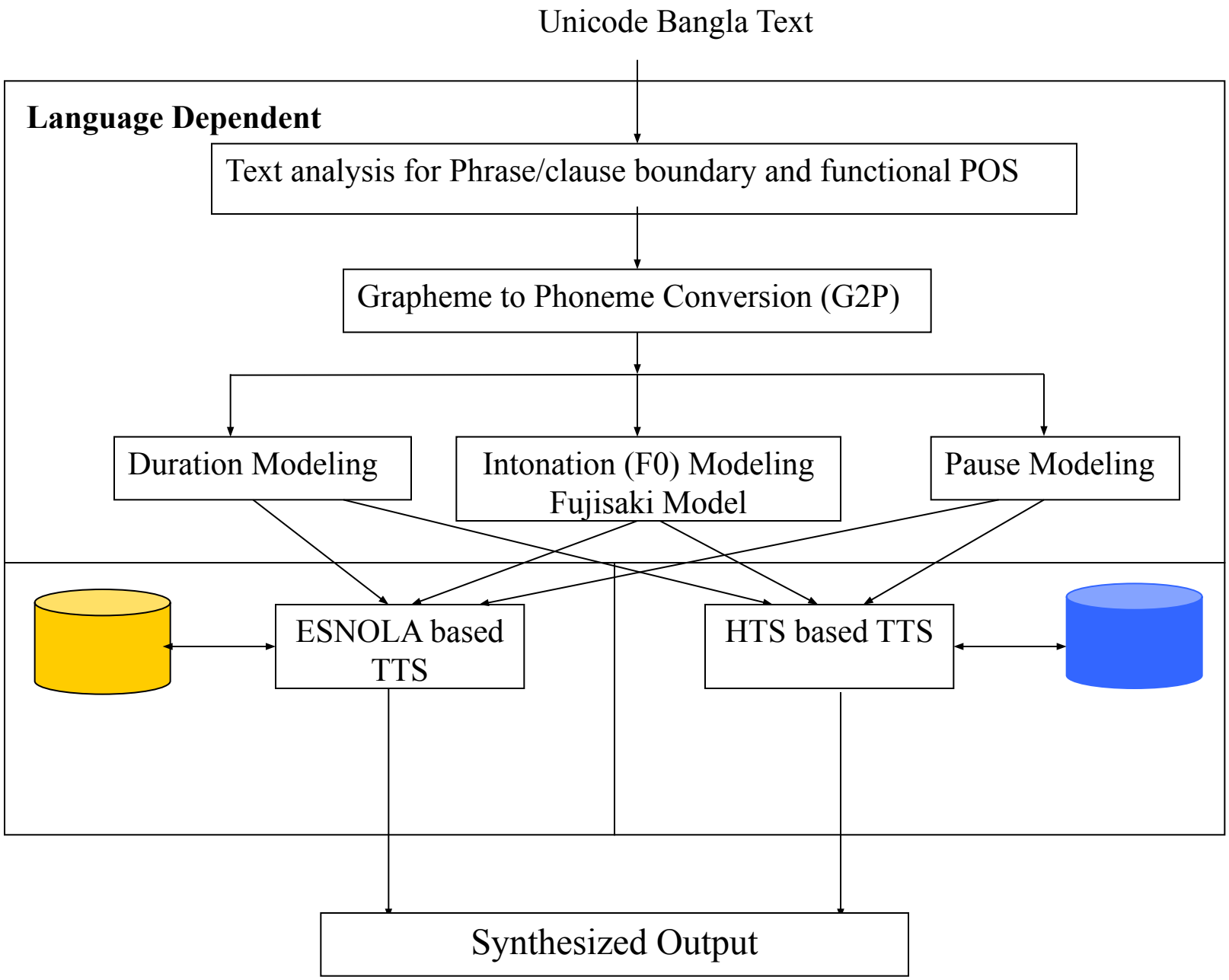

Unicode Bangla Text
Language Dependent
Text analysis for Phrase/clause boundary and functional POS
Grapheme to Phoneme Conversion (G2P)
Duration Modeling
Intonation (F0) Modeling
Fujisaki Model
Pause Modeling
ESNOLA based
TTS
HTS based TTS
Synthesized Output

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

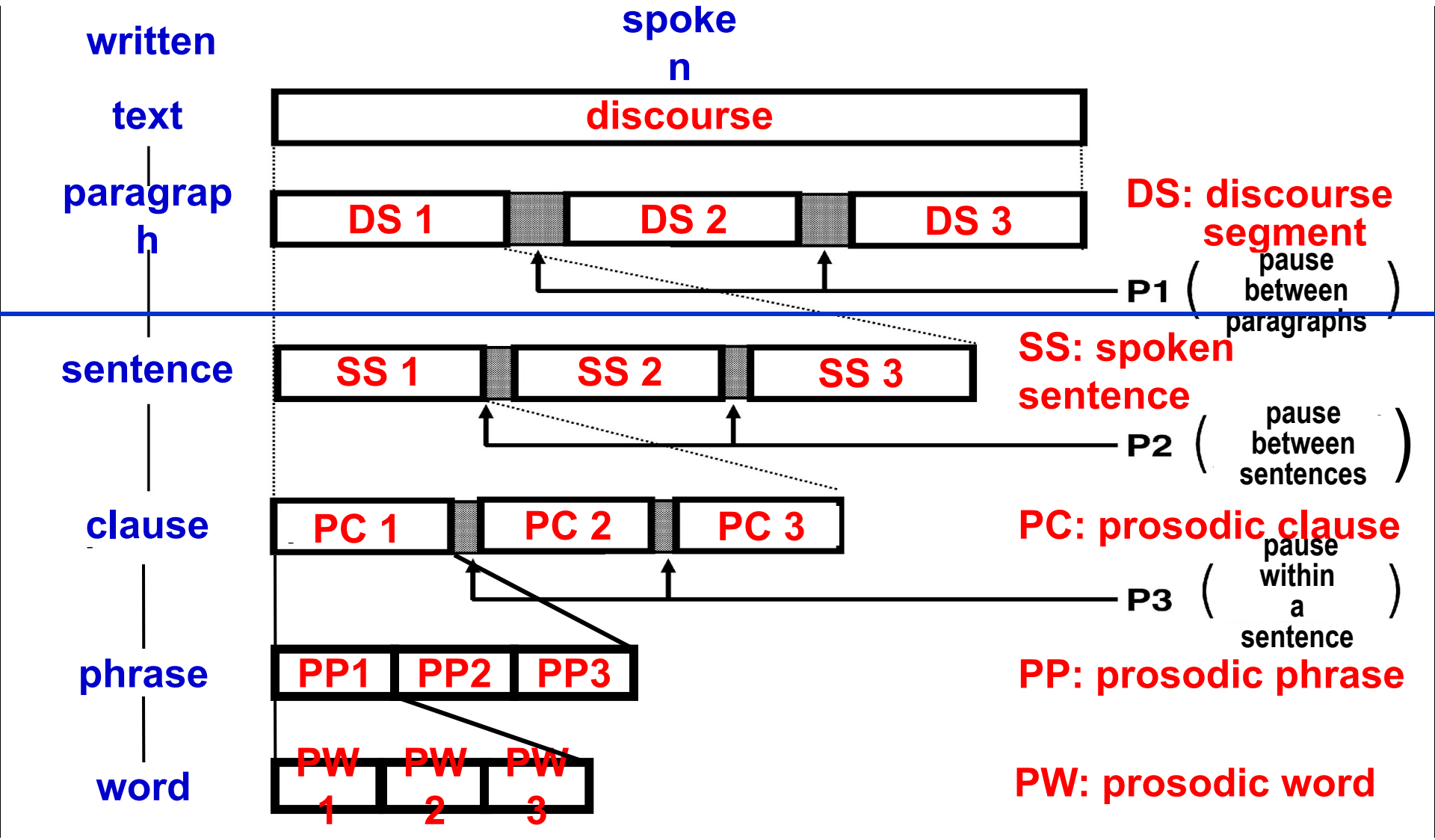
Prosody is defined as the systematic organization of individual linguistic units into an utterance, or a coherent group of utterances, in the process of speech production.

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.

Speech Parameters controlling the speech prosody

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

Prosodic Structure



Factor affecting the Occurrence and Duration of Sentence Medial Pause

a. Type of the Phrase

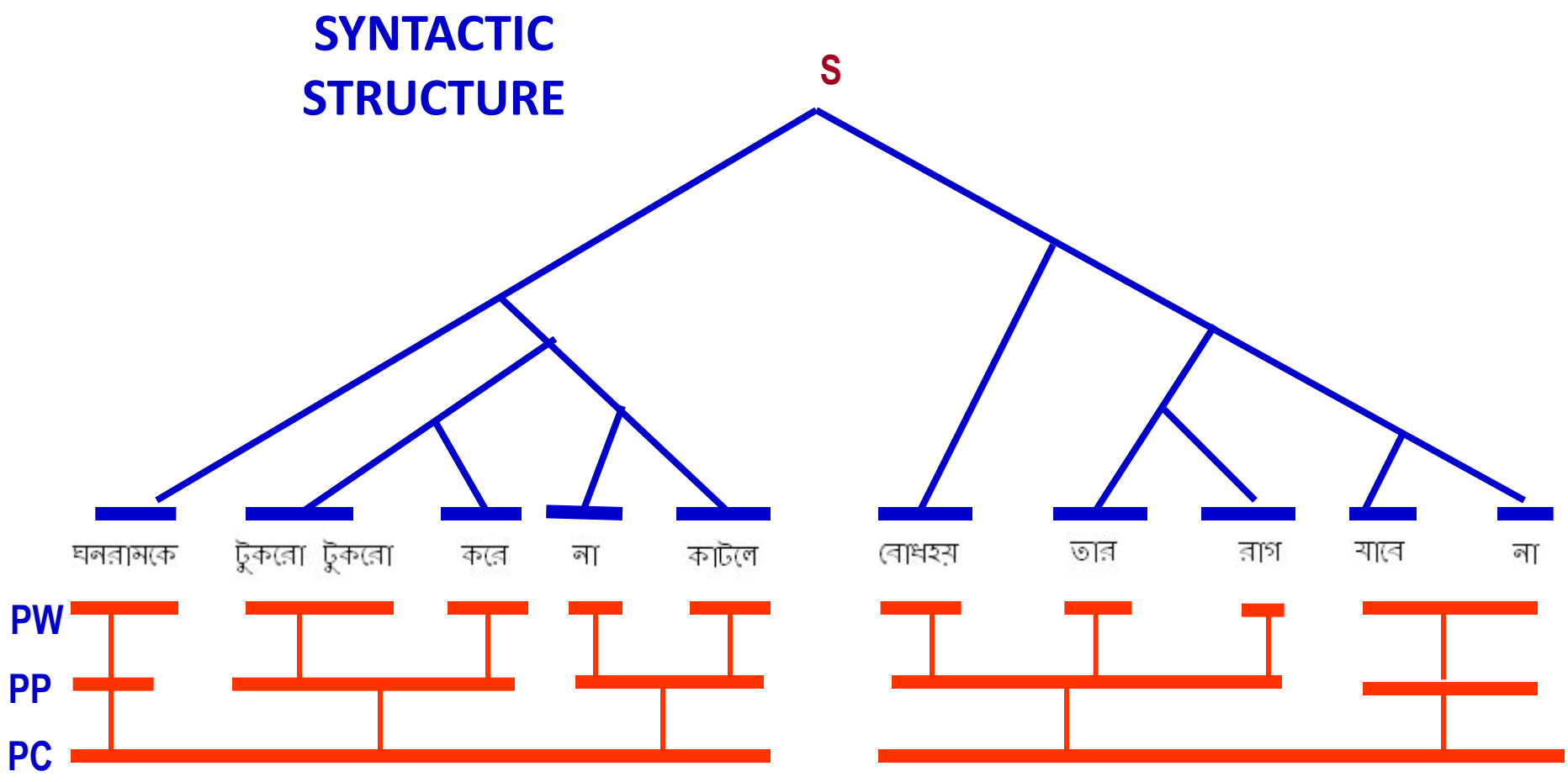
The syntactic category of the phrase (like (a) Noun Phrase (NP), (b) Adjective phrase (AJP), (c) Adverbial Phrase (ADVP), (d) Post-positional Phrase (PP) and (e) Verb Phrase (VP).

b. Phrase Length (F2)

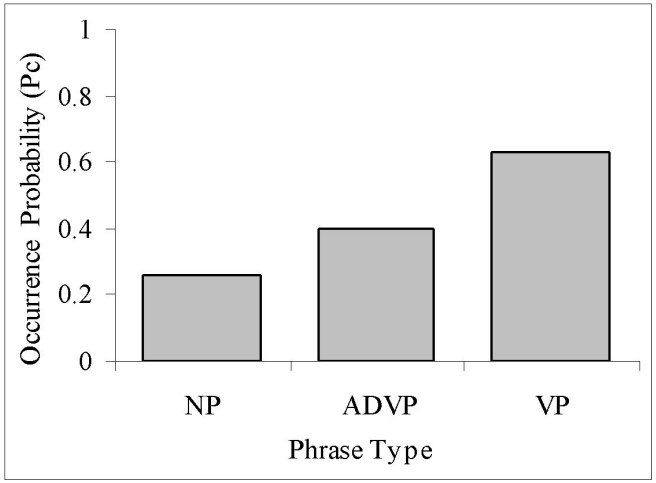
Length of the phrase in term of Syllable

c. Distance between the Current Phrase and its Dependent Counterpart

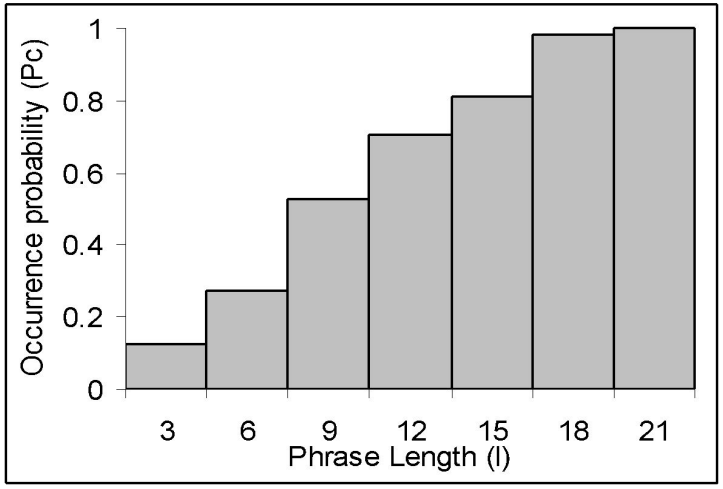
Distance (d) is calculated using the distance information (in terms of words) between the last word of the current phrase and the last word of its dependent phrase



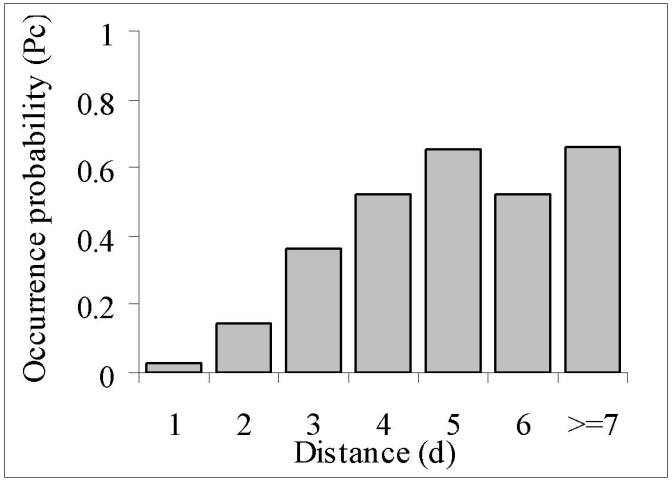
Pause Occurrence Probability of Bangla Read out Text



Effect of Phrase Types



Effect of Phrase Length



Effect of Distance

Modeling of Pause Occurrence Probability

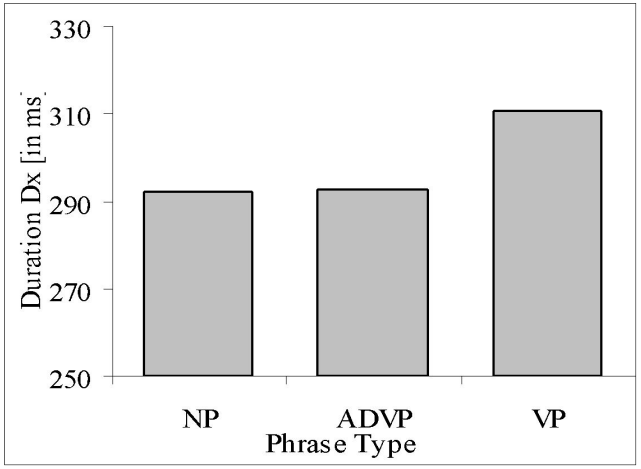
Pause Occurrence Probability can be model using a linear Model like:

$$R_x = al + bd + c \quad (X : NP, ADVP, VP)$$

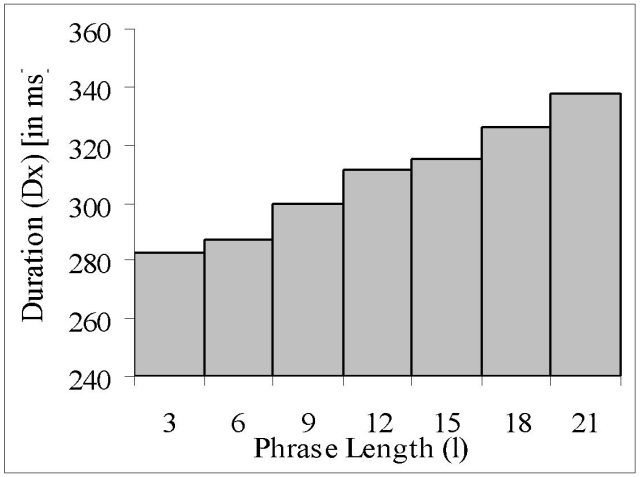
Model Parameters for Pause Occurrence Probability for Bangla

Phrase Type	Coefficients	Value	R square value
NP	a	0.077	0.902
	b	0.122	
	c	-0.436	
ADVP	a	0.072	0.813
	b	0.139	
	c	-0.486	
VP	a	0.078	0.791
	b	0.162	
	c	-0.467	

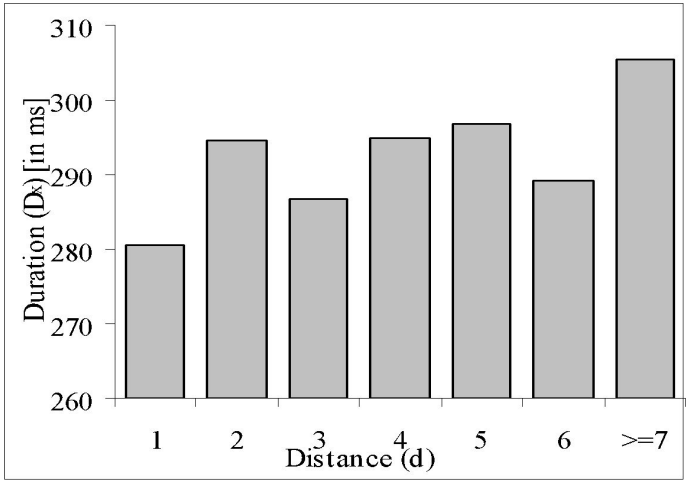
Pause Duration of Bangla Read out Text



Effect of Phrase Types



Effect of Phrase Length



Effect of Distance

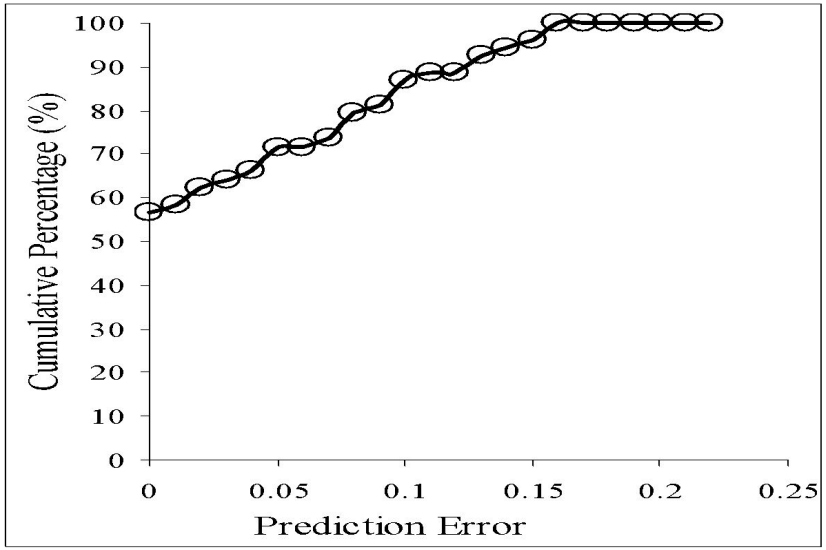
Modeling of Pause Duration

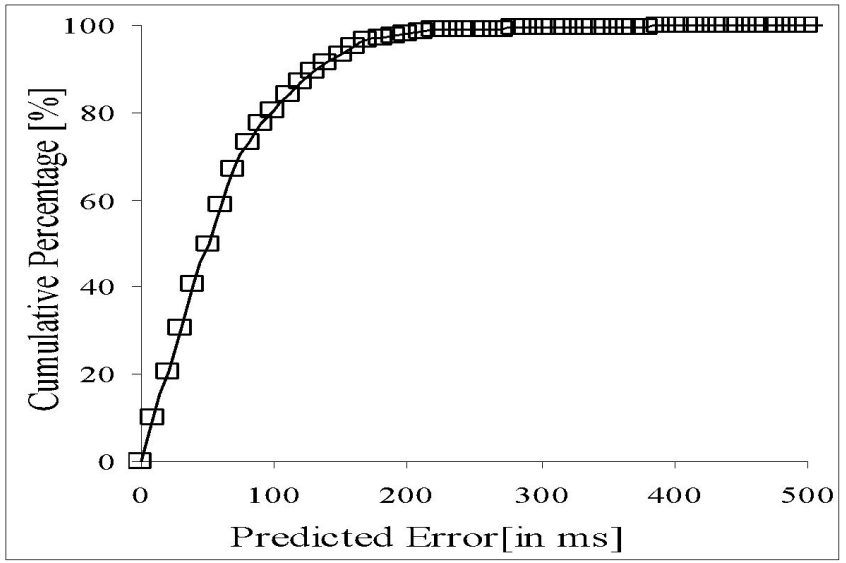
Pause Occurrence Probability can be model using a linear Model like:

$$L_X = \alpha l + \beta d + \gamma \qquad (X : NP, ADVP, VP)$$

Model Parameters for Pause Occurrence Probability for Bangla

Phrase Type	Coefficients	Value	R square value
NP	α	5.29	0.813
	β	11.06	
	γ	194.88	
ADVP	α	5.84	0.833
	β	8.46	
	γ	197.2	
VP	α	5.72	0.846
	β	6.11	
	γ	237.18	


Cumulative Percentage (%)
100
90
80
70
60
50
40
30
20
10
0
0
0.05
0.1
0.15
0.2
0.25
Prediction Error


Cumulative Percentage [%]
100
80
60
40
20
0
0
100
200
300
400
500
Predicted Error[in ms]

রাজা মহানন্দ রাজধানীতে তৈরি করেছিল শিব মন্দির ও বৈষ্ণবদের মন্দির।

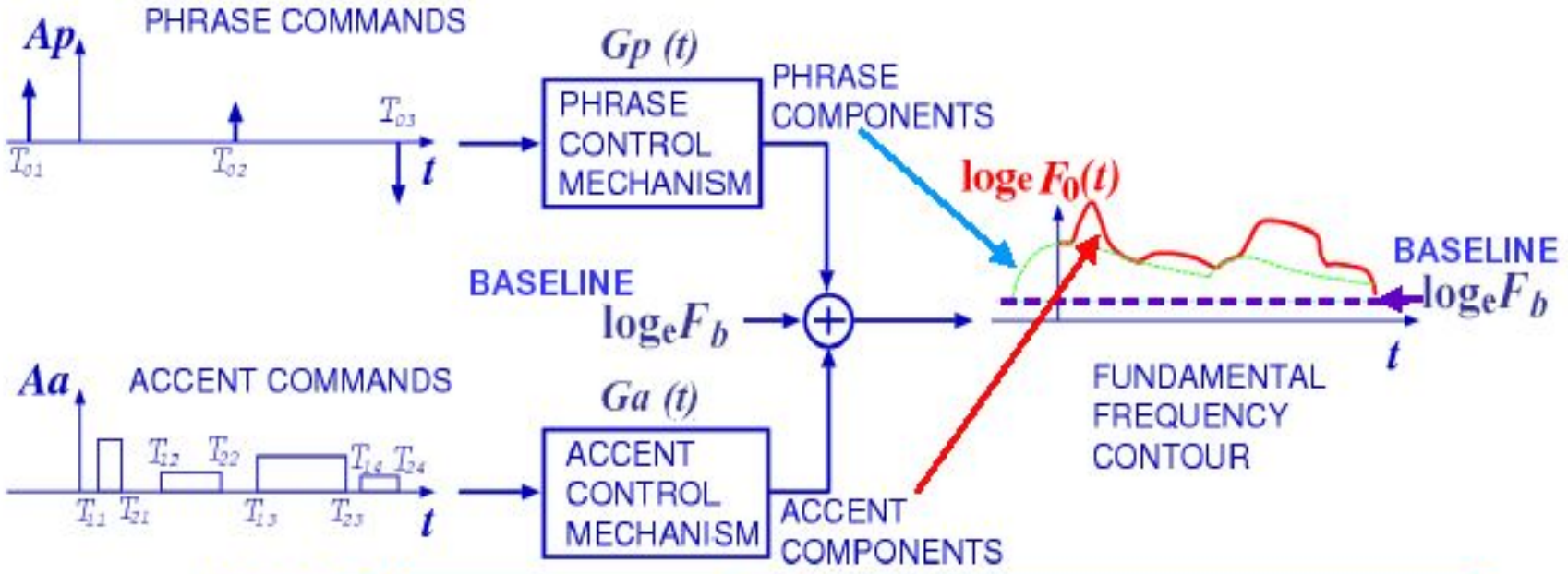
Role of Voice Fundamental Frequency (F_0) Contour

- 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

A functional model for the process of generating F0 contours by Prof. Hiroya Fujisaki

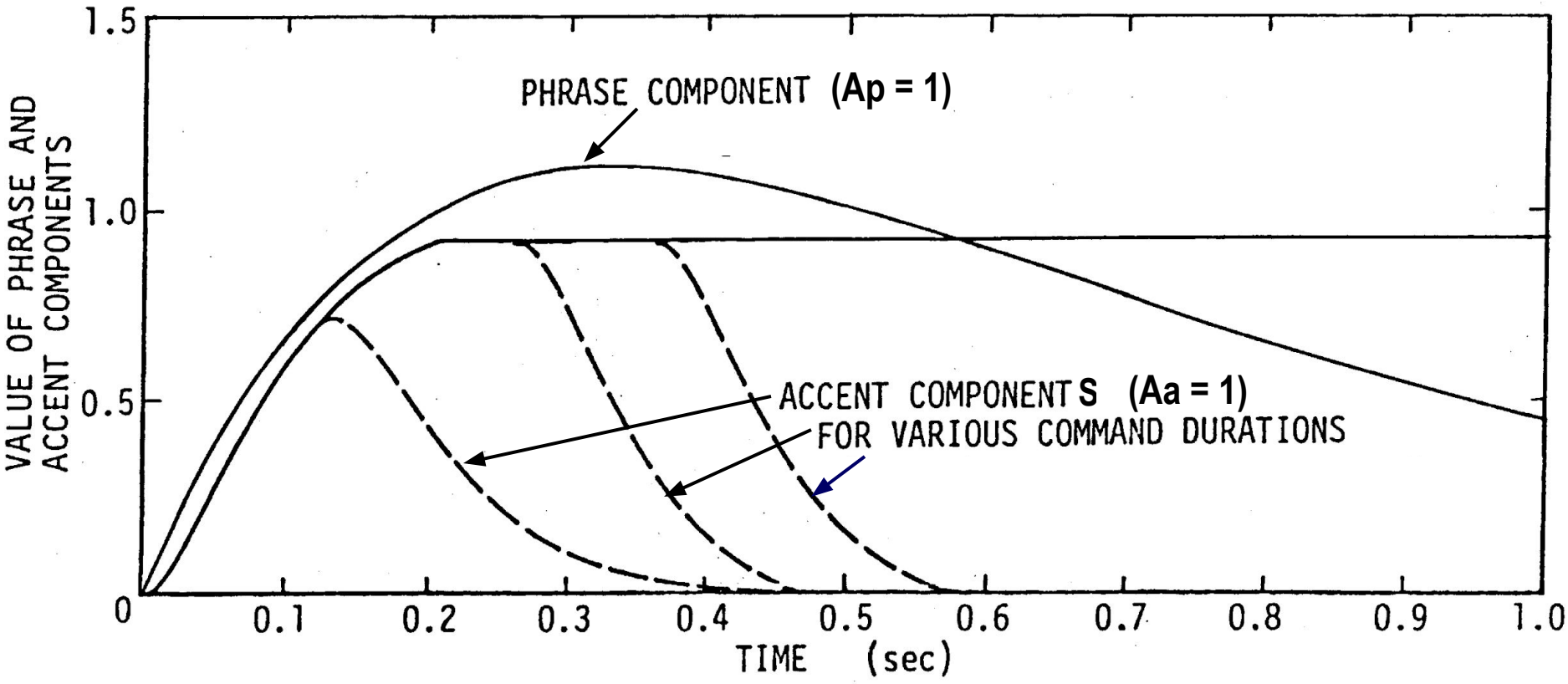


$$\log_e F_0(t) = \log_e F_b + \sum_{i=1}^{I} Ap_i Gp(t-T_{0i}) + \sum_{j=1}^{J} Aa_j\{Ga(t-T_{1j}) - Ga(t-T_{2j})\} \quad (1)$$

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

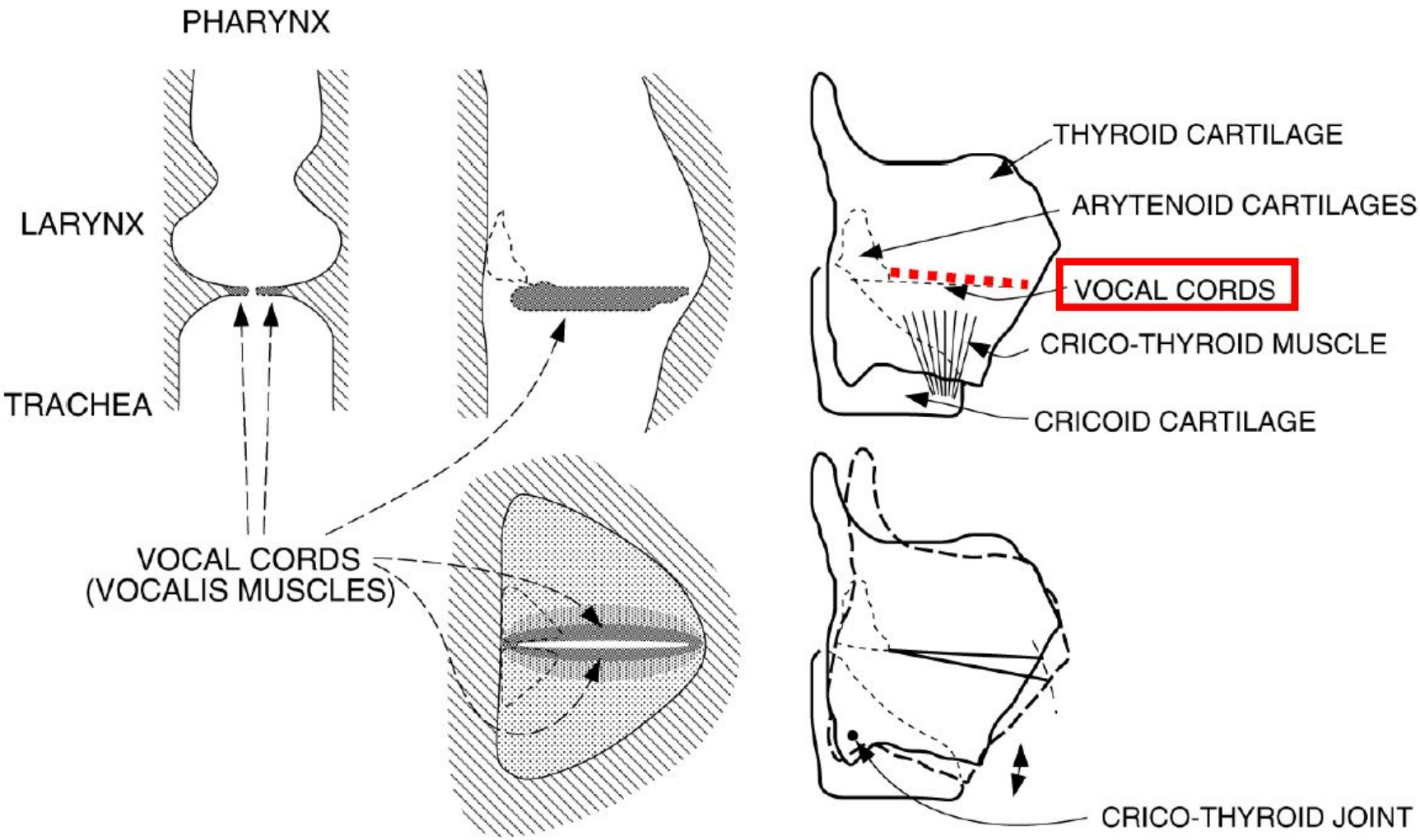
$$G_a(t) = \begin{cases} Min[1-(1+\beta t)\exp(-\beta t),\ \gamma], & t \geq 0, \\ 0, & t < 0. \end{cases} \quad (3)$$

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



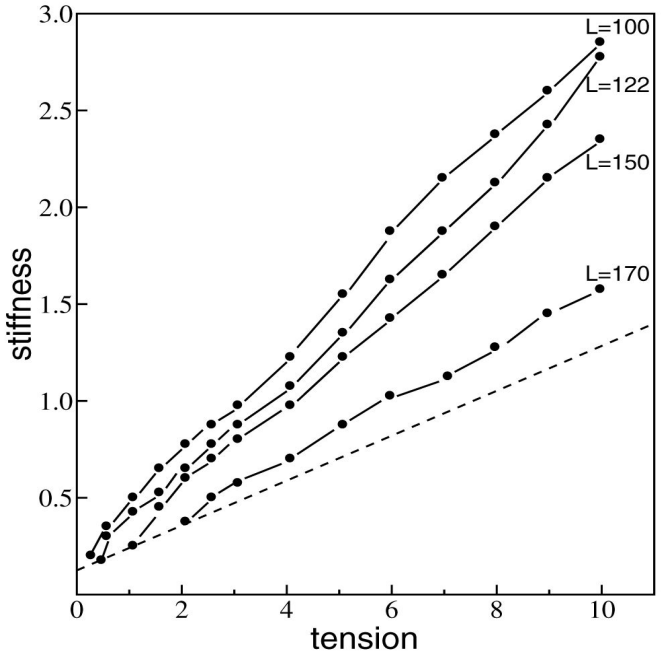
Parameter values for the phrase component: α= 3.0/s,
the accent components: β= 20.0/s, γ= 0.9.

Structure and Function of Larynx



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].



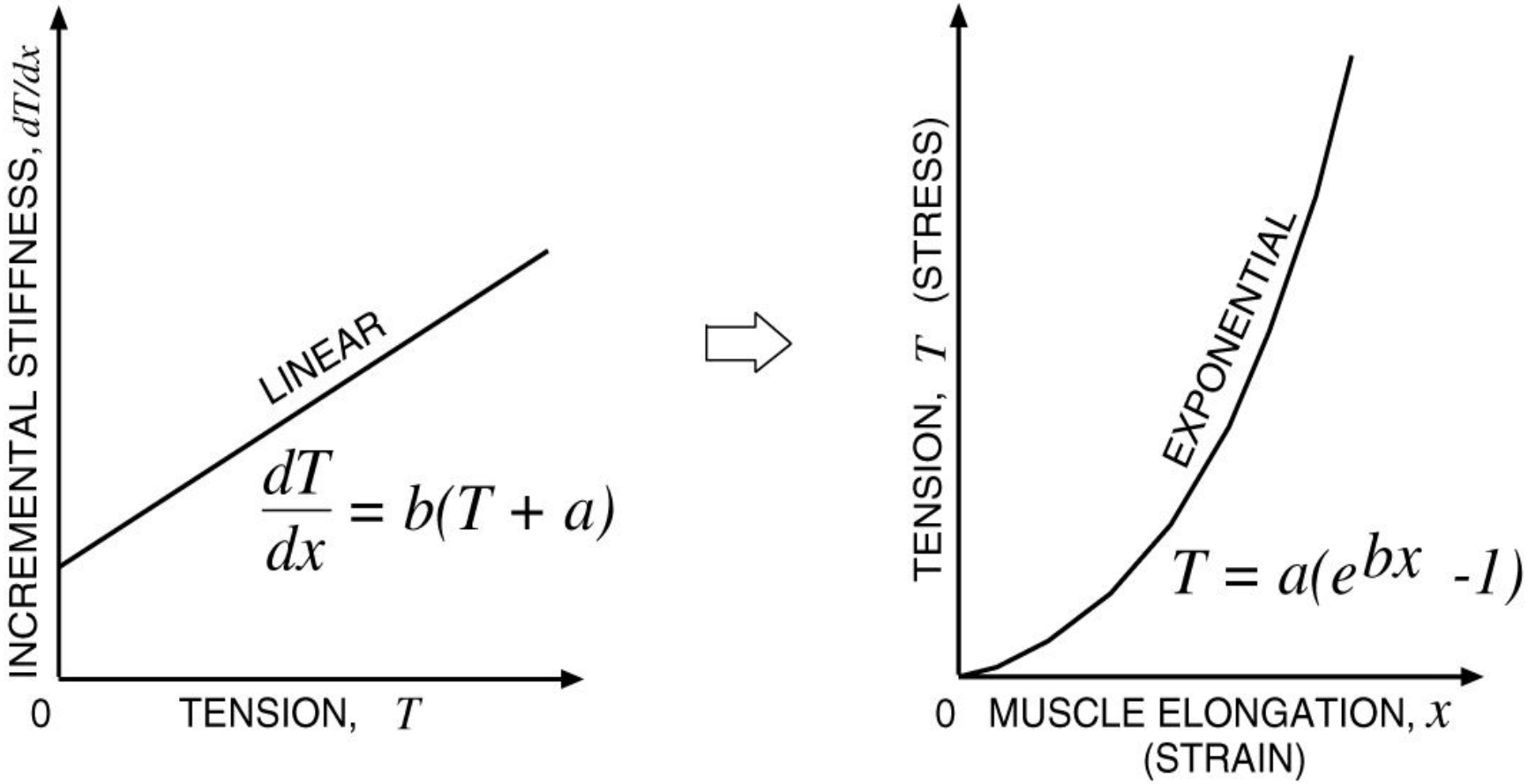
Stiffness as function of tension at rest (- - - - -) and during isometric tetanic contraction initiated at different original lengths. In the top curve contraction is initiated at a length below 100 (equilibrium length = 100).

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

Physical properties of skeletal muscles (2)



From vocal cord elongation to tension

Stress-strain relationship in a skeletal muscle

$$\frac{dT}{dl} = b(T + a) \qquad (1)$$

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

By integration $$T = (T_0 + \frac{a}{b})e^{b(l-l_0)} - \frac{a}{b} \qquad (2)$$

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

When $T_0 >> a/b$

$$T \cong T_0 e^{bx} \quad (3)$$

$$\text{where } x = (l - l_0)$$

From vocal cord tension to fundamental frequency

Frequency of vibration of an elastic membrane is given by

$$F_0 = c_0 \left(\frac{T}{\sigma}\right)^{\frac{1}{2}} \quad (4)$$ where σ is the density/unit area

From Eqs. (3) and (4)

$$\log_e F_0 = \log_e [c_0 (\frac{T_0}{\sigma})^{\frac{1}{2}}] + \frac{\mathrm{b}}{2} x \quad (5)$$

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

$$\log_e F_0 = \log_e F_{\mathrm{b}} + \frac{\mathrm{b}}{2} x(t) \quad (6)$$

$$\text{where} \quad F_b = c_0 \left(\frac{T_0}{\sigma}\right)^{\frac{1}{2}}$$

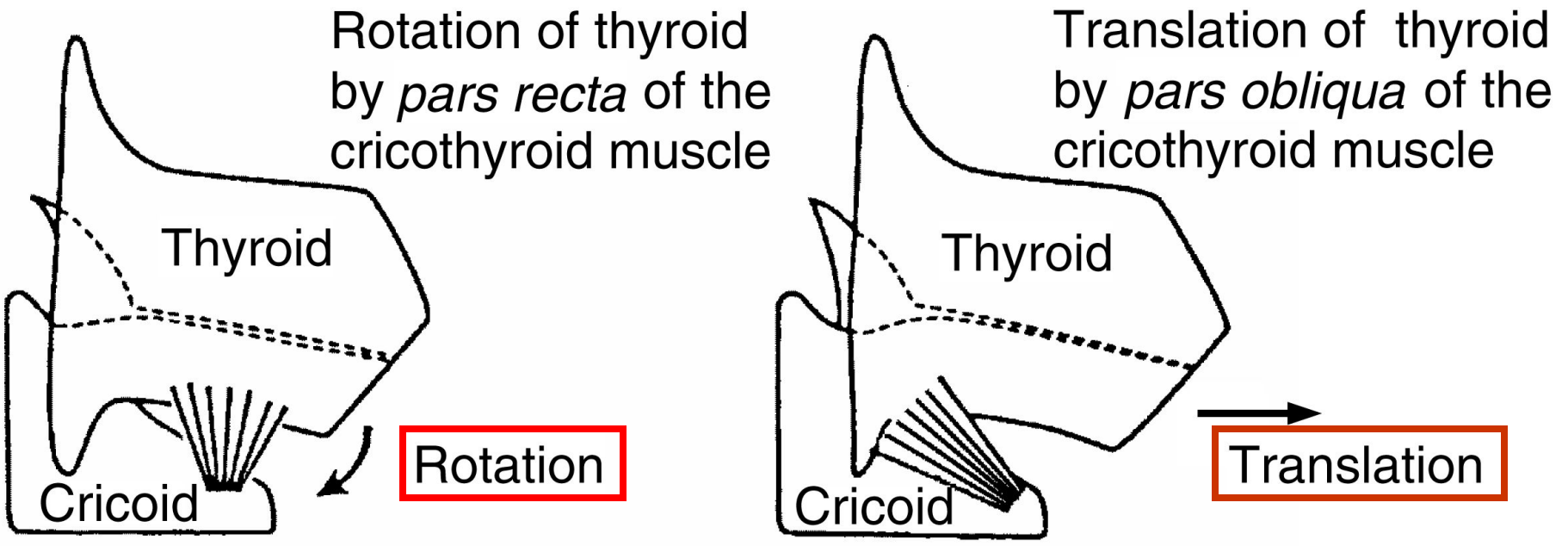
Thus an F_0 contour, when plotted in the ***logarithmic* scale** as a function of time, can be expressed as the sum of **a constant (baseline) term** and **a time-varying term**, proportional to the elongation of the vocal cord.

The role of the cricothyroid (CT) muscle

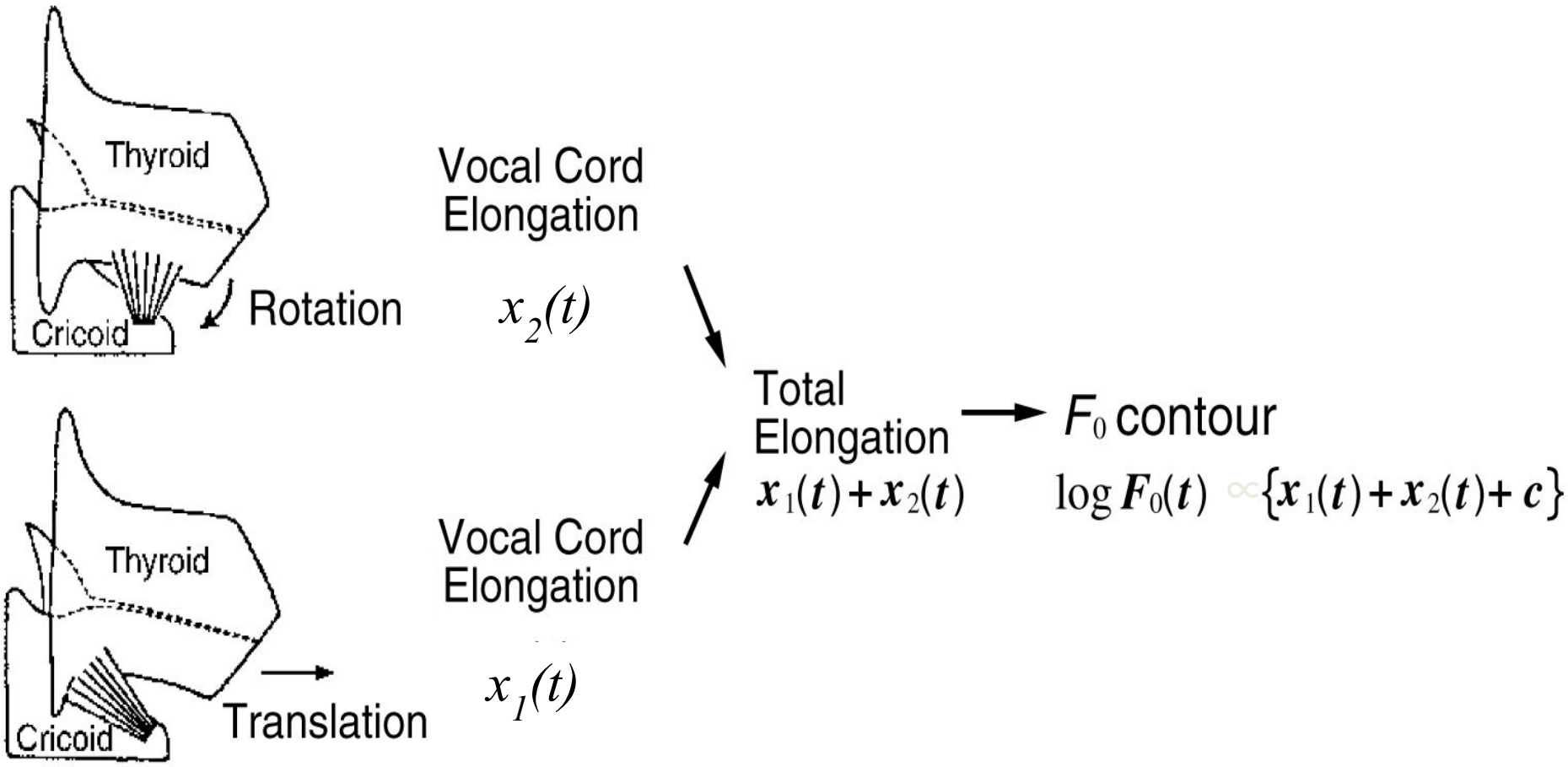
Analysis of the laryngeal structure suggests that the movement of the thyroid cartilage has two degrees of freedom [e.g., Zemlin 1968, Fink & Demarest 1978].

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.

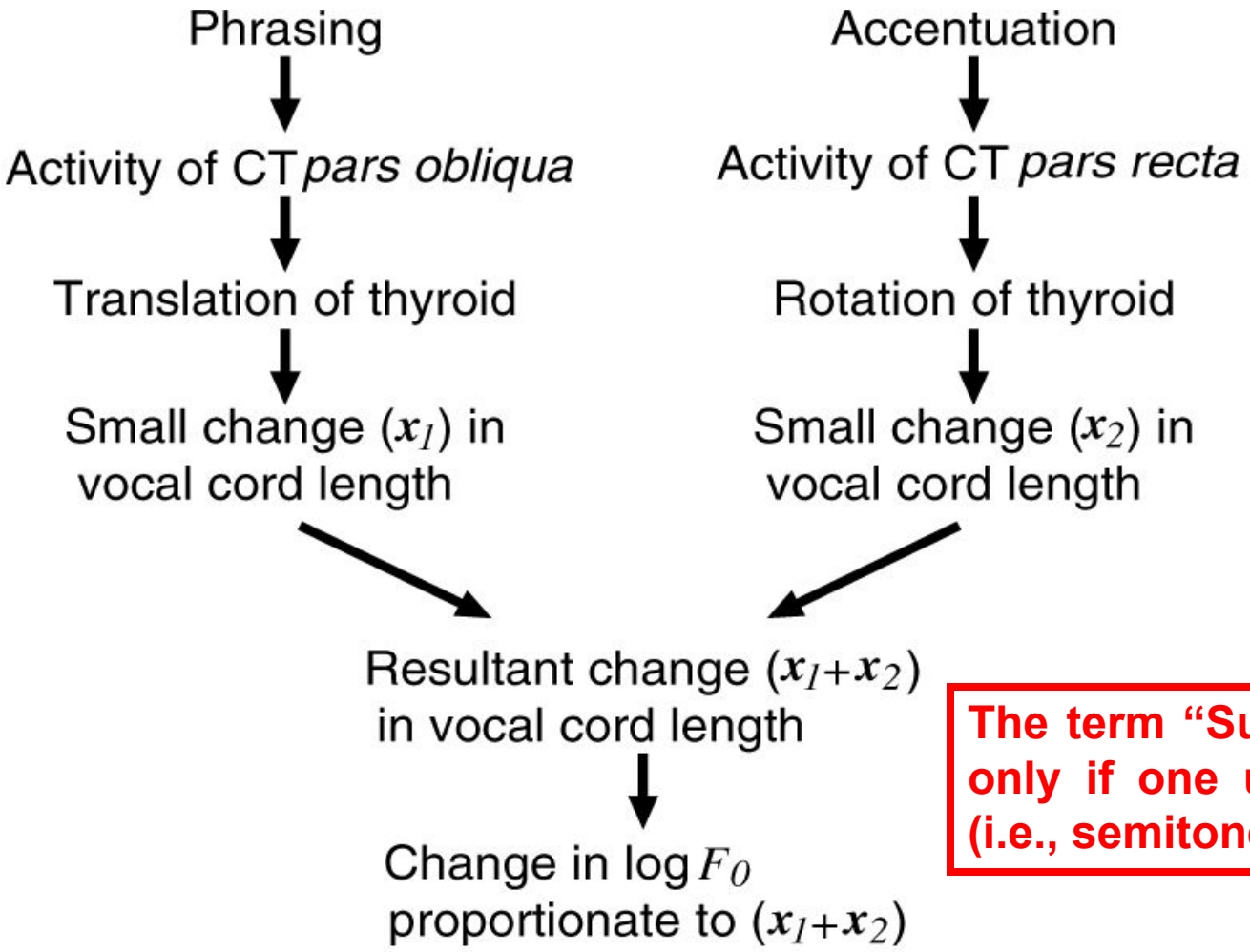
Motion of thyroid with two degrees of freedom



Additivity of phrase and accent components

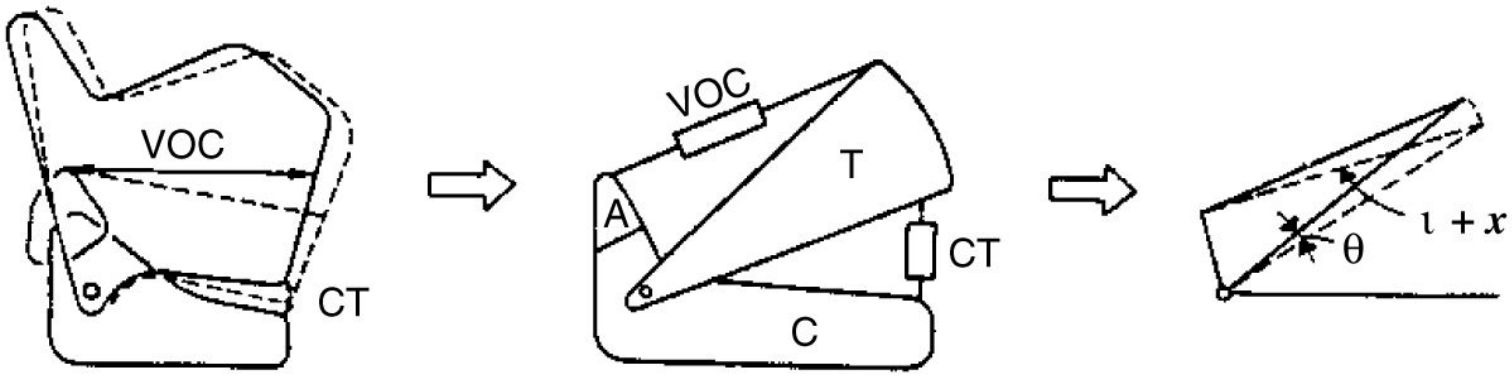


Additivity of components in logF_0 domain



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

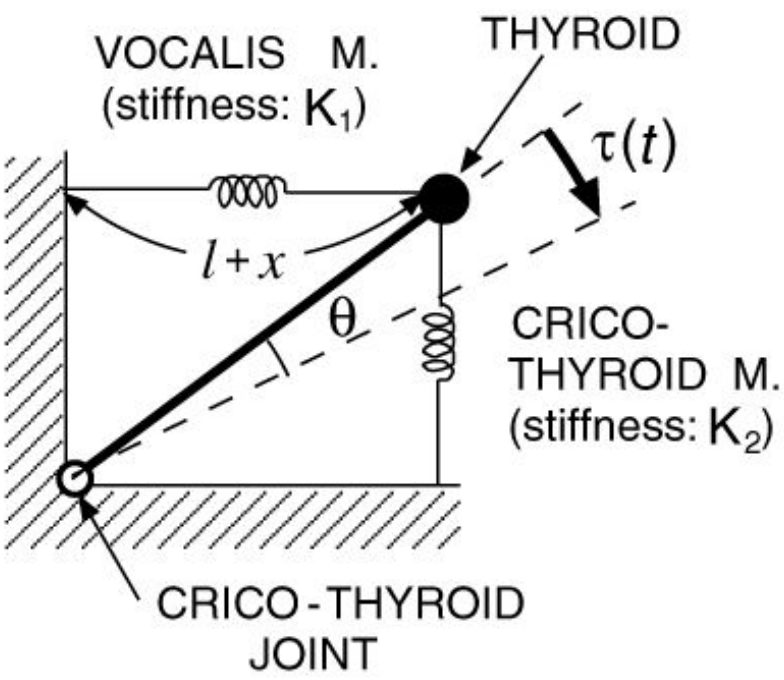


VOC : Vocalis M.
CT : Criothyroid M.

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,

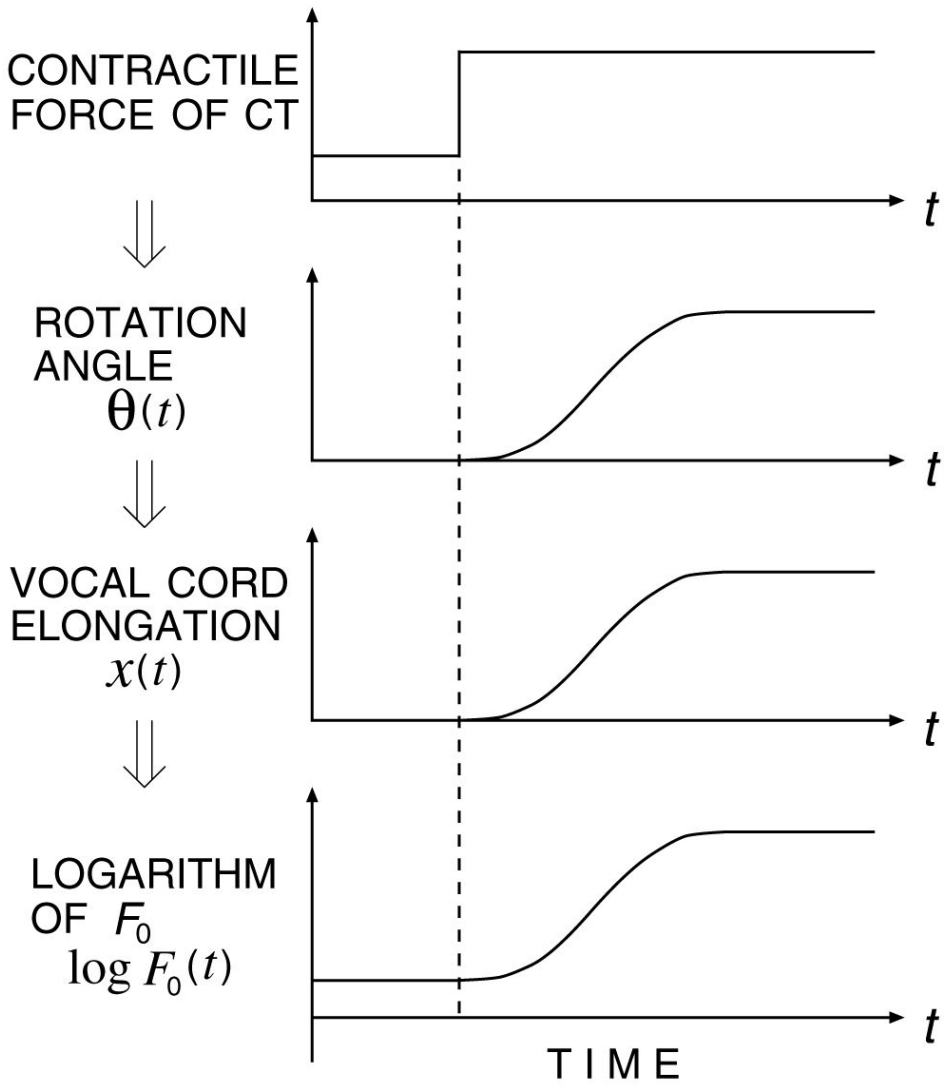
$$x(t) = C_4\theta(t) = C_5 G_a(t)$$

- Hence

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

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

From cricothyroid activity to logarithm of F_0

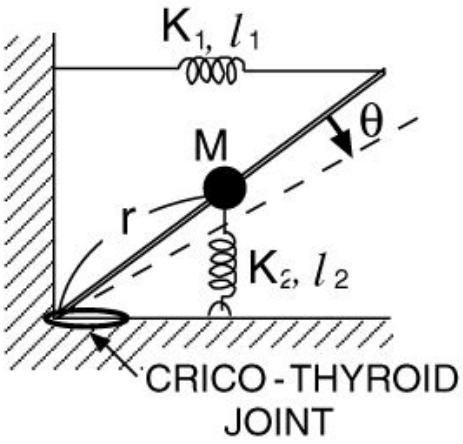


The rate of F_0 change is determined, not by the speed of contraction or relaxation of the muscle, but by the mechanical properties of the laryngeal structure.

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

Rotation and translation of the thyroid

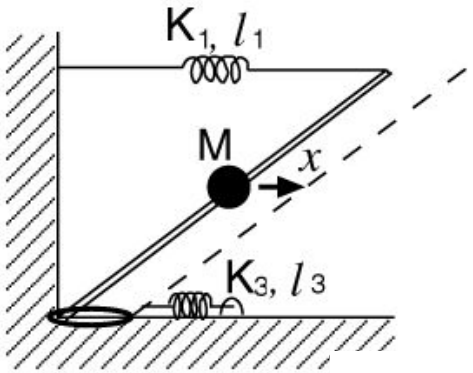
ROTATION



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

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

TRANSLATION



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

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

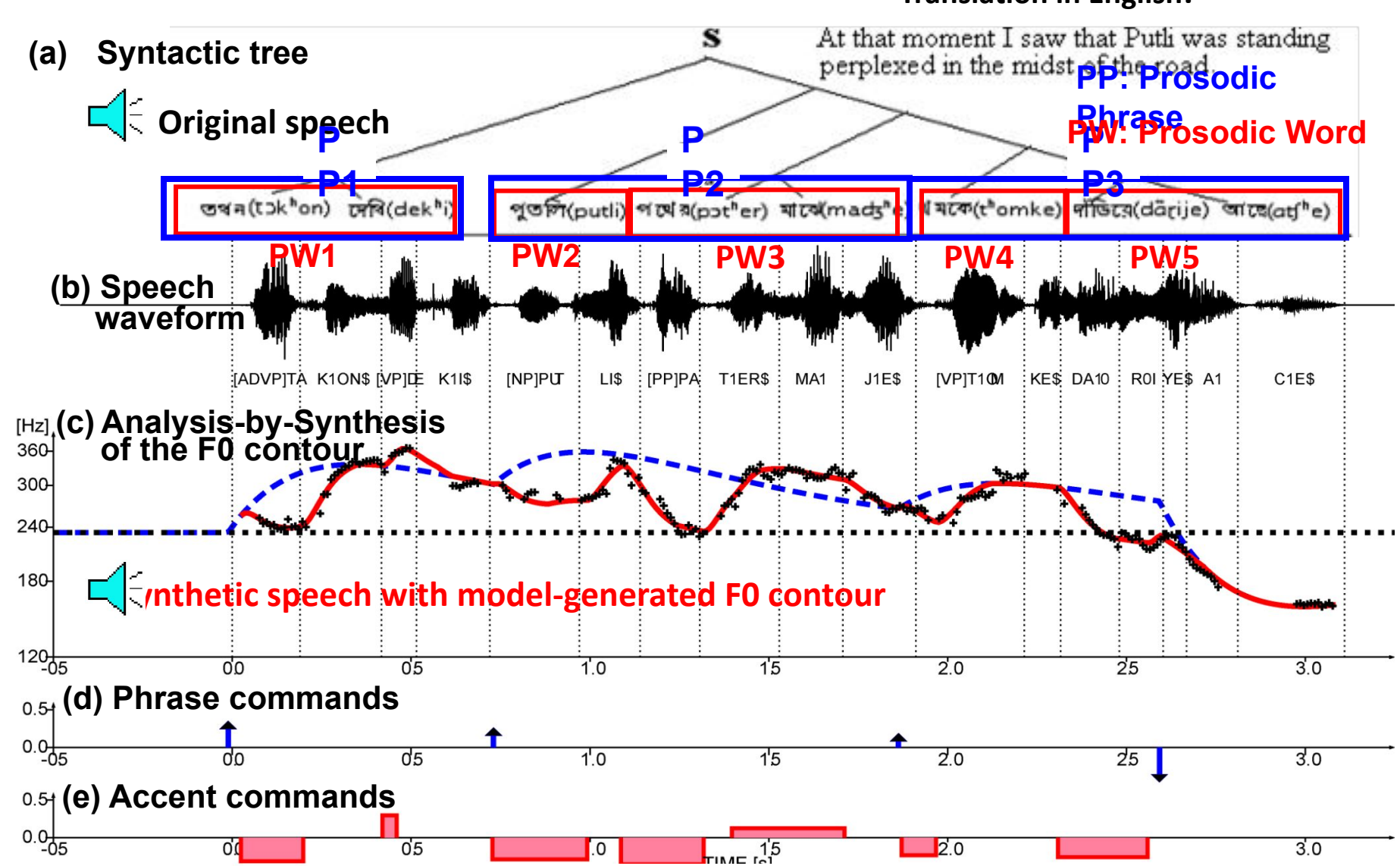
$$G_p(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & t \geq 0, \\ 0, & t < 0, \end{cases}$$

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

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

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

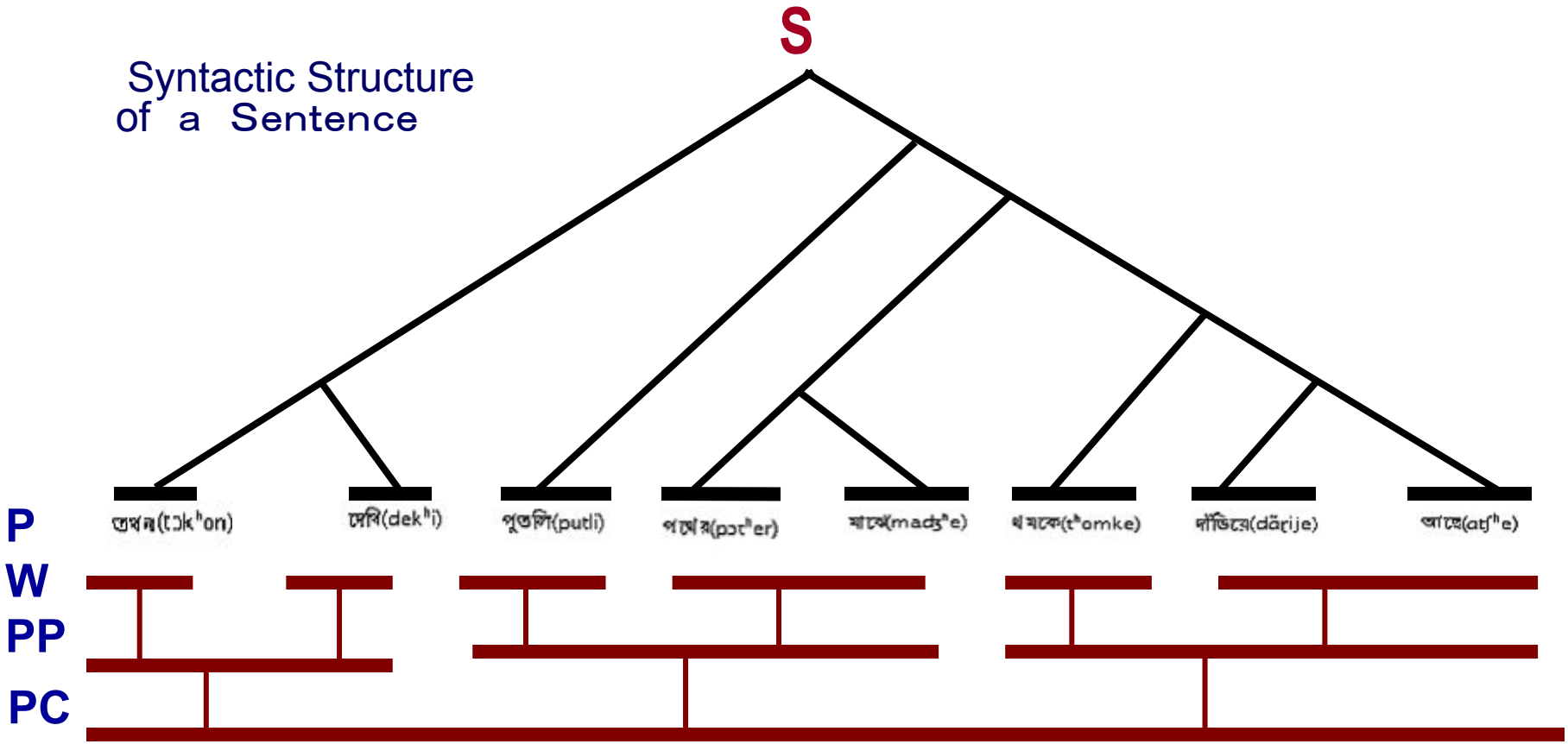
Analysis-by-Synthesis of the F_0 contour of an Utterance of a Bangla sentence



Definition of prosodic units (Fujisaki 1988)

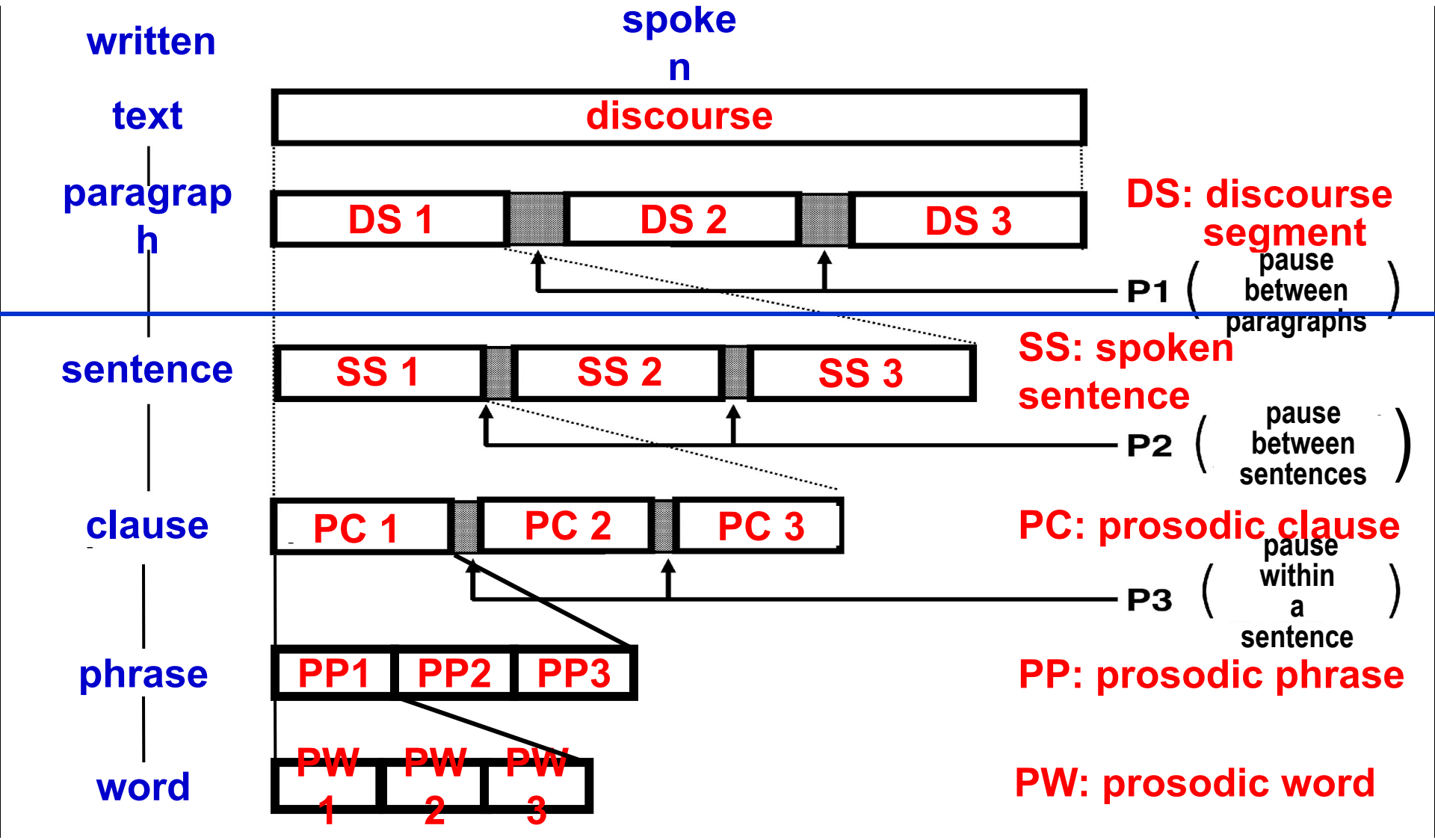
1. A **prosodic word** is defined as a part or a whole of an utterance that forms an accent type.
2. A **prosodic phrase** is defined as the interval between two successive phrase commands uninterrupted by a pause.
3. A **prosodic clause** is defined as a successive set of prosodic phrases delimited by utterance-medial/final pauses.
4. A **prosodic sentence** is defined as the utterance delimited by utterance-final pauses (except for the initial utterance of a discourse segment).

Syntactic Structure vs. Prosodic Structure (Fujisaki 1984)

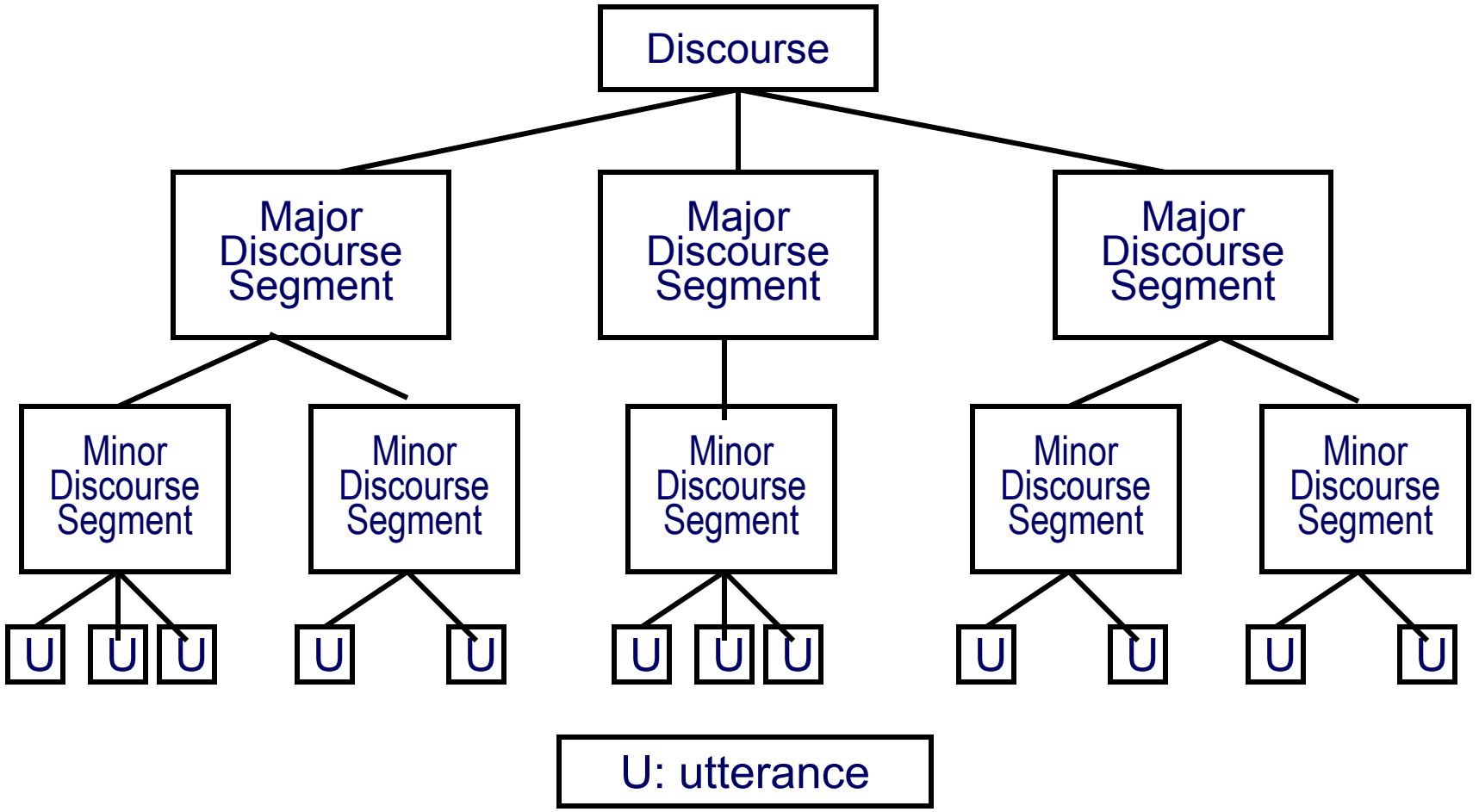


PW: Prosodic Word
PP : Prosodic Phrase
PC : Prosodic Clause

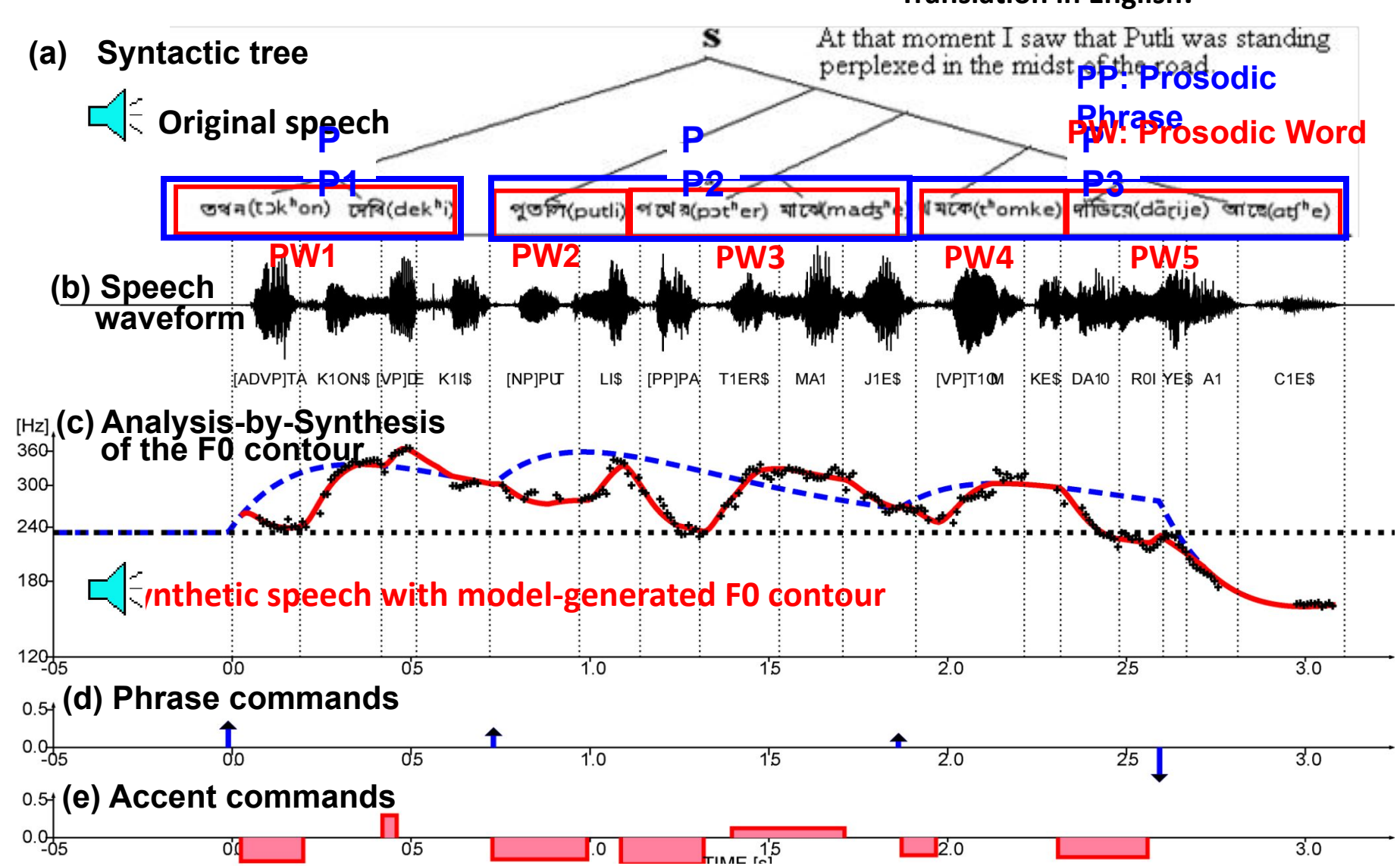
Prosodic Structure



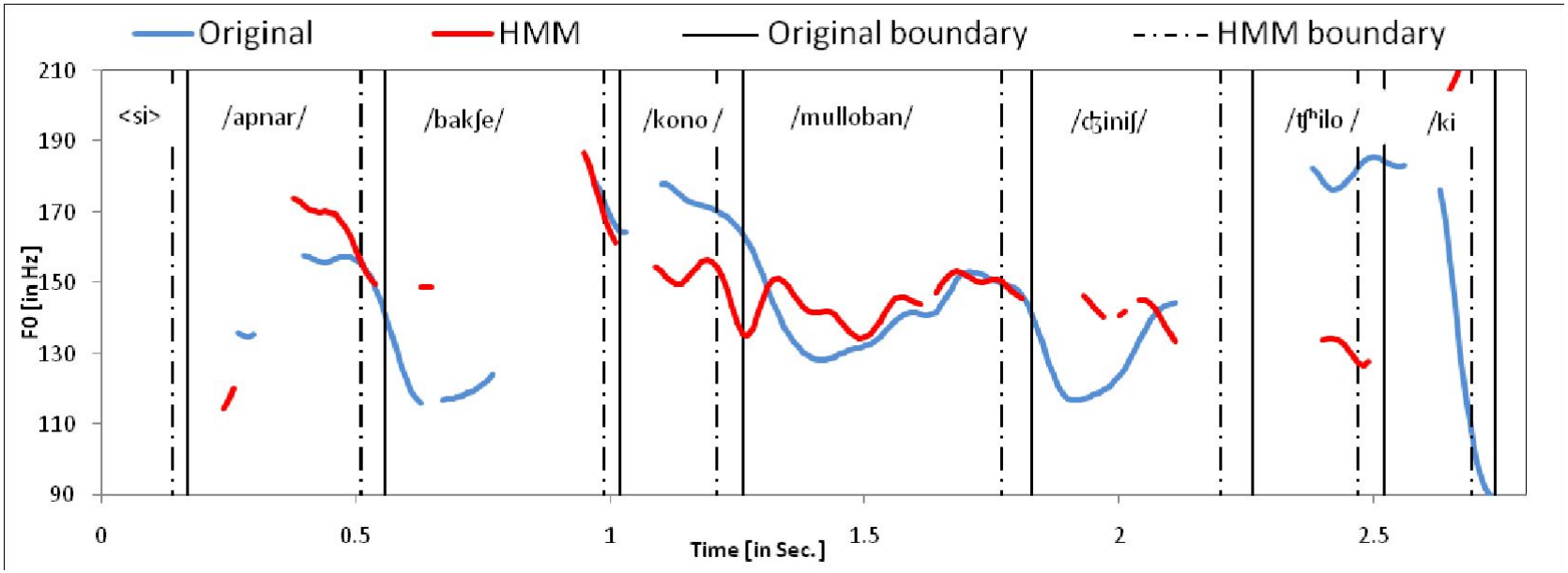
Hierarchical structure of prosody of a discourse



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



Synthesized F0 contour



Original

HTS

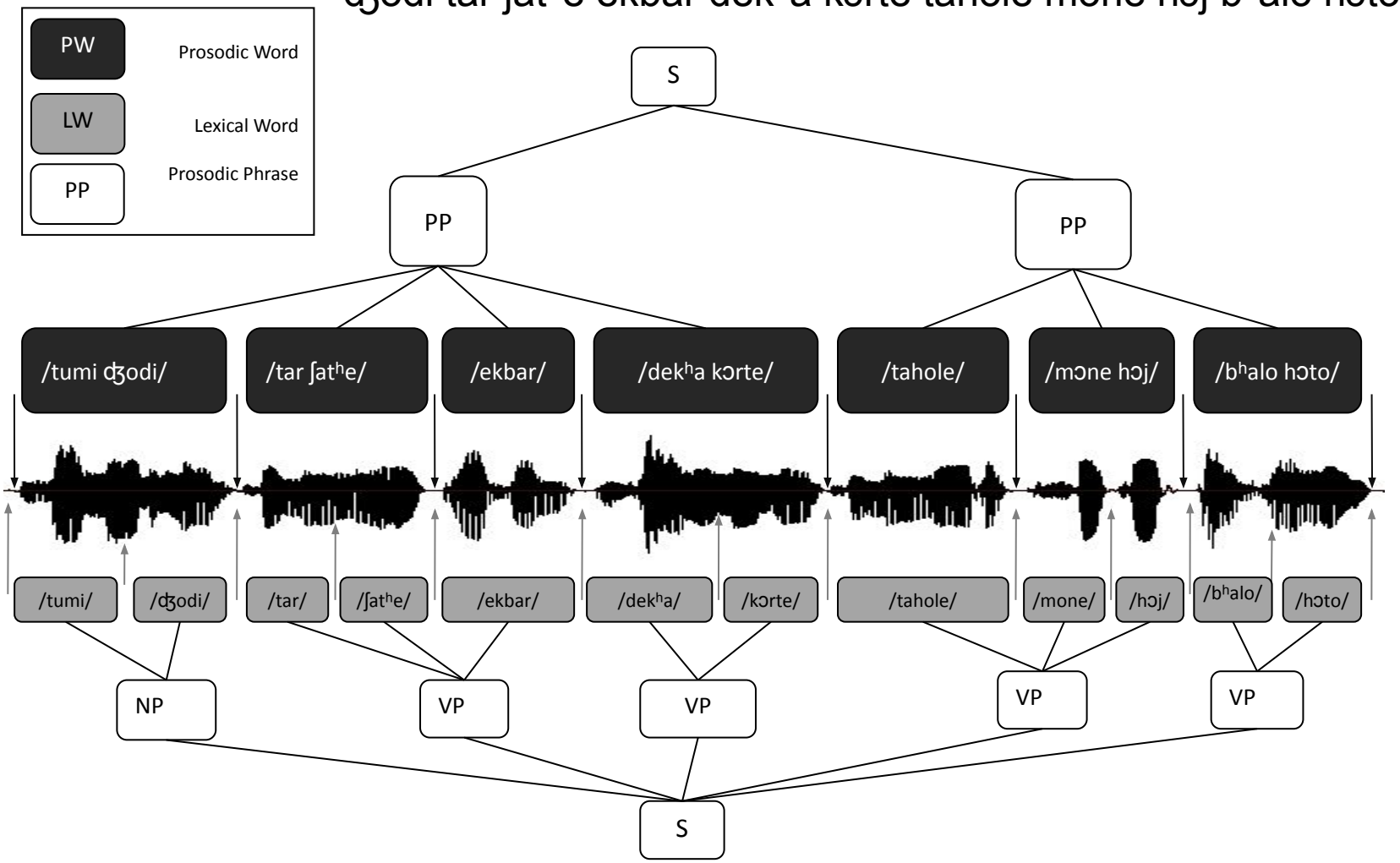
Training Difference

- ❑ Traditional HTS uses **Lexical** sentence structure

- ✔ Adaptation of **Prosodic** sentence structure

Sentence Prosodic and Lexical Structure

তুমি যদি তার সাথে একবার দেখা করতে তাহলে মনে হয় ভালো হত /tumi ʤodi tar ʃatʰe ekbar dekʰa kɔrte tahole mone hɔj bʰalo hɔto/



Improvement!

- There are more number of phonemes within small deviation in case of Prosodic-Bengali-HTS rather than Lexical-Bengali-HTS

System Type	Gender #phonemes	No. of Phonemes for which duration is within deviation				
		1%	3%	5%	10%	15%
Lexical-Bengali-HTS	Male (3895)	70	113	125	572	487
	Female (3782)	97	102	188	498	651
Prosodic-Bengali-HTS	Male (3895)	100	145	157	699	754
	Female (3782)	147	127	209	562	714

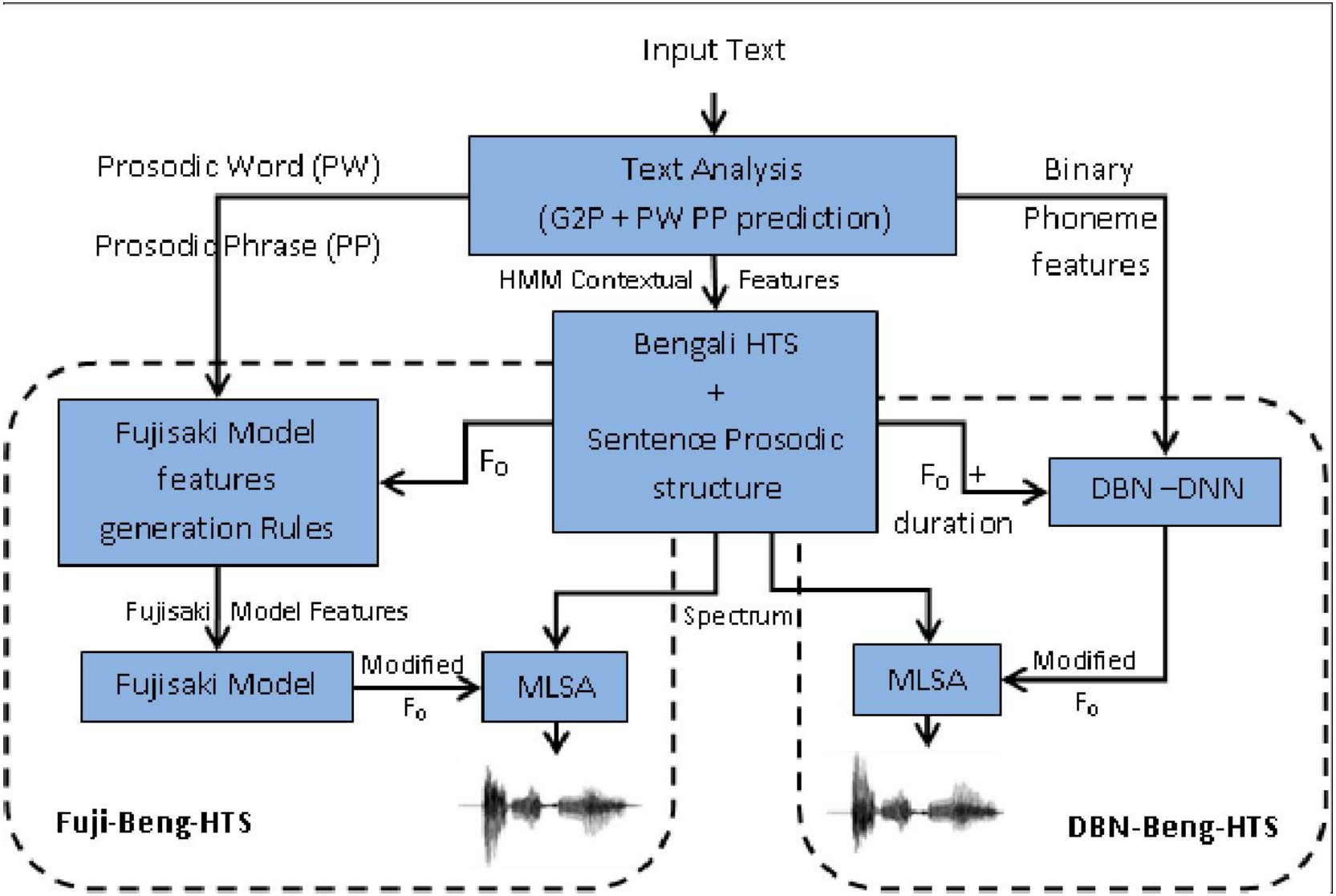
- No improvement in F0 contour

Choice of F0 modeling techniques

- Fujisaki F0 generation Process Model
  - Based on very strong mathematical background
  - Success in many languages

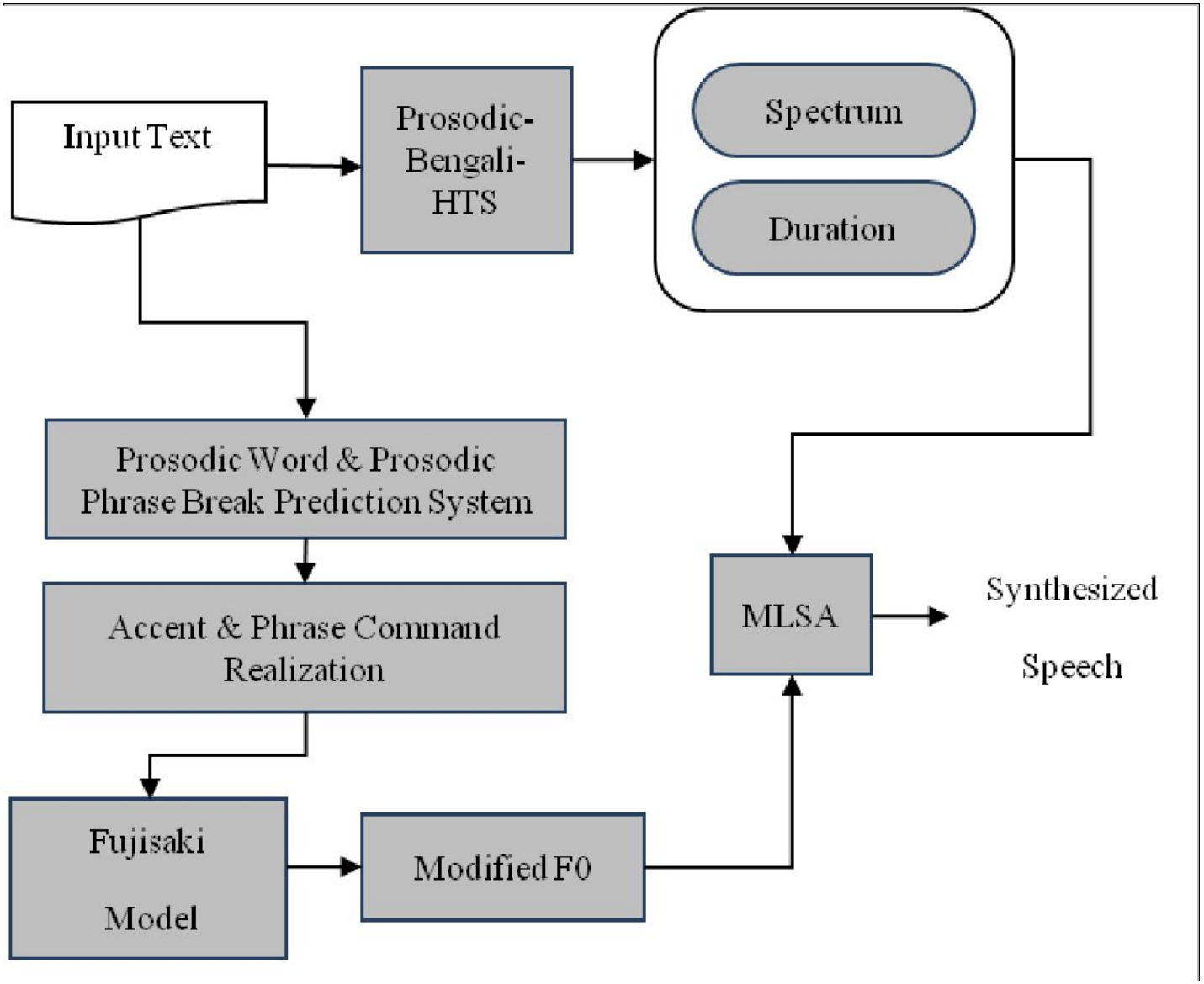
- Deep Belief Network (DBN)
  - Fujisaki Model requires lots of labeled data
  - DBN learns in unsupervised fashion

Overall TTS system



F0 Modeling using Fujisaki F0 Generation Process Model (Fujisaki Model)

Block Diagram of Proposed System



Fujisaki F0 generation Process Model (Fujisaki model)

- Fujisaki model treats F0 contour as a superposition of two components
  - accent component
  - phrase component

- Fujisaki Model Features
  - Phrase Command Magnitude
  - Phrase Lead Time
  - Accent Command Magnitude
  - Accent Command (Lead Time and Lag time)

phrase component accent component

$$F_0(t) = \ln(F_b) + \sum_{i=1}^{I} A_{pi}\, G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{aj} \{G_a(t - T_{1j}) - G_a(t - T_{2j})\}$$

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

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

F0 modeling using Fujisaki Model

- This work is inspired by

SK Das Mandal, Anal Haque Warsi, Tulika Basu, Keikichi Hirose, and Hiroya Fujisaki. *Analysis and synthesis of f0 contours for bangla readout speech*. Proc. Of Oriental COCOSDA 2010.

- Fujisaki model features are calculated by analyzing different sentence types
  - Declarative
  - Imperative
  - Exclamatory
  - Integrative etc.

- Some rules are formed according to those features in order to predict F0 contour

Phrase Command Features

Phrase Position	Lead Time (T_{oi})			Command Magnitude (A_{pi})	
	Pause>0.25s	0.1s ≤Pause≤0.25s	Others	Mean	Standard deviation
Utterance Initial	-	-	0.037	0.478	0.016
2nd phrase	0.321	0.244	0.102	0.265	0.025
3rd phrase	0.357	0.261	0.387	0.182	0.034
≥ 4th phrase	0.366	0.227	0.211	0.105	0.061
Utterance Final	-	-	0.034	-0.346	0.073

Accent Command Features

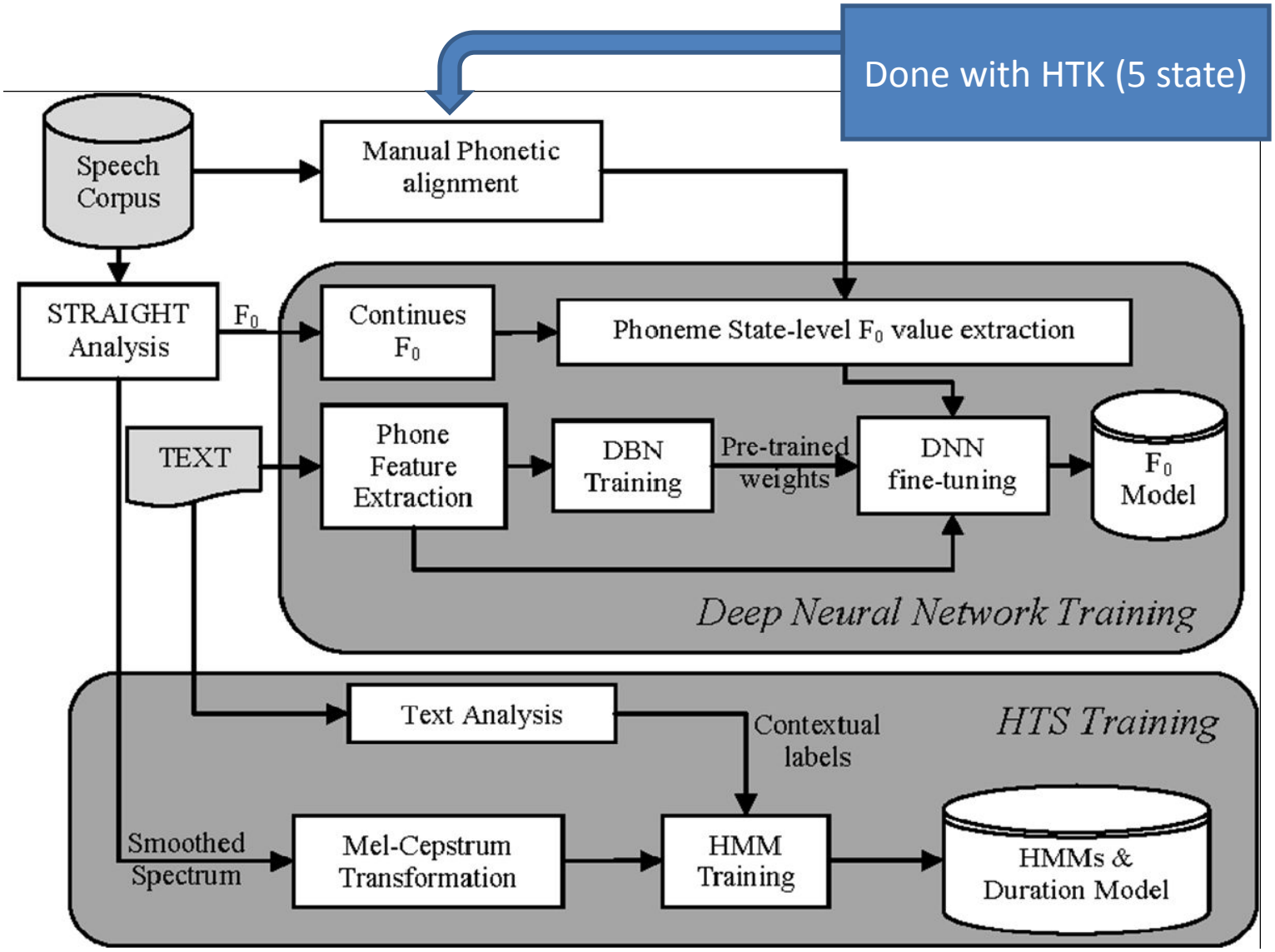
Position		Accent command amplitude (A_{ai})	
		Negative	Positive
Utterance-initial		-0.325 ($\sigma = 0.06$)	0.232 ($\sigma = 0.05$)
Utterance Medial	Phrase initial	-0.317 ($\sigma = 0.05$)	
	Phrase medial	-0.261 ($\sigma = 0.08$)	
Utterance final		-0.328 ($\sigma = 0.09$)	-

Position	t_1		
Utterance-Initial	0.148		
Utterance-medial	Pause $\geq$ 0.10 s	0.147	
	Pause $<$ 0.10 s	Voiced	0.071
		Unvoiced	0.111

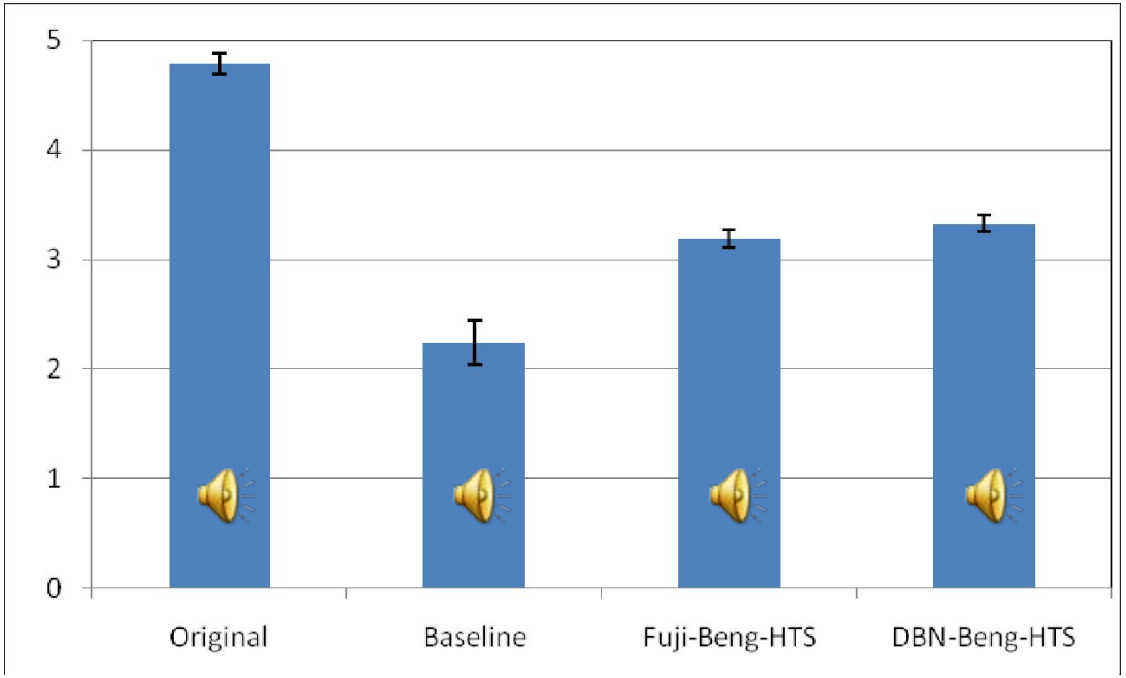
Word Length	t_2	
Monosyllabic	0.176	
Polysyllabic	Syllable Type-'V','VC','CV'	0.037
	Syllable Type-'others	0.071

F0 Modeling using Deep belief Network (DBN)

Training Stage



Mean Opinion Score Test



No. of sentences = 15

Original	4.788
Baseline	2.24
Fuji-Beng-HTS	3.19
DBN-Beng-HTS	3.33