Invited Plenary Lecture at SICOPS '96

Toward a Multi- and Inter-disciplinary Science of Human Speech Communication

Hiroya Fujisaki

Professor Emeritus, University of Tokyo
Professor, Science University of Tokyo

1. Introduction

The human abilities of thought and communication owe very much to the use of language as a unique system of codes by which the information is expressed as messages. Communication by language, however, is achieved only when the messages are converted into physical signals, i.e., either into speech as their acoustic manifestations or into characters as their optical manifestations. The primary importance of the spoken language as compared to the written language is apparent from its exclusive use by many of the primitive cultures as well as from the order of acquisition by infants.

The ultimate goal of speech science is obviously to understand the entire process of human speech communication: from the information generated by the mind of a speaker to the information retrieved by the mind of a listener. In this lecture, I will first try to describe the multi- and inter-disciplinary nature of our efforts toward this goal. This point will be illustrated by a few example from my own works. I will then present some relatively unexplored area as new frontiers of research for the coming decades, and will conclude by stating my personal view on how the ultimate goal could be attained.

The human processes involved in speech communication can be illustrated by Fig. 1. "Verbalization" is the process of putting the speaker's ideas into discrete, linguistic codes.

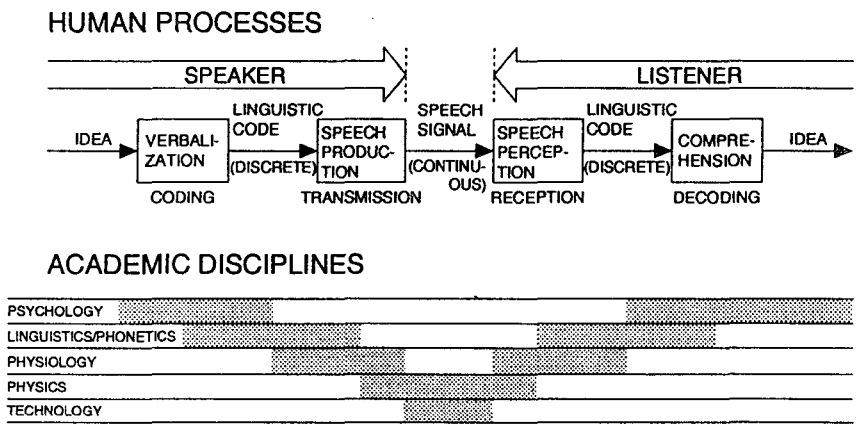

Fig. 1. Processes involved in human speech communication and related academic disciplines.

These discrete codes are then converted into speech, i.e., a continuous physical signal, by the process of "speech production." The speech signal then propagates through the medium and reaches the ear of the listener. On the part of the listener, the linguistic codes are restored from the speech signal by the process of "speech perception," and are subsequently decoded into the idea by "comprehension." The lower part of the figure shows the academic disciplines that are related to these processes, and the shaded bars indicate the stages that are most directly concerned with the respective disciplines. Thus the figure indicates that the whole process of human speech communication is related to a number of academic disciplines [1]. Understanding the whole process therefore is a multi-disciplinary problem, and the elucidation of even a small part often requires an inter-disciplinary approach, as will be shown in the following sections.

2. Language and Mind — Mathematical Formulation of the Processes of Coding and Decoding Information by Language [2, 3]

2.1 The problem

The process of communication by language may quite generally be represented by Fig. 2. In this process, a certain part of the information I_1, which one person (the sender) possesses and intends to transmit, is transformed into a linguistic expression E and is presented to another person (the receiver). Upon receiving the expression E, the receiver acquires the information I_2 which generally is an approximation to the information I_1 intended by the sender. In ordinary situations, the linguistic expression E is always converted into a physical signal for the purpose of transmission, and lends itself to quantitative measurement and objective description. The information contents I_1 and I_2, on the other hand, generally elude direct observation, since they are only represented by the psychological states of the sender and the receiver. Quantitative descriptions of the sender's coding characteristics and the receiver's decoding characteristics are not possible under these circumstances.

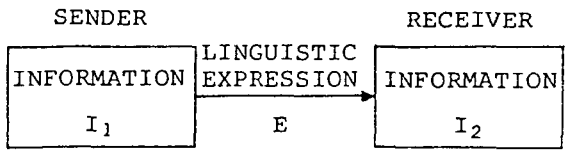

Fig. 2. The process of communication by language.

In order to bring these coding and decoding processes into more tractable forms, we shall adopt a situation where an object O_1 with a physically observable attribute is presented to a subject, who as a sender selects an expression E to describe the object O_1. The expression E is presented to another subject, who as a receiver selects, among various alternatives, an object O_2 which he considers to be implied by the expression E. The situation is schematically shown by Fig. 3. The addition of physically observable objects at both ends of the communication process thus makes it possible to measure the coding and decoding processes by the method of experimental psychology, and to describe their characteristics in quantitative terms, though it imposes certain limitations on the nature of information to be transmitted.

The study to be described here is restricted to situations where a physically measurable attribute of objects, such as the age of a person or the color of a paint, constitutes the information to be transmitted through the use of language. The linguistic expressions to be

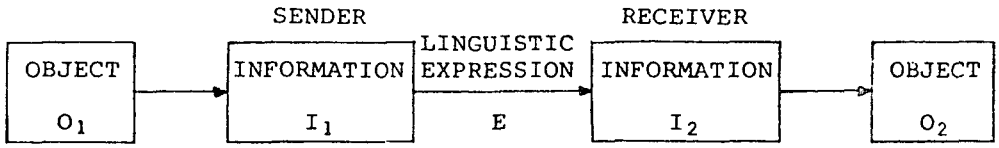

Fig. 3. The process of communication by language where physically observable objects are added to facilitate quantitative descriptions of sender and receiver characteristics.

used for transmission are also restricted to a pre-determined set of nouns. The psychological processes involved in such communication situations are shown in Fig. 4. When the physical attribute is presented as a stimulus to the sender, it is converted into a percept on a certain perceptual continuum. The selection of a linguistic expression is based on quantization and coding of the perceptual continuum. Thus the sender's process of verbalization can be considered as consisting of two sub-processes: perception of a stimulus and its expression by language. Since the intervening percept eludes direct measurement, however, these two sub-processes are treated as a single process and its input-output characteristics are defined as the sender's coding characteristics.

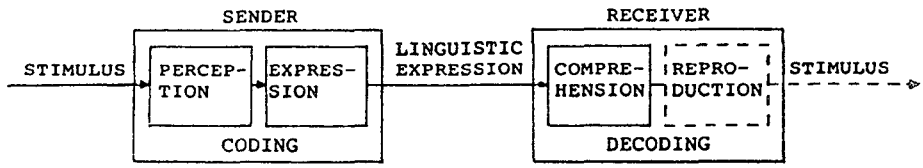

Fig. 4. Processes in transmission of meaning by language in the communication situation of Fig. 3.

The linguistic expression (a noun in this case) selected by the sender is transformed into letter strings, and is presented as a stimulus to the receiver. The perceived expression is decoded by the receiver and reconstructs a percept in the receiver's mind, which is an approximation to the one in the sender's mind. While information transmission is accomplished and communication in the ordinary sense is terminated by this decoding process, another process has to be added in order for the receiver's percept to be transformed into a measurable entity. In this study, the receiver is asked to reproduce the original stimulus which is implied by the received expression. The reproduction is accomplished by selecting a physical stimulus from a number of candidates. Thus the receiver's process of understanding a message consists of two sub-processes: comprehension and reproduction. As in the case of the sender's coding characteristics, these two sub-processes are treated as one and its input-output characteristics are defined as the receiver's decoding characteristics.

These coding and decoding processes are decision processes. Since human decisions are not exempt from statistical fluctuations caused by a number of psychological and physiological factors, characteristics of these processes may most properly be described in probabilistic terms. In other words, a sender's coding characteristics can be represented by the set of probability distributions that each one of the possible expressions displays on the continuum of stimuli presented to the sender. On the other hand, a receiver's decoding characteristics can be represented by the set of probability density functions of the stimuli reproduced by the receiver against each one of the transmitted expressions.

2.2 Experimental method

As an example of experimental determination of the coding and decoding characteristics involved in the above-mentioned model of semantic information transmission, an investigation was conducted on the relationship between the age of a person and the nouns of Japanese commonly used to designate the age. Most languages, including Japanese, provide nouns for classifying and designating age ranges, but the age ranges they designate are not necessarily equal in size, nor complementary in their distribution. Moreover, the information conveyed by these nouns is not restricted to the chronological age of a person, and the use of these nouns is influenced by a number of contextual factors. In order to minimize the effects of these extraneous factors, a set of nouns were selected such that they were nearly complementary with each other in designating age and nearly uniform in other aspects, and the measurement of coding and decoding characteristics were conducted using this pre-determined vocabulary. The vocabulary used in the major part of the following experiments consisted of the five Japanese nouns: "*yō-nen*", "*shō-nen*", "*sei-nen*", "*sō-nen*", and "*rō-nen*", corresponding roughly (but not exactly) to the English "childhood", "boyhood", "youth", "manhood", and "old age", respectively. In some part of the experiments, however, the size of the vocabulary was controlled to see the effect.

Subjects A total of nine subjects, eight male adults and one female adult, took part in the following experiments. They were all native speakers of Japanese and their ages ranged from 22 to 47. Each subject served both as a sender and a receiver. Since individual differences are naturally found in characteristics of language users, the experimental data of the nine subjects were not pooled, but were analyzed individually to extract coding and decoding characteristics of each subject.

Measurement of coding characteristics The measurement of a sender's coding characteristics was conducted by the method of constant stimuli. A randomized list of integers from 0 to 70 was presented sequentially to a subject by a digital computer. The subject was instructed to select, by forced judgment, a noun from a pre-determined vocabulary which he or she considered to be appropriate for the age represented by the integer. Each subject made at least 10 responses to each integer presented, and the individual data were analyzed to obtain the probability distributions of the words on the age scale.

Measurement of decoding characteristics The measurement of a receiver's decoding characteristics was conducted also by the method of constant stimuli. A randomized list of nouns in a pre-determined vocabulary was presented to a subject by a digital computer. The subject was instructed to select, by force judgment, an integer which he or she considered to be appropriate for the noun presented. Each subject made at least 100 judgments on each of the nouns, and the individual data were analyzed to obtain the probability density functions of the ages reproduced from the nouns.

The measurements of coding and decoding characteristics can naturally be conducted independently, but the results can be combined to study the communication process where information is transmitted between an arbitrary pair of subjects.

2.3 Results

Coding characteristics As an example of a sender's coding characteristics, the performance of one subject in the five-category coding experiment is shown in Fig. 5. It can be seen that the probability distributions of the five nouns are contiguous, i.e., that the decay of the probability of occurrence of one category is accompanied only by the rise of the probability of another category, and no more than two categories occur at any point on the age axis.

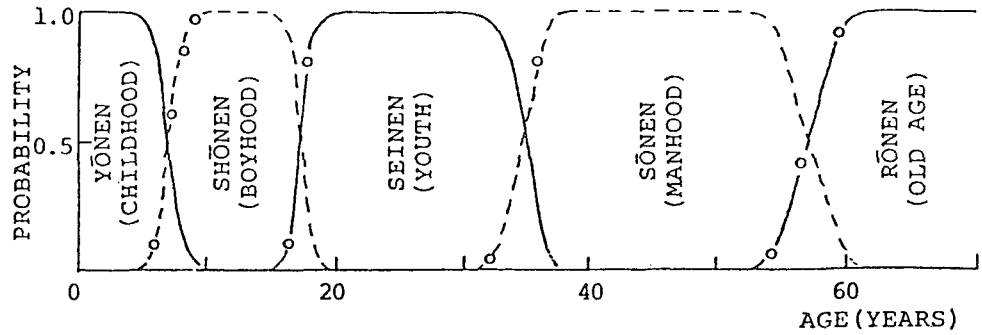

Fig. 5. An example of a sender's coding characteristics.

In this case, the categorical judgment, i.e., the choice of a particular category against a given stimulus, can be regarded as a binary threshold operation whose threshold fluctuates due to a number of psychological and physiological factors. Hence the probability distribution of a category can be approximated, in the vicinity of the category boundary, by a Gaussian distribution. The validity of the approximation is demonstrated by Fig. 6, where the probability of occurrence of "*rō-nen*" for the same subject is plotted on the normal scale against the age. Thus a transition from one category to another can be characterized by the mean θ and the standard deviation ρ, to be defined respectively as the coding boundary and the coding accuracy, of the probability distribution for one of the categories. The coding characteristics of a sender can then be characterized by the set of θ's and ρ's for all the category boundaries. Quite naturally, individual differences are to be expected in the values of these parameters.

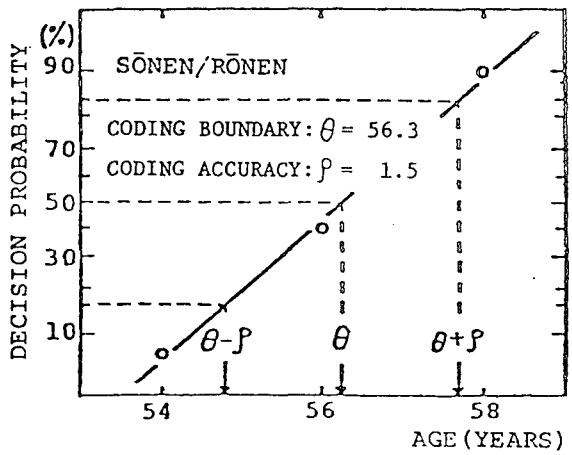

Fig. 6. Approximation of coding characteristics by a Gaussian distribution.

Decoding characteristics As an example of a receiver's decoding characteristics, the performance of one subject in the five-category decoding experiment is shown in Fig. 7.

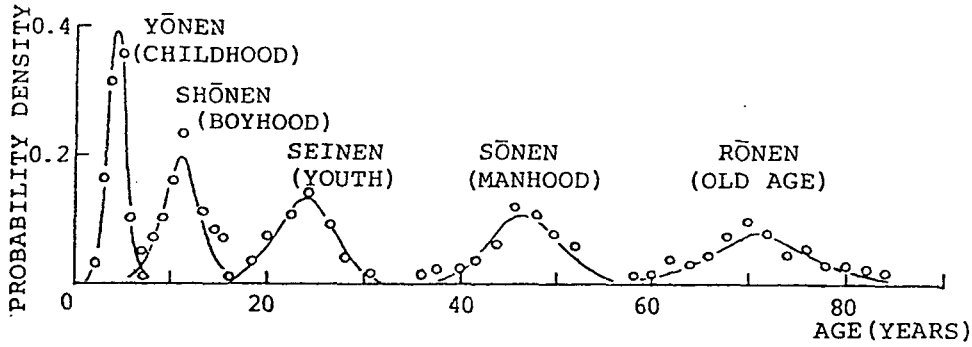

Fig. 7. An example of a receiver's decoding characteristics.

The decoding characteristics are represented by the set of five probability density functions corresponding to the five categories, and are quite different from the coding characteristics. If we assume that a receiver's response to a noun is based on a certain psychological reference but is perturbed by a number of psychological and physiological factors, the probability density function may be approximated by a Gaussian density function. The validity of the approximation is demonstrated by Fig. 8, where the cumulative probability for the response to "*sei-nen*" is calculated from the decoding data of the same subject and is plotted on the normal scale against the age. Thus a receiver's response to a particular noun can be characterized by the mean μ and the standard deviation σ, to be defined respectively as the decoding reference and the decoding accuracy, of the probability density function. The decoding characteristics of a receiver can then be characterized by the set of μ's and σ's for all the noun categories. As in the case of coding characteristics, individual differences are to be expected in the values of these parameters.

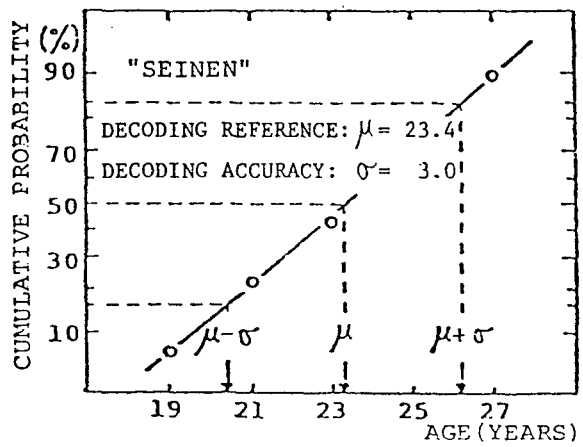

Fig. 8. Approximation of decoding characteristics by a Gaussian distribution.

2.4 Comments

To the best of the present author's knowledge, the studies by Lenneberg and his co-workers are almost the only ones which have treated the problem of language use from a probabilistic point of view [4, 5]. It may not be an overstatement to say that these studies represent a pioneering effort to introduce methods of experimental psychology into the study of the transmission of meaning. The use of unrestricted vocabulary and the pooling of data by individual subjects, however, did not allow these authors to attain a clear insight into the underlying process of the coding behavior of a language user. It should also be mentioned that their methods to associate meaning with words failed to reveal the essential characteristics of the decoding process in a receiver.

The formulation of a functional relationship between a physical variable and a linguistic expression is treated also by Zadeh, though quite qualitatively, as an example in his proposal for the concept of "fuzziness." [6] The membership function is defined, not as the probability, but as the "grade of membership" that a stimulus is considered to belong to a "fuzzy" set. The present author considers, however, that the membership function proposed by Zadeh is exactly equivalent to the probability distributions in the coding characteristics obtained by observing the sender's behavior. Despite the conceptual distinction made by Zadeh between probability and membership function, the experimental procedure of determining the membership function will be exactly the same as that of determining the coding characteristics described in the present study. It should also be noted that the concept of fuzziness corresponds only to one kind of indeterminacy in language use, i.e., the indeterminacy in regards to decoding. The formulation of indeterminacy in language use cannot be complete unless both of these factors are taken into account.

3. Phonetics and Phonology — Articulatory Description of the Korean Vowel System [7, 8]

3.1 The problem

The shape of the vocal tract, which is the primary factor contributing to the acoustic/phonetic characteristics of a vowel, is mainly determined by the shape of the tongue surface relative to the lips, palate and pharyngeal wall, which are determined, in turn, by jaw opening, shape and position of the tongue relative to the mandible, and lip rounding. The displacements/deformations of these articulators are essentially continuous variables, but will have to represent the discreteness of phonemic information. In the study to be reported here, simultaneous measurements are made of the jaw opening angle, the tongue surface shape relative to the mandible, and the area of the oral aperture in the production of sustained vowels of Korean. The results of measurement are interpreted in quantitative terms, leading to an articulatory description of the system of Korean vowels.

3.2 Method and results

An X-ray photography system and a regular camera were used for the lateral image of the subject's head, while another camera was used for the frontal image. The speech signal was recorded simultaneously with other measurements. The subject was a native speaker of standard Korean (a male adult from Seoul).

Jaw opening The measurement of the degree of jaw opening was accomplished with the implement of a pair of special angle indicators. Since the position of the mandible cannot be externally observed, photographic measurements without the aid of special instruments are hardly satisfactory for the accurate observation of degree of jaw opening. To cope with this situation, a pair of artificial gums were constructed to fit each individual subject, and a needle which projected from the mouth was attached to each of them. These artificial gums with needles allow complete closure of the jaw and cause little interference with normal articulation of vowels. The needles serve as the indicators of mandibular opening. The aperture of the jaw during articulation of an isolated and sustained vowel is measured from a lateral view photograph of the subject's head, and can be expressed in terms of the opening angle, since mandibular displacement is comparatively small in this case and thus may be considered as rotation around a fixed axis. The position of the axis can be determined from X-ray photographs taken with different degrees of jaw opening.

Figure 9 indicates the estimated position of the axis of jaw rotation on an X-ray picture, while Figure 10 shows the mean and the range of variation of 7 measurements of jaw opening angle for each of the eight Korean vowels produced in isolation.

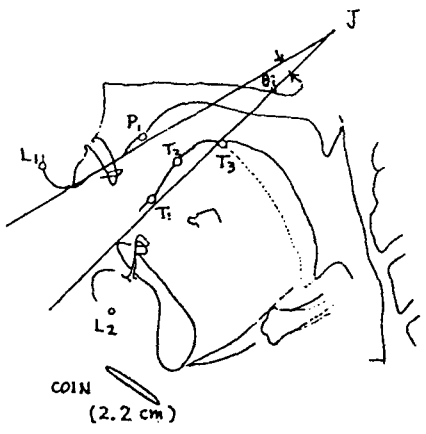

Fig. 9. Positions of lead pellets and the axis of jaw rotation.

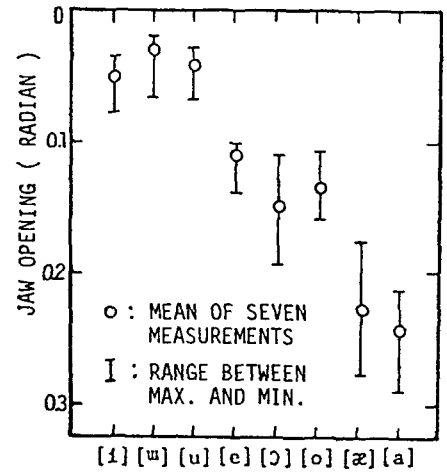

Fig. 10. Jaw opening angles of the eight Korean vowels.

These results suggest that Korean vowels can be categorized into three groups from the point of view of the degree of jaw opening. Namely, [i], [ɯ] and [u] can be considered as one group with smaller jaw openings, [e], [ɔ] and [o] as another group with medium jaw openings, [æ] and [a] as the third group with larger jaw openings. These findings were confirmed by the analysis of variance which indicated that the differences between vowels within each group failed to reach significance at the 5 % level. These results allow us to infer that the control of jaw opening for isolated vowels of Korean is basically discrete and can assume only three values, while some intra-group differences may be the secondary effects of other articulatory controls. For instance, the differences between jaw openings for [i] and [ɯ] as well as for [e]

and [o], though only significant at the 5 % level, may be the consequences of differences in tongue retraction and/or lip rounding.

Tongue position and shape Conventional methods for describing tongue articulation rely on the position of the highest point on the mid-sagittal tongue surface, but are not sufficient to describe to shape and position of the tongue. In this study, the mid-sagittal contour of the tongue relative to the mandible was traced from the X-ray picture, and five points, indicated by the position of lead pellets on the tongue, were used to calculate the closest approximation by a partial ellipse, of which the position of the center, direction and length of the major axis can be used as quantitative measures for the picture, direction and degree of tongue deformation.

The results are shown in Fig. 11, which indicate that the vowels fall broadly into two categories in regards to tongue articulation: 'front' ([i], [e], and [æ] and 'back' ([ɯ], [u], [ɔ], [o], and [a]), though the vowels [ɯ] and [u] may be regarded as a third group of 'mid' vowels.

Oral aperture The area of the oral aperture can be used as a quantitative measurement for the degree of lip constriction/rounding. Figure 12 shows the plot of the aperture area against the jaw opening angle for the eight Korean vowels. Although the figure shows that the aperture area varies almost linearly with the jaw opening angle, it also indicates that the eight vowels can be divided into two groups, viz., those with lip constriction ([u] and [o]) and those without lip constriction ([i], [e], [æ], [ɯ], [ɔ], and [a]).

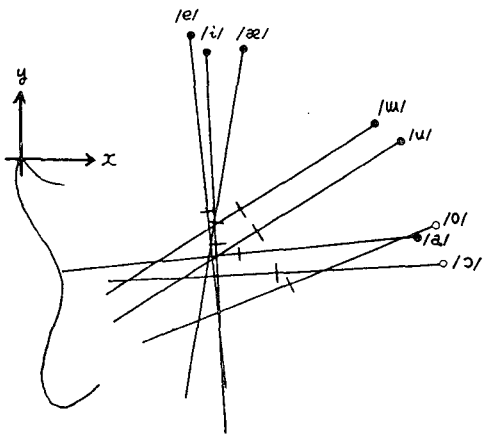

Fig. 11. Centers and major axes of ellipses approximating tongue contours of the eight Korean vowels against mandible-based co-ordinates.

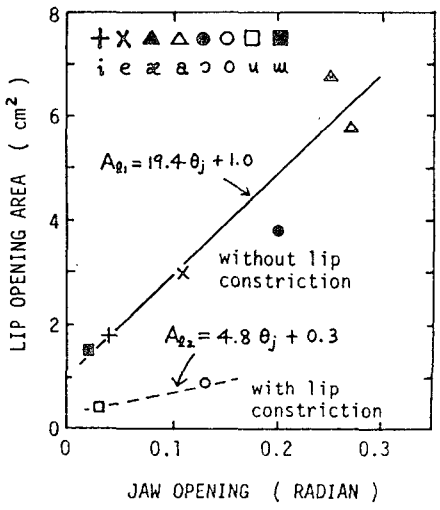

Fig. 12. Lip opening area versus jaw opening angle for the eight Korean vowels.

3.3 An articulatory description of the Korean vowel system

On the basis of these measurements and analysis of their results, a system for the specification of vowel articulation can be proposed in terms of three parameters: jaw opening, tongue retraction and lip rounding, and the articulation of the Korean vowels can be represented

by a prism-shaped structure shown in Fig. 13.

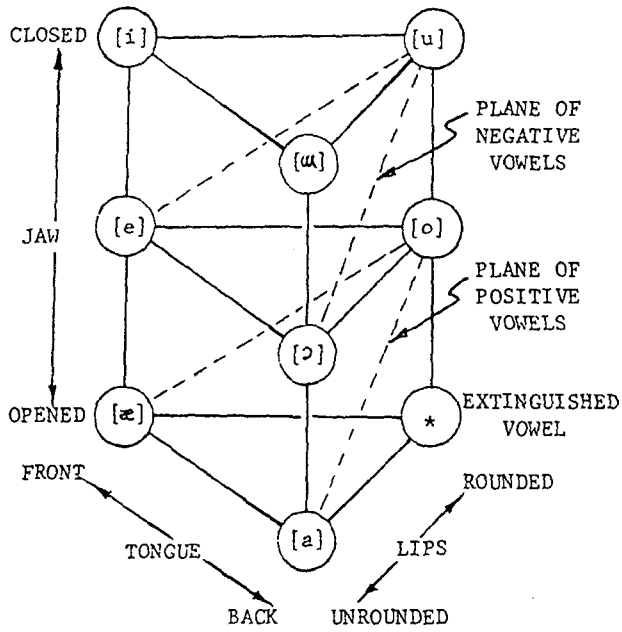

Fig. 13. Articulatory representation of the Korean vowel system.

Among the nine points shown on this prism, only eight have their corresponding single vowels in contemporary Korean. The missing vowel, indicated by *, seems actually to have existed at the time of constitution of Hangeul, and had a phonetic value somewhat between [a] and [o], but is now extinct. Its extinction may well be explained from the articulatory point of view; namely, rounding the lips under condition of wide jaw opening and tongue retraction is so difficult that it was gradually assimilated to its counterpart [a] without lip rounding.

As is well known, the Korean language is characterized by the phenomenon of vowel harmony which is a form of statistical constraints on the class of vowels that can occur within a word or can form a diphthong. The phenomenon is most commonly observed in onomatopoeic words but also in some other word categories. In traditional Korean phonetics, these constraints are named as the "positive" vowel group including [æ], [a], [o] and the "negative" vowel group including [e], [ɔ], [u]. In the vowel prism of Fig. 13, these groups constitute two triangular planes, with the jaw opening always one step larger for the vowels in the positive group than for their counterparts in the negative group. Furthermore, it is observed that a diphthong is formed as the transition from a rounded back vowel to an unrounded vowel with jaw opening that is one step larger. The increase of jaw opening in this case can be considered as necessary for the rapid release of lip rounding.

The acoustic consequences of these articulatory controls are illustrated by Fig. 14, which shows the first and the second formant frequencies of the eight vowels of the Korean speaker in an F_1–F_2 diagram. The formant frequencies were extracted by the method of 'Analysis-by-Synthesis' in the frequency domain, and were determined with an accuracy of ±5 Hz.

Comparison with the articulatory description in Fig. 13 indicates that the primary effect of opening the jaw is an increase the first formant frequency (F_1), while that of retracting the tongue is a decrease in the second formant frequency (F_2). On the other hand, rounding the lips decreases both F_1 and F_2 in vowels with medium jaw openings, but is seen to affect mostly F_2 in vowels with smaller jaw openings. The effects of these articulatory controls can best be explained in terms of coefficients of the cosine series expansion of the vocal tract area function.

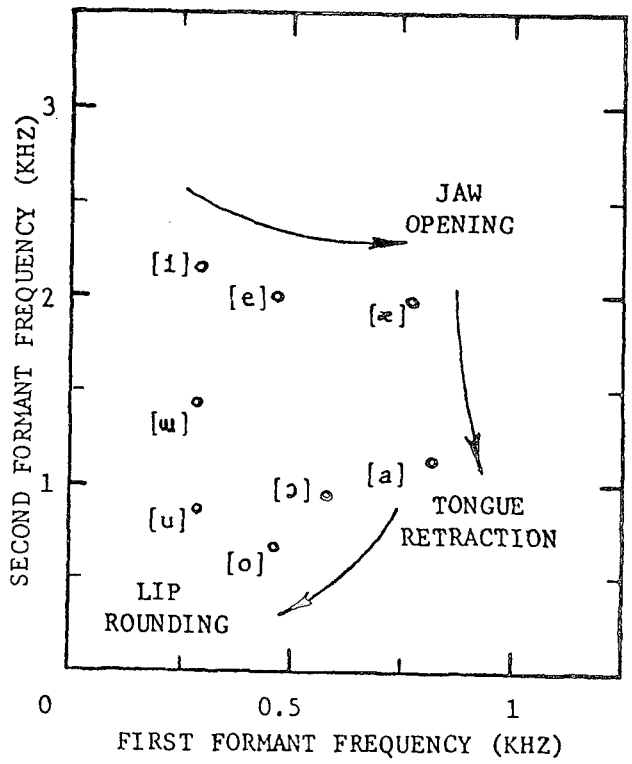

Fig. 14. F_1–F_2 diagram of Korean vowels.

4. Physiology and Physics of Tone, Accent, and Intonation [9]

4.1 A model for the process of generating F_0 contours [10~16]

It is widely accepted that F_0 contours of many languages are characterized by relatively slow undulations (henceforth global components) which roughly correspond to larger phrases, clauses, and sentences, and by relatively fast rise/fall patterns (henceforth local components) which correspond to either lexical tones of syllables or lexical accent of words. Previous works by Fujisaki and his coworkers as well as by others indicate that these two kinds of components

can be considered to be additive if one adopts a logarithmic scale (or equivalently, the semitone scale) for F_0 as a function of time. Figure 15 illustrates F_0 contours (i.e., $\log F_0(t)$) for one sentence each of the common Japanese, English, German, Chinese and Swedish.

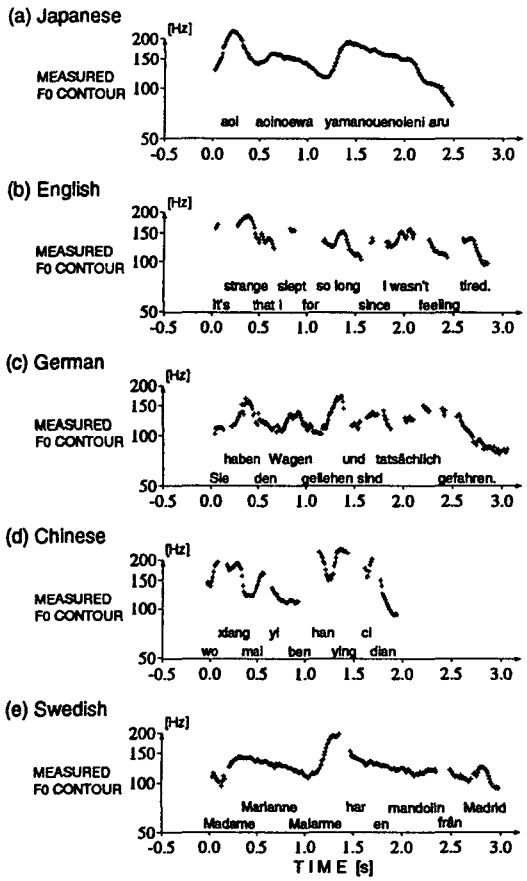

Fig. 15. Measured contours of voice fundamental frequency (F_0 contours) in the logarithmic scale for one utterance each of five languages: (a) Japanese, (b) English, (c) German, (d) Chinese, and (e) Swedish.

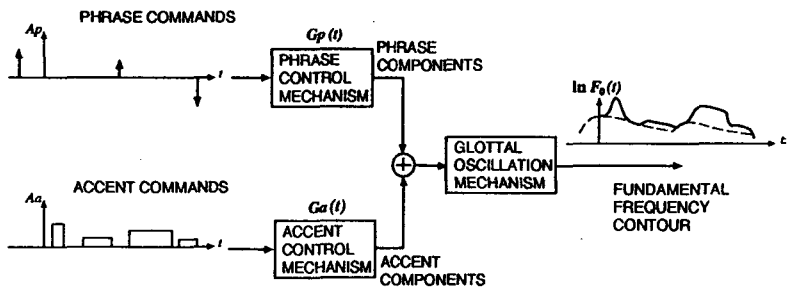

Fig. 16. A quantitative model for the process of F_0 contour generation of Japanese utterances.

A closer examination of these curves suggests that the global component has a shape similar to the impulse response of a second-order linear system (i.e., a mass-viscosity-stiffness system), and the local component has a shape similar to the step response of another second-order linear system with a shorter time constant than the former one. Thus the process can be represented functionally by the model shown in Fig. 16.

By the method of Analysis-by-Synthesis, it is possible to determine the number, magnitude and timing of the phrase and accent commands that will generate an F_0 contour which is the closest approximation to that of an actual utterance, say in the sense of the least mean squared error in the logarithmic scale of F_0. In fact, it has been shown that the above model applies quite well to F_0 contours of utterances of common Japanese, as well as to those of English and German, by assuming only stepwise input commands of positive polarity for the accent components. Figures 17(a) to (c) demonstrate the goodness of fit of the theoretical F_0 contour (solid line) to the measured F_0 contour (+ symbols) of one utterance each of these languages.

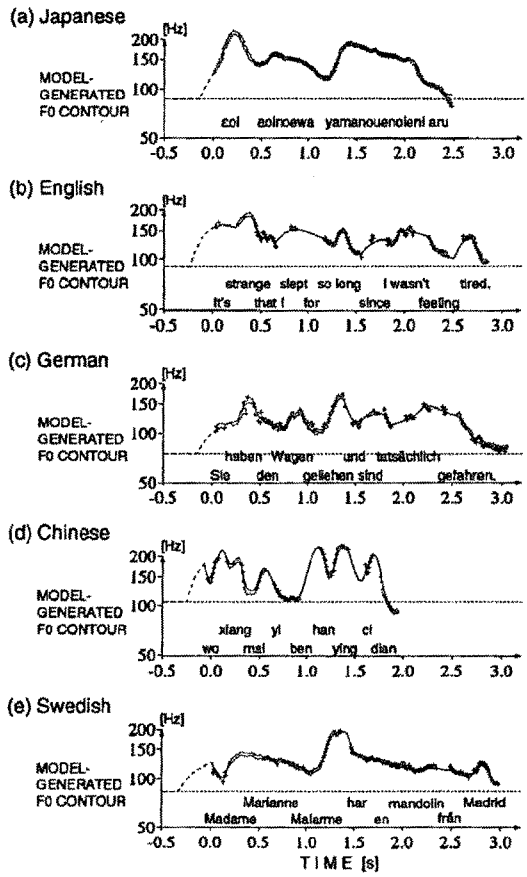

Fig. 17. Comparison of measured F_0 contours (same as in Fig. 15) and their best approximations generated by the model shown in Fig. 16.

On the other hand, the four tones in standard Chinese is conventionally classified as High (Tone 1, T_1), Rising (Tone 2, T_2), Low (Tone 3, T_3), and Falling (Tone 4, T_4). A close examination of the F_0 contours of these tone types indicates that the local component corresponding to T_1 is always positive, while those corresponding to the three other tones are partially (T_2 and T_4) or entirely (T_3) negative. This suggests that, instead of word accent commands of only positive polarity used for F_0 contours of the above-mentioned languages, we should assume tone commands that may have both positive and negative polarities. In fact, the best approximation can be obtained by assuming a positive command for T_1, a negative one followed by a positive one for T_2, a negative one for T_3, and a positive one followed by a negative one for T_4. Figure 17(d) demonstrates the goodness of fit of the theoretical F_0 contour to the measured contour of an utterance of Chinese, obtained under this assumption. Similarly, commands of positive and negative polarities are found to be necessary to generate the characteristics of grave and acute word accents in Swedish, as shown by the example of an utterance of Swedish given in Fig. 17(e).

4.2 Mechanisms for the generation of F_0 contours

The above-mentioned characteristics of the F_0 contour originate from the following properties of the larynx.

(1) A small change in the length of the vibrating part of the vocal cord produces a proportionate change in $\log F_0(t)$.

(2) The small change in the vocal cord length is caused by the addition of two components, each being the consequence of movement of a mass-viscosity-stiffness system.

Property (1) comes from the non-linear stress-strain relationship of skeletal muscles. In a skeletal muscle such as the vocalis muscle, the measured relationship between the tension and its incremental stiffness can be approximated quite well by

$$dT/dx = b(T + a), \tag{1}$$

where T indicates the tension and x indicates the amount of elongation. This leads to the stress-strain relationship

$$T = a(\mathrm{e}^{bx} - 1). \tag{2}$$

For $bx \gg 1$, this can be approximated by

$$T = a\mathrm{e}^{bx}. \tag{3}$$

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

$$F_0 = c_0\sqrt{T/\sigma}, \tag{4}$$

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

$$\log F_0 = bx/2 + c, \tag{5}$$

where, strictly speaking, c also varies slightly with x, but the overall dependency of $\log F_0$ on x is primarily determined by the first term on the right hand side. This linear relationship was confirmed by an experiment in which a stereoendoscope was used to measure the length of the vibrating part of the vocal cord.

Property (2) comes from the mechanical structure of the larynx. Anatomical and radiographic observations indicate that the relative movement of the thyroid cartilage against the cricoid cartilage has two degrees of freedom: translation and rotation around the cricothyroid joint as shown in Figs. 18(a) and (b). These movements can be represented by two separate second-order systems, and both cause small changes in vocal cord length. An incremental change $x_1(t)$ due to an instantaneous activity of *pars obliqua* of the cricothyroid muscle (henceforth CT), contributing to thyroid translation, takes the form of an impulse response function, while an incremental change $x_2(t)$ due to a sudden increase or decrease in the activity of *pars recta* of CT or other muscles, contributing to thyroid rotation, takes the form of a step response function. The resultant change in vocal cord length is obviously the sum of these two changes, as long as the two movements are small and can be considered independent from each other. Since each one of them has a proportionate counterpart in $\log F_0(t)$, this explains the superposition of the two components in the domain of $\log F_0(t)$.

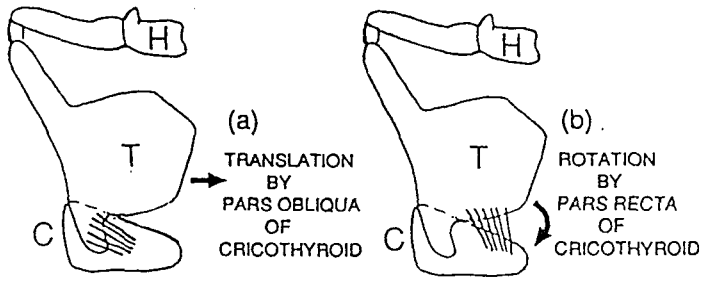

Fig. 18. The roles of *pars recta* and *pars obliqua* of the cricothyroid imuscles in laryngeal control. C: cricoid cartilage, T: thyroid cartilage.

Although the above-mentioned mechanism seem to be basically common to a number of languages, they may differ in the patterns of the commands. In some languages such as Chinese and Thai, the thyrotyoid muscle also seems to be involved. Table 1 summarizes the chain of events involved in going from word accent and sentence intonation to the corresponding F_0 contour in the case of the common Japanese.

Table 1. From linguistic information to acoustic manifestations.

Observed Phenomena	Related Academic Disciplines
Word accent and sentence intonation	linguistics and phonetics
↓ (neuromotor commands)	
Cricothyroid muscle activity	physiology of larynx
↓ (contractile force)	
Thyroid cartilage movements	physics
↓ (dynamics of rigid bodies)	
Changes in vocal cord length	mathematics
↓ (geometry)	
Changes in vocal code tension	physiology / physics
↓ (elasticity of muscles)	
Changes in fundamental frequency	physics
(vibration of elastic membrane)	

5. New Frontiers of Research [17]

5.1 Beyond linguistic information

It is true that speech is a means to convey linguistic information, i.e., lexical, syntactic, semantic, and pragmatic contents of a message, it also conveys other kinds of information. In my terminology, these can be classified into two broad categories: para-linguistic and non-linguistic. The information concerning the speaker's intentions (e.g., exhortation, question, suspicion, etc.), attitude (e.g., politeness, friendliness, etc.), and styles (e.g., fast/slow, formal/informal, etc.), which are usually under the conscious control of the speaker, can be considered to fall into the former category (para-linguistic), while the information concerning the speaker's physical states (e.g., age, gender, health, idiosyncracies, etc.) and emotional states (e.g., joy, sorrow, anger, fear, etc.), which are usually not under the conscious control of the speaker, can be considered to fall into the latter category (non-linguistic), though conscious simulation is possible, as is done by actors. Establishing a framework for the representation of these kinds of information as well as finding their articulatory, acoustic, and perceptual correlates presents new, unexplored areas of investigation which are important not only for their own sake, but also for certain practical applications such as synthesis/recognition of emotional speech as well as interpreting telephony capable of conveying subtle nuances.

5.2 Learning from human processes

Turning to the spoken language technology, it has been my belief over the decades that much can be gained by learning from nature, namely by investigating into the human processes and utilizing the principles and mechanisms that can be incorporated into technology. Nobody will deny the fact that our current technology of speech synthesis and speech recognition are far from being comparable to the human abilities of speech production and perception. In my opinion, this is mainly due to the fact that these techniques have been developed without paying due attention to the human processes.

For example, in text-analysis, which is a prerequisite to high-quality text-to-speech synthesis, little attention has been paid to the human processes of lexical access and parsing which is far more accurate and efficient than the current methods adopted in the automatic processing. In speech understanding, the widely accepted approach is to try first to recognize the smallest units — phonemes or sub-phonemic segments —, then to identify word candidates, and finally to infer the meaning of the whole sentence. Thus it is taken for granted that speech recognition is a prerequisite to speech understanding. On the contrary, a human being cannot read correctly what he/she does not understand, indicating that understanding is the prerequisite to recognition.

Still another example is the lack of the ability to acquire knowledge in almost all the systems for speech synthesis and speech recognition/understanding. These systems are invariably trained on a limited amount of speech samples or operated by rules derived therefrom, and their performances are fixed once they are in operation. On the contrary, humans acquire the skills of speaking and listening only gradually by being exposed to a large amount of data over a long period. I believe that we need to investigate the human process of spoken language acquisition and to incorporate it into artificial systems if we are to expect them to eventually possess capabilities comparable to those of a mature human being as a native speaker/listener of a language. These are only a few examples where we can profit a great deal by learning from nature.

6. From Mind to Mind — The Ultimate Goal of Speech Science and Spoken Language Technology

I have already mentioned that speech is a means to convey information, but one may naturally ask: Where does the information come from, and where does it go to ? Ultimately it comes from the mind of the speaker and goes into the mind of the listener. Thus spoken language is merely a medium of communication between two minds. In this sense, language models that are being widely used in conventional speech recognition systems only serve as crude approximations to the speaker's mind as the source of information. A more accurate way of modeling the speaker's mind should involve models of the speaker's self, of the listener, of the rest of the world as seen by the speaker, as well as of the processes of selecting a relevant piece of information, of constructing the message, and of executing the utterance. At the same time, the ultimate speech recognition system should also have a model of a listener's mind which involves models of the listener's self, of the speaker, and the rest of the world as seen by the listener, as well as of the processes of constructing the expected message from all these sources of knowledge. Likewise, the ultimate system for speech synthesis from concept should also have models of a speaker's mind and the listener's mind. I do not deny that phonetics and speech communication research have their own goals and their own ***raisons d'être***, but to me they are more exciting as means for elucidating the entire process of mind-to-mind communication.

References

1. Fujisaki, H. "Language and communication." In *Language*, University of Tokyo Press, Tokyo, pp. 1–52, 1983.
2. Fujisaki, H. "Transmission of meaning by language." *Descriptive and Applied Linguistics*, International Christian University, vol. 13, pp. 1–17, 1980.
3. Fujisaki, H. and Katagiri, Y. "Formulation and quantitative evaluation of processes of information transmission by means of words." *Transactions of the Institute of Electronics and Communication Engineers of Japan*, vol. J64-D, pp. 395–402, 1981.
4. Brown, R. W. and Lenneberg, E. H. "A study in language and cognition." *J. Abnormal and Social Psychology*, vol. 49, pp. 454–452, 1954.
5. Lenneberg, E. H. *Biological Foundations of Language*, John Wiley & Sons, New York, 1967.
6. Zadeh, L. A. "Fuzzy sets." *Information and Control*, vol. 8, pp. 338–353, 1965.
7. Kim, B.-I. and Fujisaki, H. "Measurement of mandibular control in vowels and its relevance to the articulatory description of the vowel systems of Korean and Japanese." *Proceedings of the Speech Communication Seminar – 1974, Stockholm*, vol. 2, pp. 277–284, 1974.
8. Kim, B.-I., Fujisaki, H. and Sawashima, M. "Observation of jaw, tongue and lip control in articulation of Korean vowels." *Transactions of the Committee on Speech Research, Acoustical Society of Japan*, vol. S74, no. 56, pp. 1–8, 1975.
9. Fujisaki, H. "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour." In *Vocal Physiology: Voice Production, Mechanisms and Functions* (O. Fujimura, ed.), Raven Press, New York, pp. 347–355, 1988.
10. Fujisaki, H. and Nagashima, S. "A model for the synthesis of pitch contours of connected speech." *Annual Report of Engineering Research Institute, University of Tokyo*, vol. 28, pp. 53–60, 1969.
11. Fujisaki, H. and Sudo, H. "A model for the generation of fundamental frequency contours of Japanese word accent." *Journal of the Acoustical Society of Japan*, vol. 27, pp. 445–453, 1971.
12. Fujisaki, H. and Hirose, K. "Analysis of voice fundamental frequency contours for declarative sentences of Japanese." *Journal of the Acoustical Society of Japan (E)*, vol. 5, pp. 233-242, 1984.
13. Fujisaki, H., Hirose, K., Hallé, P., and Lei, H. "Analysis and modeling of tonal features in polysyllabic words and sentences of the Standard Chinese." *Proceedings of the 1990 International Conference on Spoken Language Processing*, Kobe, vol. 2, pp. 211–214, 1990.

14. Fujisaki, H., Ljungqvist, M., and Murata H. "Analysis and modeling of word accent and sentence intonation in Swedish." *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 211–214, 1993.
15. Fujisaki, H., Ohno, S., Nakamura, K., Guirao, M. and Gurlekian, J. "Analysis of accent and intonation in Spanish based on a quantitative model." *Proceedings of the 1994 International Conference on Spoken Language Processing*, Yokohama, vol. 1, pp. 355–358, 1994.
16. Mixdorff, H. and Fujisaki, H. "Analysis of voice fundamental frequency contours of German utterances using a quantitative model." *Proceedings of the 1994 International Conference on Spoken Language Processing*, Yokohama, vol. 4, pp. 2231–2234, 1994.
17. Fujisaki, H. "Future of speech science and spoken language technology." In *European Studies in Phonetics and Speech Communication* (Bloothooft, G., Hazan, V., Huber, D. and Llisterri, J., eds.), OTS Pablications, Utrecht, pp. 15–18, 1995.