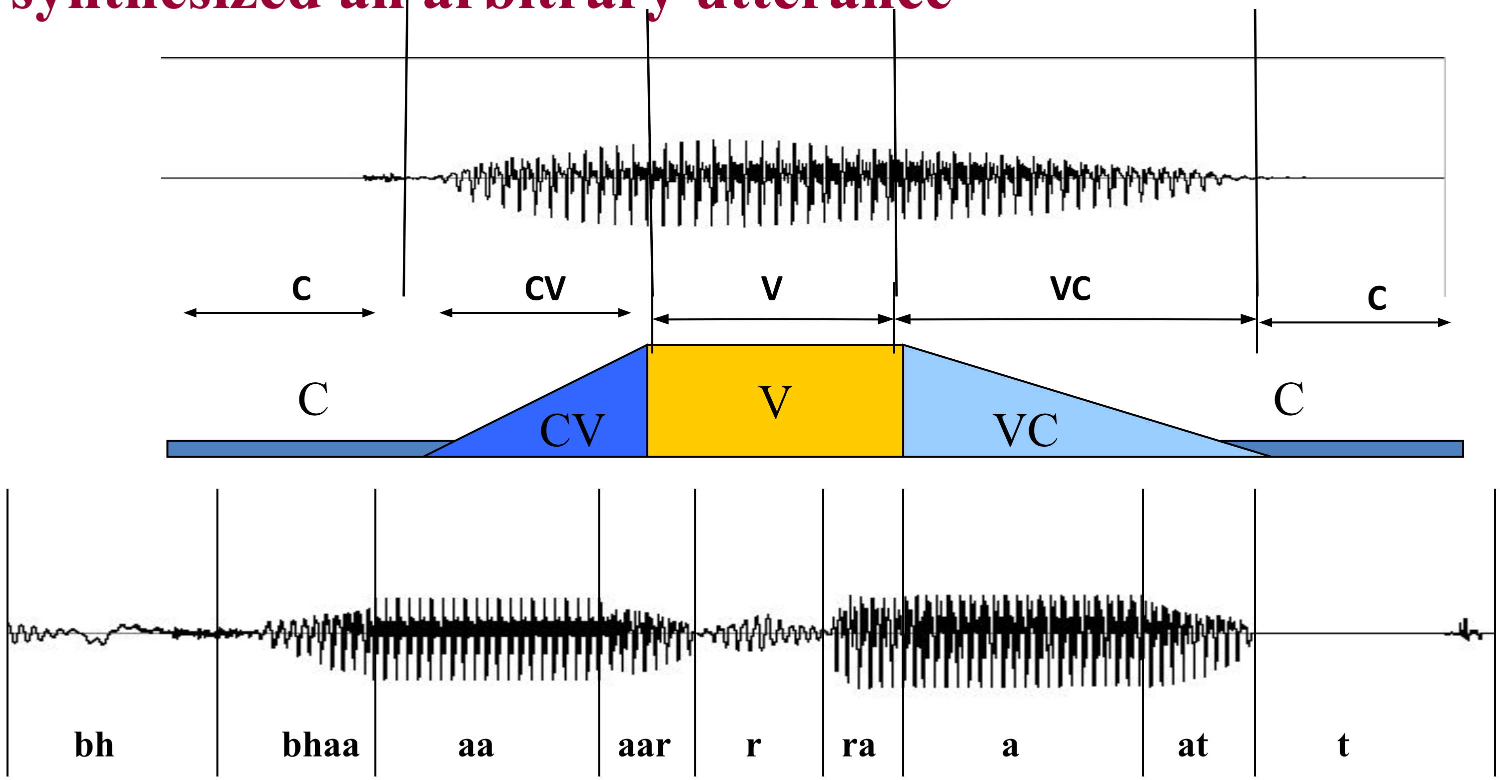
रोयो /royo/ $\longrightarrow$ $r + ro + o + oy + y + yo + o + o_o$

भजन् */bhajon/*



ESNOLA Method Based Speech Synthesis System for Bangla

Concatenative synthesis is based on putting together pieces (acoustic unit) of natural(recorded) speech to synthesized an arbitrary utterance



/bhaarat/

*** Epoch Synchronous Non-OverLapping Add**

রাজা মহানন্দ রাজধানীতে তৈরি করেছিল শিব মন্দির ও বৈষ্ণবদের মন্দির।

What is Prosody ? --- The Author's Definition (Fujisaki 1995)

Prosody is defined as the systematic organization of individual linguistic units into an utterance, or a coherent group of utterances, in the process of speech production.

Its realization involves both segmental and suprasegmental features of speech, and is influenced, not only by linguistic information, but also by para-linguistic and non-linguistic information.

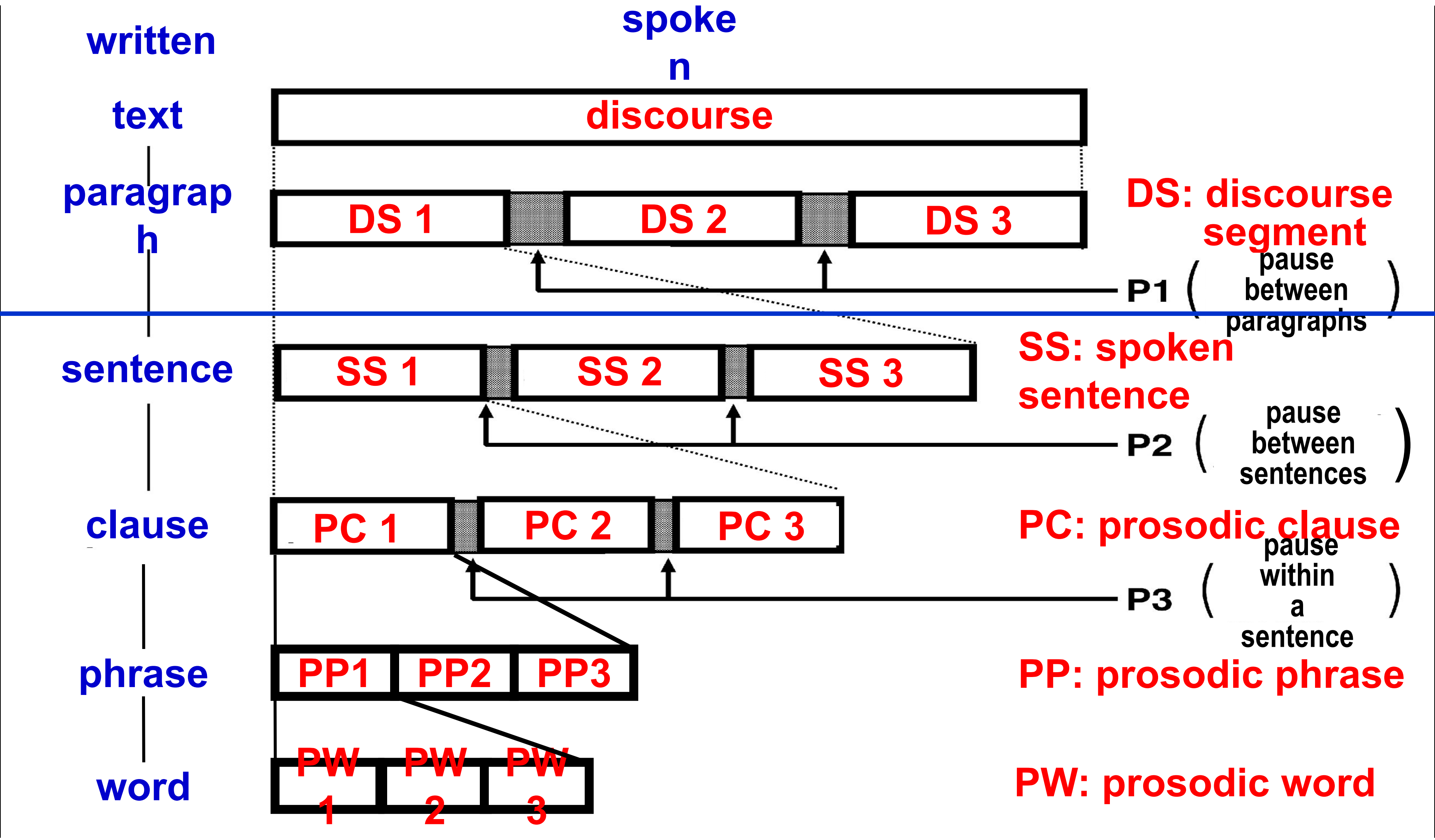
Role of Voice Fundamental Frequency (F_0) Contour

- In many languages, the pattern of temporal changes in F_0 (henceforth the F_0 contour) is used to express *tone*, *accent*, and *intonation*, and plays a major role in conveying linguistic information on the prosody (i.e., the structural organization of various linguistic units into a coherent utterance or a coherent group of utterances).
- It can convey also *para-linguistic* information concerning speaker's intention and attitude, as well as *non-linguistic* information concerning speaker's physical and mental states (such as age, emotion, etc.)

Three Approaches to the Description/Representation of F_0 Contour Characteristics

	Example	Outcome	Method	Coding/ Decoding
Labeling	ToBI	Discrete Labels	Subjective Qualitative	Irreversible
Stylization	’t Hart	Piece-wise Linear Approx.	Subjective Quantitative	Irreversible
Modeling	Fujisaki	Timing and Magnitude of Commands	Objective Quantitative	Reversible

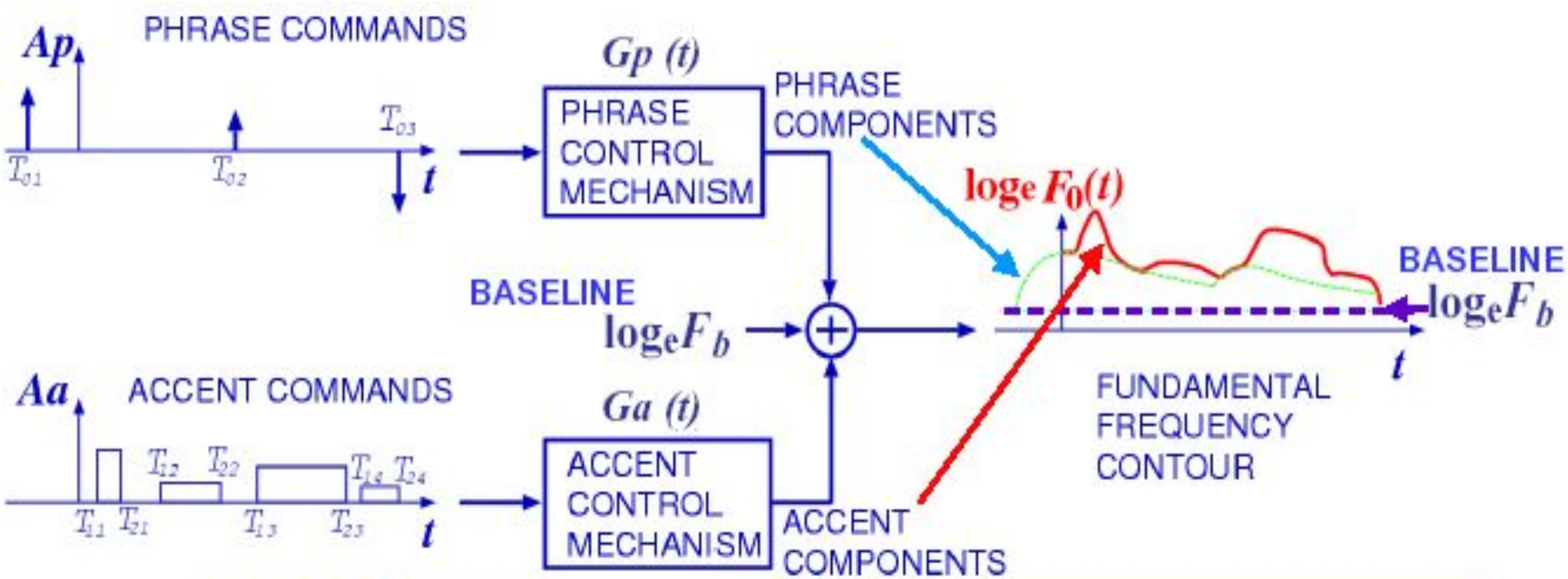
Prosodic Structure



Speech Parameters controlling the Supra-segmental Features

- Pause
- Duration
- Intonation(F0 model)
- Loudness(Amplitude model)

A functional model for the process of generating F0 contours

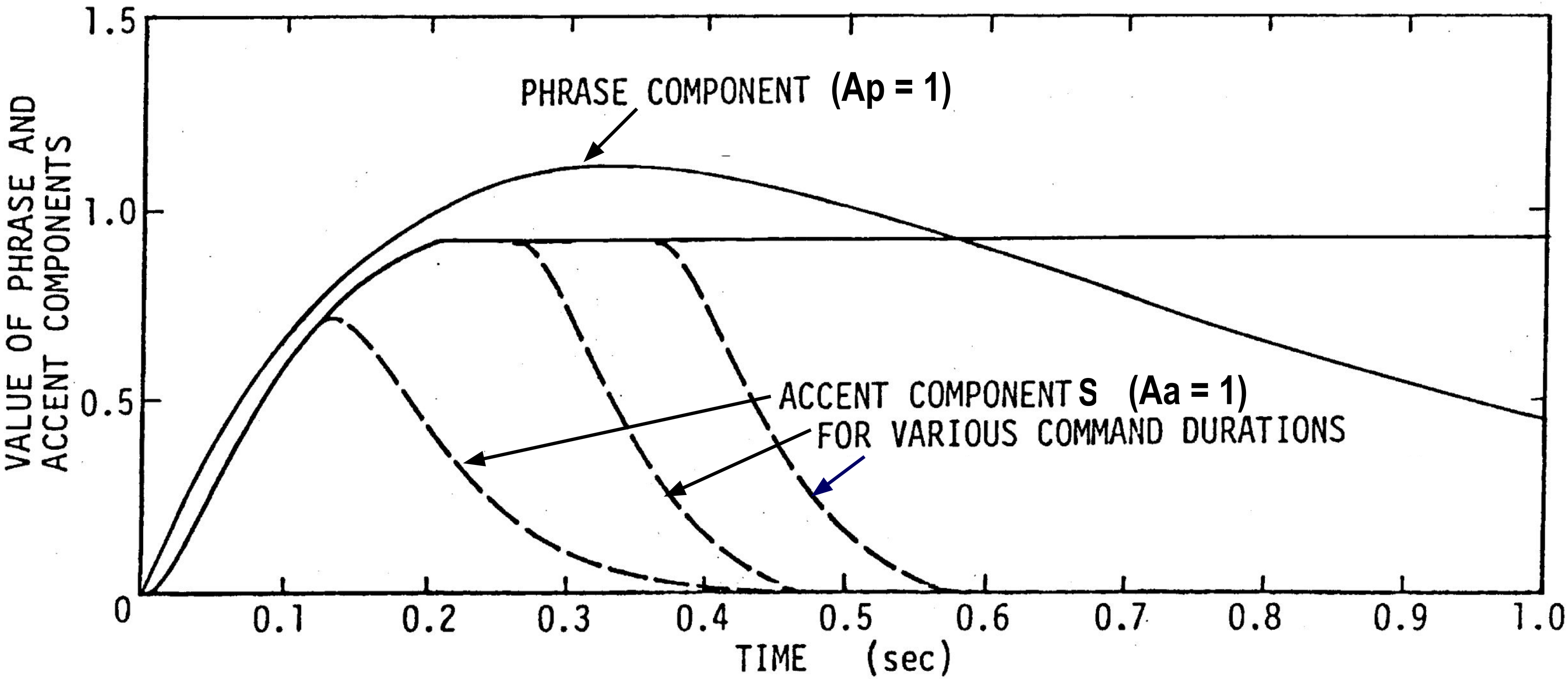


$$\log_e F_0(t) = \boxed{\log_e F_b} + \boxed{\sum_{i=1}^{I} Ap_i Gp(t - T_{0i})} + \boxed{\sum_{j=1}^{J} Aa_j \{Ga(t - T_{1j}) - Ga(t - T_{2j})\}} \qquad (1)$$

$$G_p(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & t \geq 0, \\ 0, & t < 0, \end{cases} \qquad (2)$$

$$G_a(t) = \begin{cases} Min[1 - (1 + \beta t)\exp(-\beta t),\ \gamma], & t \geq 0, \\ 0, & t < 0. \end{cases} \qquad (3)$$

Shapes of phrase and accent components with typical values of α, βand γ



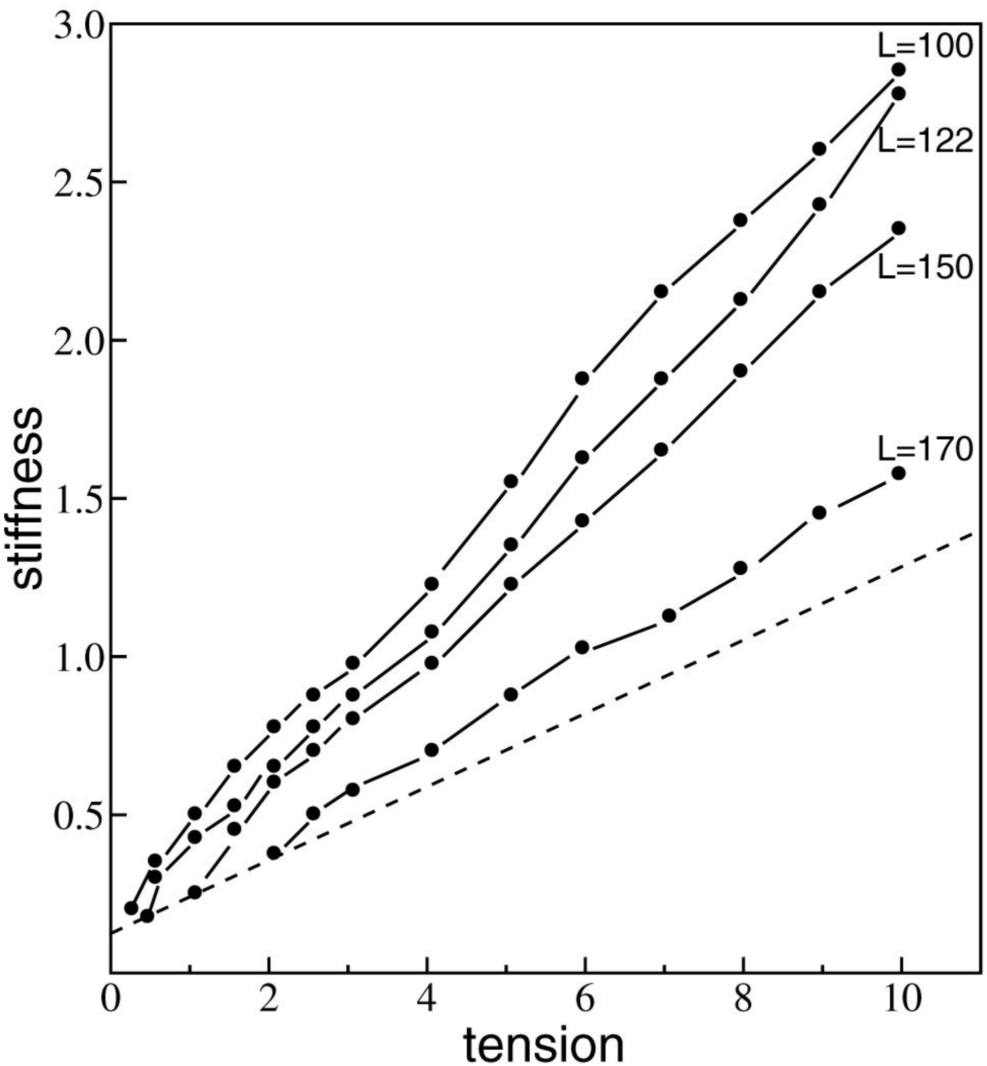
Parameter values for the phrase component: α= 3.0/s,
the accent components: β= 20.0/s, γ= 0.9.

Stress-strain relationship of skeletal muscles

The stress-strain relationship of skeletal muscles including the human vocalis muscle has been widely studied [e.g., Buchthal & Kaiser 1944, Sandow, 1958].

The next figure shows the earliest published data on the relationship between tension and stiffness of a skeletal muscle by Buchthal and Kaiser published in *Acta Physiol. Scand.*

Physical properties of skeletal muscles (1)

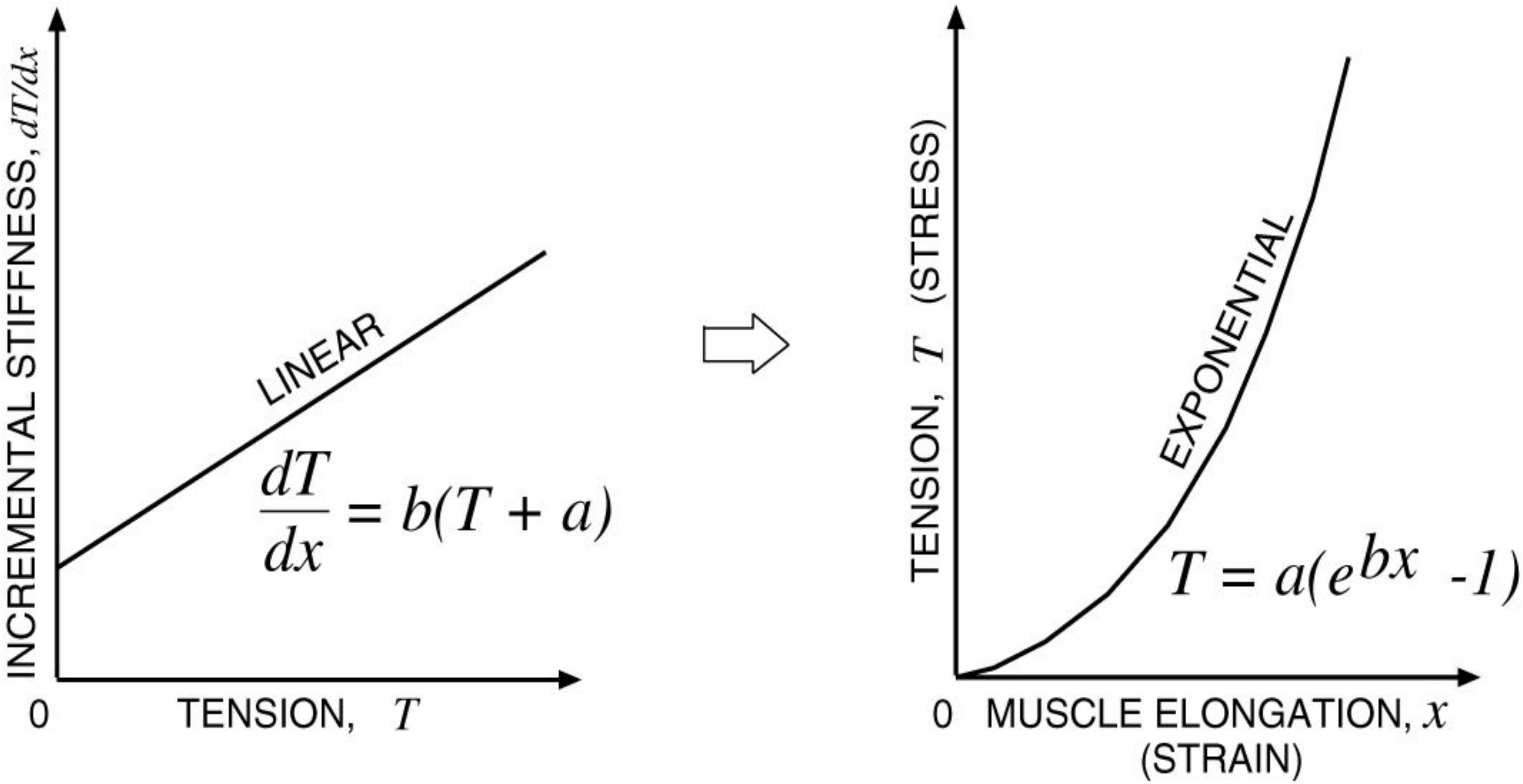


Stiffness as function of tension at rest (- - - - -) and during isometric tetanic contraction initiated at different original lengths. In the top curve contraction is initiated at a length below 100 (equilibrium length = 100).

Ordinate: stiffness in arbitrary units.
Abscissa: tension in arbitrary units.

Buchthal, F. and E. Kaiser, *Acta Physiol. Scandinavica,* vol. 8, pp. 38-74, 1944.

Physical properties of skeletal muscles (2)



From vocal cord elongation to tension

Stress-strain relationship in a skeletal muscle (i.e., vocalis)

$$dT/dl = b\,(T + a), \qquad (1)$$

where T: tension, l : length of vocalis, a: stiffness at $T = 0$.

By integration, we obtain

≈

$$T = (T_0 + a/b)\exp\{b(l - l_0)\} - a/b, \qquad (2)$$

where T_0: static tension, l_0: vocalis length at $T = T_0$.

When $T_0 >> a/b$,

where $x = l - l_0$.

$$\boxed{T \cong T_0 \exp(bx). \qquad (3)}$$

From vocal cord tension to fundamental frequency

Frequency of vibration of an elastic membrane is given by

$$\boxed{F_0 = c_0\{T/\sigma\}^{1/2}, \quad \text{where } \sigma : \text{density/unit area.} \qquad (4)}$$

From Eqs. (3) and (4) we obtain

$$\boxed{\log_e F_0 = \log_e[c_0\{T_0 / \sigma\}^{1/2}] + (b/2)x. \qquad (5)}$$

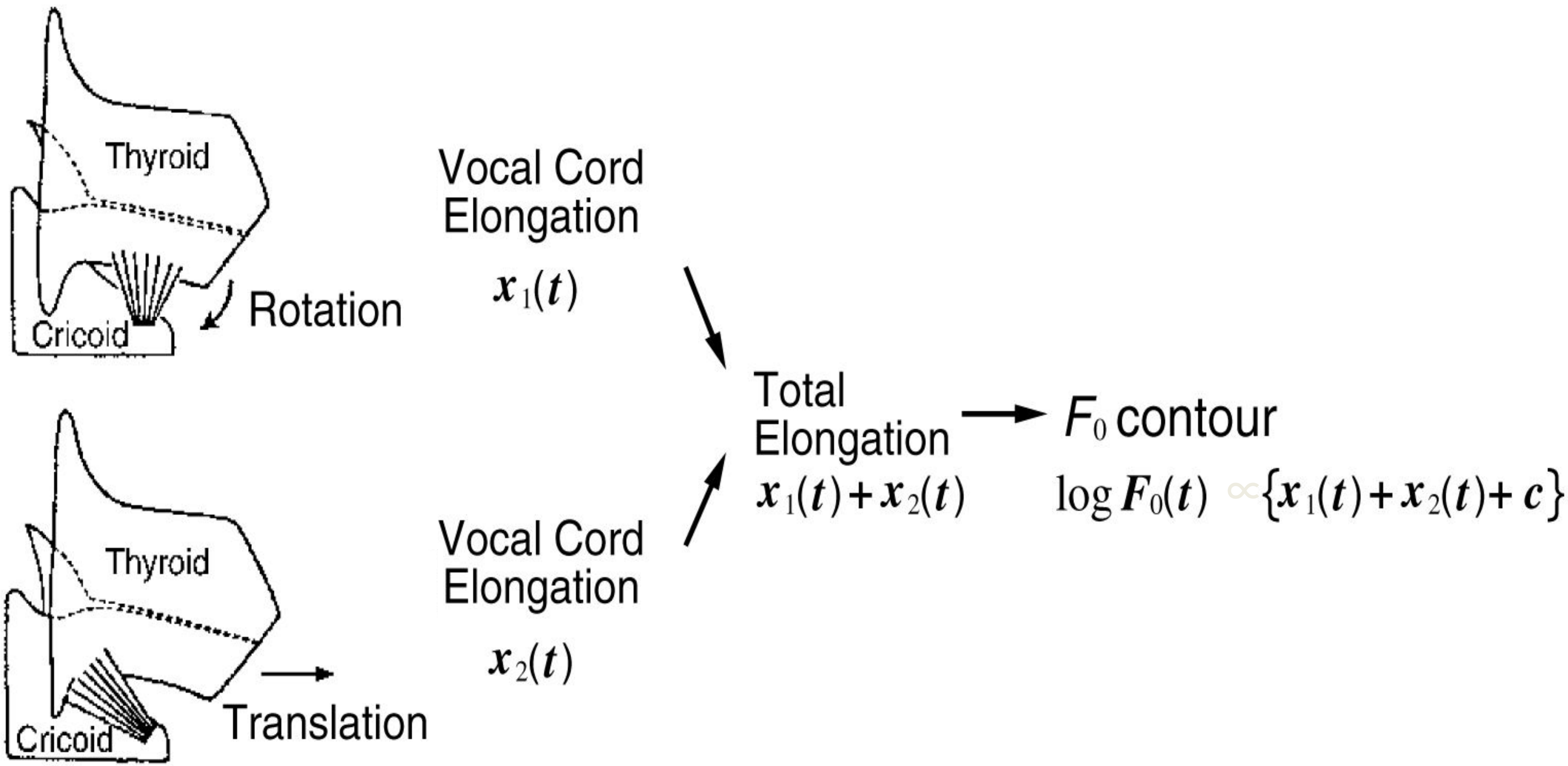
When x is time-varying, i.e., $x = x(t)$,

$$\boxed{\log_e F_0(t) = \log_e F_b + (b/2)\, x(t), \qquad (6)}$$

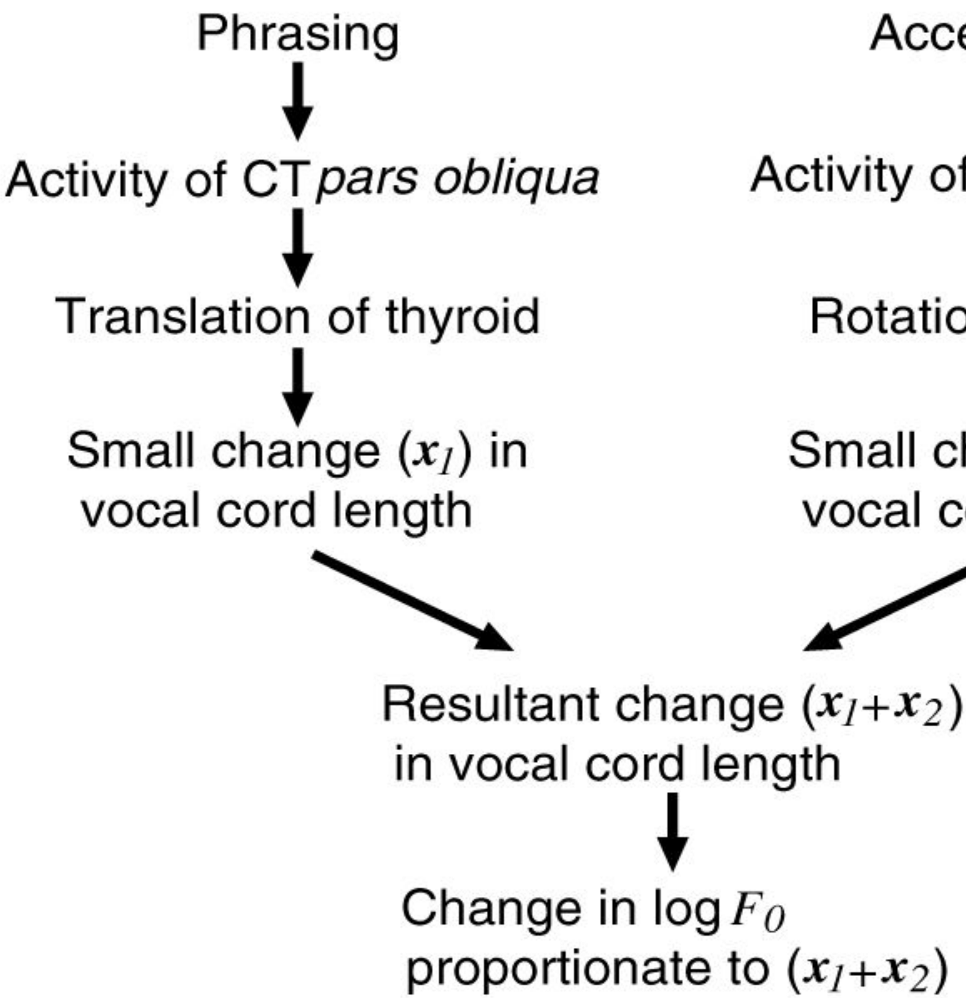
where $F_b = c_0\{T_0 / \sigma\}^{1/2}$.

Thus an F_0 contour, when plotted in the ***logarithmic* scale** as a function of time, can be expressed as the sum of **a constant (baseline) term** and **a time-varying term**, the latter being proportional to the elongation of the vocal cord.

Additivity of phrase and accent components

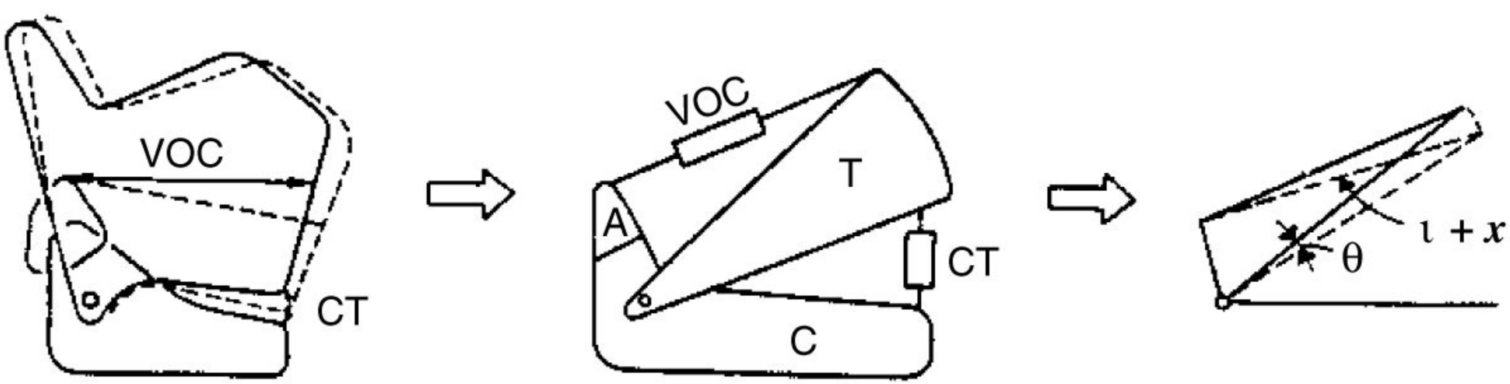


Additivity of components in logF_0 domain



The term "Superposition" applies only if one uses the logarithmic (i.e., semitone) scale for F_0.

Thyroid and Cricoid Cartilages with Vocalis and Cricothyroid Muscles Forming a Two-Mass, Two-Spring System

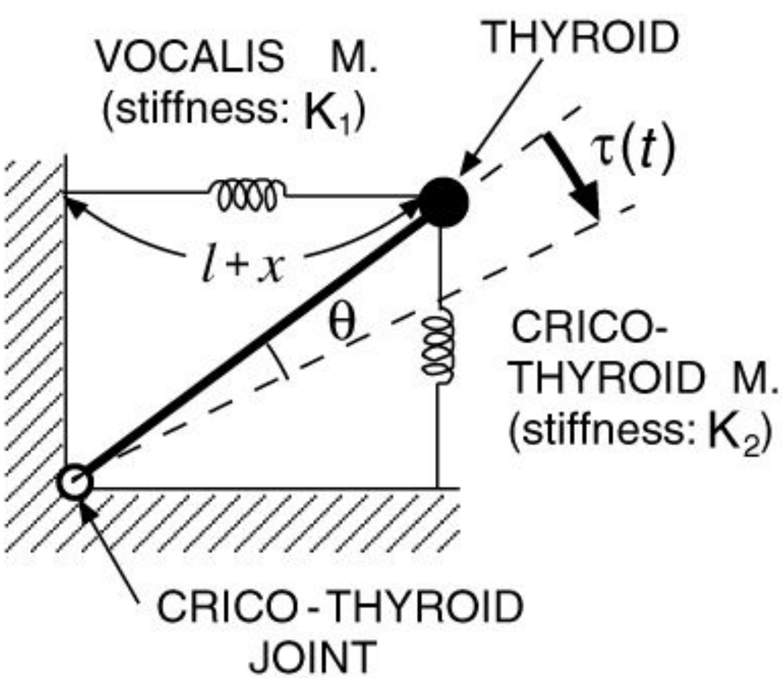


VOC : Vocalis M.
CT : Criothyroid M.

T : Thyroid
C: Cricoid
A: Arytenoid

ι : Length of vocalis
x : Elongation of vocalis
θ: Angular dispacement of thyroid

Rotation of thyroid around the crico-thyroid joint



θ : Angular displacement

x : Change in length of VOC

- Equation of Motion (Rotation)

$$I\frac{d^2\theta}{dt^2} + R\frac{d\theta}{dt} + K\theta = \tau(t),$$

where $\tau(t)$: Torque generated by contraction of CT

thus $\theta(t) = C_3 G_a(t).$

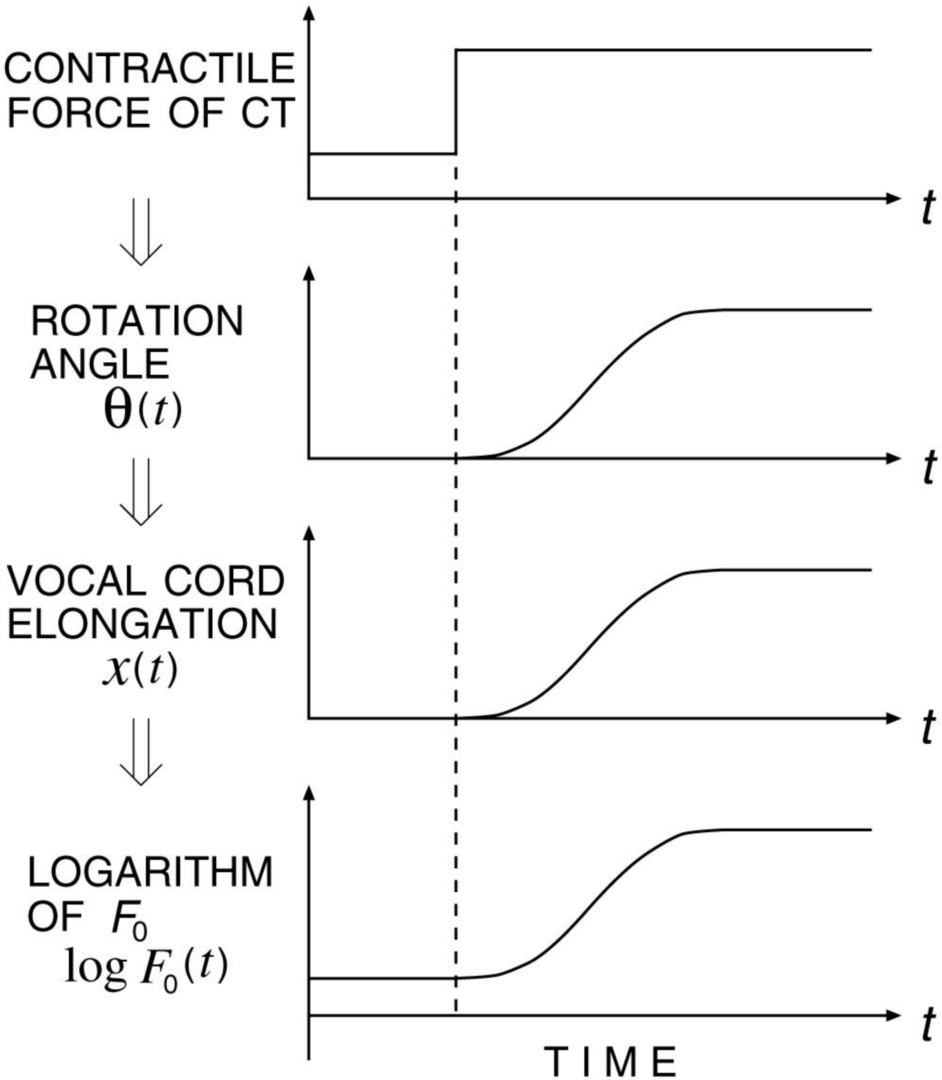
- For small θ ,

$$x(t) = C_4\theta(t) = C_5 G_a(t)$$

- Hence

$$\log_e F_0(t) = C_6 G_a(t) + C$$

From cricothyroid activity to logarithm of F_0



The rate of F_0 change is determined, not by the speed of contraction or relaxation of the muscle, but by the mechanical properties of the laryngeal structure.

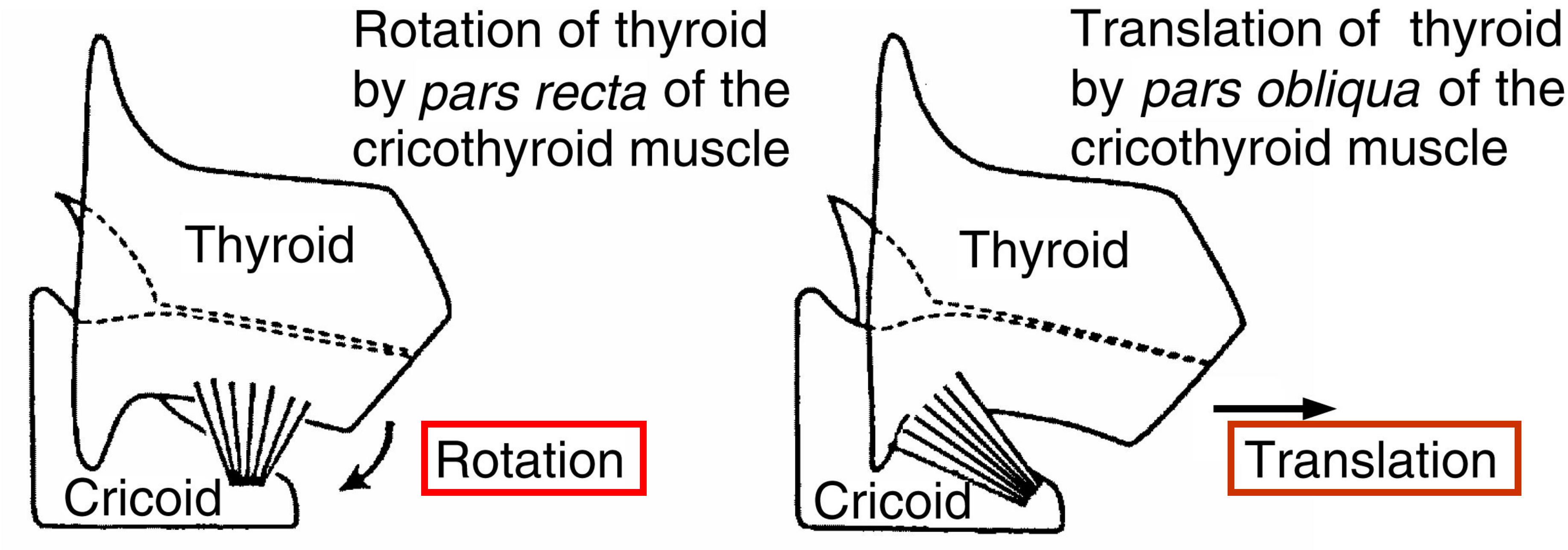
The rate of change varies with the amplitude of the command, but the time constant remains the same.
(Fujisaki, 1981)

The role of the cricothyroid (CT) muscle

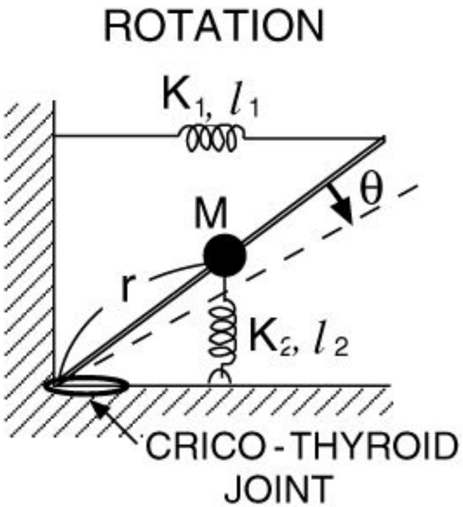
Analysis of the laryngeal structure suggests that the movement of the thyroid cartilage has two degrees of freedom [e.g., Zemlin 1968, Fink & Demarest 1978].

One is **rotation** around the cricothyroid joint due to the activities of the *pars recta* of the cricothyroid muscle (henceforth CT) and the other is **horizontal translation** due to the activities of *pars obliqua* of CT.

Motion of thyroid with two degrees of freedom

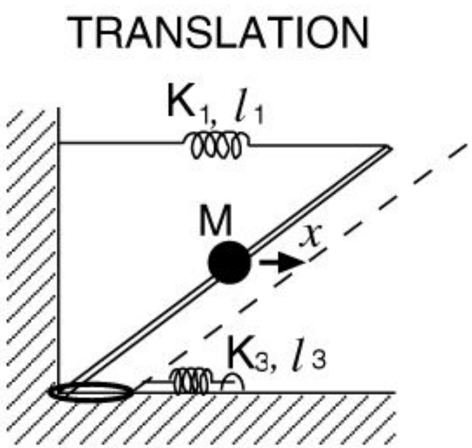


Rotation and translation of the thyroid



$$Mr^2\frac{d^2\theta}{dt^2} + R\,\frac{d\theta}{dt} + K\theta = \tau(t)$$

$\tau(t)$ **:** Torque generated by contraction of **CT *pars recta***

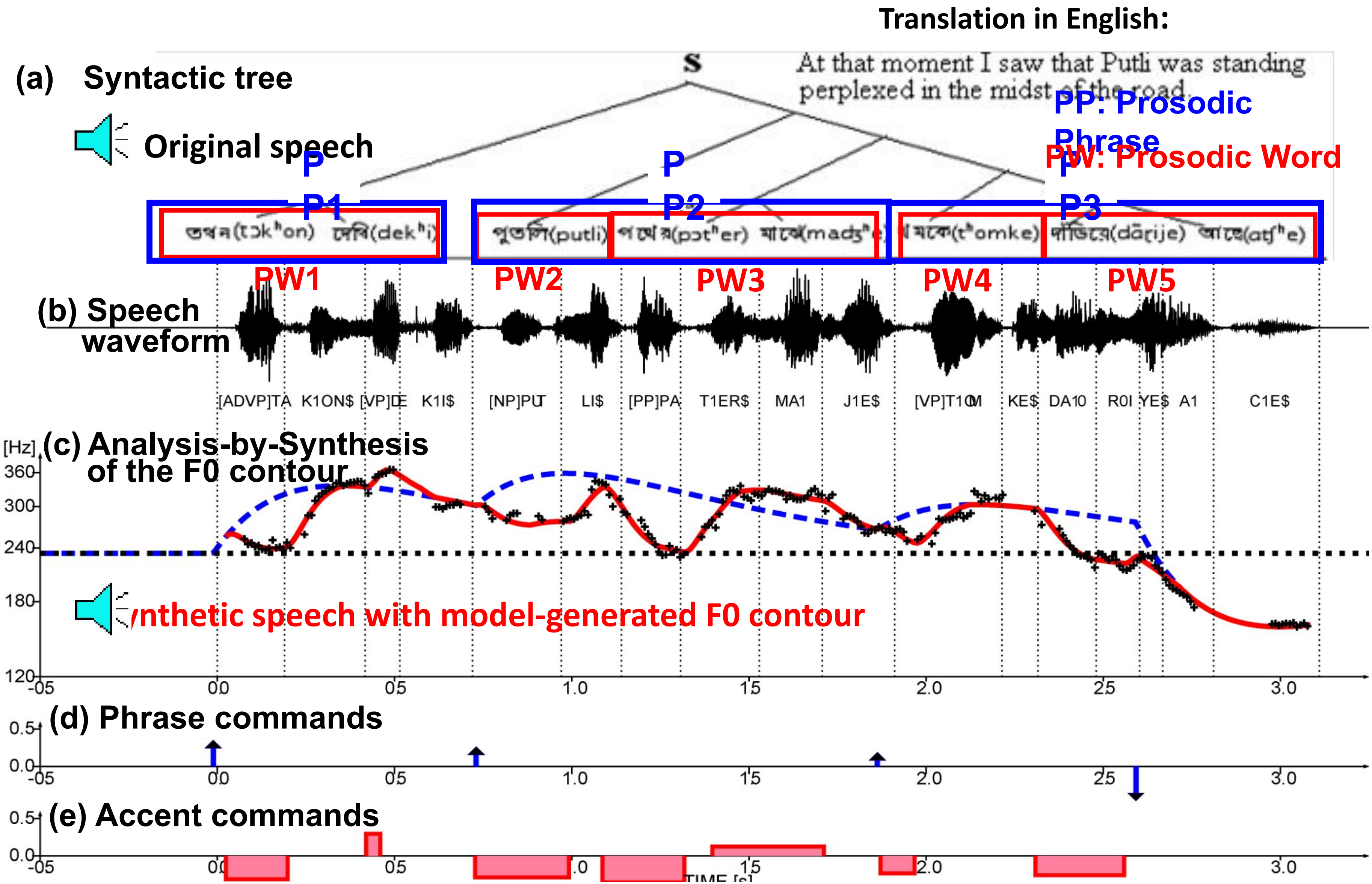


$$M\frac{d^2x}{dt^2} + R'\frac{dx}{dt} + K'x = f(t)$$

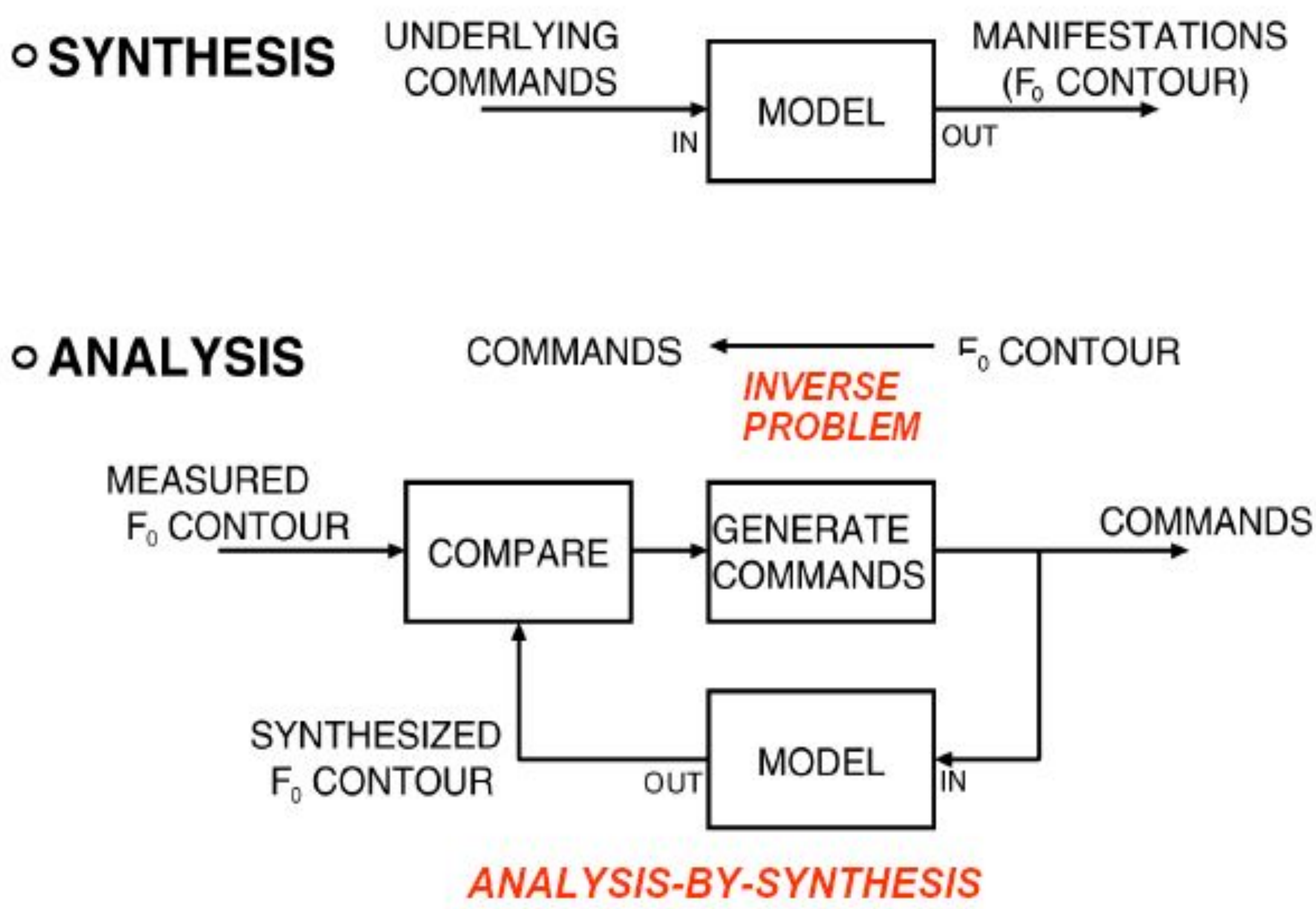
$f(t)$ **:** Force generated by contraction of **CT *pars obliqua***

Analysis and Synthesis of F_0 Contours

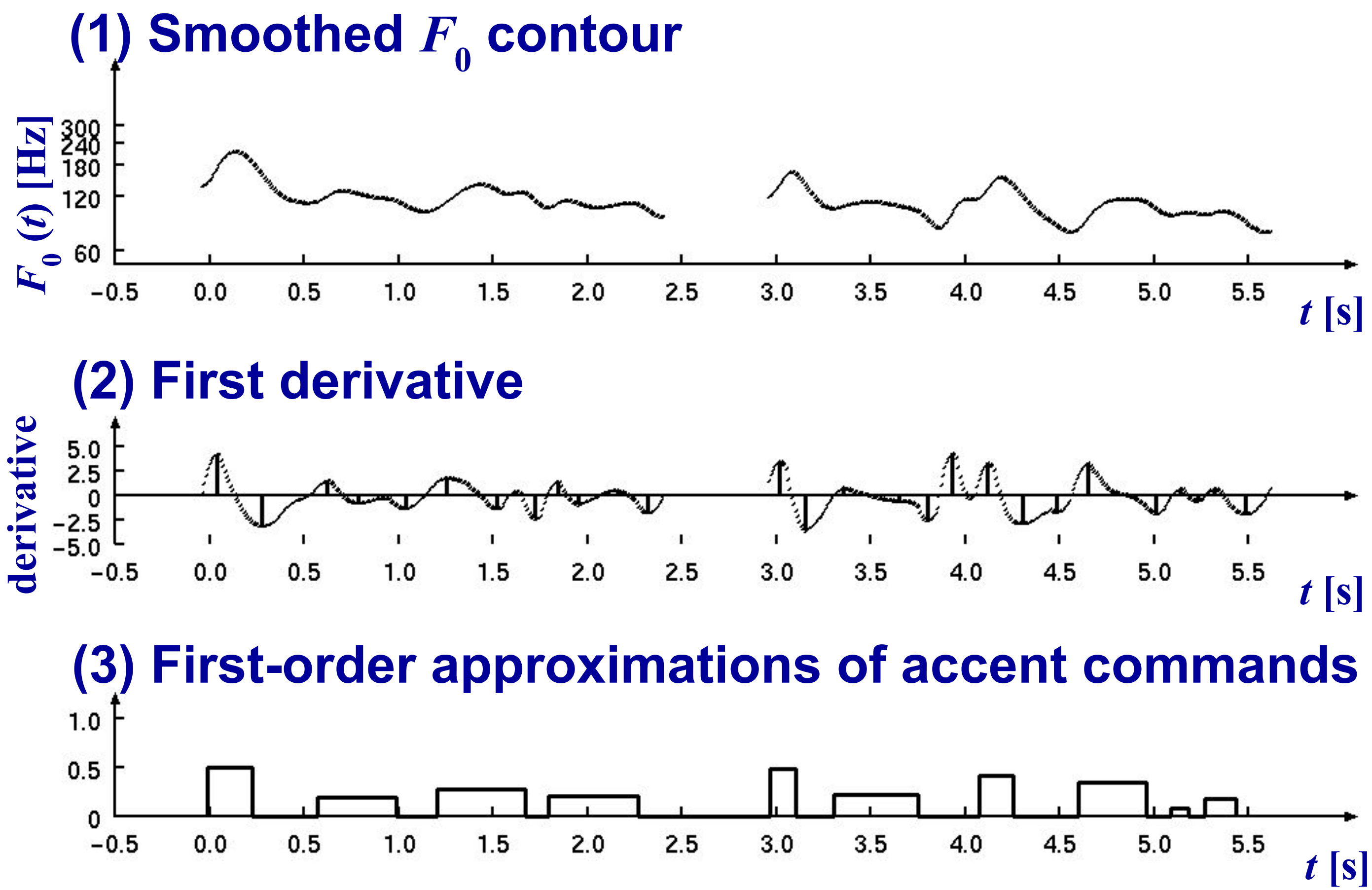
Analysis-by-Synthesis of the F_0 contour of an Utterance of a Bangla sentence



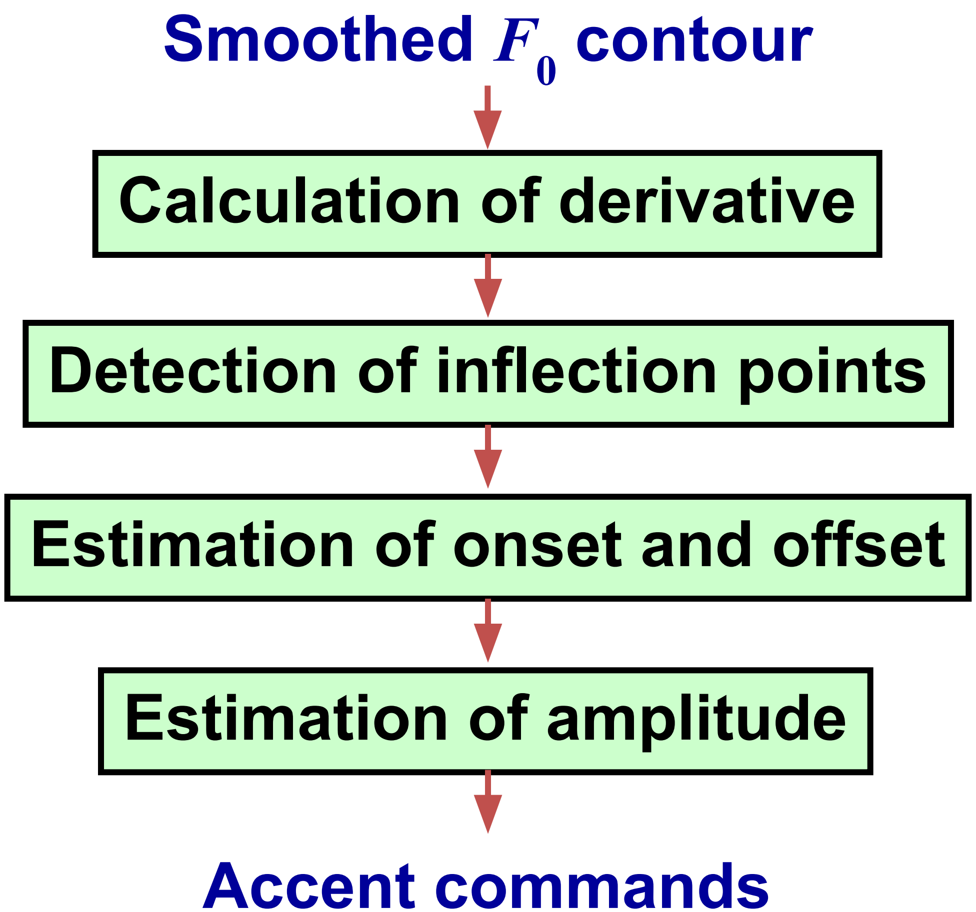
Use of a Generative Model in Synthesis and Analysis



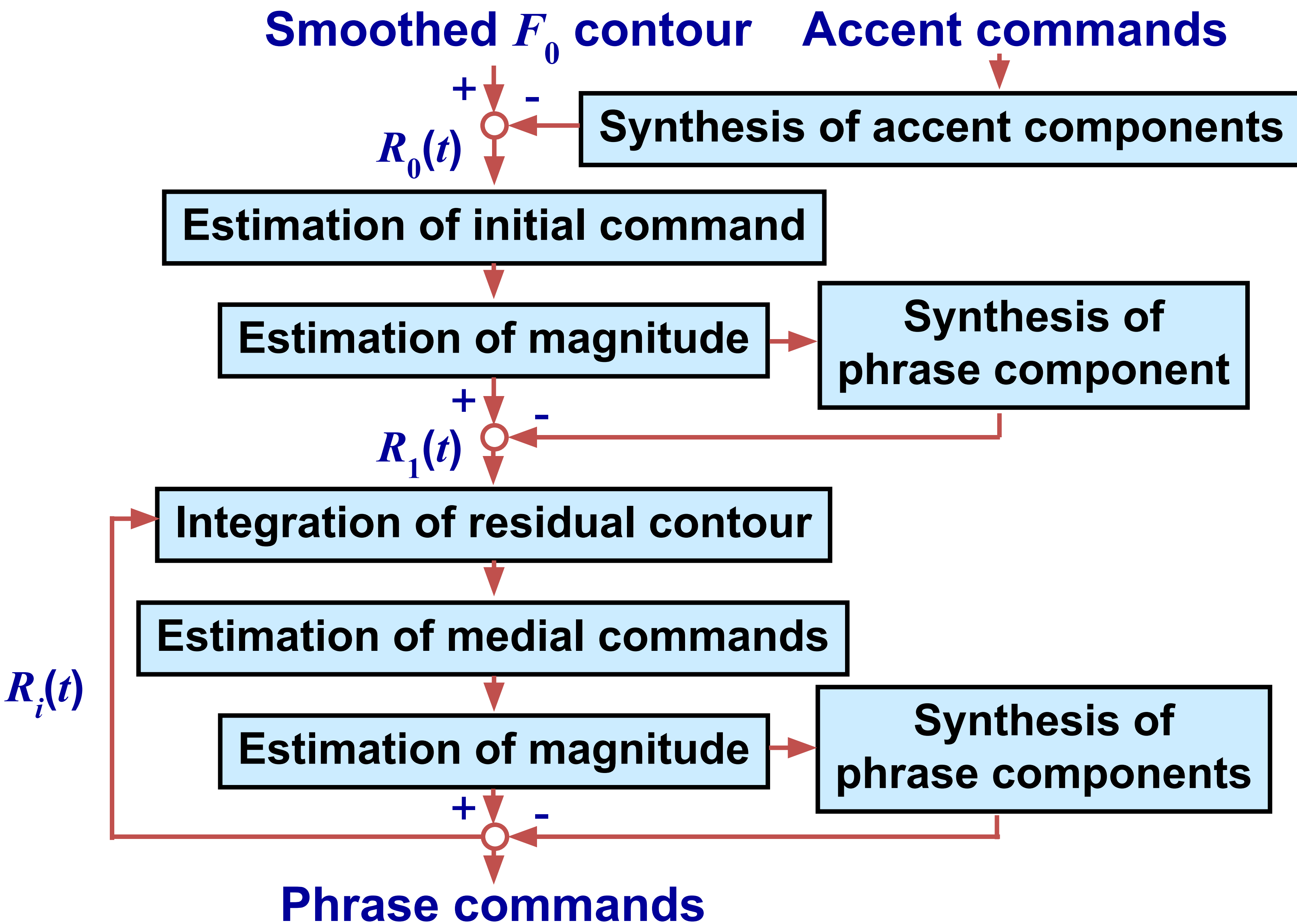
An example of estimation of first-order approximations of accent commands



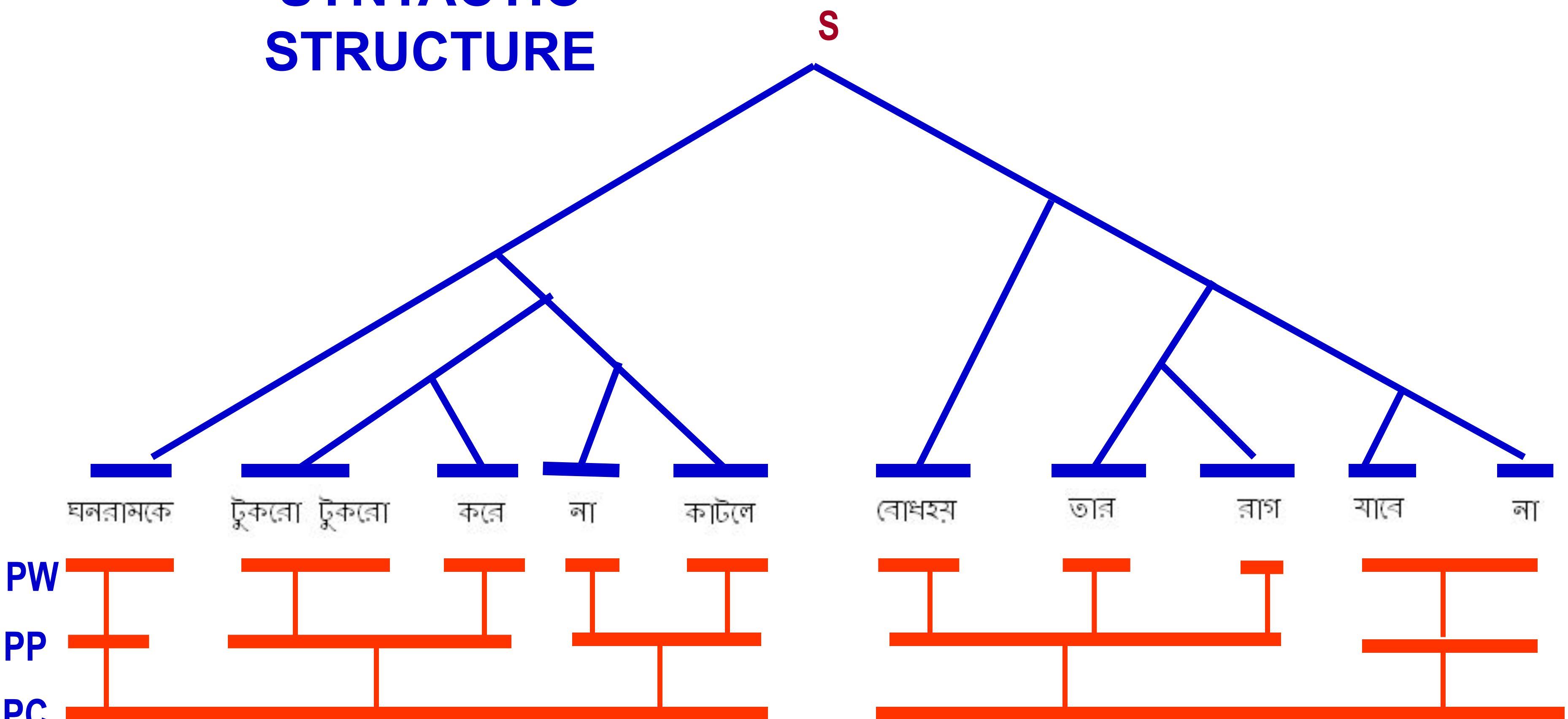
Estimation of first-order approximations of accent commands



Estimation of first-order approximations of phrase



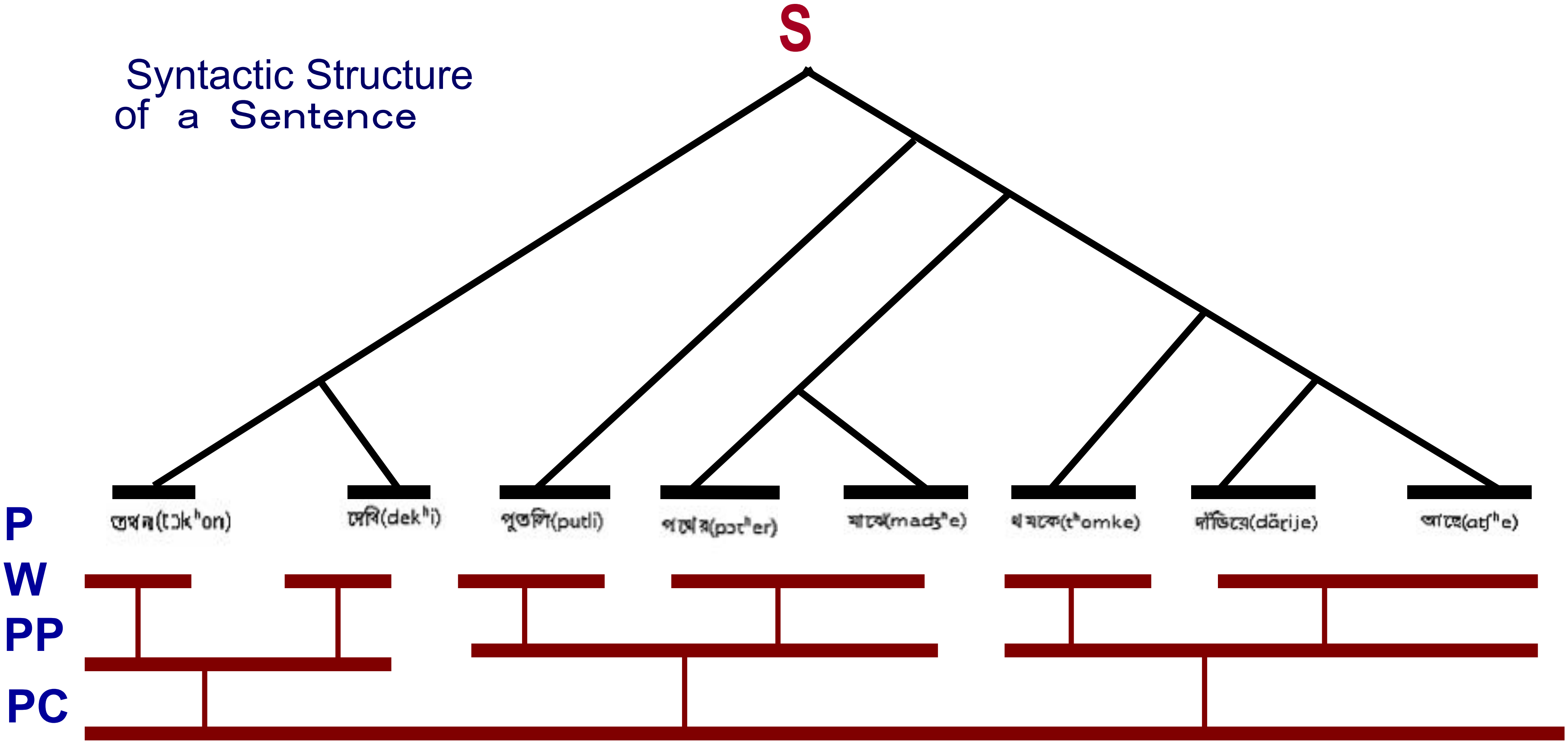
SYNTACTIC STRUCTURE



PROSODIC STRUCTURE

PW: Prosodic Word
PP : Prosodic Phrase
PC : Prosodic Clause

Syntactic Structure vs. Prosodic Structure (Fujisaki 1984)



Prosodic Structure
of an Utterance

PW: Prosodic Word
PP : Prosodic Phrase
PC : Prosodic Clause

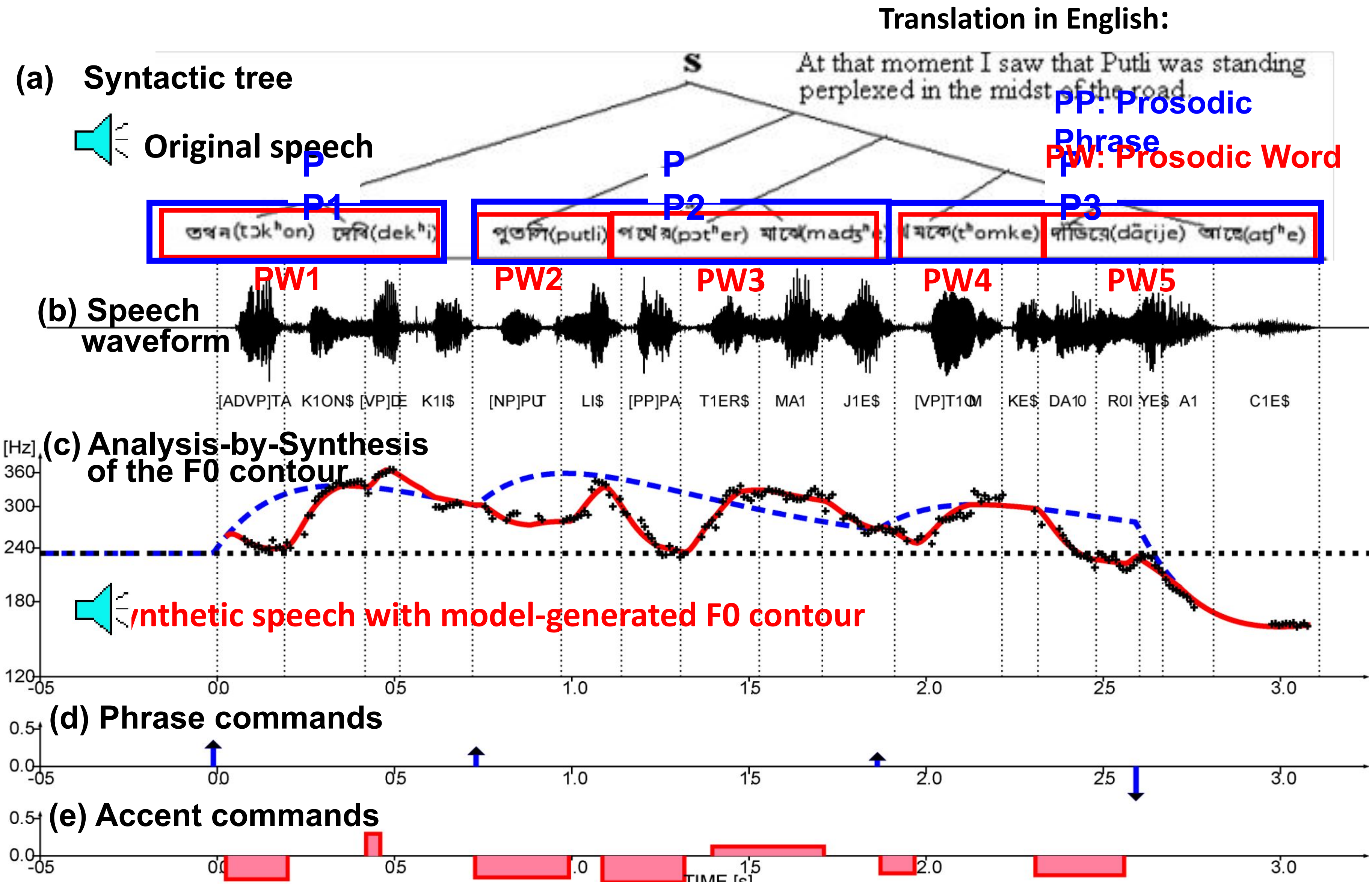
Relationship between F_0 model parameters and prosodic units

- Prosodic Phrase
  - The constituents of a prosodic phrase in Bangla can be one or more than one syntactic phrase depending on the length of the phrase
  - Prosodic phrase can be defined on the basis of the phrase command
  - The placement of phrase commands is related to the syntactic structure, but they occur mostly at deeper syntactic boundaries.
  - The magnitudes of the consecutive phrase commands within an utterance have a general tendency of decreasing
  - The initiation of the phrase command always leads the segmental onset of the corresponding prosodic phrase by a few hundred milliseconds.

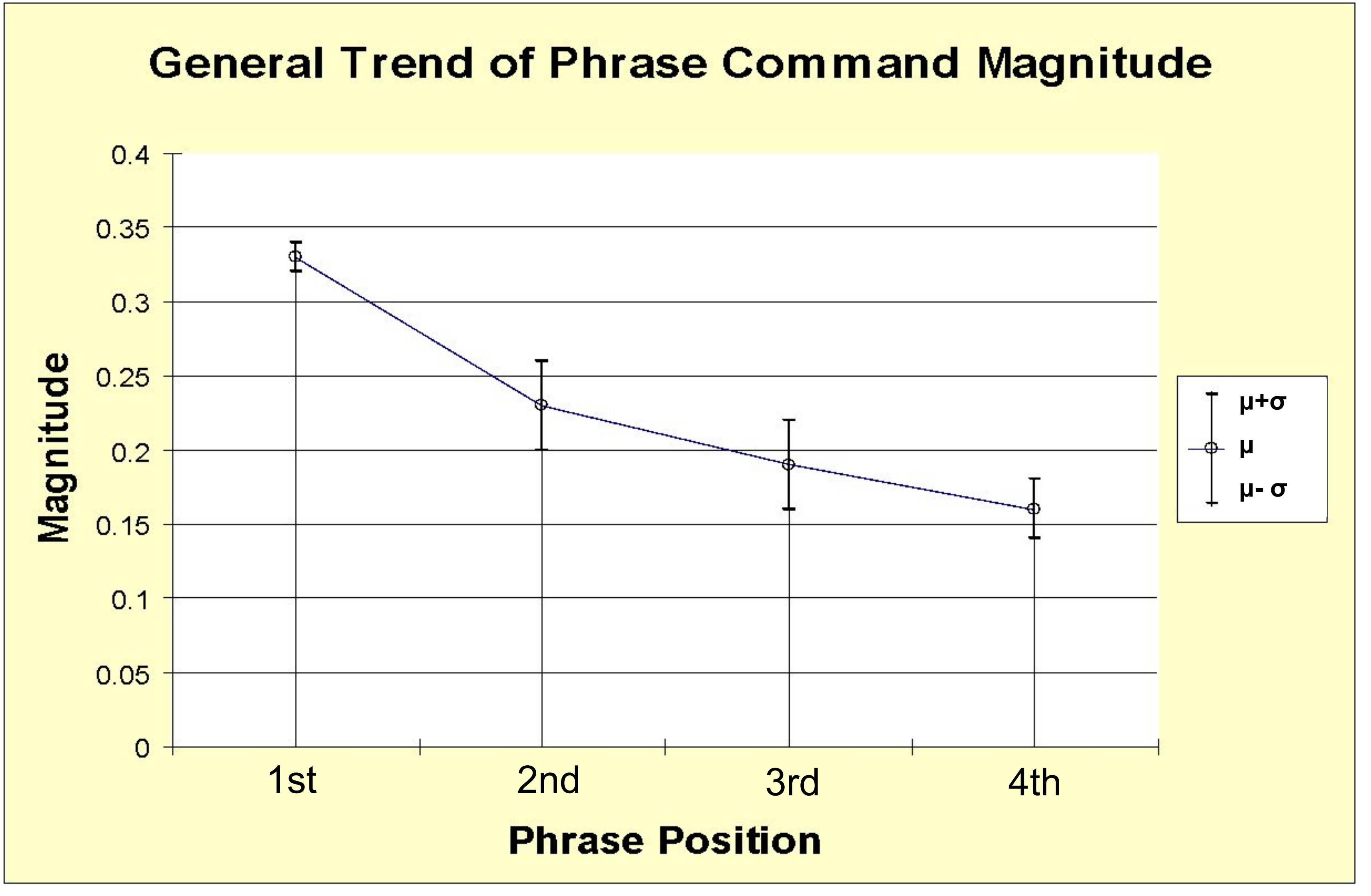
Relationship between F_0 model parameters and prosodic units (cont'd)

- Prosodic Word
  - The syntactic constituents of a prosodic word may be a sequence of one or more content words followed by one or more function words
  - Prosodic word of Bangla can be defined on the basis of the accent command
  - Most of the prosodic words in Bangla can be modeled using one negative accent command at the beginning
  - Prosodic words (polysyllabic) that are emphasized or that contain suffix can be modeled using one negative accent command followed by one positive accent command

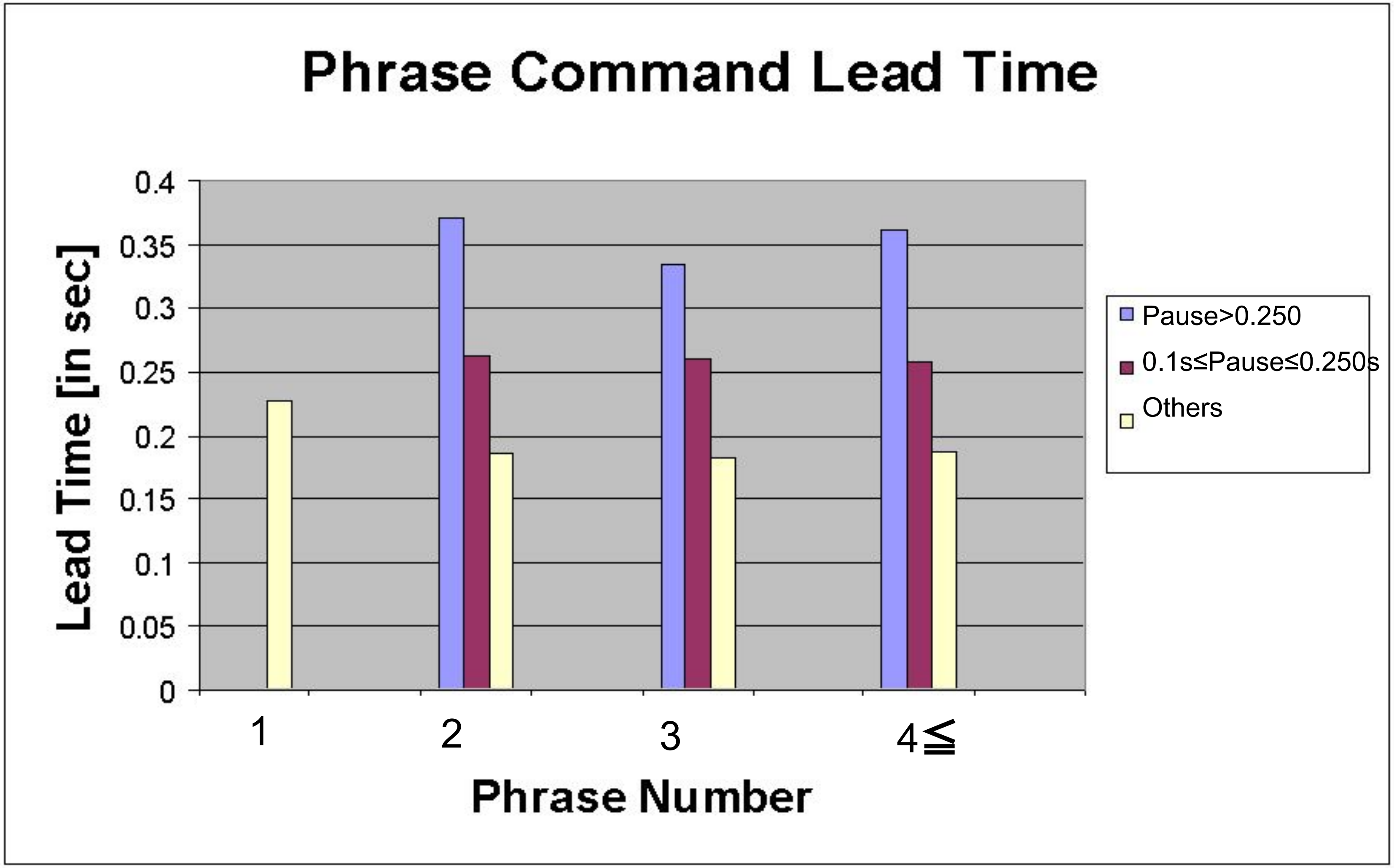
Analysis-by-Synthesis of the F_0 contour of an Utterance of a Bangla sentence



Results of Analysis of Phrase Commands



Results of Analysis of Phrase Commands (cont'd)



Results of Analysis of Accent Commands

<table>
<tr><th rowspan="2" colspan="2">Position</th><th colspan="2">Accent command amplitude (A_{aj})</th></tr>
<tr><th>Negative</th><th>Positive</th></tr>
<tr><td colspan="2">Utterance-initial</td><td>- 0.325
(σ = 0.06)</td><td rowspan="3">0.232
(σ = 0.05)</td></tr>
<tr><td rowspan="2">Utterance-medial</td><td>Phrase-initial</td><td>- 0.317
(σ = 0.05)</td></tr>
<tr><td>Phrase-medial</td><td>- 0.261
(σ = 0.08)</td></tr>
<tr><td colspan="2">Utterance-final</td><td>- 0.328
(σ = 0.09)</td><td>-</td></tr>
</table>

Results of Analysis of Accent Commands (cont'd)

<table>
<tr><th>Position</th><th colspan="3">Value of t_1 [s]</th></tr>
<tr><td>Utterance-initial</td><td colspan="3">0.148</td></tr>
<tr><td rowspan="3">Utterance-medial</td><td>Pause ≥0.10 s</td><td colspan="2">0.147</td></tr>
<tr><td rowspan="2">Pause<0.10 s</td><td>Voiced</td><td>0.071</td></tr>
<tr><td>Unvoiced</td><td>0.111</td></tr>
</table>

<table>
<tr><th>Word Length</th><th colspan="2">Value of t_2 [s]</th></tr>
<tr><td>Monosyllabic</td><td colspan="2">0.176</td></tr>
<tr><td rowspan="2">Polysyllabic</td><td>Syllable Type: 'V','VC','CV'</td><td>0.037</td></tr>
<tr><td>Syllable Type: 'others'</td><td>0.071</td></tr>
</table>

Evaluation Method

- Objective evaluation
  - The original F_0 contour is compared with the synthesized F_0 contour and the root mean squared value of the difference is 0.041.
- Subjective evaluation
  - The two sets (I.e., original and synthesized) of 50 sentences were randomly mixed.
  - These were presented to 5 subjects for giving their judgment on the naturalness on a 5-point scale

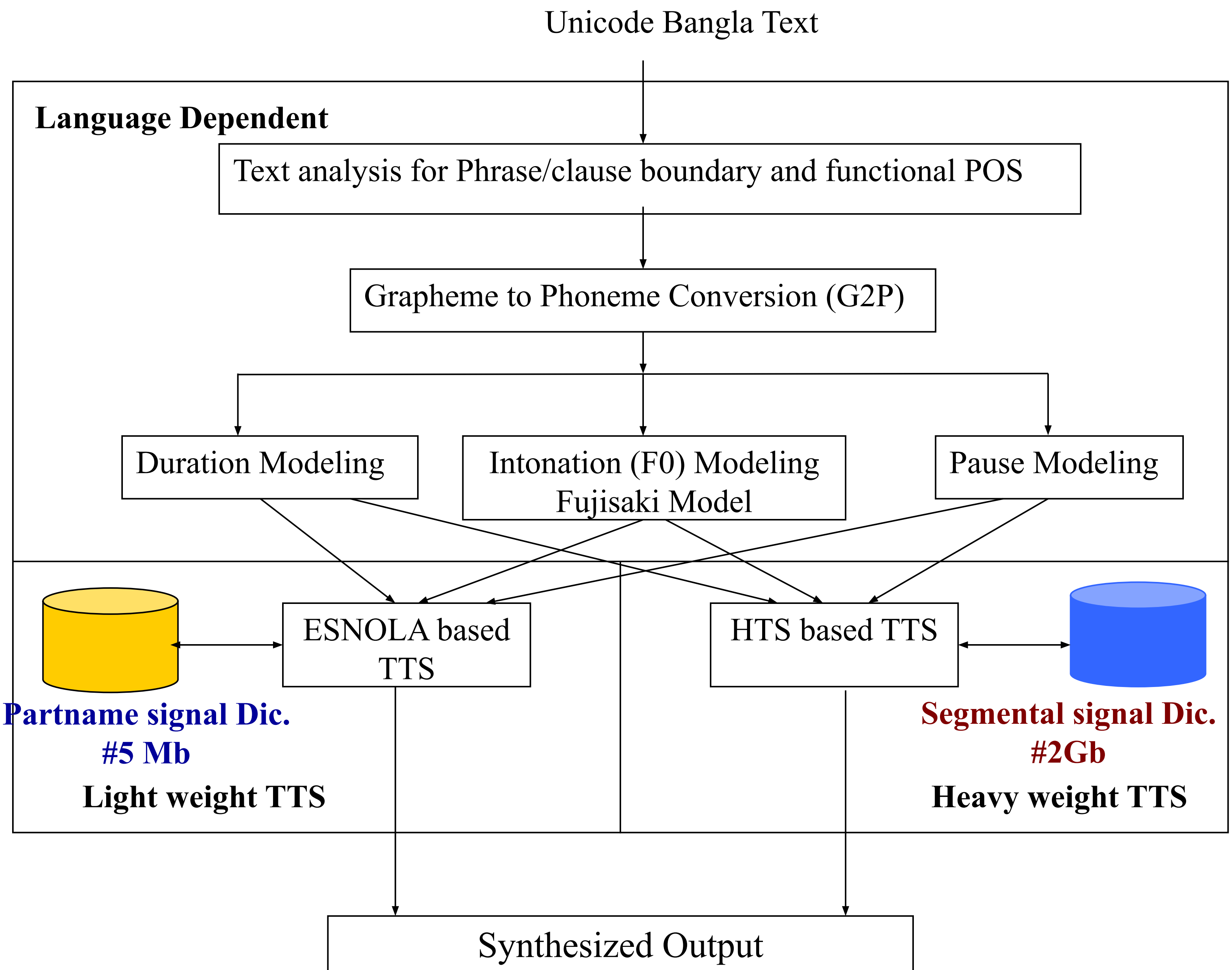
    (5: very good , 4: good, 3: neutral, 2: poor,

Evaluation Results

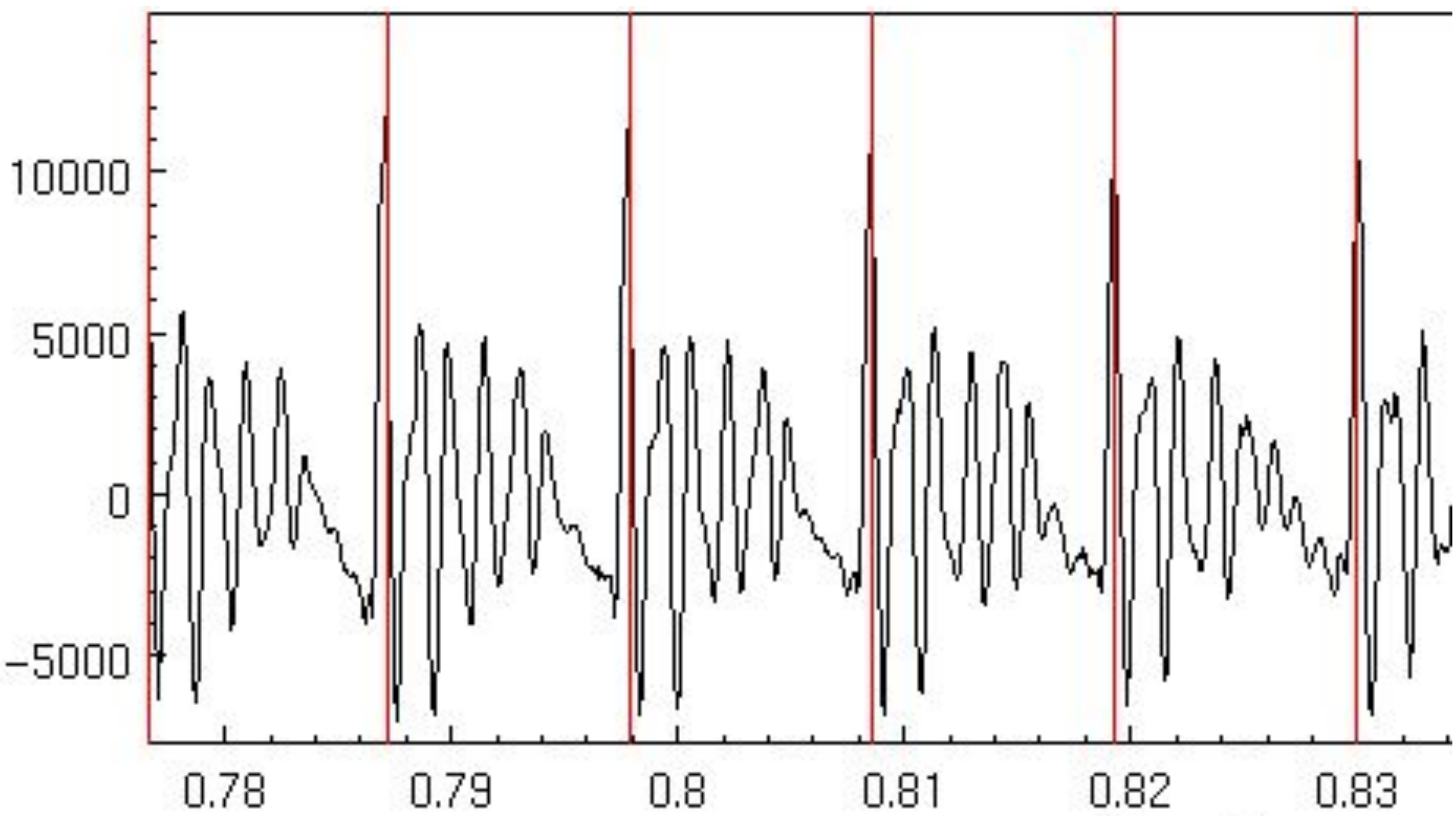
Mean opinion scores for the original and the synthesized speech

Score		Subject				
		L1	L2	L3	L4	L5
Original	Avg	4.64	4.52	4.58	4.70	4.62
	Stdv	0.56	0.58	0.67	0.54	0.49
Model-generated	Avg	4.52	4.44	4.52	4.64	4.56
	Stdv	0.71	0.68	0.71	0.62	0.67

Results of ANOVA reveals that
differences within each speaker and across the speakers are not significant at all

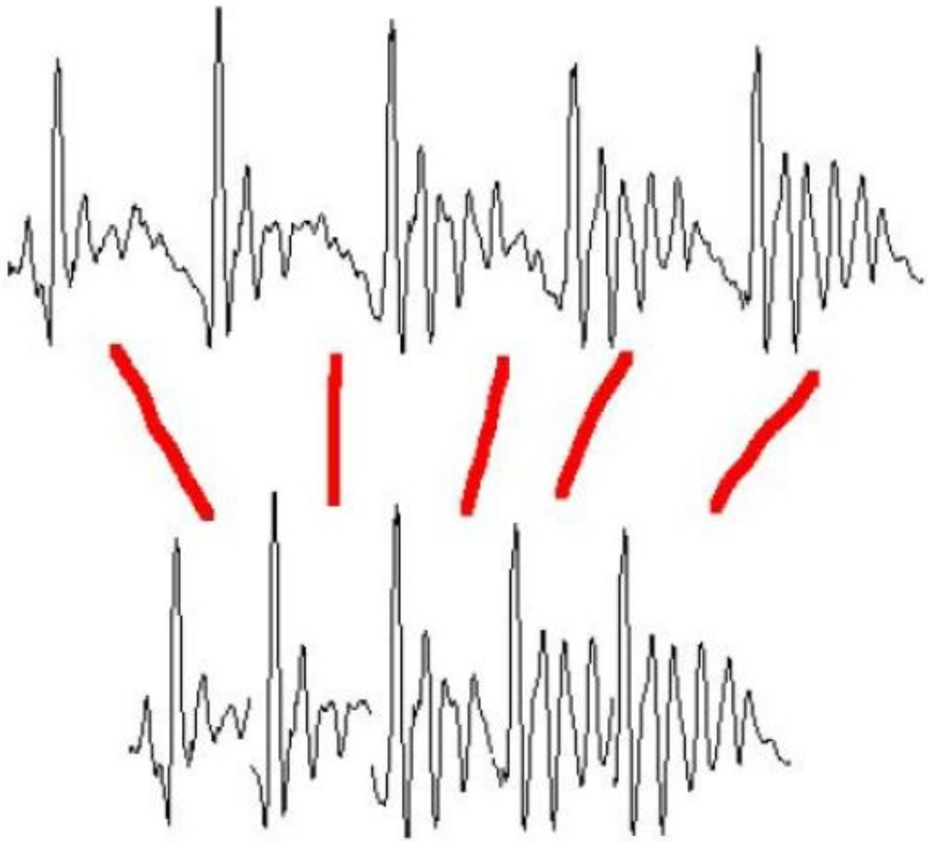

Unicode Bangla Text
Language Dependent
Text analysis for Phrase/clause boundary and functional POS
Grapheme to Phoneme Conversion (G2P)
Duration Modeling
Intonation (F0) Modeling
Fujisaki Model
Pause Modeling
ESNOLA based TTS
HTS based TTS
Partname signal Dic.
#5 Mb
Light weight TTS
Segmental signal Dic.
#2Gb
Heavy weight TTS
Synthesized Output

Speech as Short Term signals

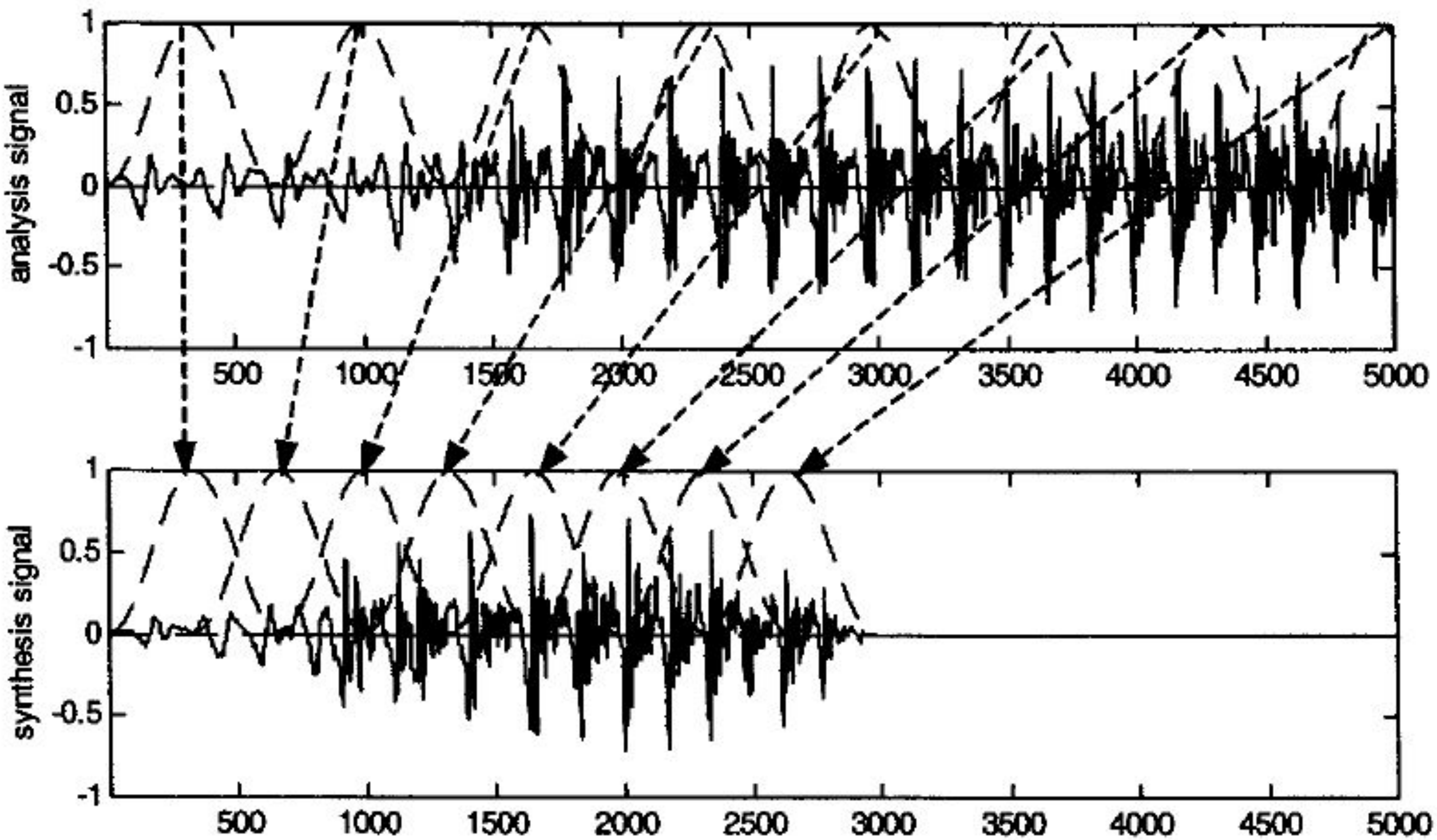


Pitch Modification

- Move short-term signals closer together/further apart



Overlap-and-add (OLA)



Overlap and Add (OLA)

- Hanning windows of length 2N used to multiply the analysis signal
- Resulting windowed signals are added
- Analysis windows, spaced 2N
- Synthesis windows, spaced N
- Time compression is uniform with factor of 2
- Pitch periodicity somewhat lost around 4th window

Huang, Acero, and Hon

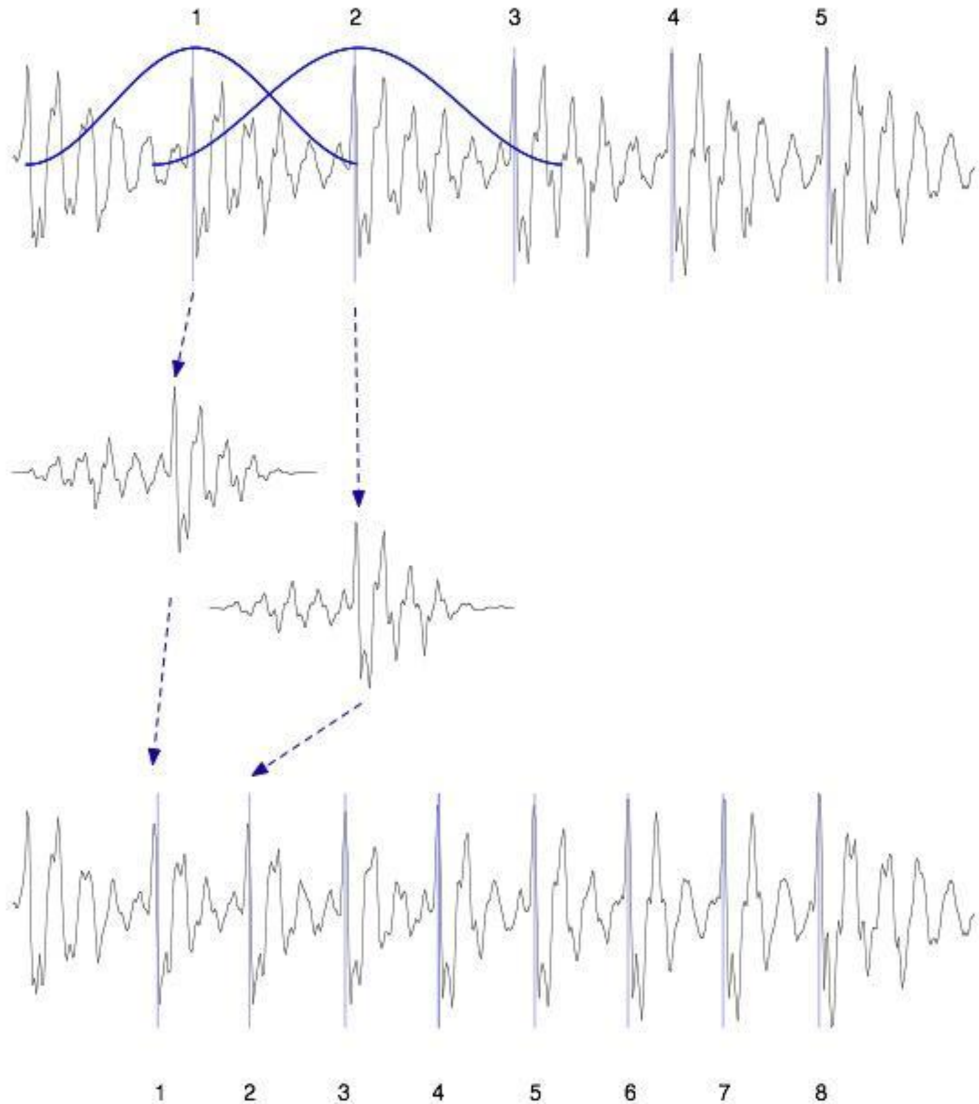
TD-PSOLA ™

- Time-Domain Pitch Synchronous Overlap and Add
- Patented by France Telecom (CNET)
- Very efficient
  - No FFT (or inverse FFT) required
- Can modify Hz up to two times or by half

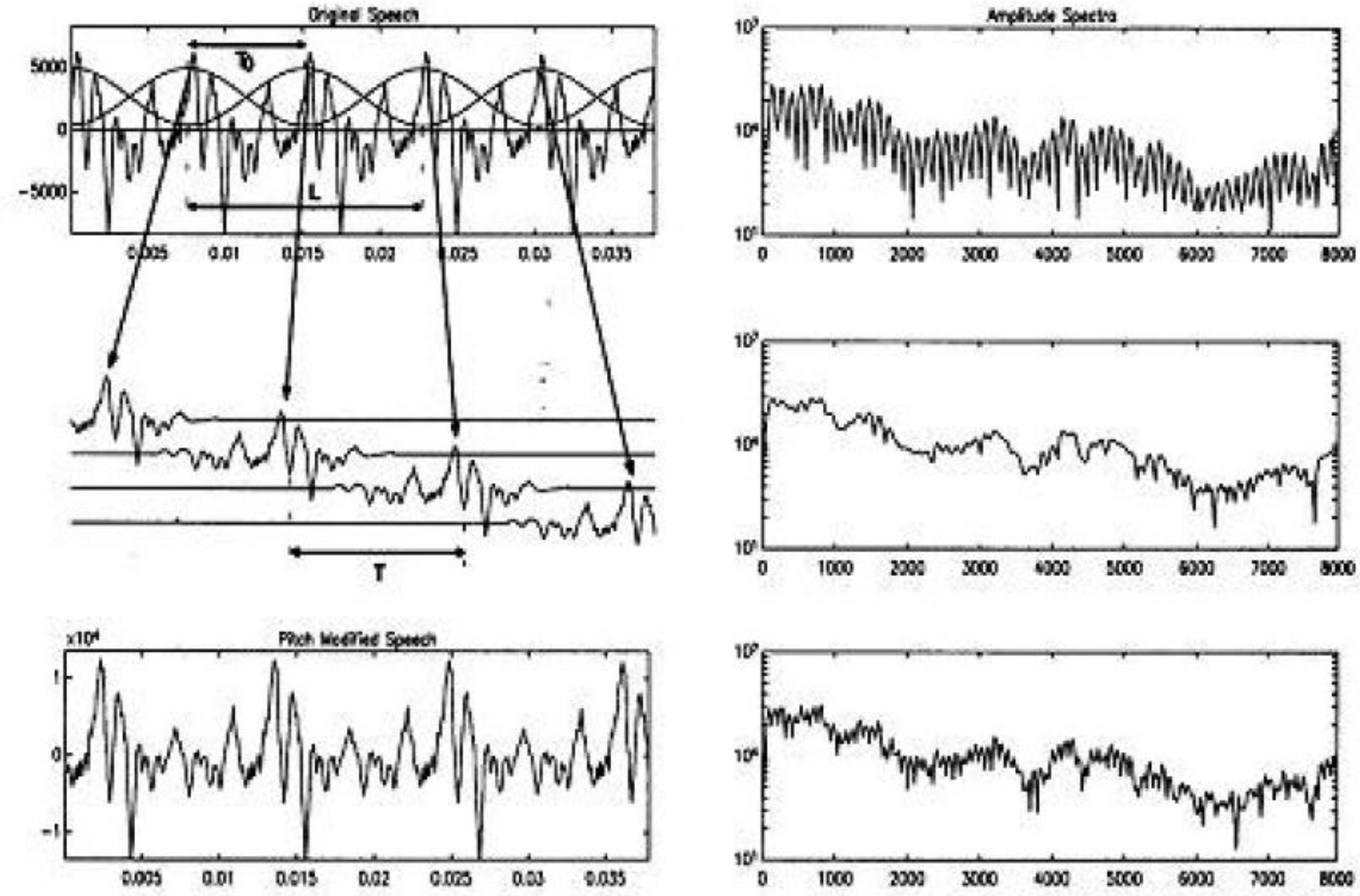
Slide from Richard

TD-PSOLA ™

- Windowed
- Pitch-synchronous
- Overlap-
- -and-add



TD-PSOLA ™


Original Speech
T0
L
T
Pitch Modified Speech
Amplitude Spectra

HMM synthesis

- A quite new technology is speech synthesis based on HMM, a mathematical concept called Hidden Markov models.
- It is a statistical method where the text-to-speech system is based on a model that is not known beforehand but it is refined by continuous training.
- The technique consumes large CPU resources but very little memory.
- This approach seems to give a better prosody, without glitches, and still produces very natural sounding, human-like speech