The Roles of Physiology, Physics and Mathematics in Modeling Prosodic Features of Speech

Hiroya Fujisaki

Professor Emeritus, The University of Tokyo, Tokyo, Japan
fujisaki@alum.mit.edu

Abstract

This paper presents the author's view on prosody, information, and models, as well as on the roles of physiology, physics and mathematics in modeling, and describes the theoretical and experimental bases of the command-response model for the mechanisms of F_0 contour generation, which has been extensively used in the analysis and synthesis of F_0 contours of utterances of various languages. Although the model represents only those factors that are inherent to the control mechanism of F_0, it allows one to identify those factors that carry communicative functions of speech as input commands and as parameters of the mechanism.

1. Introduction

In this paper, prosody is defined as the organization of various linguistic units into an utterance, or a coherent group of utterances in the process of speech production, and is represented by features of speech that are primarily (but not exclusively) related to the voice source, such as fundamental frequency (henceforth F_0), intensity, voice quality, speech rate, as well as occurrence and duration of pauses [1]. Thus these prosodic features of speech primarily carry linguistic information, but at the same time carry two other types of information: paralinguistic and nonlinguistic [1]. The former is concerned with such factors as the intention and the attitude of the speaker, and is deliberately added to or modifies the linguistic information of the message, while the latter is concerned with such factors as physical and mental states of the speaker including sex, age, physical size of the vocal tract as well as the vocal cord, health, and emotion, and is generally not under conscious control of the speaker, though it is possible to consciously simulate these factors, as it is done in the case of acting, or in the case when the speaker intends to show/exaggerate/suppress/feign his/her emotion. Thus the linguistic information is encoded into the message prior to utterance planning, but the other two types of information are utilized, either consciously or unconsciously, mainly at later stages of utterance planning and speech sound production [1]. The communicative function of prosody, in the broad sense, involves the transmission of all three types of information, regardless of whether the speaker intends to communicate them or not.

It is necessary to clarify here what is meant by a model and modeling in this paper. Since 1969, the present author has been using the word 'model' to indicate a formulation of the essential characteristics of the structure of a mechanism or the function of a process, in the context of speech production and speech perception, and the word 'modeling' to mean the act of constructing such a model [2]. In the case of speech production, the mechanism has to be driven by an external force as its input, and generates the speech signal as its output. Thus the working of a model can be described or approximated by mathematical functions representing the causal relationship between the input to the model and the output from the model. This definition will be adopted in the current paper. However, since early 70s, the word 'model' has also been used to mean a formal description of a set of rules and constraints, or the elements and structures in a theory of generative grammar. In this case, a model is described in terms of symbolic rules for the relationships between discrete units constituting both the input and the output of some grammatical operations.

In the current paper, we shall use the word 'model' in the first sense. By adopting this definition, the linguistic information is not a part of the model, but rather the input to the model, and the paralinguistic information is considered to introduce certain modifications/perturbations to the input, while nonlinguistic information is considered to affect mainly certain parameters of the model.

Furthermore, the author sets the following four requirements for the characteristics of a model:

1. Objective – it can be derived and confirmed by objective methods.
2. Quantitative – its function can be expressed and tested in quantitative terms.
3. Generative – it can (re-)generate the entire process.
4. Explicable – it can be explained in terms of its underlying mechanism.

Although it is desirable to discuss the possibility of modeling the entire process of generating prosodic features of speech, it is beyond the scope of this paper. Instead, this paper will deal with only the process concerned with the generation of the contour of the fundamental frequency of voice (henceforth the F_0 contour).

2. Roles of physiology, physics and mathematics in formulating the model

In order to serve as channels for communication of all three types of information, prosodic features must be realized through the articulatory mechanisms. (In this paper, the term 'articulatory' is meant to cover both phonation and articulation in its narrow sense.) The functions of these mechanisms are based on the physiological and physical properties of their components. Hence a proper formulation of a model, in the author's view, should be based on, and should conform to, the physiological and physical properties of the articulatory mechanism, regardless of whether the word 'articulatory' is explicit or not.

What is then the role of mathematics? In the author's view, mathematics here serves as the tool for quantitative description of physical phenomena, say the vibration of the vocal folds, as well as the motion of the thyroid cartilage within the larynx. Since mathematical formulation of physical phenomena often

requires a certain degree of simplification and approximation of the real world events, one might think that there can be many alternative ways of mathematical formulation. For instance, one can model the orbit of the earth around the sun either by a circle or by an ellipse. However, a circular model fails to explain many astronomical observations that can be explained by an elliptic model. Thus there can be one and only one model that truly conforms to the physical reality. With an elliptic model, one can represent the earth's orbit with a very high degree of accuracy if we assign proper values for its parameters, but one can never do so if we use a circular model.

3. Physiology and physics of F_0 contour generation and their mathematical formulations

3.1. Physiology and physics of F_0 determination

In the case of a normal speaker, the vocal folds are physically a pair of symmetrical elastic membranes, and the fundamental frequency of speech is the frequency of their normal mode of vibration when longitudinal tension is applied across the glottis. The tension is the primary factor determining the frequency of vibration.

The stress-strain relationship of skeletal muscles including the human vocalis muscle has been widely studied [3, 4]. Figure 1 shows the earliest published data on the relationship between tension and stiffness [3].

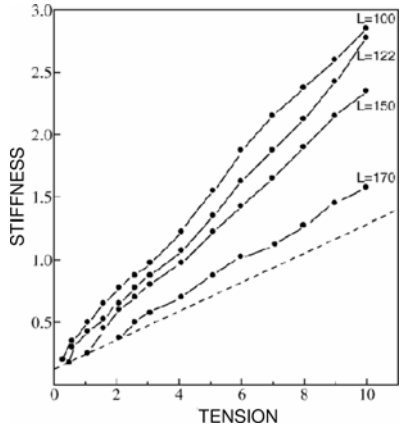

Figure 1*: Stiffness as function of tension at rest (- - - -) and during isometric tetanic contraction initiated at different original length. In the top curve contraction is initiated at a length below 100 (equilibrium length = 100). Ordinate: stiffness in arbitrary units. Abscissa: tension in arbitrary units. (From [3])*

The data shown in Figure 1 indicate the existence of a very good linear relationship between tension and stiffness over a wide range of values, and can be approximated quite well by the following equation:

$$dT / dl = a + bT , \quad (1)$$

where T indicates the tension, l indicates the length of the muscle, and a indicates the stiffness at $T = 0$. This leads to the stress-strain relationship

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

where T_0 indicates the static tension applied to the vocal cord, and l_0 indicates its length at $T = T_0$. When $T_0 >> a / b$, Equation (2) can be approximated by

$$T = T_0 \exp(bx), \quad (3)$$

where x indicates the change in vocal cord length when T is changed from T_0.

On the other hand, the fundamental frequency F_0 of vibration of an elastic membrane is given by

$$F_0 = c_0\sqrt{T/\sigma}, \quad (4)$$

where σ is the density per unit area of the membrane and c_0 is a constant inversely proportional to the size of the membrane. From Equations (3) and (4) we obtain

$$\log_e F_0 = \log_e\{c_0\sqrt{T_0/\sigma}\} + (b/2)x. \quad (5)$$

Strictly speaking, the first term varies slightly with x, but the overall dependency of $\log_e F_0$ on x is primarily determined by the second term on the right hand side. This linear relationship was confirmed for sustained phonation by an experiment in which a stereoendoscope was used to measure the length of the vibrating part of the vocal cord [5], and will hold also when x is time-varying. Thus we can represent $\log_e F_0(t)$ as the sum of a constant term and a time-varying term, such that

$$\log_e F_0(t) = \log_e F_b + (b/2)x(t), \quad (6)$$

where the constant $c_0\sqrt{T_0/\sigma}$ in Equation (5) is rewritten as F_b to indicate the existence of a baseline value of F_0 to which the time-varying term is added when the logarithmic scale is adopted for $F_0(t)$. It is to be noted, however, that the first term ($\log_e F_b$) can be regarded to be approximately constant only as long as the speaker maintains the same speaking style and emotional state. In fact, F_b is found to become appreciably higher when the speaker is excited than when he/she is not.

3.2. Role of cricothyroid muscles in changing F_0

Analysis of the laryngeal structure suggests that the movement of the thyroid cartilage relative to the cricoid cartilage has two degrees of freedom [6, 7]. One is horizontal translation due presumably to the activity of *pars obliqua* of the cricothyroid muscle (henceforth CT); the other is rotation around the cricothyroid joint due to the activity of *pars recta* of the cricothyroid muscle, as illustrated by Figure 2.

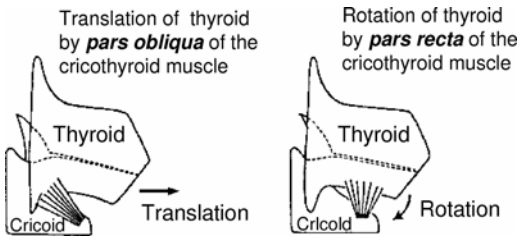

Figure 2: *The roles of pars obliqua and pars recta of the cricothyroid muscle in translating and rotating the thyroid cartilage.*

The translation and the rotation of the thyroid can be represented by separate second-order systems as shown in Figure 3, and both cause small changes in vocal cord length. An instantaneous activity of *pars obliqua* of the CT, contributing to thyroid translation, causes an incremental change $x_1(t)$, while a sudden increase or decrease in the

activity of *pars recta* of CT, contributing to thyroid rotation, causes an incremental change $x_2(t)$ in vocal cord length.

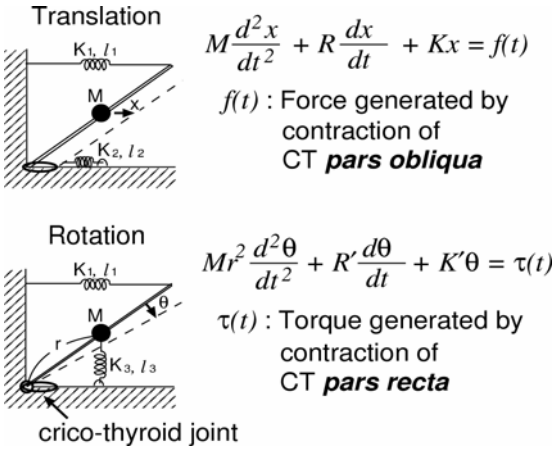

Figure 3: *Equations of translation and rotation of the thyroid cartilage.*

The resultant change is obviously the sum of these two changes, as long as the two movements are small and can be considered independent from each other. In this case, Equation (6) can be rewritten as

$$\log_e F_0(t) = \log_e F_b + (b/2)\{x_1(t) + x_2(t)\}, \qquad (7)$$

which means that the time-varying component of $\log_e F_0$(t) can be represented by the sum of two time-varying components [8]. Since the translational movement of the thyroid cartilage has a much larger time constant than the rotational movement, the former is used to indicate global phenomena such as phrasing, while the latter is used to indicate local phenomena such as word accent or syllable tone.

3.3 Polarity of local components

The foregoing analysis of physiological and physical mechanisms for controlling $F_0(t)$ provides a basis for the command-response model, proposed by the present author, for languages with only positive local components [2, 9]. In this case, a rapid increase in the activity of CT *pars recta* for a certain time interval is represented by a positive pedestal function and named 'accent command,' while a sudden activity of CT *pars obliqua* over a shorter time interval as compared to the time constant of the translational mechanism is represented by an impulse function and named 'phrase command.' The resulting changes in $\log_e F_0(t)$ caused by these commands are called 'accent component' and 'phrase component,' respectively. It is to be noted that the lowering of $\log_e F_0$ in this case occurs due to the sudden decrease in the activity of CT *pars recta*, and does not require an increase in the activity of other muscles. For the rest of the paper, we shall use the word 'F_0 contour' to indicate $\log_e F_0(t)$.

Analysis of F_0 contours of several languages including Mandarin, Thai and Swedish, however, indicates that the local components (associated with tones in the case of Mandarin and Thai) are not always positive but can be both positive and negative. In other words, it is necessary in these languages to posit commands of both positive and negative polarities for the local components, the latter causing active lowering of $\log_e F_0$ below the phrase component.

3.4 Role of extrinsic laryngeal muscles

Although several hypotheses have already been presented on the possible mechanisms for the active lowering of F_0, none seems to be satisfactory since these hypotheses do not take into account the activities of muscles that are directly connected to the thyroid cartilage and are antagonistic to CT *pars recta* in rotating the thyroid cartilage in the opposite direction.

Several EMG studies have shown that the sternohyoid (henceforth SH) muscle is active when the F_0 is lowered in Mandarin [10, 11], the five tones of Thai [12] as well as of the grave accent of Swedish [13], but the mechanism itself has not been made clear since SH is not directly attached to the thyroid cartilage, whose movement is essential in changing the length and hence the tension of the vocal cord.

On the basis of an earlier study on the production of tones of Thai, the present author suggested the active role of the thyrohyoid (henceforth TH) muscle in F_0 lowering in these languages [14]. Figure 4 shows the relationship between the hyoid bone, thyoid and cricoid cartilages, and TH in their lateral and frontal views, and Figure 5 shows their relationships with three other muscles: VOC (thyrovocalis muscle), CT, and SH.

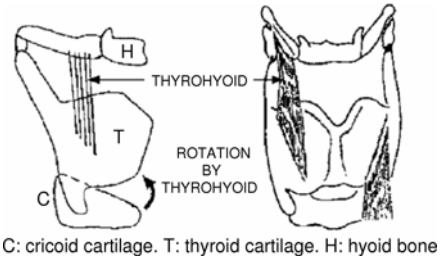

Figure 4*: Role of thyrohyoid in laryngeal control.*

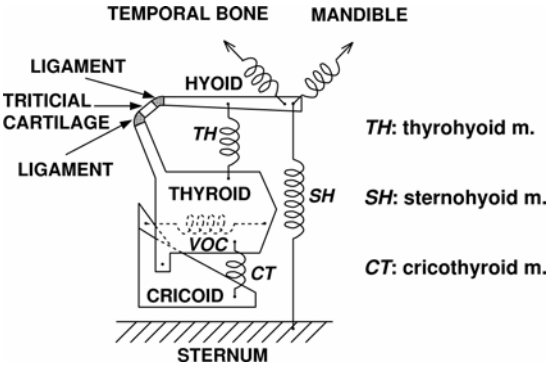

Figure 5: *Mechanism of F_0 lowering by activities of TH and SH.*

The activity of SH stabilizes the position of the hyoid bone, while the activity (hence contraction) of TH causes rotation of the thyroid cartilage around the crico-thyroid joint, in a direction that is opposite to the direction of rotation when CT is active, thus reducing the length of the vocal cord and thereby reducing its tension, and eventually lowering F_0. This is made possible by the flexibility of ligamentous connections between the upper ends of the thyroid cartilage and the two small cartilages (triticial cartilages) and also between these cartilages and the two ends of the hyoid bone, as in Figure 5.

4. Mathematical representation of F_0 contours of tone languages with positive and negative local components

The foregoing analysis leads to a model for the generation process of F_0 contours of tone languages from phrase commands and tone commands of positive and negative polarities, as shown in Figure 6.

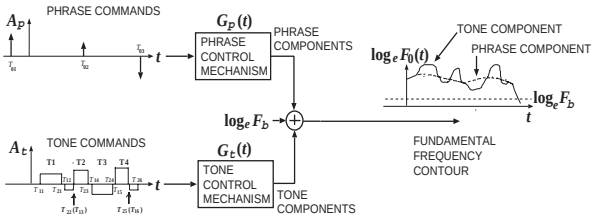

Figure 6: *The F_0 contour generation model for Mandarin.*

In this model, the F_0 contour can be given by the following mathematical formulation:

$$\log_e F_0(t) = \log_e F_b + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{tj} \{ G_t(t - T_{1j}) - G_t(t - T_{2j}) \}, \quad (8)$$

where

$$G_p(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & t \geq 0, \\ 0, & t < 0, \end{cases} \quad (9)$$

$$G_t(t) = \begin{cases} \min[\, 1 - (1 + \beta_1 t) \exp(-\beta_1 t), \gamma_1 \,], & t \geq 0, \\ 0, & t < 0, \end{cases}$$

(for positive tone commands), (10)

$$G_t(t) = \begin{cases} \min[\, 1 - (1 + \beta_2 t) \exp(-\beta_2 t), \gamma_2 \,], & t \geq 0, \\ 0, & t < 0, \end{cases}$$

(for negative tone commands),

where $G_p(t)$ represents the impulse response function of the phrase control mechanism and $G_t(t)$ represents the step response function of the tone control mechanism. The symbols in these equations indicate

F_b : baseline value of fundamental frequency,
I : number of phrase commands,
J : number of tone commands,
A_{pi} : magnitude of the ith phrase command,
A_{tj} : amplitude of the jth tone command,
T_{0i} : timing of the ith phrase command,
T_{1j} : onset of the jth tone command,
T_{2j} : end of the jth tone command,
α : natural angular frequency of the phrase control mechanism,
β_1: natural angular frequency of the tone control mechanism to positive tone commands,
β_2: natural angular frequency of the tone control mechanism to negative tone commands,
γ_1 : relative ceiling level of positive tone components,
γ_2 : relative ceiling level of negative tone components.

Although both β and γ should take different values depending on the polarity of commands as in Equation (10), the use of a common value for both β and γ irrespective of command polarity was found to be acceptable in almost all cases. The model has been successfully applied to the analysis of F_0 contours of a number of tone languages [15-19].

5. Conclusions

This paper has presented the author's view on prosody, information, and models, as well as on the roles of physiology, physics and mathematics in quantitative modeling, and has described the theoretical and experimental background for the command-response model for F_0 contour generation. It should be noted that the model represents only those factors that are inherent to the control mechanism of F_0, but allows one to identify those factors that carry communicative functions of speech as input commands and as parameters of the mechanism. The processes of deriving the input commands from linguistic information, as well as modifying the commands and model parameters by para- and non-linguistic information have been dealt with in our separate studies on several languages, but are not discussed here since they are much more language-specific.

6. References

[1] Fujisaki, H., 1997. Prosody, Models, and Spontaneous Speech. In *Computing Prosody,* Y. Sagisaka et al. (eds.) New York: Springer-Verlag, 27-42.

[2] Fujisaki, H.; Nagashima, S., 1969. A model for the synthesis of pitch contours of connected speech. *Annual Report of Engineering Res. Inst., Univ. of Tokyo* 28, 53-60.

[3] Buchthal, F.; Kaiser, E., 1944. Factors determining tension development in skeletal muscles. *Acta Physiol. Scand.* 8, 38-74.

[4] Sandow, W., 1958. A theory of active state mechanisms in isometric muscular contraction. *Science* 127, 760-762.

[5] Honda, K.; Hibi, S.; Kiritani, S.; Niimi, S.; Hirose, H., 1980. Measurement of the laryngeal structure during phonation by use of a stereoendoscope. *Ann. Bull. Res. Inst. Logoped. Phoniatr. Univ. Tokyo* 14, 73-78.

[6] Zemlin, W. R., 1968. *Speech and Hearing Science, Anatomy and Physiology*. New Jersey: Prentice Hall, Inc.

[7] Fink, B. R.; Demarest, R. J., 1978. *Laryngeal Biomechanics*. Harvard Univ. Press.

[8] Fujisaki, H., 1988. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In *Vocal Physiology, Voice Production, Mechanisms and Functions*, O. Fujimura (ed.). New York: Raven Press, 347-355.

[9] Fujisaki, H.; Hirose, K., 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *J. Acoust. Soc. Jpn (E)* 5 (4), 233-242.

[10] Sagart, L.; Hallé, P.; De Boysson-Bardies, B.; Arabia-Guidet, C., 1986. Tone production in modern Standard Chinese: an electromyographic investigation. *Cahiers de Linguistique Asie Orientale.* 15, 205-211.

[11] Hallé, P.; Niimi, S.; Imaizumi, S.; Hirose, H., 1990. Modern Standard Chinese four tones: electromyographic and acoustic patterns revisited. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics, Univ. of Tokyo* 24, 41-58.

[12] Erickson, D., 1976. A Physiological Analysis of the Tones of Thai, *PhD. Dissertation*, Univ. of Connecticut.

[13] Gårding, E., 1970. Word tones and larynx muscles. *Working Papers, Dept. of Linguistics, Lund Univ.* 3, 20-46.

[14] Fujisaki, H., 1995. Physiological and physical mechanisms for tone, accent and intonation. *Proc. XXIII World Congress of Int'l Assoc. Logoped. & Phoniat.*, Cairo, 156-159.

[15] Fujisaki, H., Hallé, P., Lei, H., 1987. Application of F_0 contour command-response model to Chinese tones. *Reports of Autumn Meeting, Acoust. Soc. Jpn.* 1, 197-198.

[16] Fujisaki, H., Ohno, S., Luksaneeyanawin, S., 2003. Analysis and synthesis of F_0 contours of Thai utterances based on the command-response model. *Proc. 15th ICPhS*, Barcelona, 1129-1132.

[17] Mixdorff, H., Nguyen, B., Fujisaki, H., Luong, M., 2003. Quantitative analysis and synthesis of syllabic tones in Vietnamese. *Proc. EUROSPEECH 2003*, Geneva, 177-180.

[18] Gu, W., Hirose, K., Fujisaki, H., 2004. Analysis of F_0 contours of Cantonese utterances based on the command-response model. *Proc. ICSLP 2004*, Jeju, 781-784.

[19] Gu, W., Hirose, K., Fujisaki, H., 2004. Analysis of Shanghainese F_0 contours based on the command-response model. *Proc. ISCSLP 2004*, Hong Kong, 81-84.