Information, Prosody, and Modeling
— with Emphasis on Tonal Features of Speech —

Hiroya Fujisaki

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

Abstract

Starting from the author's view on the process of information manifestation in the tonal features of speech, this paper emphasizes the importance of objective and quantitative modeling in the study of these features. It then describes a model for the process of fundamental frequency control of speech that has been originally proposed and established for Japanese, and explains the physiological and physical evidences on which the model is based. Application of the model for generation of F_0 contours of languages other than Japanese is then described, indicating how the original model can be modified and extended to cover those features that are not found in Japanese. The underlying mechanisms responsible for production of these features are also discussed.

1. Introduction [1]

In the first place, we shall look into the process by which certain kinds of information intended by a speaker are manifested in the segmental and suprasegmental features of speech.

The information expressed by speech can be regarded to fall into three categories: linguistic, para-linguistic, and non-linguistic, though their boundaries may not always be clear. Here I define *linguistic information* as the symbolic information that is represented by a set of discrete symbols and rules for their combination. It can be represented either explicitly by the written language, or can be easily and uniquely inferred from context. Linguistic information thus defined is discrete and categorical. For example, the information concerning the accent type of a Japanese word is discrete in the sense that it specifies one out of a finite number of possible accent types.

On the other hand, *paralinguistic information* is defined as the information that is not inferable from the written counterpart but is deliberately added by the speaker to modify or supplement the linguistic information. A written sentence can be uttered in various ways to express different intentions, attitudes, and speaking styles which are under the conscious control of the speaker. Paralinguistic information can be both discrete and continuous. For example, the information regarding whether a speaker's intention is an assertion or a question is discrete, but it can also be continuous in the sense that a speaker can express the degree within each category.

Nonlinguistic information concerns such factors as the age, gender, idiosyncracy, physical and emotional states of the speaker, etc. These factors are not directly related to the linguistic and paralinguistic contents of the utterances and cannot generally be controlled by the speaker, though it is possible for a speaker to control the way of speaking to intentionally convey an emotion, or to simulate an emotion, as is done by actors. Like paralinguistic information, nonlinguistic information can be discrete as well as continuous. Gradation within a category is a common feature of para- and nonlinguistic information, as opposed to the essentially discrete and categorical nature of linguistic information.

The relationship between these three types of information and the acoustic-phonetic manifestation of prosody as the organization of various linguistic units (i.e., segments, syllables, words, etc.) is schematically shown by Fig. 1.

How does a speaker organize various linguistic units systematically into a meaningful and expressive utterance? Although various languages may differ in the details, it is generally done by three means: accentuation, phrasing, and pausing.

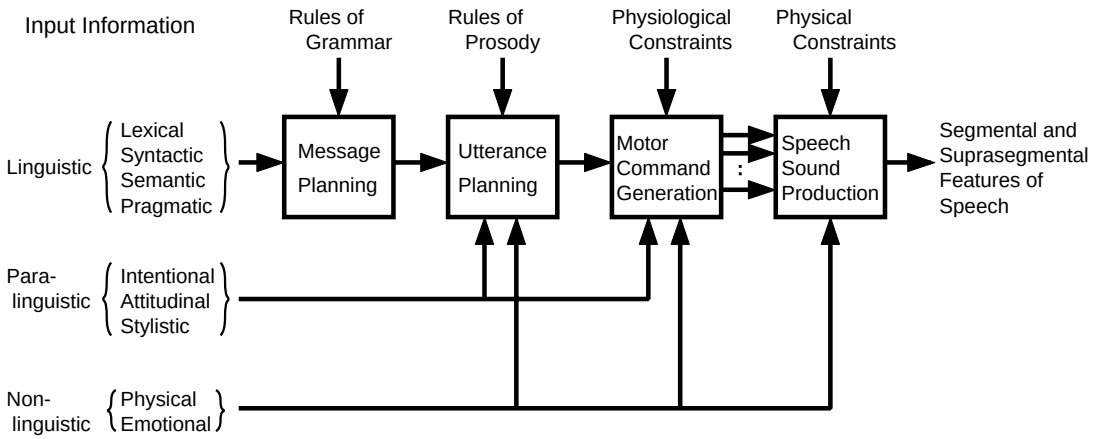

Figure 1: *Processes by which various types of information are manifested in the segmental and suprasegmental features of speech.*

'Accentuation' means changing (generally increasing) the relative prominence of a syllable within a word or a group of words. It is usually accomplished by controlling the fundamental frequency, duration, and intensity, though languages differ in the way they use these features. 'Phrasing' means grouping together a string of words into a perceptually coherent constituent. It is generally done by controlling the fundamental frequency and the local speech rate. 'Pausing' literally means putting a pause after a constituent (which can be one word or more in length), to indicate that the constituents at both sides of the pause should be processed separately. Thus it is often accompanied by 'phrasing,' but not *vice versa*.

The prosodic organization of linguistic units is influenced, however, by para- and non-linguistic information. For instance, the speaker consciously increases the speech rate in the case of emergency, with the intention to transmit the emergency of the situation and thereby urging the listener to react quickly. Also, the speaker may do so unconsciously in the case of anger. On the other hand, the speech rate is usually dropped unconsciously in the case of depression/sadness. Furthermore, there is another dimension in the suprasegmental features of speech, *viz.*, voice quality, which plays a significant role in the expression of both paralinguistic and non-linguistic information. These factors exert their influences mostly at the stages of utterance planning, motor command generation, and speech sound production.

The schematic diagram of Fig. 1 shows the complex, multi-stage and multi-dimensional nature of the process of conversion from information to speech sound features, and explains why it is difficult to find clear and unique correspondence between physically observable characteristics of speech and the underlying information contained in an utterance or a sequence of utterances. In this paper, therefore, I will restrict myself only to one of the features, namely the contour of the voice fundamental frequency (henceforth the F_0 contour), with full awareness of various interactions between the F_0 contour and other features.

In order to infer the underlying information from the observed characteristics of speech, it is logical to go through the following two steps:

1. Inferring the commands from the acoustic characteristics of speech,
2. Inferring the units and structures of prosody from these commands,
3. Retrieving various kinds of information from their influences on the speech characteristics and the underlying commands.

Since step 1 is the inverse operation of the process of speech sound production, it can be most accurately and objectively conducted if we have a quantitative model of the production process [2]. Such a model has been presented initially for the F_0 contours of Japanese utterances, and has since been shown to apply, with certain language-specific modifications, to F_0 contours of utterances of many other languages. Also, the scope of this paper will be restricted to step 1, though considerable amount of work has been done on steps 2 and 3.

2. A Quantitative Model for Generating F_0 Contours of Words and Sentences of Japanese [3, 4]

By way of illustration, the F_0 contour of a Japanese declarative sentence:

Aoi aoinoewa yamanouenoieni aru. (The picture of the blue hollyhock is in a house on top of the hill.)

is shown in Fig. 2 as a function of time on the logarithmic scale of fundamental frequency.

The informant is a male speaker of Common Japanese uttering the sentence with a declarative intonation. Examination of this and a number of other F_0 contours of sentences suggests that an F_0 contour of a sentence, represented on the logarithmic scale of F_0, can be considered to consist of two kinds of elements. One is a slowly varying component which may or may not show a slight initial rise and then gradually decay toward an asymptotic baseline, but may be resumed or reinforced at certain syntactic boundaries, at least in the case of Japanese sentences. The others are local humps (peaks or plateaus) closely corresponding to the accent patterns of words constituting the sentence. The humps may differ in their height.

For a quantitative formulation, we set up the following assumptions:

(1) The phrase commands are a set of impulses and the phrase components are the response of a critically-damped second-order linear system to these commands.

(2) The accent commands are a set of stepwise functions and the accent components are the response of another critically-damped second-order linear system to these commands.

(3) The phrase and accent components are superimposed and produce a proportionate change in the logarithm of F_0. Although these two systems may not be exactly critically-damped, preliminary analysis of F_0 contours suggests that the assumption is appropriate.

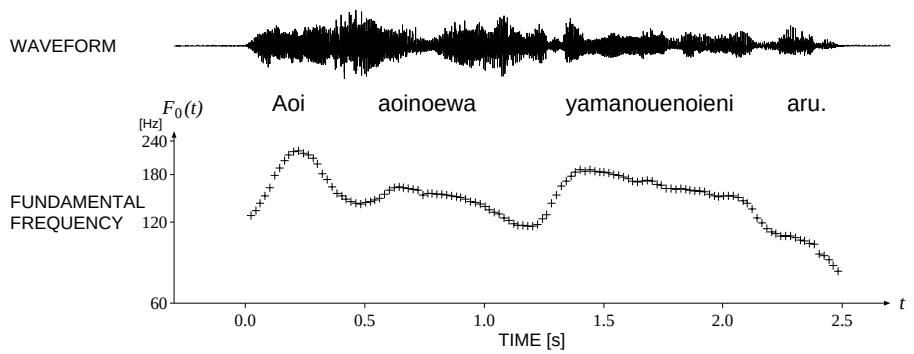

Figure 2: *An example of a measured F_0 contour of the declarative sentence "Aoi aoinoewa yamanouenoieni aru." (The picture of the blue hollyhock is in a house on top of the hill.) of Japanese.*

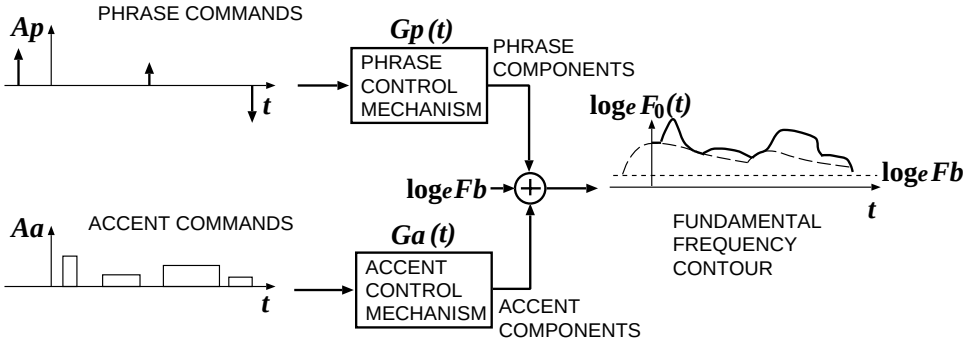

Figure 3: *A functional model for the process of generating F_0 contours.*

For the rest of this paper we shall re-define an F_0 contour to be the contour of the logarithm of F_0(t), viz. $\log F_0$(t).

Based on these assumptions, a model is constructed for the generation process of the F_0 contours of utterances of Common Japanese, and is shown in Fig. 3.

In this model, the F_0 contour can be expressed by

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

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

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

where $Gp(t)$ represents the impulse response function of the phrase control mechanism and $Ga(t)$ represents the step response function of the accent control mechanism. The symbols in these equations indicate

- F_b : baseline value of fundamental frequency,
- I : number of phrase commands,
- J : number of accent commands,
- Ap_i : magnitude of the ith phrase command,
- Aa_j : 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,
- γ : relative ceiling level of accent components.

Parameters α and β are assumed to be constant at least within an utterance, while the parameter γ is set equal to 0.9. A rapid downfall of F_0, often observed at the end of a sentence and occasionally at a clause boundary, can be regarded as the response of the phrase control mechanism to a negative impulse for resetting the phrase component.

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 shown in Fig. 4.

The two positive phrase commands correspond to the subject phrase and the predicate phrase, respectively, while the negative phrase command toward the end of the utterance corresponds to the utterance-final fall in F_0. The accent commands, which are always positive in the case of Common Japanese, correspond to the prosodic words. The model-generated F_0 contour is so close to the measured F_0 contour that they are perceptually indistinguishable in synthetic speech.

Thus the model can predict and generate from a set of commands, not just a few points on the F_0 contour such as its peaks and valleys subjectively selected, but the entire contour. Moreover, the close agreement of the model's output with the mea-

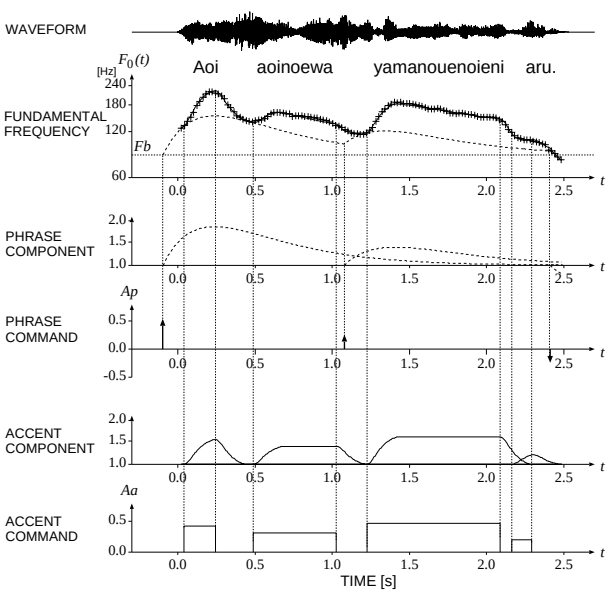

Figure 4: *Analysis-by-Synthesis of an F_0 contour of the Japanese declarative sentence: Aoi aoinoewa yamanouenoieni aru. The figure illustrates the optimum decomposition of a given F_0 contour into the phrase and accent components, and also shows the underlying commands for these components.*

sured F_0 contour, found in this as well as in a number of speech samples analyzed, attest the validity of the model.

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 to 50 msec before the onset of the vowel of a subjectively high mora and to end also at 40 to 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. Our analysis also shows that the variations in the values of natural angular frequencies α and β are quite small from utterance to utterance as well as from one individual to another.

Thus the model allows one to separate those factors that are closely related to linguistic and paralinguistic information as the magnitude and timing of the commands, from the factors that are related to physiological and physical mechanisms of phonatory control as the response characteristics, i.e., as the shapes of phrase and accent components.

3. Physiological and Physical Mechanisms Underlying the Model [5]

The ability of the aforementioned model to produce very accurate approximations to observed F_0 contours has its basis in the physiological and physical mechanisms of the larynx.

3.1. Structure of the larynx

Figure 5 shows the sections of the human larynx: (a) anterior-posterior section, (b) median section, and (c) horizontal section, while Fig. 6 shows the two cartilages, *i. e.*, the thyroid cartilage and cricoid cartilage, whose relative positions are changed by the activity of the crico-thyroid (henceforth CT) muscle, causing a change in the length of the vocal cord.

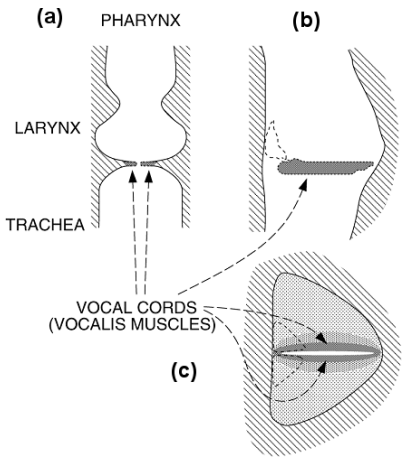

Figure 5: *Sections of the human larynx.*
(a) anterior-posterior section,
(b) median section,
(c) horizontal section.

3.2. Stress-strain relationship of skeletal muscles

The stress-strain relationship of skeletal muscles including the human vocalis muscle has been widely studied [6, 7]. Figure 7 shows the earliest published data on the relationship between tension and stiffness [6].

The data shown in Fig. 7 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:

$$dT/dl = a + bT, \tag{4}$$

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

$$T = (T_0 + a/b)\exp\{b(l - l_0)\} - a/b, \tag{5}$$

where T_0 indicates the static tension applied to the vocal cord, and l_0 indicates its length at $T = T_0$. When $T_0 \gg a/b$, Eq. (5) can be approximated by

$$T = T_0 \ \exp(bx), \tag{6}$$

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

$$F_0 = c_0\sqrt{T/\sigma}, \tag{7}$$

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 Eqs. (3) and (7) we obtain

$$\log_e F_0 = \log_e\{c_0\sqrt{T_0/\sigma}\} + (b/2)x. \tag{8}$$

Strictly speaking, the first term varies slightly with x, but the overall dependency of $\log_e F_0$ on x is primarily determined

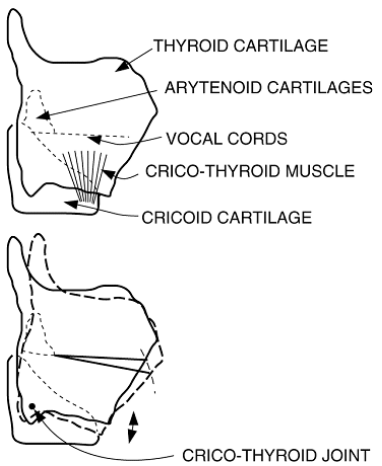

Figure 6: *Changes in the relative positions of the thyroid cartilage and the cricoid cartilage due to the activity of the crico-thyroid muscle, causing a change in the vocal cord.*

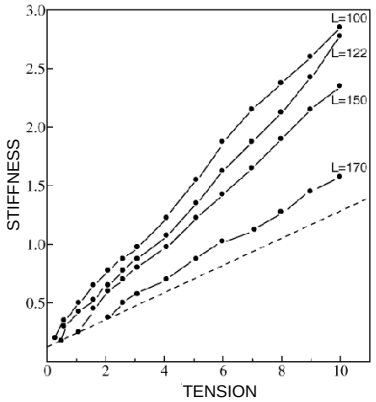

Figure 7: *Stiffness as function of tension at rest (- - - - -) and during isometric tetanic contraction initiated at different original length. In the top curve contraction is initiated at a length below 100 (equilibrium length = 100).*

Ordinate: stiffness in arbitrary units.

Abscissa: tension in arbitrary units.

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 [8], 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

$$\log_e F_0(t) = \log_e F_b + (b/2)x(t), \tag{9}$$

where the constant $c_0\sqrt{T_0/\sigma}$ in Eq. (8) 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 $F_0(t)$.

3.3. Role of cricothyroid muscle

Analysis of the laryngeal structure suggests that the movement of the thyroid cartilage relative to the cricoid cartilage has two degrees of freedom [9, 10]. 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 Fig. 8. The translation and the rotation of the thyroid can be represented by separate second-order systems as shown in Fig. 9, 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. 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, Eq. (9) can be rewritten as

$$\log_e F_0(t) = \log_e F_b + (b/2)\{x_1(t) + x_2(t)\}, \tag{10}$$

which means that the time-varying component of $\log_e F_0(t)$ can be represented by the sum of two time-varying components. Since the translational movement of the thyroid cartilage has a much larger time constant than the rotational movement, the

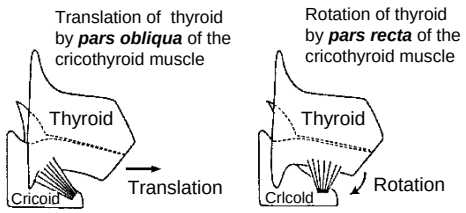

Figure 8: *The roles of pars obliqua and pars recta of the cricothyroid muscle in translating and rotating the thyroid cartilage.*

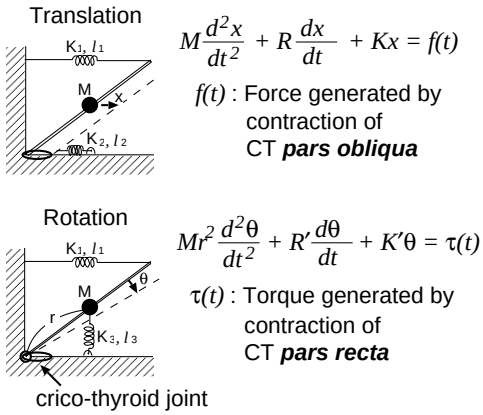

Figure 9: *Equations of translation and rotation of the thyroid cartilage.*

former is used to indicate global phenomena such as phrasing, while the latter is used to indicate local phenomena such as word accent.

4. Applicability of the Model to F_0 Contours of Various Other Languages

4.1. Languages with only positive local commands

It has been shown elsewhere [11] that the units of prosody of spoken Japanese can be clearly and objectively defined in terms of the phrase and accent components, and the rules for derivation of these units from text have also been constructed. The model has been widely used in text-to-speech systems because of the extremely high naturalness it can provide [12].

The success of the model on F_0 contours of Japanese suggests its applicability to other languages, considering the fact that the model captures essential characteristics of the dynamics of the human larynx which is similar, if not completely identical, among speakers of different languages. It is of course possible that some languages may require certain ways of laryngeal control which are not required by other languages. This point can be easily appreciated by the fact that a multilingual speaker can produce native intonation patterns of various languages using the same larynx.

From this point of view, analysis has been made of F_0 contours of a number of utterances of each of the following languages: English [13], Estonian [14], German [15], Greek [16], Korean [17], Polish, and Spanish [18]. The six panels (a)∼(f) in Fig. 10 show examples of results of F_0 contour analysis for one utterance each of these six languages. The texts of these utterances are:

(a) English: *It's strange that I slept so long since I wasn't feeling tired.*

(b) German: *Sie haben den Wagen geliehen und sind tatsächlich gefahren*. (They have rented the car and have actually driven away.)

(c) Greek: *Είναι το σύμβολο όχι μόνο της Αθήνας, αλλά και της Ελλάδας*. (It is the symbol not only of Athens, but also of Greece.)

(d) Korean: *Onwrwn barami ma:nido bu:mnida.* (It is very windy today.)

(e) Polish: *Norwid tworzył w dziewiętnastym wieku.* (Norwid wrote in the 19th century.)

(f) Spanish: *La buena abuela de Lola le da un melon a la nena.* (The good grandmother of Lola gives a melon to the baby.)

These results indicate that both the polarities of the input commands and the response mechanisms are essentially the same as for those of Common Japanese, showing that the identical model can be applied to F_0 patterns of all these languages.

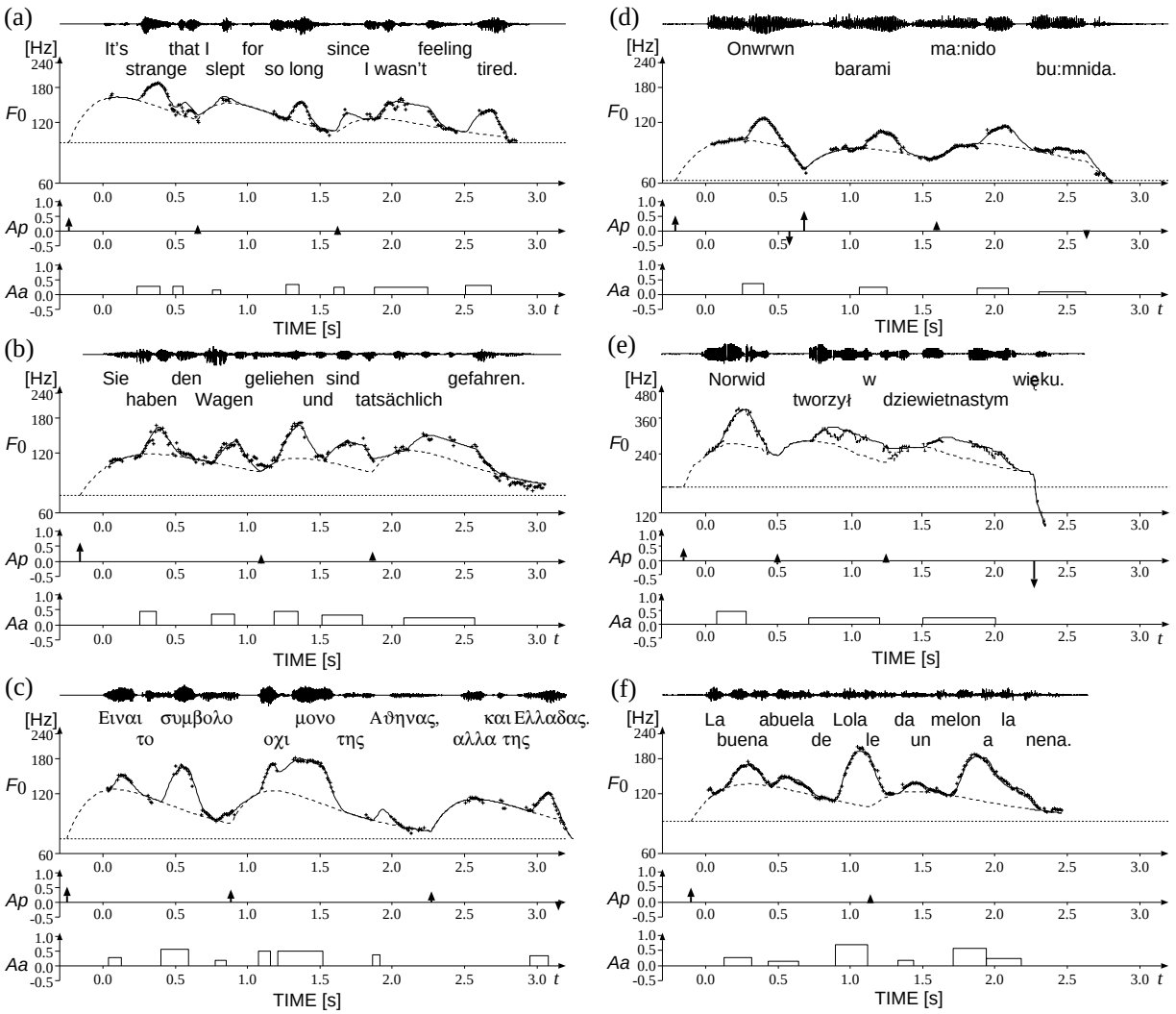

Figure 10: *Examples of Analysis-by-Synthesis of F_0 contours of various languages. (a) English, (b) German, (c) Greek, (d) Korean, (e) Polish, and (f) Spanish.*

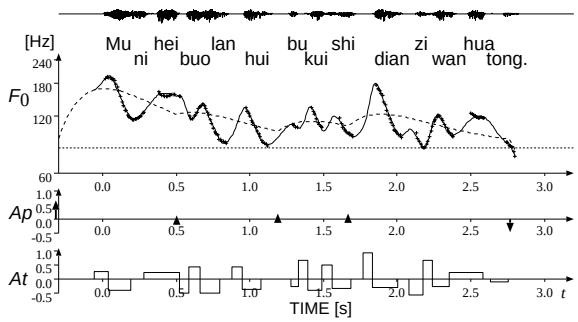

Figure 11: *An example of Analysis-by-Synthesis of the F_0 contour of an utterance of the Standard Chinese.*

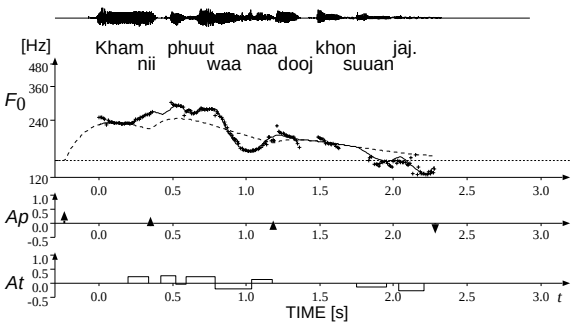

Figure 12: *An example of Analysis-by-Synthesis of the F_0 contour of an utterance of Thai.*

4.2. Languages with both positive and negative local commands

Analysis of F_0 contours of several languages including Standard Chinese, Thai and Swedish, however, indicates that the local components (associated with tones in the case of Standard Chinese 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 to obtain good approximations to the observed F_0 contours. For example, the F_0 contours of the four tones in Standard Chinese, conventionally classified as High (Tone 1, T1), Rising (Tone 2, T2), Low (Tone 3, T3), and Falling (Tone 4, T4), can be approximated quite well by assuming a positive tone command for T1, a negative one followed by a positive one for T2, a negative one for T3, and a positive one followed by a negative one for T4 within a syllable [19, 20].

Figure 11 illustrates the analysis of the F_0 contour of the utterance of the Standard Chinese:

Mu4 ni2 hei1 buo2 lan3 hui4 bu2 kui4 shi4 dian4 zi3 wan4 hua1 tong3. (The Munich exposition is really an electronic kaleidoscope.)

The figure indicates that the modified model with tone commands of both polarities can generate an F_0 contour which perfectly matches the measured contour.

Analysis-by-Synthesis of F_0 contours was conducted on a number of utterances of Standard Chinese, and the results indicate that the mean value of β for negative tone commands is significantly smaller than that for positive tone commands, suggesting that the lowering of F_0 in T2, T3, and T4 is caused by an additional mechanism involving muscles other than CT.

Thai is another tone language which is known to have five tone types, conventionally classified as Mid (Tone 0), Low (Tone 1), Falling (Tone2), High (Tone 3), and Rising (Tone 4). Note that the numbering of Thai tones is different from that of Standard Chinese. Our preliminary analysis of the F_0 contours of Thai utterances indicated that the basic structure of the model can be regarded as the same for tones of the Standard Chinese, but commands for some of the Thai tones are different [21]. Namely, the mid tone has no tone command, and the high tone of Thai has a positive tone command only at the later part of the syllable, while the low, falling and rising tones of Thai have essentially similar commands as the corresponding tones of Standard Chinese.

Figure 12 illustrates the analysis of the F_0 contour of the utterance of Thai:

Kham0 nii3 phuu2 waa2 naa4 dooj0 khon0 suuan1 jaj1. (This word is pronounced as naa by most speakers.)

Similar switching of polarity is found to occur also in Swedish [22], in which a disyllabic word of Accent 1 is characterized by a positive F_0 excursion for the first syllable followed by a negative excursion for the second syllable. The order of polarity is reversed in a disyllabic word of Accent 2. For example, the dissyllabic '*anden*' can be two different words depending on the lexical accent: '*ánden*' (with Accent 1) (the wild duck) and '*ànden*' (with Accent 2) (the spirit). It was shown that the F_0 contours of these words are generated by a pair of accent commands of different polarities; a positive one followed by a negative one for Accent 1, and a negative one followed by a positive one for Accent 2.

Figure 13 illustrates the analysis of the F_0 contour of the Swedish utterance:

Av invånarna i Sverige beräknas för närvarande cirka 10% vara av utländsk härkomst. (Of the inhabitants of Sweden, about 10% are estimated to be born in foreign countries.)

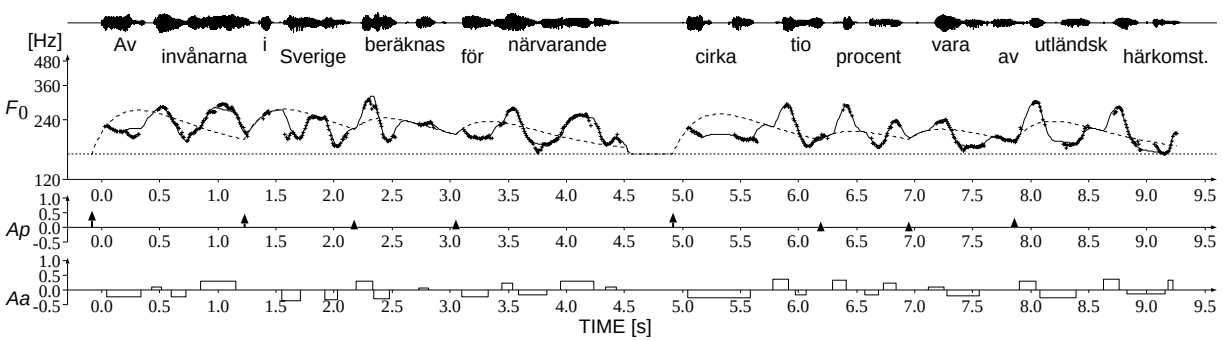

Figure 13: *An example of Analysis-by-Synthesis of the F_0 contour of an utterance of Swedish.*

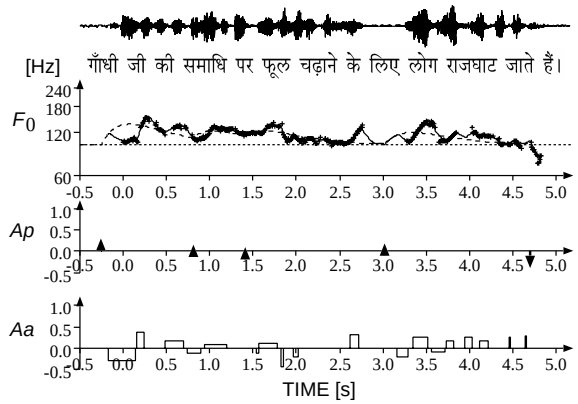

Figure 14: *Examples of Analysis-by-Synthesis of the F_0 contours of Hindi utterances.*

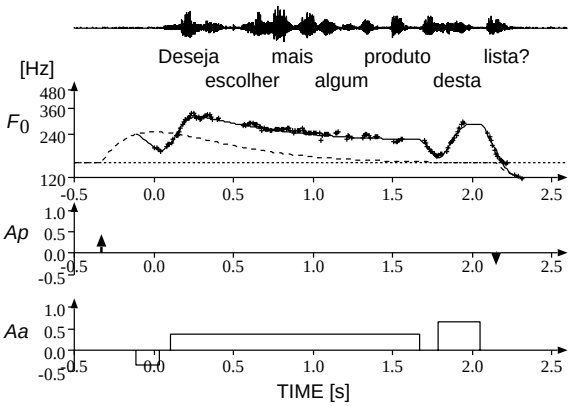

Figure 15: *Examples of Analysis-by-Synthesis of the F_0 contours of Portuguese utterances.*

In certain languages including Hindi [23] and Portuguese [24], the positive accent commnads are used to indicate word accent, while negative accent commands are occasionally used to signal emphasis, especially at phrase-initial positions.

Figure 14 illustrates the analysis of the F_0 contour of the Hindi utterance:

गाँधी जी की समाधि पर फूल चढ़ाने के लिए लोग राजघाट जाते हैं।

(Everybody is going to Rajghat to offer flower tributes to the tomb of Sir 'Gandhi'.)

Note that a negative accent command is used on the initial syllable of the first word Gandhi for emphasis.

Figure 15 illustrates the analysis of the F_0 contour of the Portuguese utterance:

Deseja escolher mais algum produto desta lista? (Do you want to choose one more product from this list?)

This is a yes/no question. Note that there is a large negative accent component on the initial syllable of the word 'Deseja.'

In some languages, a negative local command is used occasionally to replace a positive one. For instance, local dips in the F_0 contour occur more often in British English than in American English. Their occurrences are somewhat idiosyncratic — some speakers use them more often than others. Figure 16 illustrates three utterances in which they occur. The utterances are:

(a) *Where are you going?*

(b) *Have you made up your mind?*

(c) *You always try to do everything at the last minute.*

These utterances were from three different speakers. As shown in these panels, the dips in the F_0 contours can be modeled very well by the corresponding negative accent components.

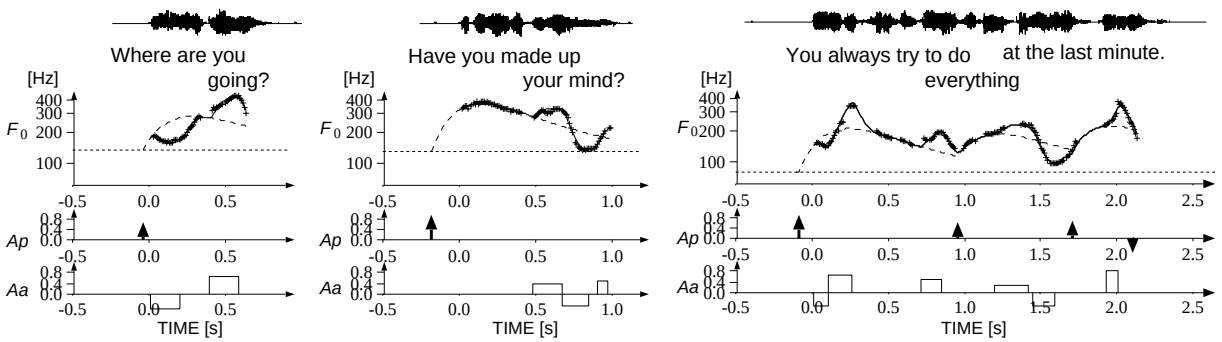

Figure 16: *Examples of Analysis-by-Synthesis of the F_0 contours of British utterances.*

4.3. Roles 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 catilage 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 Standard Chinese [25, 26], the five tones of Thai [27] as well as of the grave accent of Swedish [28], 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 [29]. Figure 17 shows the relationship between the hyoid bone, thyoid and cricoid cartilages, and TH in their lateral and frontal views, and Fig. 18 shows thier relationships with three other muscles: VOC (thyrovocalis muscle), CT, 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 shown in Fig. 18.

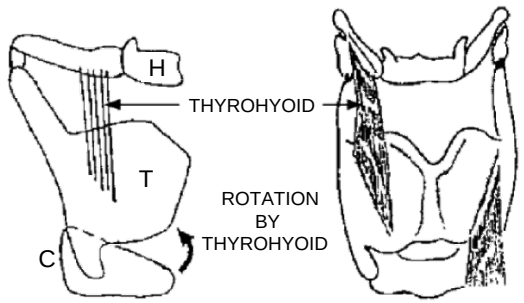

Figure 17: *Role of thyrohyoid in laryngeal control.*

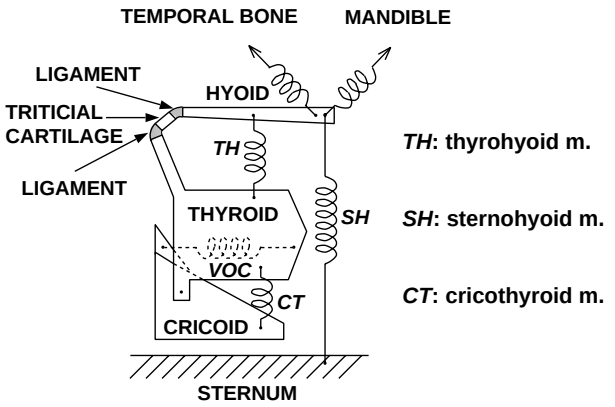

Figure 18: *Mechanism of F_0 lowering by activities of TH and SH.*

4.4. Application to multilingual speech synthesis

The results of analysis of a number of utterances for each of the 14 languages thus far investigated by the present author suggest that these languages fall broadly into two groups, depending on whether the command for the local F_0 component is always positive, or can be both positive and negative, as shown in Table 1.

Table 1: *Grouping of languages on the basis of tone/accent command polarity.*

Group	Polarity of tone/ accent commands	Languages
1	positive only	English*, Estonian, German, Greek, Japanese, Korean, Polish, Spanish
2	positive and negative	Hindi, Portuguese, Swedish, Chinese†, Thai†

* Certain speakers of English occasionally use negative accent commands, especially in order to express paralinguistic information.

† Tone languages.

As it was shown elsewhere, parameters α and β are found to be nearly constant across speakers and languages, while the baseline frequency F_b is speaker-dependent [30]. These parameters characterize the physiological and physical mechanisms of the speaker's larynx. The results of analysis shown above indicate that the command-response model is capable of producing F_0 contours of natural intonation from the same mechanism but with various language-specific patterns of input commands.

5. Conclusions

This paper has presented the author's view on the classification of the types of information conveyed by speech, and the necessity of a quantitative model for the process of their manifestation in the acoustic/phonetic features. It has then demonstrated such a model developed by the author and his coworkers for the process of controlling the voice fundamental frequency (i. e., generating the F_0 contour) for spoken Japanese. The model has been shown to be capable of generating very close approximations to a large number of actually observed F_0 contours from a limited number of meaningful parameters. Physiological and physical mechanisms underlying the model have also been elucidated. Finally, the applicability of the model to F_0 contour generation for a number of other languages has been demonstrated, indicating the usefulness of the model for multi-lingual speech synthesis.

6. References

[1] Fujisaki, H., 1996. Prosody, Models, and Spontaneous Speech. In *Computing Prosody* (Sagisaka, Y., Campbell, N., and Higuchi, N., eds.), Springer-Verlag, 27–42.

[2] Öhman, S., 1967. Word and sentence intonation: A quantitative model. *Speech Transmission Laboratory Quarterly Progress and Status Report, KTH*, STL-QPSR 2-3/1967, 20–54.

[3] Fujisaki, H.; Nagashima, S., 1969. A model for the synthesis of pitch contours of connected speech. *Annual Report of the Engineering Research Institute, University of Tokyo*, 28, 53–60.

[4] 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.

[5] 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* (Fujimura, O., ed.), Raven Press, 347–355.

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

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

[8] Honda, K.; Hibi, S.; Kiritani, S.; Niimi, S.; Hirose, H., 1980. Measurement of the laryngeal structure during phonation by use of a stereoendoscope. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics, University of Tokyo*, 14, 73–78.

[9] Zemlin, W. R., 1968. *Speech and Hearing Science, Anatomy and Physiology*. Prentice Hall, Inc.

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

[11] Fujisaki, H.; Hirose, K.; Takahashi, T., 1993. Manifestation of linguistic information in the voice fundamental frequency contours of spoken Japanese. *IEICE Trans. Fundametal. Electro. Comm. Comp. Sci.*, A E76, 1919–1926.

[12] Hirose, K.; Fujisaki, H.; Kawai, H.; Yamaguchi, M., 1989.

[13] Fujisaki, H.; Ohno, S., 1995. Analysis and modeling of fundamental frequency contours of English utterances. *Proc. EUROSPEECH'95*, 2, 985–988.

[14] Fujisaki, H; Lehiste, I., 1982. Some temporal and tonal characteristics of declarative sentences in Estonian. *Preprints of Working Group on Intonation, the 13th International Congress of Linguists*, Tokyo, 121–130.

[15] Mixdorff, H.; Fujisaki, H., 1994. Analysis of voice fundamental frequency contours of German utterances using a quantitative model. *Proc. 1994 Int'l Conf. Spoken Language Processing*, 4, 2231–2234.

[16] Fujisaki, H.; Ohno, S.; Yagi, T., 1997. Analysis and modeling of fundamental frequency contours of Greek utterances. *Proc. EUROSPEECH'97*, 1, 465–468.

[17] Fujisaki, H., 1996. Analysis and modeling of fundamental frequency contours of Korean utterances — A preliminary study —. In *Phonetics and Linguistics — in honour of Prof. H. B. Lee*, 640–657.

[18] Fujisaki, H.; Ohno, S.; Nakamura, K.; Guirao, M.; Gurlekian, J., 1994. Analysis of accent and intonation in Spanish based on a quantitative model. *Proc. 1994 Int'l Conf. Spoken Language Processing*, 1, 355–358.

[19] 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.

[20] Fujisaki, H.; Hirose, K.; Hallé, P.; Lei, H., 1990. Analysis and modeling of tonal features in polysyllabic words and sentences of the Standard Chinese. *Proc. 1990 Int'l Conf. Spoken Language Processing*, 2, 841–844.

[21] 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 International Congress of Phonetic Sciences*, 2, 1129–1132.

[22] Fujisaki, H.; Ljungqvist, M.; Murata, H., 1993. Analysis and modeling of word accent and sentence intonation in Swedish. *Proc. 1993 Int'l Conf. Acoust., Speech, & Signal Processing*, 2, 211–214.

[23] Fujisaki, H.; Ohno, S., 2003. Modeling the generation process of fundamental frequency contours of Hindi utterances. *Reports of Fall Meeting, the Acoustical Society of Japan*, 1, 217–218.

[24] Fujisaki, H.; Narusawa, S.; Ohno, S.; Freitas, D., 2003. Analysis and modeling of F_0 contours of Portuguese utterances based on the command-response model. *Proc. 8th European Conference on Speech Communication*, 3, 2317–2320.

[25] 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.

[26] 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, University of Tokyo*, 24, 41–58.

[27] Erickson, D., 1993. Laryngeal muscle activity in connection with Thai tones. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics, University of Tokyo*, 27, 135–149.

[28] Gårding, E., 1970. Word tones and larynx muscles. *Working Papers, Dept. of Linguistics, Lund University*, 3, 20–46.

[29] Fujisaki, H., 1995. Physiological and physical mechanisms for tone, accent and intonation. *Proc. the XXIII World Congress of the International Association of Logopedics and Phoniatrics*, 156–159.

[30] Ohno, S.; Fujisaki, H.; Hara, Y., 1998. On the effects of speech rate upon parameters of the command-response model for the fundamental frequency contours of speech. *Proc. 1998 Int'l Conf. Spoken Language Processing*, 3, 659–662.