

**Digital Speech Processing**  
**Prof. S. K. Das Mandal**  
**Centre for Educational Technology**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 43**  
**Fundamental Frequency Contour Modeling**

(Refer Slide Time: 00:21)

***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.)

© Hiroya Fujisaki IWSLPB-2009, Kolkata 13

So let us start with the  $F_0$  modeling. So, if you see what kind of information; if  $F_0$  is carrying in many languages, the pattern temporal change in  $F_0$  is phonemic which is called tonal language if you see the tone variation of tone can change the meaning. So, tonal languages; so, temporal variation of  $F_0$  can change the meaning of the words. Similarly if I say the trace main parameter for defining trace is  $F_0$ , if I say emphasize word; that means, if you found I have increased the  $F_0$  there. So, change of  $F_0$  define the tone define whether that words is emphasis or not and also in case of English if you see there is a contrasting trace language. So, it trace depends on the context. So, if I put arbitrary trace in the words, then I am going to say that my pronunciation of English is not native like.

The problem in here also so much; I am in Bengali on Indian language; most of the Indian languages are bound trace language; that means, the trace is defined at the beginnings level of every prosodic word not linguistic words not in the I can say the written word for every prosodic word beginning trace is defined that is why it is called

bound trace language, but English is contrastive trace language. So, you can say the F<sub>0</sub> is convey the linguistic information also like the tone it can be phonemic also so; that means, variation of the tone change the word meaning then F<sub>0</sub> is convey the prosodic meaning of the sentence if I put the trace in different location the meaning may be change the old man and women if I say the old man and women the old man and women then I say the man is old, but women is not.

But if I said the old man and women then man and women both are old. So, F<sub>0</sub> linguistic information is carrying not only linguistic information F<sub>0</sub> also carry the para-linguistic information and non-linguistic information, para-linguistic information and non linguistic information also convey by F<sub>0</sub>. If it is male speaker you know whether I am male of female speaker that depends on the variation of the F<sub>0</sub>, if it is male speaker you know the average F<sub>0</sub> cannot goes a 200 after 180 hertz, but if it is female speaker it is started from 200 hertz; it can goes up to 300 hertz if it is child speech you can say that it start can started as 250 hertz and goes to 350 hertz.

So, age all kinds of non linguistic information and para-linguistic information all are carried by F<sub>0</sub> attitude emotion and then speaking style all can be found in F<sub>0</sub> modeling.

(Refer Slide Time: 04:00)

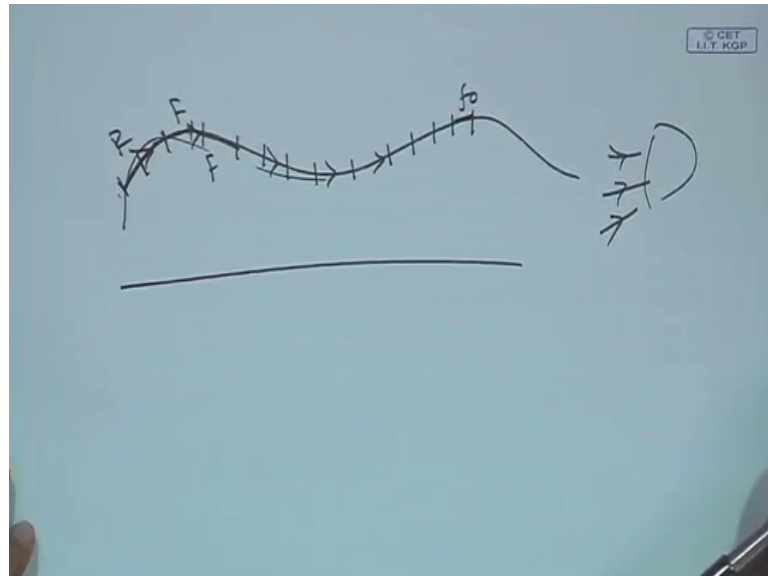
| <b>Three Approaches to the Description/Representation of F<sub>0</sub> Contour Characteristics</b> |                 |                                  |                         |                   |
|--|-----------------|----------------------------------|-------------------------|-------------------|
|  | Example         | Outcome                          | Method                  | Coding/Decoding   |
| Labeling   | ToBI            | Discrete Labels                  | Subjective Qualitative  | Irreversible      |
| Stylization  | 't Hart         | Piece-wise Linear Approx.        | Subjective Quantitative | Irreversible      |
| Modeling   | <b>Fujisaki</b> | Timing and Magnitude of Commands | Objective Quantitative  | <b>Reversible</b> |

© Hiroya Fujisaki  
IWSLPB-2009, Kolkata  
14

So, if you see; there is a lot of F<sub>0</sub> modeling and you see that this slides are taken from the Fujisaki F<sub>0</sub> modeling. So, that is why copyright Fujisaki is given. So, 3 approaches

of description and representation of F0 contour. So, I can what I said in beginning that for a n utterance has an F0 movement.

(Refer Slide Time: 04:20)



So, there is a F0 movement. So, these movement of the F0 can be modeled by 3 things one is called labeling which is called Tobi stylization and modeling.

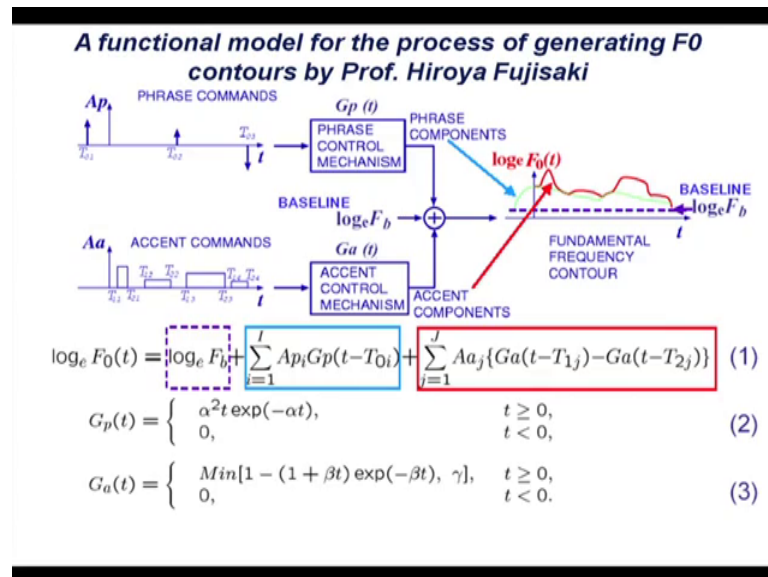
So, I can model generate some model we can generate this contour or I can say I can I can assign some level. So, it is a rising then I can say flat then I can say falling F. So, rising, flat, falling, soft rising, slow rising, slow falling, soft falling all kinds of level I can assign and I can model that F0 things that is called Tobi model.

Similarly, I can do some stylization means I can linearly approximate this F0 contour. So, I can say whole F0 contour is nothing, but a sum of some linear line within this line F0 is fixed. So, this is called piecewise linear approximation or I can generate a mathematical model by which changing the parameter I can generate this contour which is called Fujisaki command response model which we will discuss details. So, whole sentence F0 can be model using 3 model one is called Tobi stylization and modeling

So, Tobi is nothing, but a assigning some label on F0, it is may be in syllable it may be let syllable wise. So, I can say this is rising this is flat this is falling then I can define the pattern of rising it is sharp rising it is flat rising. So, those kind of things is used in Tobi model then stylization I can approximate whole F0 contour is some set of linear line or

linear segment, then I can model that things that is called stylization or I can generate some mathematical model and that model has a some parameter and I can vanish those parameter to generate this F 0 contour that is called generation process Fujisaki generation process modeling. So, we will discuss details on generation process modeling.

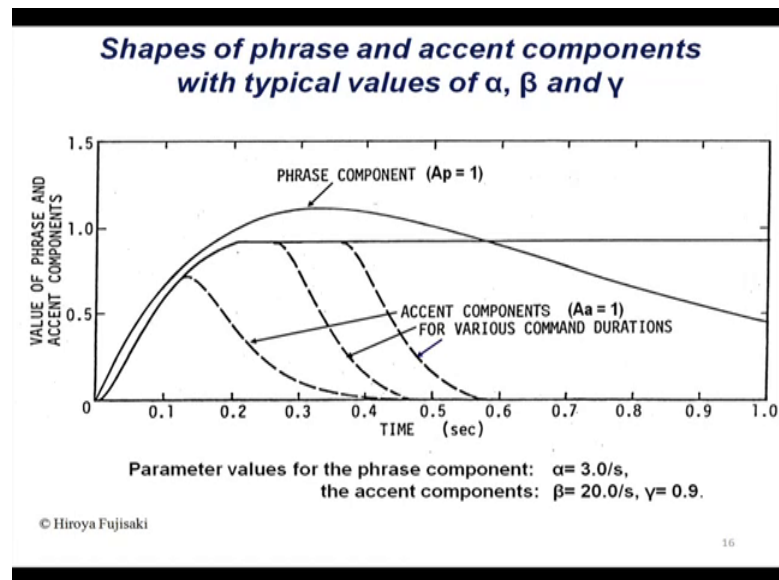
(Refer Slide Time: 06:59)



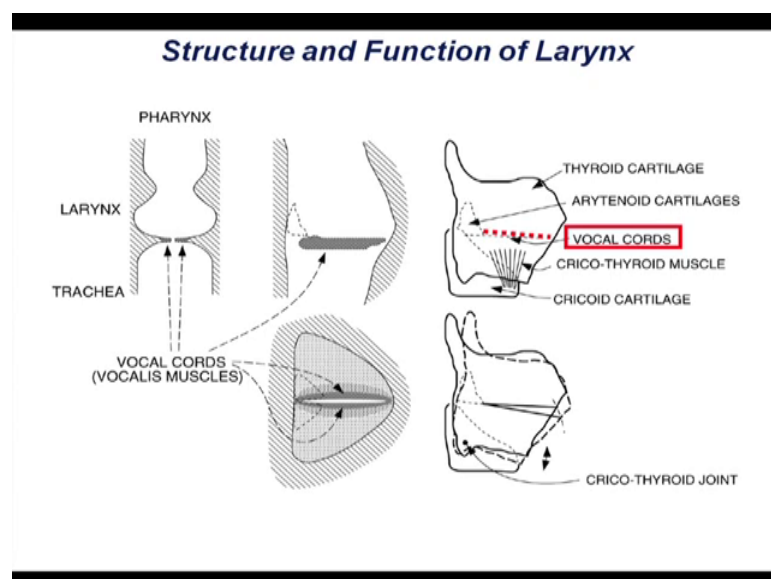
So, this is called Fujisaki generation process F 0 contour modeling that if you remember we have said F 0 has 2 kinds of variation one is called local variation another one is call global variation if you see here the green line is the global variation and red line are local variation. So, the local variation are called accent and global variation are called phrase. So, that is why is called it is command response model which is control by the phrase control and accent control mechanism. So, one is control the phrase command phrase command control the phrase part accent command control the accent part and there is a base line and Fujisaki model mathematical model is this.

So, let us try to derive this mathematical model as or I can describe this model as per the Fujisaki papers at this I will come later on.

(Refer Slide Time: 08:09)



(Refer Slide Time: 08:12)

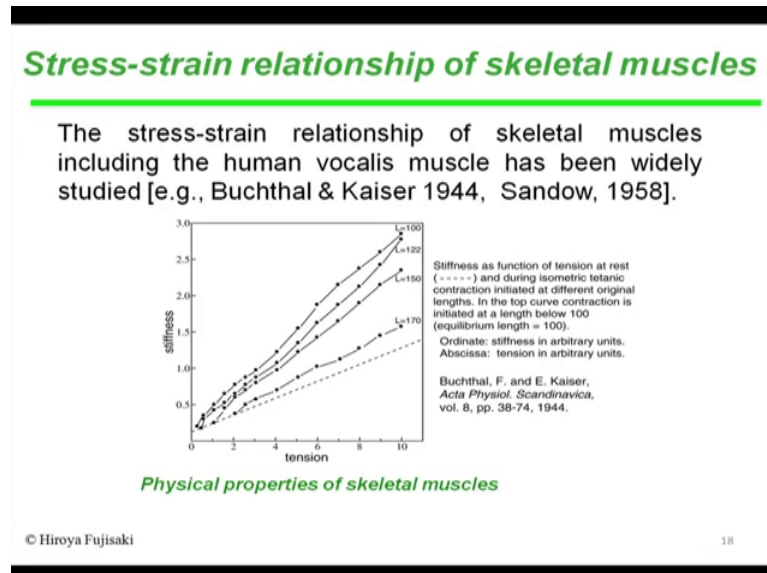


So, if you see that what is; how this are vocal cords are there this is the section of vocal cords. So, vocal cords are here; here there is a vocal cords and if you remember that vocal cords are closed in one end and open in one end this is closing and opening closing and opening and this is housed on a some muscle that muscle control the closing and opening. So, those are the muscle structures.

Now, if you see this muscle this of the vocal cord housed in some muscle of the some bone and the muscle control bone structures where it can move. So, that muscles can

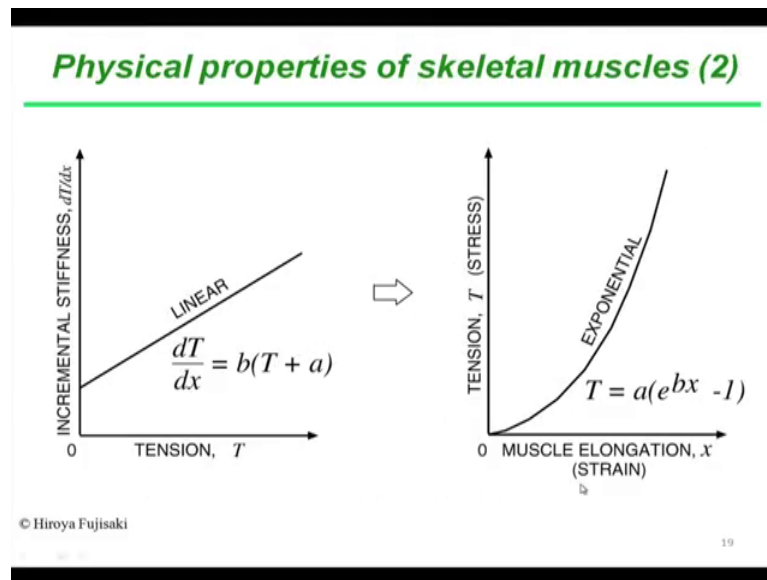
move that bone and that can change the tension of the vocal cords and that create different  $F_0$ . So, details I will describe that create different  $F_0$ .

(Refer Slide Time: 09:19)



So, let us first see stress strain relationship of skeletal muscles that is already studied and that the relationship is there. So, this is called physical property of skeletal muscle the muscles which is the bone which is house the vocal cords are connected by a muscles or you can house by a muscles and those muscles has a property that property is skeletal muscles properties that is already studied. Then physical properties of skeletal muscle then how it is varies if you see this first graph x axis is tension y axis is incremental tension  $dT$  by  $dx$ .

(Refer Slide Time: 09:46)



So,  $dT/dx$  is equal to  $b(T + a)$  where  $T$  is a tension,  $b$  and  $a$  are constant.

(Refer Slide Time: 10:15)

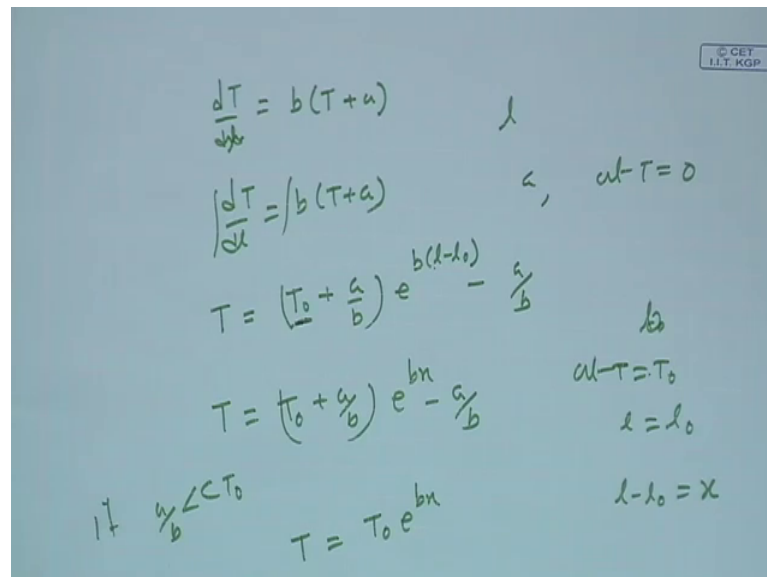
Handwritten notes on a blue background. At the top left, the differential equation  $\frac{dT}{dx} = b(T+a)$  is written. Below it, the variables  $x, T$  are written. To the right, the equation  $T = a(e^{bx} - 1)$  is written. At the bottom, the equation  $T = P(T)$  is written with a large 'P' that has a horizontal line through it, indicating it is crossed out.

So, I can say  $dT/dx$  is equal to  $b(T + a)$  where  $T$  is a tension,  $b$  and  $a$  are constant, then muscle elongation and tension  $T$ . So, I can say how much muscle has to be elongated to produce certain tension that is required. So, that is the relationship.

So, the elongation is  $x$  and tension is  $T$  relationship is  $T$  is equal to  $a(e^{bx} - 1)$ . So, elongation if I elongate the muscles how

much tension will be created that depends on this equation if I increase incremental tension  $dT$  by  $dx$  is equal to  $b(T+a)$ , now from there stress strain skeletal muscle. So,  $dT$  by  $dx$  is equal to  $b(T+a)$  now if I integrate both side. So, what we will get I will get  $T$  is equal to  $T_0$  sorry;  $T$  is.

(Refer Slide Time: 11:44)



Handwritten derivation of the Hill equation for skeletal muscle tension:

$$\frac{dT}{dx} = b(T+a)$$

$$\int \frac{dT}{dx} = \int b(T+a)$$

$$T = \left(T_0 + \frac{a}{b}\right) e^{b(l-l_0)} - \frac{a}{b}$$

$$T = \left(T_0 + \frac{a}{b}\right) e^{bx} - \frac{a}{b}$$

At  $T = T_0$ ,  $l = l_0$ ,  $l - l_0 = x$

$$T = T_0 e^{bx}$$

So, I can say  $dT$  by  $dx$  is equal to  $b(T+a)$  where  $T$  is the tension or I can say  $x$  instead of  $l$ ;  $l$  is the length of the vocalic  $T$  is the tension and  $a$  is the stiffness at  $T$  is equal to 0 then the relationship is  $dT$  by  $dl$  is equal to  $b(T+a)$ . So,  $b$  is a constant  $a$  is a stiffness at  $a$  is the stiffness at  $T$  equal to 0 and  $T$  is the tension  $l$  is the length; length of the vocalic. Now if I integrate both side what I will get I get  $T$  is equal to  $T_0$  plus  $a$  by  $b$  into  $e$  to the power  $b$  into  $l$  minus  $l_0$  minus  $a$  by  $b$  if I integrate both side now  $T$  is equal to I can say  $T_0$  what is  $T_0$  static tension what is  $l_0$  static length and so  $T$  is equal to  $T_0$ . So, at  $T$  equal to  $T_0$  the  $l$  is equal to  $l_0$ . So, I can say  $T_0$  or at static tension is  $T_0$ .

So, static tension is  $T_0$  then I can say  $l$  minus  $l_0$   $l$  minus  $l_0$  is nothing, but a change of length which I can defined as  $x$ . So, I can say  $T_0$  plus  $a$  by  $b$   $e$  to the power  $b$   $x$  minus  $a$  by  $b$ .



(Refer Slide Time: 14:09)

### From vocal cord elongation to tension

Stress-strain relationship in a skeletal muscle

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

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

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

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

When  $T_0 \gg a/b$

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

where  $x = (l - l_0)$

© Hiroya Fujisaki

20

Now if  $a/b$  is much much less than  $T_0$  then I can say  $T$  is equal to  $T_0 e$  to the power  $b \times T_0 e$  to the power  $b \times$  which is there in here which is equation 3.

(Refer Slide Time: 14:15)

### From vocal cord tension to fundamental frequency

Frequency of vibration of an elastic membrane is given by

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

From Eqs. (3) and (4)

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

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

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

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

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.

© Hiroya Fujisaki

21

Now if it is  $T_0 e$  to the power  $b \times$ , then frequency of vibration of elastic membrane  $F_0$  is depends on  $C_0$  into  $T$  by  $\sigma$  to the power half where  $\sigma$  is the density per unit area unit area density per unit area this is relationship between the elastic membrane vibration membrane vibrations.

(Refer Slide Time: 14:23)

$$F_0 = c_0 \left( \frac{I}{\sigma} \right)^{1/2}$$

$$\log_e F_0 = \log_e c_0 \left( \frac{I}{\sigma} \right)^{1/2}$$

$$= \log_e c_0 \left( \frac{T_0}{\sigma} \right)^{1/2} + \log_e e^{\frac{1}{2} b x}$$

$$= \log_e c_0 \left( \frac{T_0}{\sigma} \right)^{1/2} + \frac{1}{2} b x$$

$$\log_e F_0 = \log_e F_b + \left( \frac{1}{2} b x(t) \right)$$

$T = T_0 e^{b x}$   
 $c_0 \left( \frac{T_0}{\sigma} \right)^{1/2}$   
 $\downarrow$   
 $F_b$   
 $x = l - l_0$

Now, if I take log e of  $F_0$  is nothing, but a log e of  $c_0$  into  $T$  by sigma to the power half now I put the value of  $T$  this is  $T$  is equal  $T_0 e$  to the power  $b \times \log e$ ; I can say log e  $c_0 T_0$  by sigma is one term to the power half is one term and  $e$  to the power  $b \times$  and log e that is plus log e;  $e$  to the power  $b \times$ . So, I can say log e  $c_0 T_0$  by sigma to the power half plus log e  $b \times$ . So, I can say an log e  $b \times$  to the power half. So, half will be there. So, root half  $b \times$ .

Ok. So, now, if you see I can say  $c_0 T_0$  by sigma to the power half is nothing, but a constant. So, I can say let's it is defined as  $F_b$ . So, if it is  $F_b$  then; I can say log of  $e F_0$  is nothing, but a log of  $F_b$  plus half of  $b \times$ . So,  $x$  is nothing, but a  $l$  minus  $l_0$  now if it is half of  $b \times$ . So, if  $x$  is time dependent I can write it is  $x T$ . So, I can say the  $F_0$  contour or log  $F_0$  contour one plotted in logarithmic scale as a function of time can be expressed it is the sum of constant baseline this is the constant and term and time varying term a time varying term which will be changed and a constant baseline.

So, if you see here this is the baseline constant baseline  $F_b$  and then the time varying term which is half of  $b \times T$ . So, how this time varying term is generated how this time varying term is generated this time varying term is generated due to the muscle movement or due to the movement while the vocal cord is vocal cord is housed in a muscle stage or you can say bones cage which is connected by a muscle and that movement is defined it  $x T$  what kind of movement it is?

(Refer Slide Time: 18:02)

### 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.

© Hiroya Fujisaki 22

It is a 2 kind of movement one is called rotation around the cricothyroid joint due to; if I see this.

(Refer Slide Time: 18:10)

### Motion of thyroid with two degrees of freedom

Rotation of thyroid by *pars recta* of the cricothyroid muscle

Translation of thyroid by *pars obliqua* of the cricothyroid muscle

© Hiroya Fujisaki RWSLPR-2009, Kolkata 23

This is kind of housing is there. So, there is a 2 kind of movement one is rotation another is translation rotation and translation. So, if you see in here vocal cords I get 2 kind of movement one is translation and that can change the change the x 1 minus 1 0 and 1 is rotation. So, suppose I have; now rope in here if I this end if I rotate the tension on the rope will be increases or if I translate it then also tension will be increases. So, these 2

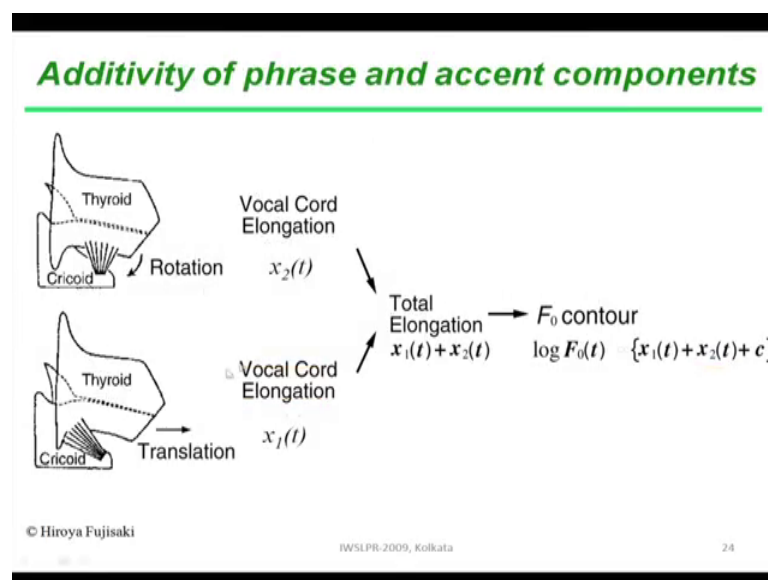
kind of movement increase the tension and once it is increase the tension that change the tension change the  $F_0$  that is change the  $F_0$  contour that is x T part time dependent part.

So, this kind of rotation and translator movement it change the fundamental frequency if you see somebody; say a man is mimicking the female voice how he does it because there is vocal cords mass and tension and there. So, once he talk normally his  $F_0$  may be in around 100 hertz or 120 hertz, but when he mimicking he can change it to somewhere else female voice.

So, how he does it by moving this by practicing this type of movement rotation and translation movement he practice it and he can change the  $F_0$  to that scale anyhow we can change  $F_0$  movement is there is 1 octa movement is not natural if by base  $F_0$  is 80 hertz, I can change up to 160 hertz because that is practice when you are singing when I singing what you are doing we are practicing the movement of the  $F_0$ . So, you are actually rotation and translations are practiced.

How much movement and this rotation and translation can change the  $F_0$  then how this will be used in prosody or how this will be using continuous movement of the  $F_0$  is an important one, so how it is used?

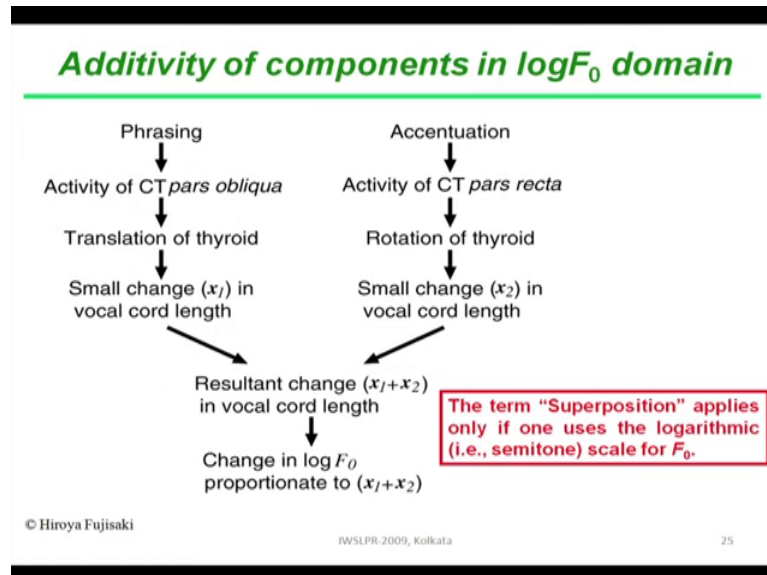
(Refer Slide Time: 20:33)



If you see that rotation is providing a component which  $x_2(t)$  and translation is provide a component  $x_1(t)$ . So, total change is  $x_1(t)$  plus  $x_2(t)$  which is in  $\log F_0$  contour. So, this

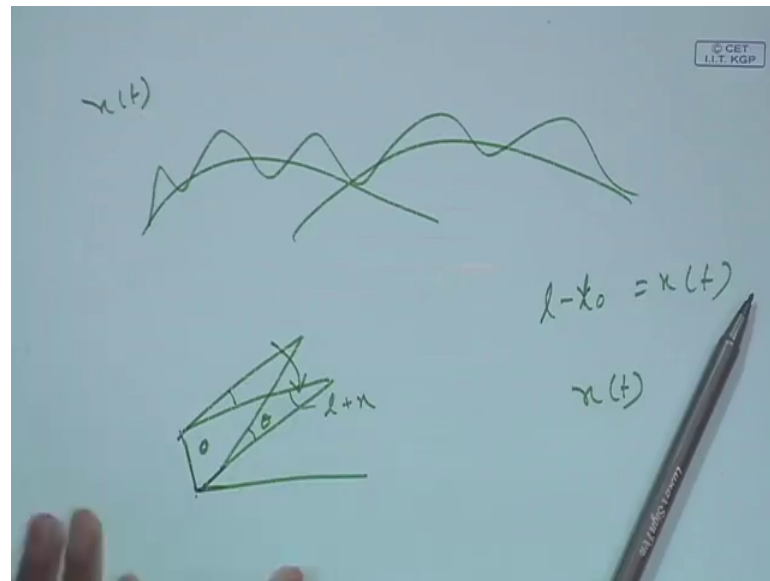
will be additive in  $\log F_0$  contour. So,  $x_1$  and  $x_2$  is added and that change the  $\log F_0$  it is additive.

(Refer Slide Time: 20:59)



Now if I say this relation with that language. So, if I say the phrasing and accentuation. So, there is a local variation and there is a global variation. So, if I say the global variation is fascination and local variation is accentuation accent local accent. So, then I can say accent is control by rotation of thyroid muscle and phrasing is controlled by translation. So, translation mechanism controls the global variation. So, translation effect is time if it is  $x$  t.

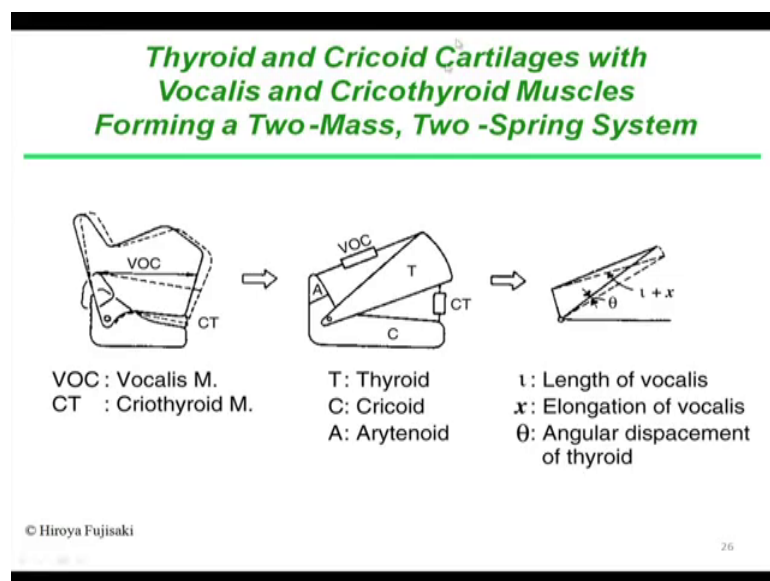
(Refer Slide Time: 21:38)



So, translation will change the  $F_0$  in roughly change the  $F_0$  and rotation will change the local variation.

So, global variation plus local variation in logarithmic  $F_0$  it is a superposition, but in if it is non logarithm, then it is a multiplication. So, the superposition; so, they are super-imposed with normal  $F_0$  with the accent component and phrase component. So, that way we are changing the  $F_0$ . So, translation and rotation details are is there.

(Refer Slide Time: 22:17)

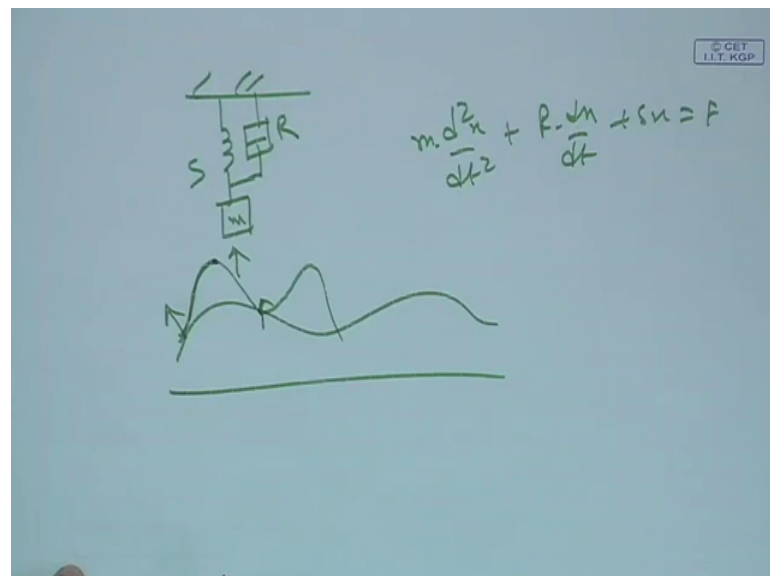


So, you can see this is this kind of system. So, this is there is a membrane. So, it can rotate if CT is this; this is constricted if this portion is constricted CT is constricted and then I can say it is rotatory movement and if this portion is strength, then it is a translatory movement.

So, if I say this is my membrane I am not drawing it correctly. So, this is the lets this is my things. So, if I move it like this way then it is rotation. So, if I move it lets this position then I can say this angle is nothing, but a theta. So, this is rotation and if I translate it then it is l plus x if I move this direction then it is l plus x. So, l is the length of the vocalic x is the extension of elongation of vocalic theta is a angular displacement of the thyroid.

So, if I do the angular displacement there also be a length change if I do the translation there also be a change of length and that is l minus l 0 which is produced by which is represented by x T and this x T is summation of this 2 change in the logarithmic domain one is rotation another is translation and this rotation and translations are nothing. But the spring mass movement say if it is a spring mass movement there is a second order differential equation.

(Refer Slide Time: 23:55)



If you remember that spring mass movement of a body mechanical vibration spring mass movement if you remember that m S and R. So, you know that d 2 m into d 2 x by d T square plus r into d x d T plus S x is equal to force the force which is applied.

So, I can say  $I \ddot{\theta}$  is in term of  $\ddot{\theta} = \frac{k}{I} \theta + \frac{r}{I} \dot{\theta}$  where  $k$  is the stiffness  $r$  is the mechanical resistance and  $I$  is the mass; mass of the muscle. So, mechanical resistance mass and stiffness then I can find out the relationship the movement variation of  $\theta$ . So, I if I solve this second order equation. So, it is nothing, but a exponential solution. So, it is nothing, but a exponential solution. So, I can say  $\theta(T)$  which can be expressed as a constant multiplied by a exponential curve which is represented by minima of  $1 - \frac{1}{\beta T} e^{-\beta T}$  or  $\gamma$ .

So, this is experimentally verified and find out by professor Fujisaki and he can find out the value for  $\beta$  and  $\gamma$  for Japanese language and we have verified it for Bangla language this  $\beta$  and  $\gamma$  is working fine. So, if rotation give me the accent component then I can say  $\theta$  represent the accent component with a multiply by a constant which actually content the amplitude of the; that component.

So, if I see that the rotation angle  $\theta(T)$  is varying like this elongation will be delayed then logarithmic  $F_0$  also will be like varying like this. So, if you see in accent component. So, suppose this is the global component and the accent component it will vary like. So, even if I apply the command in here it will be take time to reach the half and again it will be going down because of elasticity.

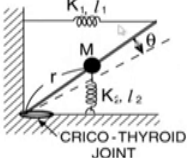
Then again if I apply I command in here it will be again up somewhere else because it require some rising time.



(Refer Slide Time: 26:33)

### Rotation and translation of the thyroid

**ROTATION**

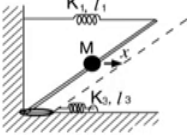


CRICO-THYROID JOINT

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

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

**TRANSLATION**



CRICO-THYROID JOINT

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

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

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

© Hiroya Fujisaki 29

So, this is Fujisaki's slide I have taken, then you can say the translation also same equation differential equation and translation is defined by G p t.

(Refer Slide Time: 26:44)

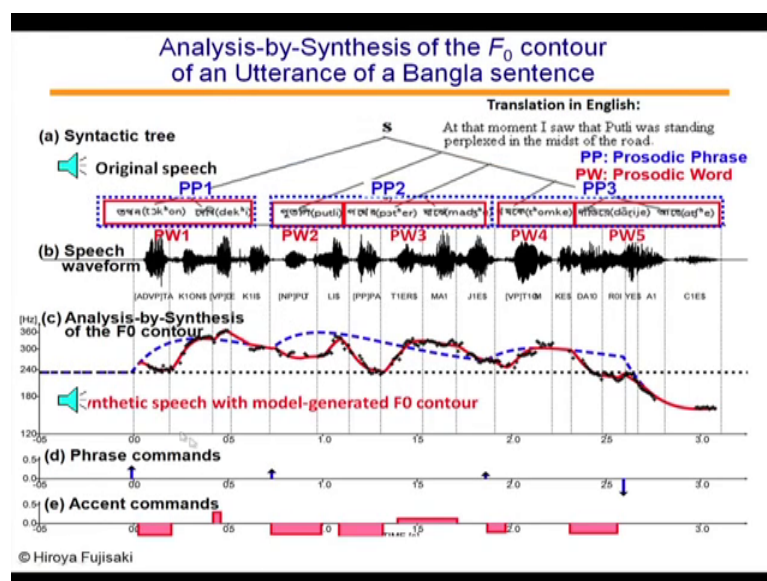
$$\log_e F_0(t) = \log_e F_b + \sum_{i=1}^I A p_i G_p(t - T_{0i}) + \sum_{j=1}^J A a_j \{G_a(t - T_{1j}) - G_a(t - T_{2j})\} \quad (1)$$

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

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

So, you can say the total equation is like this well you can say i equal to 1 to capital I what is this; this capital I is nothing, but a number of phrase command.

(Refer Slide Time: 26:57)



So, suppose I have a sentence I can say this is here this is the Bengali sentence if you see this dotted line are original  $F_0$  extracted from the speech. So, this is align actually this  $F_0$  contour is aligned with the original speech or Bangla sentence one bangle sentence is there if you see the Bengali sentence are written.

Now, if you see the black dots are original red one if you see the red one are generated  $F_0$  contour and blue ones are the phrase command blue one is the phrase contour and this accent command is created after the phrase command is placed then the accent command is put this red line is generated if you see the red line almost follow the black dots. So, it is almost possible to model the  $F_0$  contour using this phrase command and accent. So, if you see there is a number of phrase.

So, here I represent the number of phrase. So, I equal to 1 to capital I. So, if there is a 3 phrase in here if you see how many phrase are there 1, 2, 3. So, 3 phrases are there one 2 3 and four last one is negative amplitude is negative. So, it will be negative direction. So, 3 phrases are there 1, sorry, 4; 1 to 4 A p i are the amplitude of the phrase command and this is the variation then accent command A j equal to 1 to capital J number of accent command how many number of accents command are there; 1, 2, 3, 4, 5, 6, 7. So, there may be a 7 accent command and A; this is nothing, but a accent command amplitude. Now, if you see this one the red line is completely follow the green line dotted line. So, I

can say it is possible to completely generate the original F 0 contour using this phrase command and accent command.

Let us today I stop here to this lecture I stop here and tomorrow I will. So, I will play this voice sound. So, let I arrange this playing of the voice sound I show you how close this original speech and synthesized speech.

Thank you.