Ann.Bull.RILP(1987)No.21, 165-175

A NOTE ON THE PHYSIOLOGICAL AND PHYSICAL BASIS FOR THE PHRASE AND ACCENT COMPONENTS IN THE VOICE FUNDAMENTAL FREQUENCY CONTOUR*

Hiroya Fujisaki**

1. INTRODUCTION

It is a common observation of many researchers, including the present author, that the contour of the fundamental frequency of the voice (henceforth the F_0 contour) of an utterance in many spoken languages, such as Japanese, Dutch, Swedish, English, Italian, French, etc., is characterized by the presence of more or less local, relatively fast, rise-fall components superposed on a global, relatively slow, declining baseline. This is true for a very short utterance such as a word uttered in isolation, as well as for longer utterances such as spoken sentences, as illustrated in Figs. 1 and 2.

The physiological mechanism for generating such components, however, is not well understood. In an earlier article on the analysis of pitch control in singing[1)], the author presented an explanation of a possible mechanism for producing a trajectory of F_0 that looks like a step response of a second-order linear system. The purpose of this paper is to present a further explanation of the possible physiological and physical mechanisms that give rise to the two types of components in the F_0 contour.

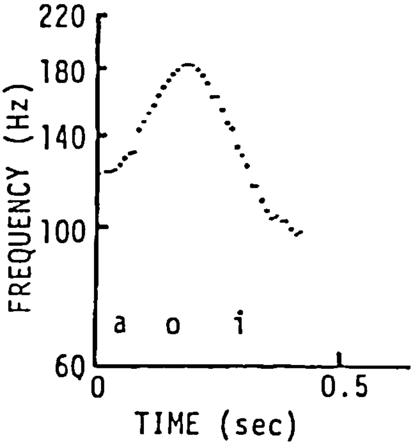

Fig. 1. An example of a measured F_0 contour of the word /aoi/ ("blue") in the Tokyo dialect of Japanese. The dots indicate fundamental frequencies extracted at intervals of 10 msec.

---

* A version of this paper was presented at the fifth Vocal Fold Physiology Conference, Tokyo, January 1987.

** Department of Electronic Engineering, Faculty of Engineering, University of Tokyo

2. EXPERIMENTAL DATA AND THEIR ANALYSIS

The analysis of experimentally observed F_0 contours of a large number of Japanese utterances, conducted by the author and his co-workers2),3), has revealed that, if we plot the F_0 contour on a logarithmic scale of fundamental frequency versus time, it can always be approximated very closely by the sum of two types of components: 1) those usually accompanying prosodic words, and representing the local rise and fall of F_0 due to lexical word accent, and 2) those corresponding to larger syntactic units such as phrases, clauses, and sentences, and representing the global rise and decay of the whole contour. Our studies4),5) have revealed further that the shapes of these two types of components can be approximated respectively by the step response of a second-order linear system with a comparatively short rise time, and by the impulse response of another second-order linear system with a comparatively long rise time. From these observations we have proposed a functional model for the control mechanism of F_0, which generates an F_0 contour from two kinds of commands, viz. the accent commands having idealized step-wise waveforms and the phrase commands having idealized impulse waveforms, as illustrated by the block diagram of Fig. 3. Although the two subsystems —the accent control mechanism and the phrase control

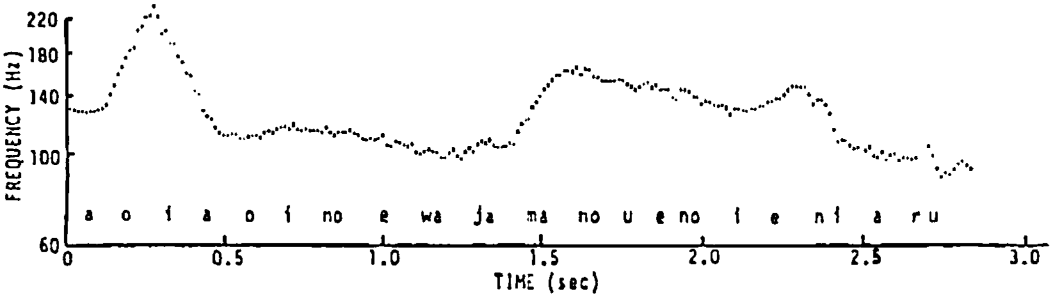

Fig. 2. An example of a measured F_0 contour of the declarative sentence /aoiaoinoewajamanouenoieniaru/("The picture of the blue hollyhock is in a house on top of the hill.") of Japanese.

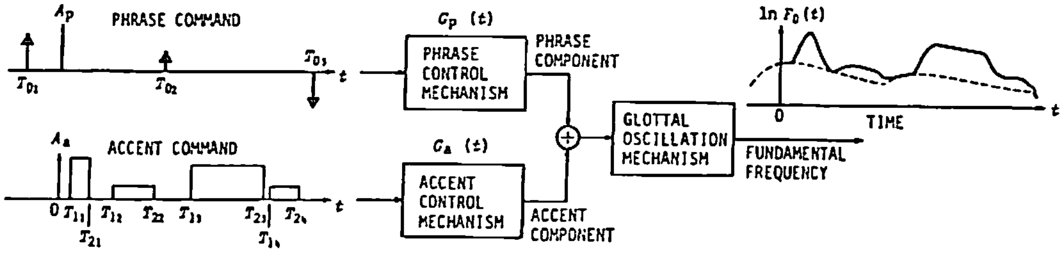

Fig. 3. A functional model for the process of generating sentence F_0 contours.

mechanism— of Fig. 3 may not be exactly critically-damped, our analysis indicates that the assumption of critical damping is practically valid for both systems. The behavior of the model can then be expressed by

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

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

$$\text{and} \quad G_a\,(t) \begin{cases} = 1 - (1 + \beta\ t) \exp(-\beta\ t), & \text{for } t \geq 0, \\ = 0, & \text{for } t < 0, \end{cases} \tag{3}$$

F_{min}: asymptotic value of fundamental frequency in the absence of accent components,
I : number of phrase commands,
J : number of accent commands,
A_{pi} : magnitude of the ith phrase command,
A_{aj} : amplitude of the jth accent command,
T_{0i} : timing of the ith phrase command,
T_{1j} : onset of the jth accent command,
T_{2j} : end of the jth accent command,
α : natural angular frequency of the phrase control mechanism,
β : natural angular frequency of the accent control mechanism.

By the technique of Analysis-by-Synthesis, it is possible to decompose a given F_0 contour into its constituents, i.e., the phrase components and the accent components, and estimate the magnitude and timing of their underlying commands by deconvolution, as illustrated by the example in Fig. 4. In particular, the timings of these commands are found to be closely related to the linguistic contents of the utterance. The accent command is found to start at 40~50 msec before the segmental onset of a subjectively "high" mora, and to end also at 40~50 msec before the segmental ending of a "high" mora. The phrase command, on the other hand, is found to be located approximately 200 msec before the onset of an utterance, and also before a major syntactic boundary such as the boundary between the subject phrase and the predicate phrase. In general, the phrase command is largest at the sentence-initial position, and is smaller at sentence-medial positions, so that the overall shape of an F_0 contour, disregarding local rises and falls due to accent components, shows a decay from the onset toward the end of the whole utterance. There are cases, however, where pragmatic factors call for the occurrence of a large phrase command at a sentence-medial position, which may be the reason why some people deny the presence of F_0 declination in running speech. Results of our analysis clearly indicate, however, that the existence of phrase components, each having the shape of a decaying response, is responsible for the so-called F_0 declination.

Our analysis also shows that the rate of rise, as indicated by the natural angular frequency β of the accent component, is approximately equal to 20/sec, while that of the phrase

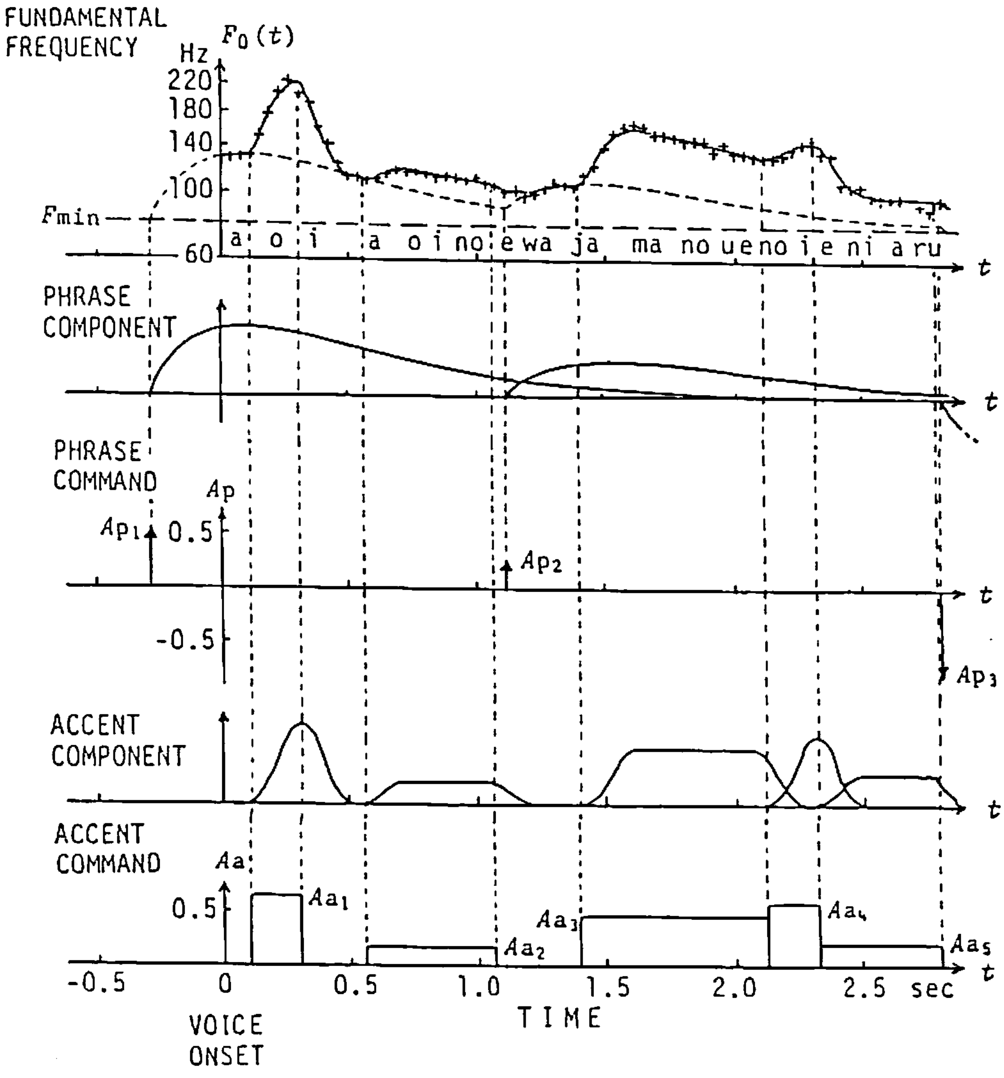

Fig. 4. Analysis-by-Synthesis of an F_0 contour of the Japanese declarative sentence: /aoiaoinoewajamanouenoieniaru/. The figure illustrates the optimum decomposition of a given F_0 contour into the phrase and accent components, and also shows the underlying commands for these components.

component, as indicated by the natural angular frequency α, is approximately equal to 3/sec. The variations in the values of these natural frequencies are found to be quite small from utterance to utterance, as well as from one individual to another.

3. POSSIBLE MECHANISMS

Since the approximation by the above-mentioned formula applies very well for all the observed F_0 contours, we are led to believe that the approximation is not fortuitous, but the contour of $\ln F_0(t)$ actually reflects the dynamic behaviors of some components of the laryngeal structure whose characteristics may be considered as those of second-order linear systems. In an effort to provide an explanation for the F_0 transition in singing, the following hypotheses have been presented:

Hypothesis (1). The logarithm of the fundamental frequency varies linearly with the strain, i. e. elongation, of the vocal cord.

Hypothesis (2). The strain of the vocal cord reflects the mechanical motion of a mass element coupled with some stiffness and viscous resistance elements which, to a first-order approximation, can be regarded as a second-order linear system.

We shall briefly review the theoretical considerations presented earlier, in order to facilitate the understanding of further discussion.

In order to test the validity of Hypothesis (1), we shall look at the stress-strain relationship of skeletal muscles for which there already exist a number of measurements. Although we do not have data from the human vocalis muscle *in vivo*, the following experimental relationship is known to apply between the tension T and the elongation x of skeletal muscles in general (for example, Buchthal and Kaiser6); Sandow7)):

$$T = a\ (e^{bx} - 1). \tag{4}$$

Here we deal with the elastic properties of the vocalis muscle, and regard x as its elongation caused by some laryngeal mechanisms. If $e^{bx} \gg 1$, the above equation can be approximated by

$$T = a\ e^{bx} \tag{5}$$

On the other hand, the frequency of vibration of an elastic membrane varies in proportion to the square root of its tension (for example, Slater and Frank8)). Since the vocal fold can also be regarded as an elastic membrane to a first-order approximation, the frequency of its vibration F_0 can be given by

$$F_0 = c_0\sqrt{T} \tag{6}$$

From Eqs. (5) and (6) we obtain

$$\ln F_0 = \frac{b}{2} x + \ln(\sqrt{a} \cdot c_0) \qquad (7)$$

where, strictly speaking, c_0 also varies slightly with x, but the overall dependency of $\ln F_0$ on x is primarily determined by the first term on the right-hand side of Eq.(7). Equation (7) shows the static relationship between the vocal cord strain and the logarithm of the fundamental frequency as predicted by Hypothesis (1).

Assuming that the static relationship actually holds, we shall next turn to Hypothesis (2) and seek evidence for the dynamic properties of some elements of the laryngeal structure that will produce dynamic changes in the elongation x of the vocal cord in such a way as we observe in the F_0 contour.

The role of the cricothyroid muscle in controlling F_0 has been known widely, and an explanation has been presented in our aforementioned article for the possible mechanism, i. e., rotation of the thyroid cartilage around the cricothyroid joint. Since a small angular displacement of the thyroid cartilage produces a proportionate small elongation of the vocalis muscle, the dynamic behavior of the rotational system, consisting of one mass element (the thyroid) supported by two stiffness elements (the vocalis and the cricothyroid muscles), will be directly reflected in the contour of $\ln F_0(t)$. This can explain, however, only one of the two components of the F_0 contour. Physiological studies of EMG activity suggest that thyroid rotation due to cricothyroid activity is related to the accent components. In order to be able to account for both the accent components and the phrase components, we need independent movements along two degrees of freedom of motion which would both affect the vocal cord length. If the two movements both contribute to the elongation, the resultant strain will be the sum of the strains due to each one of the movements, and the consequences of these two movements on the F_0 contour will also be additive.

Although the anatomical and physiological observations are quite limited on the actual movements of the thyroid, there exists at least one reference9) which, on the basis of radiographic observations, suggests the existence of two degrees of freedom of motion for the thyroid cartilage as shown in Fig. 5.

(1) rotation around the cricothyroid joint due to the activity of the *pars recta* of the cricothyroid muscle, and

(2) forward translation due to the activity of the *pars obliqua* of the cricothyroid muscle.

Assuming both the rotation and the forward/backward translation to be very small, the resultant strain of the vocal cord can be regarded as the sum of strains due to each one of the causes, as shown in Fig. 6.

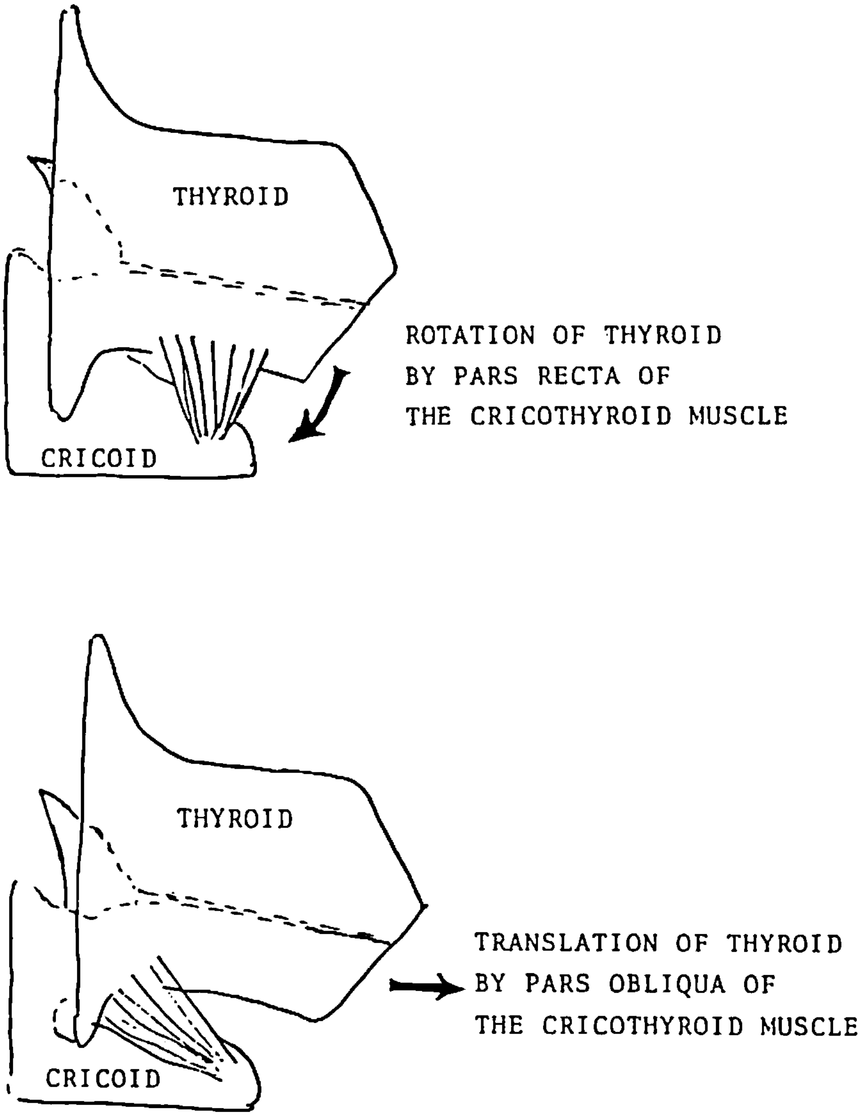

Fig. 5. The roles of pars recta and pars obliqua of the cricothyroid muscle in rotating and tranlating the thyroid cartilage.

The schematic drawing in Fig. 6 represents an oversimplification of the actual mechanism. By considering only two stiffness elements at a time both in rotation and in translation, and taking the radius of rotation r of the thyroid into account, we obtain Fig. 7.

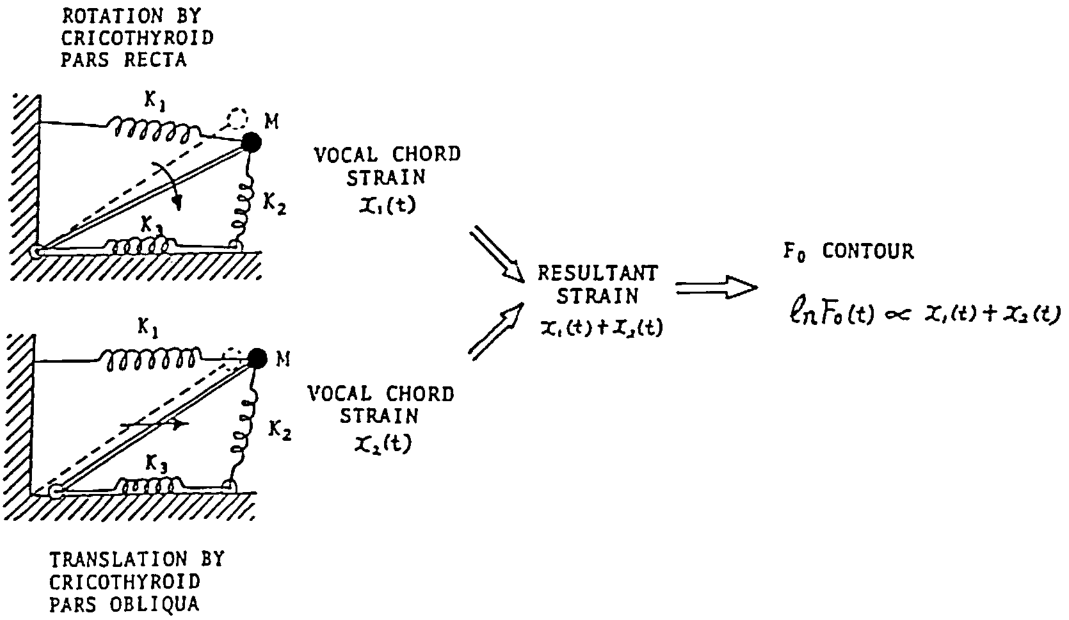

Fig. 6. Mechanical equivalent circuit for the rotation and the translation of the thyroid cartilage against the cricoid cartilage.

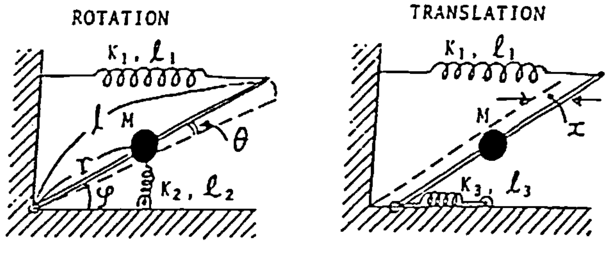

Fig. 7. Further simplification of thyroid movements by the crico-thyroid muscle.

The equation of motion for rotation can be expressed by

$$M r^2 \ddot{\theta} + R \dot{\theta} + \left\{ K_1 l^2 \sin^2\varphi + K_2 r^2 \cos^2\varphi \right\} \theta = \tau(t) \qquad (8)$$

and the natural angular frequency is given by

$$\beta_r = \sqrt{\frac{K_1 \ell^2 \sin^2\varphi + K_2 r^2 \cos^2\varphi}{M r^2}} \tag{9}$$

where the symbols indicate
θ : small angular displacement of thyroid,
M : mass of the thyroid,
R : viscous loss related to rotation,
r : radius of rotation of the thyroid around the cricothyroid joint,
ℓ : distance between the cricothyroid joint and the anterior end of the vocalis muscle,
K_1 : incremental stiffness of the vocalis muscle,
K_2 : incremental stiffness of *pars recta* of the cricothyroid muscle,
φ : angle between the thyroid and the cricoid,
$\tau(t)$: torque generated by a change in the contractile force of *pars recta* of the cricothyroid muscle.

On the other hand, the equation of motion for the translation can be expressed by

$$M\ddot{x} + R'\dot{x} + (K_1 + K_3)x = f(t) \tag{10}$$

and the natural angular frequency is given by

$$\beta_t = \sqrt{\frac{K_1 + K_3}{M}} \tag{11}$$

where the symbols represent
x : small translation of the thyroid,
K_3 : incremental stiffness of *pars obliqua* of the cricothyroid muscle,
R' : viscous loss related to translation,
$f(t)$: change in the contractile force of *pars obliqua* of the cricothyroid muscle.

The ratio of natural angular frequencies for rotation and for translation is thus given by

$$\beta_r/\beta_t = \sqrt{\frac{K_1 \sin^2\varphi (\ell^2/r^2) + K_2 \cos^2\varphi}{K_1 + K_3}} \tag{12}$$

If the restoring forces against a small displacement of the thyroid by the two balancing stiffness elements are equal both in rotation and in translation,

$$\beta_r/\beta_t = \left(\frac{\ell}{r}\right)\sin\varphi \qquad (13)$$

Equations (12) and (13) clearly show that the response time can be different for rotation and for translation, even if both movements involve the same mass element. Since no numerical data is available on the incremental stiffness of the related muscles nor on the mass and the moment of inertia of the thyroid, the actual ratio of the two β's cannot be obtained. Judging from the shape of the thyroid cartilage, however, it would be natural to assume that the ratio β_r / β_t could be of the order of 2 or 3. It thus seems to be natural to associate thyroid rotation with the accent component, and associate thyroid translation with the phrase component. Whether or not the 7 to 1 ratio commonly observed for β / α in the analysis of actual F_0 contours can be fully accounted for by the suggested mechanism, calls for further study. It also remains to be shown, that the two parts of the cricothyroid muscle actually work independently from each other and differ in the temporal pattern of their activities, *pars recta* being responsible for producing the accent components, and *pars obliqua* being responsible for generating the phrase components.

4. CONCLUDING REMARKS

An explanation has been presented for the possible mechanisms of generating the accent and the phrase components of the F_0 contour. It has been suggested that the two components might correspond to two different ways of producing vocal cord strain, by using two degrees of freedom of motion of the laryngeal structure, especially the thyroid. Calculations based on a simplified model of the glottal structure have indicated the possible difference in the natural angular frequencies of rotation and translation of the thyroid. Although the present model is based only on acoustic analysis and not on physiological observations, it is hoped that the model would at least provoke an interest in the search for the physiological mechanisms to express linguistic information in the form of the F_0 contour.

REFERENCES

1) Fujisaki, H., M. Tatsumi and N. Higuchi: Analysis of pitch control in singing. Chapter 23 in Stevens and Hirano eds., *Vocal Fold Physiology*, Tokyo: University of Tokyo Press, 1981.
2) Fujisaki, H. and S. Nagashima: A model for synthesis of pitch contours of connected speech. *Annual Report, Engineering Research Institute, Faculty of Engineering*, University of Tokyo 28, 53-60, 1969.
3) Fujisaki, H. and H. Sudo: A model for the generation of fundamental frequency contours of Japanese word accent. *Journal of the Acoustical Society of Japan* 27, 445-453, 1971.

4) Fujisaki, H. and K. Hirose: Modeling the dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation. *Preprints of Papers, Working Group on Intonation, The XIIIth International Congress of Linguists*, Tokyo, 57-70, 1982.
5) Fujisaki, H. and K. Hirose: Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan*, (E) 5, 233-242, 1984.
6) Buchthal, F. and E. Kaiser: Factors determining tension development in skeletal muscle. *Acta Physiologica Scandinavica* 8, 38-74, 1944.
7) Sandow, A: A theory of active state mechanisms in isometric muscular contraction. *Science* 127, 760-762, 1958.
8) Slater, J. C. and N. H. Frank: *Introduction to Theoretical Physics*. New York: McGraw-Hill, 1933.
9) Fink, B. R. and R. J. Demarest: *Laryngeal Biomechanics*. Cambridge, Mass., Harvard University Press, 1978.