말소리와 음성과학 제 1 권 제 1 호 (2009)
pp. 9~22

Invited Paper

In Search of Models in Speech Communication Research

Fujisaki, Hiroya [1)]

ABSTRACT

This paper first presents the author's personal view on the importance of modeling in scientific research in general, and then describes two of his works toward modeling certain aspects of human speech communication. The first work is concerned with the physiological and physical mechanisms of controlling the voice fundamental frequency of speech, which is an important parameter for expressing information on tone, accent, and intonation. The second work is concerned with the cognitive processes involved in a discrimination test of speech stimuli, which gives rise to the phenomenon of so-called categorical perception. They are meant to illustrate the power of models based on deep understanding and precise formulation of the functions of the mechanisms/processes that underlie observed phenomena. Finally, it also presents the author's view on some models that are yet to be developed.

Keywords: models, modeling, fundamental frequency of speech, F_0 contour generation, identification, discrimination test, categorical perception

1. Introduction

In my view, there are three stages in our understanding of the nature: namely, ideas, theories, and models. 'Ideas' are not necessarily expressed in scientific terms and not testable, 'Theories' are more or less well expressed in scientific terms, but are still abstract and not necessarily testable. On the other hand, 'models' are clearly formulated and testable. I wish to be excused if my use of these terms is not conventional [1].

The past decades have witnessed a great progress in our understanding of human speech communication. In my opinion, this is largely due to the introduction of powerful models on various aspects of speech and language. They include both structural and functional models of the vocal organs and the sources of excitation, models of their dynamic characteristics, models of speech perception, as well as models of language as a source of information.

As necessary features of a good model, I would require that it should be

(1) objective – should be derived by objective means,
(2) quantitative – should be expressed in quantitative terms,
(3) generative – should be able to generate/predict the entire phenomenon in question [2, 3], and
(4) elucidative – should not be an arbitrary mathematical approximation to what is observed, but should be able to elucidate the relationship between the observed phenomenon and the underlying mechanisms/processes.

In the following sections, I will describe two models from my own research, to illustrate the elucidatuve power of quantitative models. I will then briefly discuss on models that we still do not have, but are believed to be indispensable for a deep understanding of the process of human speech communication, as well as for the utilization of speech in human-machine communication.

1) The University Tokyo
(This paper is based on the author's keynote speech at INTERSPEECH2008, Brisbane, Australia)

Received: Feb. 29 2009
Accepted: Mar. 5 2009

2. A model for generating fundamental frequency contours of speech

2.1 The model and its mathematical formulation

In many languages of the world, the contour of the fundamental frequency of voice (henceforth the F_0 contour) plays an important role in expressing information on tone, accent and intonation. As an example, the F_0 contour of an utterance of the Japanese declarative sentence

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

is shown in <Figure 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.

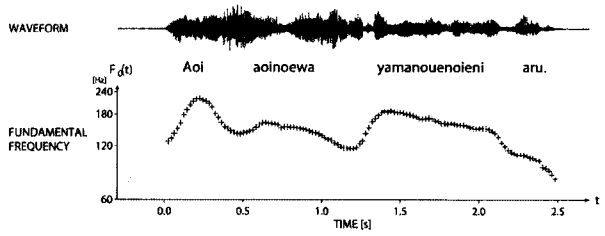


Figure 1. An example of a measured F_0 contour of the declarative sentence "Aoi aoinoewa yamanouenoieni aru." (The picture of the blue hollyhock is in a house on the top of a hill.) of Japanese.

Examination of this and a number of other F_0 contours of utterances suggests that an F_0 contour of an utterance of a sentence, represented on the logarithmic scale of F_0, can be considered to be consisting of two kinds of elements. One is a slowly varying component that 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) 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 [4, 5]:

(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.

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 <Figure 2>.

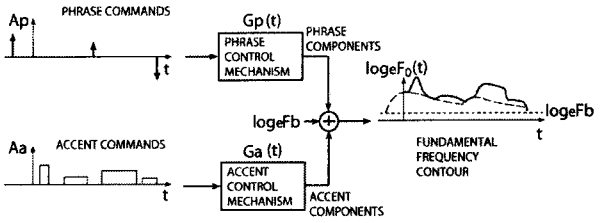


Figure 2. A functional model for the process of generating F_0 contours.

In this model, the F_0 contour can be expressed by

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

where

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

and

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

where $G_p(t)$ represents the impulse response function of the phrase control mechanism and $G_a(t)$ 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,
A_{pi} : magnitude of the i th phrase command,
A_{aj} : amplitude of the j th accent command,
T_{0i} : timing of the i th phrase command,
T_{1j} : onset of the j th accent command,
T_{2j} : end of the j th accent command,
α : natural angular frequency of the phrase control mechanisn,
β : natural angular frequency of the accent control mechanisn,
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 [6], 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 <Figure 3>.

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.

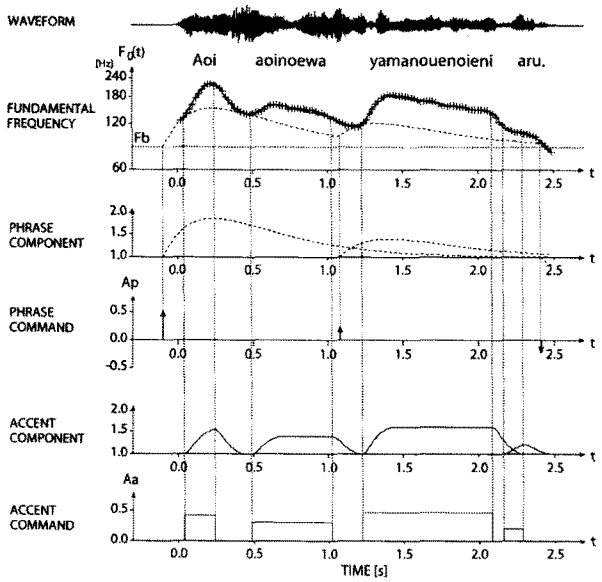


Figure 3. Analysis-by-Synthesis of an F_0 contour of the Japanese declarative sentence: Aoi aoinoewa yamanouenoieni aru. The figure illustrates the optimum decomposition of a given F_0 contour into the phrase and accent components, and also shows the underlying commands for these components.

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 measured 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 content 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.

The model, also referred to as the command-response model, was first shown to apply to F_0 contours of utterances of Common Japanese and was also used for speech synthesis from text, because of its ability to produce quite natural intonation even when the parameters of the generated F_0 contour are not exactly equal to those of a natural utterance.

The model has since been shown to apply to F_0 contours of utterances of a number of other non-tone languages, including English [7], Estonian [8], German [9], Greek [10], Korean [11], Polish and Spanish [12], in which accent commands are almost always positive.

Although the accent commands shown in <Figure 2> are all positive, the model itself is applicable to languages that have both positive and negative accent commands, such as Hindi [13], Portuguese [14], Russian [15] and Swedish [16]. Furthermore, it has also been shown that the model applies, with some minor modifications, quite well to utterances of tone languages such as

Mandarin [17, 18], Cantonese [19], Thai [20], and Vietnamese [21].

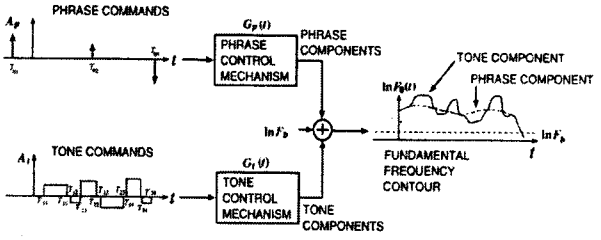


Figure 4. The model for tone languages, where each syllable may have more than one tone command, whose polarity can be positive or negative.

<Figure 4> shows the model for tone languages in which the accent commands in <Figure 3> are replaced by tone commands of both positive and negative polarities. In most tone languages, it is sufficient to assume up to two tone commands for a syllable. For example, in the case of four tones of Mandarin, Tone 1 has a positive tone command, Tone 2 has a negative tone command followed by a positive one, Tone 3 has a negative tone command, and Tone 4 has a positive tone command followed by a negative one. In certain languages, however, we have to assume more than two tone commands [22].

Exactly speaking, parameters α and β are different for positive tone connands and negative tone commands [17], but our experiments on both analysis and synthesis of F_0 contours of tone languages indicate that common values are acceptable for practical applications.

It may be worthwhile to mention here that the tone components and the phrase components of the current model are mathematical and quantitative representations of what Y. R. Chao called small ripples and large waves, respectively [23], while the baseline value ($\log F_b$) corresponds to the keynote of the phrasal contour which varies with attitudinal or emotional changes in the speaker, as indicated by Z. Wu [24].

2.2 Physiological and physical mechanisms underlying the model [25, 26]

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.

2.2.1 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 <Figure 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.

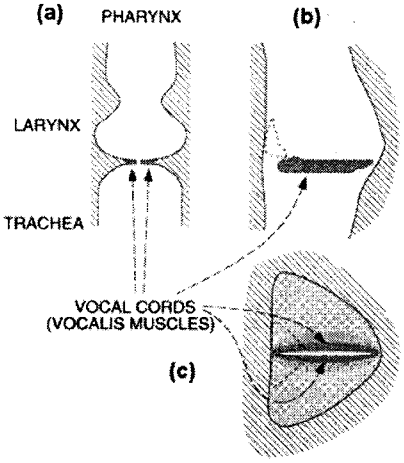


Figure 5. Sections of the human larynx.
(a) anterior-posterior section,
(b) median section,
(c) horizontal section.

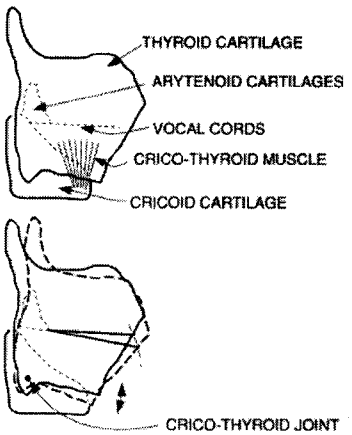


Figure 6. Changes in the relative positions of the thyroid cartilage and the cricoid cartilage due to the activity of the crico-thyroid muscle, causing a change in vocal cord length.

2.2.2 Stress-strain relationship of skeletal muscles

The stress-strain relationship of skeletal muscles including the human vocalis muscle has been widely studied [27, 28]. <Figure 7> shows the earliest published data on the relationship between tension and stiffness [27].

The data shown in <Figure 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 $T_0 >> a/b$, 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.

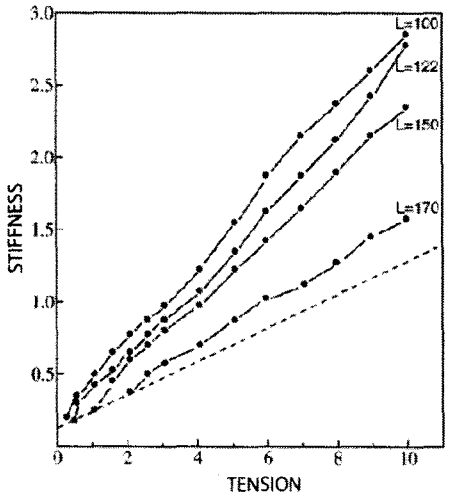


Figure 7. Stiffness as function of tension at rest (- - - - -) and during isometric tetanic contraction initiated at different original length. In the top curve contraction is initiated at a length below 100 (equilibrium length = 100). Ordinate: stiffness in arbitrary units. Abscissa: tension in arbitrary units [27].

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 by the second term on the right hand side. This linear relationship was confirmed for sustained phonation by an experiment in which a stereo-endoscope was used to measure the length of the vibrating part of the vocal cord, 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 $F_0(t)$.

2.2.3 Role of the 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 [29, 30]. One is horizontal translation due 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 8>.

The translation and the rotation of the thyroid can be represented by separate second-order systems as shown in <Figure 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 to be 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 former is used to indicate global phenomena such as phrasing, while the latter is used to indicate local phenomena such as word accent.

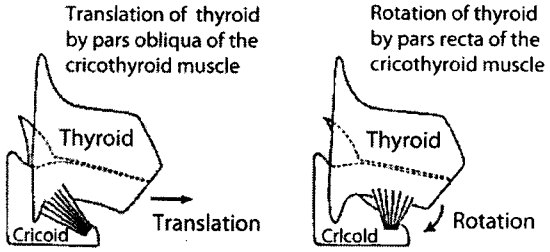


Figure 8. The roles of pars obliqua and pars recta of the cricothyroid muscle in translating and rotating the thyroid cartilage.

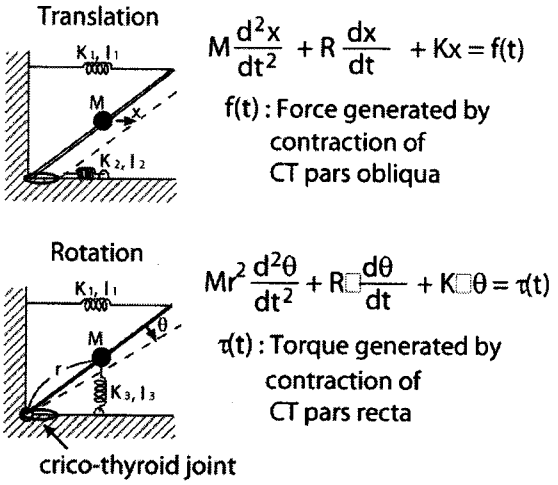
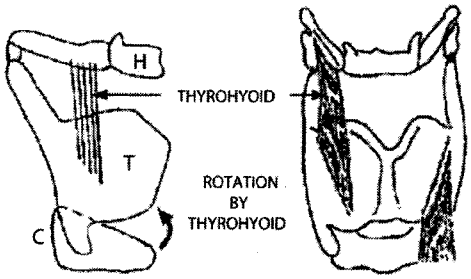


Figure 9. Equations of translation and rotation of the thyroid cartilage.



C: cricoid cartilage. T: thyroid cartilage. H: hyoid bone

Figure 10. Role of the thyrohyoid muscle (TH) in laryngeal control.

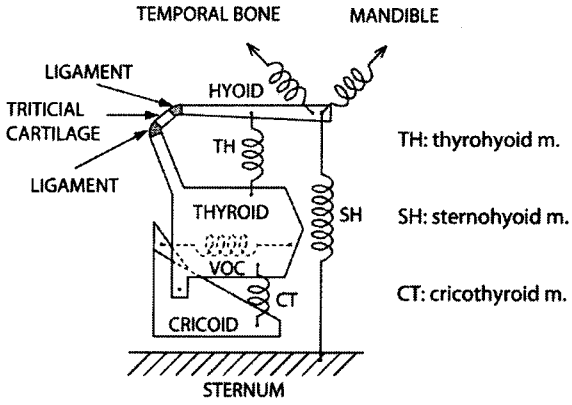


Figure 11. Mechanism of F_0 lowering by activities of TH and SH.

2.2.4 Roles of extrinsic laryngeal muscles [26]

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 Mandarin [31, 32], the five tones of Thai [33] as well as of the grave accent of Swedish [34], 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 [35]. <Figure 10> shows the relationship between the hyoid bone, thyoid and cricoid cartilages, and TH in their lateral and frontal views, and <Figure 11> shows their 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 <Figure 11>.

2.3 Some implications to phonetics and phonology

2.3.1 Typological classification of languages based on the polarity of local commands

The command-response model has already been shown to apply to the F_0 contours of utterances of more than 20 languages, in a number of studies by the present author and his coworkers, as well as by others [36-39]. The results of these studies indicate that these languages fall broadly into the following two groups:

(1) languages in which only positive local commands are

commonly used

(2) languages in which both positive and negative local commands are commonly used

The second group can further be divided into two sub-groups:

(2a) those in which the use of negative commands is lexically determined and thus is not optional (this sub-group includes both non-tone languages and tone languages)

(2b) those in which the use of negative commands is more or less optional, but is rather common.

<Table 1> shows the classification of the languages thus far studied by the author and his coworkers.

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

Group	Polarity of accent/ tone commands	Languages		
1	Positive only	English*, Greek, Polish,	Estonian, Japanese, Spanish	German, Korean,
2	Positive, zero and negative	Hindi, Swedish, Cantonese†, Vietnamese†	Portuguese, Mandarin†,	Russian, Thai†,

* Certain speakers of English (both American and British) occasionally use negative accent commands, especially in order to express paralinguistic information.

† Tone languages.

It is to be noted, however, that the classification is never exact in reality. Even in languages of Group (1), para-linguistic factors such as expression of incredibility, etc., may induce the speaker to invert the polarity of the accent command from positive to negative. Also, different dialects of a language may differ significantly in such use of negative commands.

2.3.2 Phonological structure of the tone system of a tone language

Let us first look at the local (*i.e.*, tone) commands of the four tones of Mandarin. As shown in <Figure 12>, the tone commands for the four tones are: a single positive tone command for the most part of the final (*viz.* the vowel plus the coda) in Tone 1, a negative tone command for the earlier part but then switched to a positive tone command for the later part of the final in Tone 2, a single negative tone command for the most part of the final in Tone 3, and a positive local command for the earlier part but switched to a negative tone command for the later part of the final in Tone 4.

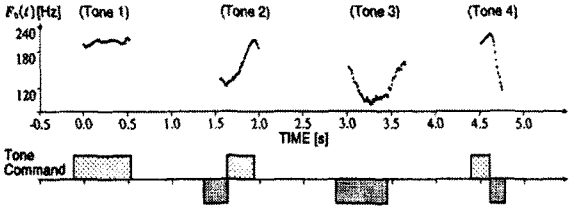


Figure 12. The F_0 contours and the underlying tone command patterns for the isolated words "yi" of four lexical tones of Mandarin [18].

If we assume, however, the existence of two tone commands for each tone, by regarding the single positive command of Tone 1 as a pair of positive tone commands of the same amplitude occurring respectively at the earlier part and the later part of the final, and by regarding the single negative tone command of Tone 3 as a pair of negative tone commands of the same amplitude occurring respectively at the earlier part and the later part of the final, we get the qualitative constellation of the four tones of Mandarin Chinese shown in <Figure 13>, where the polarity of the local command for the earlier part of the final is indicated on the horizontal axis, while that for the later part of the final is indicated on the vertical axis.

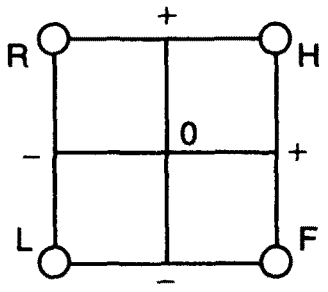


Figure 13. Phonological structure of the tone system of Mandarin. H: High (Tone 1), R: Rising (Tone 2), L: Low (Tone 3), and F: Falling (Tone 4).

In the case of Mandarin, the polarities of the tone commands are always positive or negative, but in other tone languages the tone commands can be partially or entirely null for certain tones.

<Figure 14> shows the constellations of tones of six tone languages including Mandarin already shown in <Figure 13>. A

single circle indicates a tone, while a double circle indicates overlapping of two tones, such as an entering tone and its non-entering counterpart [40]

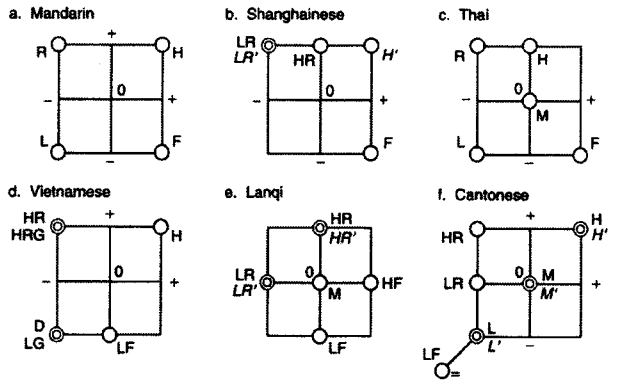


Figure 14. Phonological structure of the tone systems of six tone languages on the basis of timing and polarity of tone commands. For the tone system of Lanqi, see [41].

The figure shows the phonological structures of the tone system of these tone languages in a clear and compact manner, as compared with the traditional ways of representing the tone systems, either in terms of binary features or in terms of 5-tone levels. It is to be noted, however, that this system of representation is not meant to be final and complete. For instance, the tone system of Suzhou dialect requires more dimensions for representation [22].

3. A model of the cognitive processes involved in the task of discrimination

3.1 Background

The systematic use of synthetic speech as stimuli in the study of speech perception was first introduced by the researchers at Haskins Laboratories [42], and has since disclosed a number of important facts about speech perception. In particular, the existence of the categorical effect in discrimination, namely, the increase of discriminability between a pair of speech stimuli at phoneme boundaries [43-46], has become a subject of considerable interest and has given rise to various interpretations such as the motor theory of speech perception [47], the existence of a specific speech mode as contrasted to the normal mode of auditory perception [45], or the existence of a special decoding mechanism operating only in the perception of certain classes of speech sounds [48].

Although these interpretations have been based on an insight into the underlying mechanisms, they remained rather qualitative and speculative, and failed to show the quantitative mechanism that produces the categorical effect. This may be partially ascribable to the fact that the quality of synthetic speech stimuli as well as the accuracy of their parameters were not sufficient to guarantee reliability of results obtained in some of the earlier experiments. These difficulties, however, can be completely overcome by using high-quality stimuli generated on the basis of the acoustic theory of speech production, and by adopting digital computer techniques both for the synthesis and the compilation of the stimuli to obtain accuracy of stimulus parameters. Thus a more quantitative analysis of the perceptual phenomena and a deeper understanding of their underlying mechanisms are now possible by accurate and carefully designed experiments.

As described in our previous reports [49, 50], experimental investigations of various classes of speech sounds including vowels, fricatives, voiced stops as well as semivowels and liquids have been conducted with particular emphasis on the accuracy of stimulus parameters and the choice of experimental conditions such as duration of stimuli and context. The results of these investigations have indicated, contrary to the results of some earlier experiments [43, 44], that the so-called continuous and categorical modes of perception are not absolutely dichotomous, and the categorical effect manifests itself, though at varying degrees, in the discrimination of all the classes of speech sounds examined. On the basis of these results, we presented a model for the cognitive processes involved in the task of discrimination tests of speech sounds, which lends itself to a quantitative interpretation of the so-called phenomenon of categorical perception [51].

3.2 Analysis of the cognitive processes involved

The existence of the categorical effect in the discrimination of a wide range of classes of speech sounds strongly suggests that neither the categorical judgment nor the comparative judgment of stimuli dominates the discrimination of speech, but their interaction characterizes the so-called "speech mode" of perception. For the quantitative analysis and interpretation of experimental results, however, it is necessary to construct a model of the processes involved in a discrimination test.

Throughout this paper, we assume that the ABX procedure is

adopted for the measurement of discriminability, though similar formulations are possible also for other procedures. In the ABX procedure, three stimuli A, B and X are presented successively to the subject, where stimulus X is selected to be identical either to A or to B. The subject is asked to determine by forced choice whether X is the same as A or as B. The followings are three basic postulates for the model.

(1) A sound stimulus above threshold produces sensation of timbre regardless of whether it is speech or nonspeech. In the case of speech sounds, however, phoneme identification takes place immediately thereafter by a process of categorical judgment based on the sensation of timbre. The immediate judgment of linguistic category is what characterizes the speech mode. Thus in a discrimination test of speech stimuli by the ABX procedure, identification of individual stimuli always precedes discrimination judgment, which takes place only after the reception of all the three stimuli.

(2) When a set of proper dimensions are adopted for the specification of physical characteristics of the stimulus, the mapping characteristic for the auditory mechanism can be considered to be continuous and monotonic over the range between two phonemes that are physically adjacent on the stimulus continuum. Thus a properly selected monotonic division of such a range produces an interval scale on the perceptual continuum.

(3) In a discrimination test of speech stimuli by the ABX procedure, preference is given to categorical judgment over comparative judgment as long as the results of the former are useful for discrimination. Namely, the discrimination is based solely on the results of phonemic identification when the stimuli A and B are judged to belong to different categories. Only when the results of categorical judgment are useless for discrimination, subjects rely on comparative judgment of the three stimuli on the perceptual continuum of timbre.

3.3 The model and its mathematical formulation

On the basis of these postulates, a model can be constructed for the mechanisms and processes of discriminating speech sounds by the ABX procedure, and is shown in <Fig. 15>. The first block ① represents the auditory process for mapping the stimulus continuum onto the perceptual continuum of timbre, and its output is stored in a short-term memory ② for timbre but at the same time is fed to the next stage of phoneme identification ③, where immediate categorical judgment is made on the basis of phoneme boundaries stored in a long-term memory ④. The output of block ③ is a phonemic symbol and is stored in a separate short-term memory ⑤. After all the three stimuli A, B and X are perceived by these processes, the results are utilized at the next stage of forced discrimination ⑥. According to the above-mentioned Postulate (3), the discrimination is based on the short-term memory for phonemic symbols ⑤ when stimuli A and B have been identified as different phonemes, but is based on the short-term memory for stimulus timbre ② when A and B have been identified as the same phoneme. The difference in quality of the two short-term memories is to be noted. Namely, the stimulus timbre stored in the short-term memory ② is an analog and continuous form of information and is liable to lapses or fluctuations during the course of retention and retrieval, while the phonemic symbol stored in the short-term memory ⑤ is a discrete and encoded form of information and remains quite stable at least for a period of several seconds which is of interest in the ABX procedure.

For the purpose of quantitative description of the entire process of discrimination, the following characteristics of the model have to be formulated mathematically.

(a) Uncertainty in the process ① of mapping from the stimulus continuum to the perceptual continuum. As a first approximation, this can be represented by a Gaussian random noise in the process of mapping. The variance of the Gaussian noise can be considered to be the same over the entire range of stimulus timbre under examination.

(b) Uncertainty in the phoneme boundary. Fluctuations in the long-term memory ④ for the phoneme boundary occur by the influence of long-term factors and short-term factors. The former can be approximated by slow Gaussian random variations of the phoneme boundary, which can be considered constant for a period of several seconds neces-sary for a single discrimination judgment, but considered to be statistically distributed over the entire discrimination test. The latter can be further divided into fast random variations and the so-called context effect, which is a temporary shift in the phoneme boundary due mainly to the influence of the immediately preceding stimulus.

(c) Uncertainty in the short-term memory for timbre. The stimulus timbre, after it has undergone the process of retention and

retrieval through the short-term analog memory ②, is subject to random fluctuations and mutual interactions. As a first approximation, however, these variations can be considered as random Gaussian noise superposed independently on the timbre of each stimulus.

Although it is possible to formulate all of these factors in the description of a complete model of the discrimination process, the numerical calculation becomes prohibitively complex as the number of statistical factors involved in the model is increased. It is therefore reasonable to find those factors that have dominant influences on the behavior of the model and to formulate an approximate model. On the other hand, it is rather difficult to make separate estimate of the effect of each factor from the results of discrimination tests.

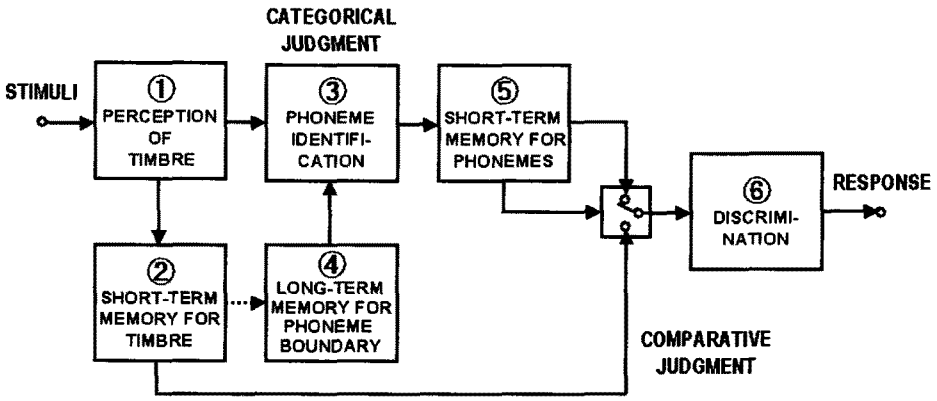


Figure 15. A model for the cognitive processes involved in the task of discriminating speech sounds.

An alternative method for determining the dominant factors is to construct several approximate models, taking into account a few factors at a time, to analyze the experimental data by the method of analysis-by-synthesis based on these models, and to select the model which gives the closest approximation to the human behavior represented by the data.

As an example of such an approximate model, our preliminary investigations suggest the following assumptions:

(b')Uncertainty in the phoneme boundary can be characterized by long-term Gaussian variations of the boundary with a mean value μ and a standard deviation σ_0 on the continuum of timbre.

(c') Uncertainty in the short-term memory for timbre can be characterized by short-term independent Gaussian variations of the stimulus timbre around its value at the instant of mapping with a standard deviation σ_i, where i indicates the order of the stimulus in an ABX triad.

(a') All other factors, including uncertainty in the process of mapping, can be neglected as compared to the above two factors.

Denoting by x_1 and x_2 the respective positions of stimuli A and B at the instant of mapping the continuum of timbre, x, the discriminability P_{ABX} of a pair of stimuli as measured by the ABX method can be given by,

$$P_{ABX} = P_1 + (1 - P_1)P_2, \tag{11}$$

$$P_1 = \int_{x_1}^{x_2} \phi(x;\mu,\sigma_0)dx,$$

$$P_2 = \frac{1}{2}\left[\int_{-\infty}^{\infty}\phi(u;x_1,\sigma_1)du\int_{u}^{\infty}\phi(v;x_2,\sigma_2)dv \times\left\{\int_{-\infty}^{\frac{u+v}{2}}\phi(w;x_1,\sigma_3)dw + \int_{\frac{u+v}{2}}^{\infty}\phi(w;x_2,\sigma_3)dw\right\} + \int_{-\infty}^{\infty}\phi(u;x_1,\sigma_1)du\int_{-\infty}^{u}\phi(v;x_2,\sigma_2)dv \times\left\{\int_{\frac{u+v}{2}}^{\infty}\phi(w;x_1,\sigma_3)dw + \int_{-\infty}^{\frac{u+v}{2}}\phi(w;x_2,\sigma_3)dw\right\}\right],$$

where $\phi(t; m, s)$ denotes a normal probability density function on a variable t with mean m and standard deviation s. The first term in Eq (11) represents the contribution of categorical judgment, while the second term represents that of comparative judgment of timbre.

For the purpose of testing the validity of the model and the assumptions described above, a vowel discrimination test was designed and performed. In view of the fact that the methods of previous experiments fell short of accuracy in the stimulus parameters as well as of resolution in the measurement of a

discrimination curve, special effort was made to achieve both accuracy and resolution sufficient for the quantitative analysis of the results. Namely, the stimuli were synthesized by simulating a terminal-analog synthesizer on a digital computer. Formant frequencies and bandwidths were specified exactly in steps of 1 Hz, thus achieving an accuracy not attainable by analog synthesizers. Furthermore, the stimuli were read out from the computer with an exact timing according to random numbers, and recorded on an analog magnetic tape for off-line discrimination tests, thus eliminating both the labor and the inaccuracy unavoidable in dubbing and splicing of recorded tape segments.

On the other hand, higher resolution of the discrimination curve was obtained by a finer division of the same stimulus range, and by using stimulus pairs that overlap each other. The resulting increase in the number of triads and their compiling work was considerable, but could be handled by a computer within a reasonable computation time. The details of stimuli and experimental procedures are described elsewhere [51].

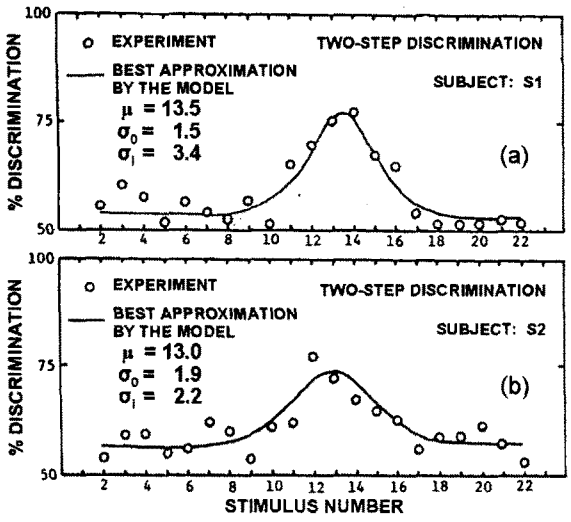


Figure 16. Results of two-step discrimination test of vowels [i]-[e] for two subjects and their analysis-by-synthesis based on the model.

Because of differences in the phoneme boundary as well as in other parameters that characterize individual performance, the performance of each subject requires separate analysis. <Figure 16> shows the performances of two subjects in the ABX test, and each point represents the average of about 400 judgments. The results, as well as those of other subjects, clearly indicate the categorical effect at the phoneme boundary. For the sake of quantitative analysis and interpretation of these results, a computer program was prepared to determine the optimum values of parameters of a model by the method of analysis-by-synthesis. The curves in <Figure 16> are theoretical discrimination curves based on the model giving the closest approximations to the measured data in the sense of the least mean squared error. The values of parameters of the model for each subject are listed in units of the stimulus scale. The σ_i's were assumed to be the same for $i = 1, 2$ and 3.

The fact, that the model is capable of predicting the actual performance of a human subject with a high accuracy, testifies that its underlying assumptions are essential in characterizing the whole process of discrimination of speech sounds. The model-based analysis of the results not only indicates the overall difference between performances of the two subjects, but also tells precisely where and how they differ. Namely, in addition to a small difference in the position of category boundary μ between the two vowels, the subjects differ both in σ_0 and in σ_i: Subject S1 is lower in σ_0 but is higher in σ_i than Subject S2. In other words, S1 is more precise in phoneme identification (because of a lower noise in categorical judgment) but is less accurate in comparative judgment of timbre (because of a higher noise in the short-term memory of timbre) as compared with Subject S2.

The close agreement between the experimental data and the predictions by the model of the discrimination process can be regarded as corroboration of the three basic postulates concerning the perception of speech presented in this study, namely the uniformity of the auditory mapping from the acoustic continuum of stimulus to the perceptual continuum of timbre preceding any linguistic judgment, the existence of categorical judgment of speech sounds immediately after the mapping, and the dominance of categorical judgment over comparative judgment of timbre in discrimination. It also validates the mathematical formulation of the discrimination process deduced from these postulates.

At the time when this study was conducted (*i.e.*, 1969-1971) it was not possible to obtain direct neurophysiological evidences for these inferences. However, findings in the neurophysiology of the human brain at least suggest that the site of the auditory mapping ① and the short-term memory for timbre ② is the auditory center in the temporal lobes of right and left cerebral hemispheres, while the site for the phoneme identification ③ as well as its retention ⑤ is the auditory speech area in the temporal lobe of the left hemisphere. The site for the discrimination judgment based on

these short-term memories, however, is the association area in the frontal lobes. This study, then, is to be regarded as an attempt for the quantitative description of successive stages of information processing performed in the cerebral cortex in the task of discriminating speech sounds.

It should be noted, however, that the present model is not specific to speech, but is applicable to perception of any stimuli that are subject to categorical judgment, regardless of whether or not they are generated by some human motor process. In other words, the categorical effect is always present, at varying degrees, when one conducts a discrimination test on a stimulus continuum that spans over a category boundary of any kind. Thus the categorical effect cannot serve as an evidence supporting the motor theory of speech perception. Also, in speech communication of real life, what is essential is the ability of identification and not the ability of discrimination.

4. Concluding Remarks

The foregoing sections were meant to illustrate the elucudative ability of quantitative models that are based on mathematical formulations of essential characteristics of the underlying human mechanisms and processes. It so happens that the first model is concerned with speech production, and is deterministic in the sense that it deals with the mechanism by which a definite F_0 contour is generated from a set of well-defined input commands. On the other hand, the second model is concerned with speech perception, and is stochastic in the sense that it does not model the process by which a specific output is generated against a specific input (in this case a set of three stimuli), but gives only the probability that a certain decision takes place on the average. They will demonstrate the advantages and disadvantages of the respective approaches.

It seems to me that efforts on the part of the speech research community have been largely successful in modeling various lower-level human functions of speech production and perception, but less so in modeling higher-level functions of generation and comprehension of spoken messages. This may be quite natural, because these higher-level functions are less accessible to direct observations and to mathematical formulations, but also because the NLP research community has tended to treat language as an abstract entity separate from human beings as its users. However, just as production and perception of speech cannot be modeled without understanding the human mechanisms and processes, generation and comprehension of linguistic messages cannot be modeled properly without paying due attention to the related human mechanisms and processes.

In my humble opinion, the source and destination of spoken messages are the minds of speaker and listener. Our attempts to understand and simulate the entire process of speech communication will never be complete unless we try and succeed in modeling the ultimate source and destination of information, namely the speaker's mind and the listener's mind [52, 53]. I believe that it must be a major target of research by the spoken language research community in the future.

This paper is based on the author's keynote speech at INTERSPEECH 2008, Brisbane, Australia.

References

[1] H. Fujisaki, "Prosody, models, and spontaneous speech. In *Computing Prosody"*, Y. Sagisaka, N. Campbell, N. Higuchi (eds.). Springer, New York, pp.27-42, 1997.

[2] S. Öhman, J. Lindqvist, "Analysis-by-synthesis of prosodic pitch contours". *Speech Transmission Laboratory Quarterly Status and Progress Report (STL-QPSR), KTH,* 4/1965, pp. 1-6, 1965.

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

[4] H. Fujisaki, S. Nagashina, "A model for synthesis of pitch contours of connected speech", *Annual Report of the Engineering Research Institute, University of Tokyo,* 28, pp. 53-60, 1969.

[5] H. Fujisaki, 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 13th International Congress of Linguists*, H. Fujisaki, E. Gårding (eds.), Tokyo, pp. 57-70, 1982.

[6] G. Bell, H. Fujisaki, J. Heinz, A. House, K. Stevens, "Reduction of speech spectra by analysis-by-synthesis techniques", *J. Acoust. Soc. Am.*, 33, pp. 1725-1736, 1961.

[7] H. Fujisaki, S. Ohno, "Analysis and modeling of fundamental frequency contours of English utterances", *Proc. 4th European Conference on Speech Communication and Technology,* Madrid, 4, pp. 2231-2234, 1995.

[8] H. Fujisaki, I. Lehiste, "Some temporal and tonal characteristics of declarative sentences in Estonian", *Preprints*

of Papers, Working Group on Intonation, the 13th International Congress of Linguists, H. Fujisaki, E. Gårding (eds.), Tokyo,pp. 121-130, 1982.

[9] H. Mixdorff, H. Fujisaki, "Analysis of voice fundamental frequency contours of German utterances using a quantitative model", *Proc. 1994 Int'l Conf. on Spoken Language Processing*, Yokohama, 4, pp. 2231-2234, 1994.

[10] H. Fujisaki, S. Ohno, T. Yagi, "Analysis and modeling of fundamental frequency contours of Greek utterances", *Proc. 5th European Conference on Speech Communication and Technology*, Rhodos, 1, pp. 465-468, 1997.

[11] H. Fujisaki, "Analysis and modeling of fundamental frequency contours of Korean utterances－A preliminary study－", In *Phonetics and Linguistics － in Honour of Prof. H. B. Lee*, Seoul, pp. 640-657, 1996.

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

[13] H. Fujisaki, S. Ohno, "Analysis and modeling of fundamental frequency contours of Hindi utterances", *Proc. International Conference on Speech Science and Technology (INTERSPEECH) 2005*, Lisbon, pp. 1413-1416, 2005.

[14] H. Fujisaki, S. Narusawa, S. Ohno, D. Freitas, "Analysis and modeling of F_0 contours of Portuguese utterances based on a quantitative model", *Proc. 8th European Conference on Speech Communication and Technology*, Geneva, 3, pp. 2317-2320, 2003.

[15] H. Fujisaki, S. Ohno, "Analysis of fundamental frequency contours of Russian utterances using the command-response model", *Proc. 2005 Autumn Meeting, Acoust. Soc. Jpn*, pp.337-338, 2005.

[16] H. Fujisaki, M. Ljungqvist, H. Murata, "Analysis and modeling of word accent and sentence intonation in Swedish", *Proc. ICASSP 93*, Minneapolis, 1, pp. 211-214, 1993.

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

[18] H. Fujisaki, C. Wang, S. Ohno, W. Gu, "Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command-response model", *Speech Communication*, 47, pp. 59-70, 2005.

[19] W. Gu, K. Hirose, H. Fujsaki, "Analysis of F_0 contours of Cantonese utterances based on the command-response model", *Proc. INTERSPEECH 2004*, Cheju,pp. 481-784, 2004.

[20] H. Fujisaki, S. Ohno, S. Luksaneeyanawin, "Analysis and synthesis of F_0 contours of Thai utterances based on the command-response model", *Proc. Int'l Congr. of Phonetic Sciences*, Barcelona, 2,pp. 1129-1132, 2003.

[21] H. Mixdorff, H. Nguyen, H. Fujisaki, C. Luong, "Quantitative analysis and synthesis of syllabic tones in Vietnamese", *8th European Conference on Speech Communication and Technology*, Geneva, 1, pp. 177-180, 2003.

[22] W. Gu, K. Hirose, H. Fujisaki, "Modeling the tones in Suzhou and Wujiang dialects on the basis of the command-response model", *Proc. TAL 2006*, La Rochelle, pp. 59-62, 2006.

[23] Y. Chao, *A Grammar of Spoken Chinese*, Univ. of Calif. Press, pp. 121-134, 1968.

[24] Z. Wu, "A new method of intonation analysis for Standard Chinese: frequency transposition processing of phrasal contours", In *Analysis, Perception and Processing of Spoken Language*, G. Fant, K. Hirose, S. Kiritani (eds.), Elsevier Science B. V., Amsterdam, pp. 255-268, 1996.

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

[26] H. Fujisaki, "Information, prosody, and modeling －with emphasis on the tonal features of speech－", In *From Traditional Phonology to Modern Speech Processing*, G. Fant, H. Fujisaki, J. Cao, Y. Xu (eds.). Foreign Language and Teaching Research Press, Beijing, pp. 111-128, 2004.

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

[28] W. Sandow, "A theory of active state mechanisms in isometirc muscular contraction", *Science*, 127, pp. 760-762, 1958.

[29] W. R. Zemlin, *Speech and Hearing Science, Anatomy and Physiology*, Prentice-Hall, New York, 1968.

[30] B. R. Fink, R. J. Demarest, *Laryngeal Biomechanics*. Harvard University Press, Cambridge, Mass.,1988.

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

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

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

[34] E. Gårding, "Word tones and larynx muscles. *Working*

Papers, Dept. of Linguistics and Phonetics, Lund University, 3, pp. 20-46, 1970

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

[36] B. Möbius, M. Pätzold, W. Hess, "Analysis and synthesis of German *F*0 contours by means of Fujisaki's model", *Speech Communication,* 13, pp. 53-61, 1993.

[37] H. Mixdorff, N. Amir, "The prosody of modern Hebrew", *Proc. Speech Prosody 2002*, Aix-en-Provence, pp. 511-515, 2002.

[38] H. Mixdorff, M. Vainio, S. Werner, J. Järvikivi, "The manifestation of linguistic information in prosodic features of Finnish", *Proc. Speech Prosody 2002*, Aix-en-Provence, pp. 516-519, 2002.

[39] E. Navas, I. Hernáez, A. Armenta, B. Exebaria, J. Salaberria, "Modelling Basque intonation using Fujisaki's model and CARTs", *Proc. Seminar on State of the Art in Speech Synthesis*, London, 3/1-3/6, 2002.

[40] H. Fujisaki, W. Gu, "Phonological representation of tone systems of some tone languages based on the command-response model for F_0 contour generation", *Proc. TAL 2006*, La Rochelle pp. 59-62, 2006.

[41] K. Richard, "*An acoustic-phonetic descriptive analysis of Lanqi citation tones",* Unpublished Honours Thesis, Australian

National University, 2005.

[42] F. S. Cooper, P. Delattre, A. M. Liberman, J. Borst, L. J. Gerstman, "Some experiments on the perception of synthetic speech sounds", *J. Acoust. Soc. Am.*, 24, pp. 597-606, 1952.

[43] A, M. Liberman, K. S. Harris, H. S. Hoffman, B. C. Griffith, "The discrimination of speech sounds within and across phoneme boundaries. *J. exp. Psychol,*. 54, pp. 358-367, 1957.

[44] D. B. Fry, A. S. Abramson, P. F. Eimas, A. M. Liberman, "The identification and discrimination of synthetic vowels", *Lang. Speech,* 5, pp. 171-188, 1962.

[45] K. N. Stevens, "On the relations between speech move-ments and speech perception", *Zeitschrift Phon. Sprach. Kommunikationsforschung,* 21, pp. 102-106, 1968.

[46] K. N. Stevens, A. M. Liberman, M. Studdert-Kennedy, S. E. G. Öhman, "Crosslanguage study of vowel perception", *Lang. Speech,* 12, pp. 1-23, 1969.

[47] A. M. Liberman, F. S. Cooper, K. S. Harris, P. F. MacNeilage, "A motor theory of speech perception", *Proc. Speech Communication Seminar, Stockholm,* Vol. II, Paper D3, 1962.

[48] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, M. Studdert-Kennedy, "Perception of the speech code. *Psychol. Rev.*, 74, pp. 431-461, 1967.

[49] H. Fujisaki, T. Kawashima, "On the modes and mecha-nisms of speech perception", *Annual Report of the Engineering Research Institute, University of Tokyo,* 28, pp. 67-73, 1969.

[50] H. Fujisaki, T. Kawashima, "Some experiments on speech perception and a model for the perceptual mechanism", *Annual Report of the Engineering Research Institute, University of Tokyo,* 29, pp. 207-214, 1970.

[51] H. Fujisaki, T. Kawashima, "A model of the mechanisms for speech perception — Quantitative analysis of categorical effects in discrimination —", *Annual Report of the Engineering Research Institute, University of Tokyo,* 30, pp. 59-68, 1971.

Also, H. Fujisaki, "On the modes and mechanisms of speech perception — Analysis and interpretation of categorical effects in discrimination", In *Frontiers of Speech Communication Research*, B. Lindblom, S. Öhman (eds.), Academic Press, London, pp. 177-189, 1979.

[52] H. Fujisaki, "Communication between minds — the ultimate goal of speech communication and the target of research for the next half-century", *Proc. 16th Congress on Acoustics,* Seattle, *pp.* 2399-2400, 1998.

[53] H. Fujisaki, "Retrospects and prospects of speech communication research", *Proc. 18th Congress on Acoustics,* Kyoto, 5, pp. 3763-3768, 2004

Hiroya Fujisaki is Professor Emeritus at the University of Tokyo, where he held professorship not only at the Faculty of Engineering, but also at the Research Institute of Logopedics and Phoniatrics, Faculty of Medicine and at the Department of Linguistics, Graduate School of Humanities. He has been engaged in a wide range of problems in both Speech Science and Speech Technology, published more than 600 articles and authored/edited more than 30 books. He also led several important projects in Japan on Speech Science and Technology, chaired major international conferences including ICASSP 86 and the ASA-ASJ Joint Meeting, 1988, and founded and chaired ICSLP in 1990, which defined the interdisciplinary field of Spoken Language Processing. His international responsibilities include: member of Permanent Council for ICPhS, life member of International Advisory Council of ISCA, and member of International Advisory Board of ASSTA. For his academic works and technical leadership activities, he received a number of awards from ASA, IEEE, IEEJ, IEICE and ISCA, including the Third Millennium Medal from IEEE and the Medal for Scientific Achievement from ISCA, and was named Person of Merit in Science and Technology by the mayor of Tokyo. He is an honorary member of ASJ, fellow of ASA and IEICE, life member of IEEE, member of ISCA, member of the Engineering Academy of Japan, and corresponding member of G ▪ ttingen Academy of Sciences.

E-mail: fujisaki@alum.mit.edu

URL: http://homepage3.nifty.com/hiroya_fujisaki/