TIMBRE STUDY OF VOCALIC VOICES

YOSHIYUKI OCHIAI and TERUO FUKUMURA

Electrotechnical Department

(Received October 31, 1953)

This is some study of attempt concerning what is called timbre quality. Vocalic voice is brought here to the front as its home question. This study does not mean therefore one-sided analysis of vowel sound only physically viewed. This is not the mere faithful description of the harmonic components contained in sustained vowel. In order to understand in full the vocalic timbre, that is the most important in the interpretation of transmission quality of speech-transmission systems, phonal structure and vocal pattern are obtained and discussed here. As what is available for the connection of these two qualities, invariant formant comes to the surface. For representation of vocalic timbre, the terms, such as vocal alps, vocal summit and vocal glen are used. For the same purpose, white phone and white voice are also provided for. Whispered vowels and devocalized voices are also investigated. It will be worth noting that this timbre study of vowels, starting from the conceptional analysis of transmission quality, was carried out conforming to the experimental plan drawn up three years since. We have led some conclusion and have discussed thereupon, basing on the material of 1287 oscillograms obtained for analysis.

CONTENTS

Introduction

Of all the sounds well known, natural or artificial, nothing can equal the human voices in respect of their meaningfulness in our actual lives. The voices seem to stand high on the list of sounds that are important for us. They come home to us so truly. The problem of voices is, therefore, not only the mere problem of acoustics but the very problem that keeps in touch with human being itself. In the present stage of study, difficulties in their treatment seem to be derived, for the most part, from the side of human nature, not from their physical side. Our studies of voice and vowel, coming from the viewpoint of a quality study in telecommunication, can date back, so to speak, to the subjective problem of humanity, and will induce the most active and inquiring intellect to further researches.

Conceptional Study

Timbre conception

The idea of timbre in physical acoustics is certainly evident. According to the conception given by Helmholtz, the timbre is one attribute of sound which does not belong to other two elements, that is, pitch and intensity. Sound analysis in various senses can be derived from this viewpoint. Up to now analytic study in voice and vowel was also carried out in the sense of Helmholtz's definition and Ohm's law. Analysis and synthesis of vowel were studied actively in every country and about every language, and seem to continue to be studied more. Analysing apparatus itself went through so much development and refinement, that the results acquired have become sensibly precise and fine. The number of formant therein, for example, is, and will be, on the increase. The so-called bars in sonagram by visible-speech method mount up now to four or five. Are we, hereupon, content with the interpretation of the formant thus obtained as their fine structure? Are we enough in the knowledge of bars or the understanding of bars, that will increase in number according to the dexterity of analysis? For what reason must so many formants crowd in our vowels? What does the formant stand for in timbre quality? Timbre study in general, seems to us, must make a start again beginning with the sharp criticism on the conception itself as to the timbre. So far as the quality of transmission in telecommunication is concerned, the mere physical aspect of timbre will be of little use: the timbre problem in speech transmission considered from the viewpoint of quality aspect, cannot get quit of without touching upon the subjective problem of human being: the quality conception in communication will not fail to connect reasonably with the subjective side of correspondents who want to communicate: therefore the quality considered purely in the sense of physical objectivity, will cease to pretend to its own positiveness as a quality phenomena in communication: for instance, as for the data given by the analysis that is carried out from the pure physical standpoint, how can we know

the meaning of a group of formants of vowels as in communication signal? We must, therefore, call in question the timbre problem in communication from the subjective side of hearing and judgement. Timbre structure should be treated as a subjective phenomena, seasoned with the nature of hearing. After all, the timbre study must be executed as being proposed from the side of human subjectiveness.

Static structure of pitched timbre

In quality theory, we have to lay so much stress upon the pitch element as next to timbre. We must, therefore, take a standpoint of pitch in a classification of timbre: pitched timbre and pitchless (or unpitched) timbre. By pitched timbre we mean the timbre with definite pitch: voiced vowel is an example of pitched timbre, and whispered vowel is an embodied illustration of pitchless timbre.

It will be needless to restrict the timbre only to the stationary timbre. However, it is convenient to take first the stationary timbre for studying the timbre structure. Sustained vowels are thus employed for the purpose of structural study of static timbre.

Basic viewpoint in quality aspect

To proceed with our quality study as one of the most basic theory of communication, we must call in question the speech signal, as is natural, in the subjective side of human sensation. Under the name of transmission system, we understand merely the effect of transmission-distortion upon the speech signal which must lie influenced by the operation of sensation in order to bear the adequate impression as a communication signal. Denote by $\mathbf{D}$ the physical operation of distortion by transmission system, and by H the hearing or judging operation. Denote also by X the physical signal, and by Ξ the subjective signal. Then the conception of subjective fidelity can be represented by the notation

$$H\cdot\mathbf{D}(X):H(X) \quad \text{or} \quad H\cdot\mathbf{D}(X)/H(X),$$

where the sign : or / means the procedure of some comparison of subjective nature.

We can denote consequently the physical fidelity by the representation

$$\mathbf{D}(X):X \quad \text{or} \quad \mathbf{D}(X)/X.$$

Now consider the following expression

$$\frac{H\cdot\mathbf{D}(X)}{H(X)} : \frac{\mathbf{D}(X)}{X},$$

which means evidently the comparison of two sorts of conceptions on fidelity, and from which we can lead the form

$$\frac{H\cdot\mathbf{D}(X)}{\mathbf{D}(X)} : \frac{H(X)}{X},$$

wherein the former term implies the corresponding relation between the physical distortion and subjective one; and the latter term implies also the correspondence between the physical speech and subjective one.

We can define the conception of *transmission quality* by the expression

$$H\cdot\mathbf{D}(X),$$

and also the conception of *speech quality* by

$$H(X).$$

If we succeed in quantification of quality conception, we will be able to express by the same notation as above even the quantified quality itself, we could define concretely the degree of subjective quality of transmission by the ratio form of transmission quality and speech quality; thus we have

$$\Gamma = \frac{H \cdot \mathrm{D}(X)}{H(X)},$$

and in the same manner, as the degree of physical quality of transmission, we can give

$$G = \frac{\mathrm{D}(X)}{X}.$$

Two qualities primary and essential

If the subjective signal

$$H(X) = \Xi$$

receives thereby exclusively the operation of timbre judgment, we can employ the expression

$$H_t(X) = \Xi(\tau)$$

where the suffix t of the operation H signifies that the attitude of judgment depends mainly on the timbre element. There exist two kinds of timbre judgments on speech; the first is the articulation judgment, the second the naturalness one. According to the same manner of expression, we have

$$A = H_{t_1} \cdot \mathrm{D}(X) : H_{t_1}(X)$$
$$N = H_{t_2} \cdot \mathrm{D}(X) : H_{t_2}(X)$$

wherin A denotes specific articulation and N specific naturalness.

When the speech quality is subject to the operation of timbre judgment as to the clearness aspect, we can give the expression

$$H_{t_1}(X) = \Xi(\tau_1)$$

and when the speech experiences the operation of timbre judgment as to the naturalness aspect, we can have the form

$$H_{t_2}(X) = \Xi(\tau_2)$$

where $\Xi(\tau_1)$ and $\Xi(\tau_2)$ denote the timbre elements in subjective expression which are concerned respectively with clearness property and natural property of speech timbre. As it will be beyond our power to obtain at once the concrete structure of subjective timbre, the principal aim of this timbre study consists in obtaining the two forms of signals in its physical expression which essentially correspond respectively to $\Xi(\tau_1)$ and $\Xi(\tau_2)$.

Vocal and phonal qualities in vowel timbre

We have seen that the quality conception in communication differenciates into

two qualities which can be reduced respectively to the articulation- and naturalness-aspect. Though these two qualities naturally can reveal themselves as well in pitch as in volume phenomena in speech transmission, their effective appearance is also found in the timbre field. By confining ourselves to the timbre field, we will try to trace the two qualities which are properly of complicated character. Needless to say here that we can of course find two quality phenomena in the transient state of timbre. It will be easy, and therefore also will be without much interest to make a study of two qualities in the field of transient timbre. The most delicate and the most interesting problem lies, without fail, in the study of two qualities as to the stationary timbre, although such a static study of sustained vowel looks, at first sight, too primitive, too obsolete, and too characterless. Not as banal frequency analysis of vowel, but as static structural-study of timbre, with such a notice of timbre conception as mentioned above, we must proceed on our way.

Denote now the stationary vowel-signal as V, or $V(p, l, t)$, which is characterised as pitched timbre with pitch p, with level l, and also with timbre t. According to the attitude of hearing, we give the following relations

$$H_p(V) = \pi_V,$$
$$H_l(V) = \lambda_V,$$
$$H_t(V) = \tau_V,$$

where π_V, λ_V mean respectively subjective pitch and subjective loudness of vowel as harmonious complex sound, and τ_V also subjective timbre of vowel. So far as the timbre is directly concerned, excluding both pitch and level for the time being out of the direct mark, we can denote V in the sense of timbre mark, or instead of T_V by $V(p, l)$, which implies the timbre of vowel at the pitch p and on the level l.

It is now our problem to find out the articulation quality and naturalness quality in the timbre construction of T_V or $V(p, l)$. We can have the following forms

$$H_{t_1}(V) = \tau_1 = \tau_a,$$
$$H_{t_2}(V) = \tau_2 = \tau_n,$$

where τ_a signifies subjective timbre quality as to the articulation, and τ_n also subjective timbre quality as to the naturalness.

As it is very difficult to obtain directly the subjective timbre structure, we must go on to search the objective timbre structures T_1, T_2, which correspond respectively to the subjective ones τ_1, τ_2.

Now turn our attention a little to the retrospective facts that we are doing actually but unconsciously in the discrimination of vocalic voice. We can identify the same vowel in spite of the difference of pitch and level. Therefore we have to obtain the timbre quality as to the phoneme value as such that is independent of pitch and level. On the other hand, we can discriminate the personal voice notwithstanding the difference of vowel, which will be in some extent dependent on pitch and level, is consistent with the following facts. The voices uttered forcibly too high or too low come to lose their personal shades. For example, the whispered vowel as a critical timbre of one end, is usually very difficult to judge whose voice it is, and the so-called *falsetto* as a critical timbre of another end, is also considerably difficult to judge its individuality. Reflecting upon these facts, we can

think of the procedure to discover two timbre qualities. As to the phoneme value of vowel, we will be able to prepare the process for finding out the phonal quality by way of (1) *depitching* process, (2) pitchless uttering method as an approximation method of (1). As to the vocal quality, we can make use of (1′) throat pick-up method (2′) *devocalizing* method as an approximation method of (1′). Denote these ideas by symbol. On the assumption of good level matching on the given level (l), the phonal quality will be obtained by

$$\begin{aligned} T_1(V) &= \sum\sum V(p_1) + \sum\sum V(p_2) + \sum\sum V(p_3) + \dots \\ &= \sum_p \sum\sum V(p), \end{aligned} \tag{1}$$

where $\sum_p$ implies depitching process as juxtaposition of harmonic components, and $\sum$ the statistical smoothing process. The phonal quality may be represented by

$$T_1(V) \fallingdotseq \sum \overset{V}{\underset{0}{}}, \tag{2}$$

where $\overset{V}{\underset{0}{}}$ means the whispered pitchless state of vowel V. On the same assumption of strict level balancing, the vocal pattern will be obtained by

$$T_2(p) = \sum Z(p), \tag{1′}$$

where $Z(p)$ means white voice* picked up at the throat part of body vibration by voicing at the pitch p, that has nearly the same wave-form notwithstanding the kind of oral vowels, and as an approximation of (1′) by

$$\begin{aligned} T_2(p) &\fallingdotseq \sum\sum V_1(p) + \sum\sum V_2(p) + \sum\sum V_3(p) + \dots \\ &\fallingdotseq \sum_{\tilde{p}} \sum\sum V(p), \end{aligned} \tag{2′}$$

where $\sum_{\tilde{p}}$ implies the devocalizing process, that is, the superposition process of all oral vowels in the language concerned. In résumé, the phonal quality in timbre is given as frequency structure which is obtained by depitching process, and the vocal quality in timbre is given as pattern build up on the basis of pitch which is obtained by devocalizing process. These two qualities in timbre can take part in the discriminating function of hearing, revealing themselves in the difference of frequency structure or of harmonic configuration of timbre.

Measurement and Treatment

The voices are liable to vary. The vowels are apt to change. It is very difficult to keep constant the manner of utterance and pronunciation. In the strict sense, there will be no voices which are quite equal even for the same person and under the same condition. The objects of our measurement are subject to sharp fluctuation. On the other hand, the timbre problem, as we have shown already, has a close connection with a most subtle sensation. When we come to deal with the formant structure or vocal pattern as timbre construction, it will be expected to bring about some information which proves to be essential and important for the accomplishment of timbre perception. It should be called for that the so-called timbre construction, whatsoever it may be, must have some such fine and full details that they can throw light on the vital point of complicated problem

* "White voice" does not mean *voce bianca* in vocal music.

such as timbre discrimination. In amount, the most precise measurement is to be demanded for obtaining the timbre construction of voices which are doomed to exist far from the constancy. We stand in need of some fineness for our measurement, some certainty for our treatment.

Experimental procedure

Five voices of five subjects, of whom two are males, aged 27 and 24, and two are females, aged 20 and 17, and a boy aged 12, are analyzed by the magneto-striction-resonater type wave-analyzer of which the functioning-diagram and frequency-characteristics are shown in Fig. 1 and 2 respectively. The condenser microphone used at the distance of 25 cm from the lips of subjects has the frequency-response characteristics, which is shown in Fig. 3, and its membrane is 25 mm in diameter. The vibrations on the body-surface at the throat (just under the *prominentia laryngis*) are picked up piezo-electrically through a rod contact. If it is set in vibration

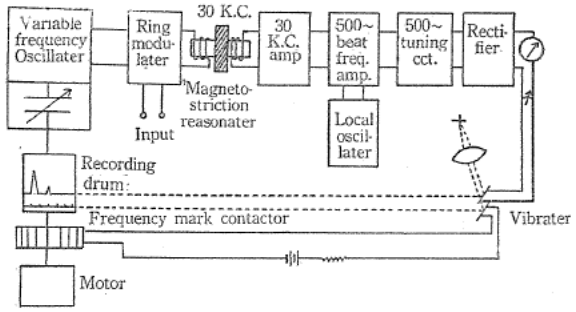


FIG. 1. Functioning-diagram of frequency analyzer.

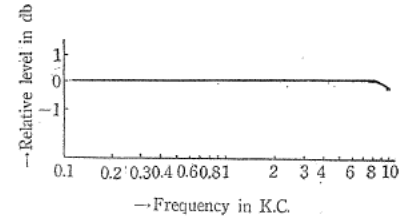


FIG. 2. Frequency-response characteristics of frequency analyzer.

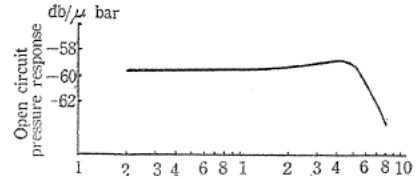


FIG. 3. Frequency-response characteristics of condenser microphone.

electrically, this pick-up has the flat response in the frequency range below 3,600 c.p.s. Under the working condition, where its rod is to be attached to the body-surface as in this experiment, its working response is unknown. But it is evident that its non-linear distortion comes to nothing. As the D type vibrater of electro-magnetic oscillograph, sensitivity of which is about 2×10^{-6} Amp/mm, is used, and the calibration curve, the deflection of vibrater in mm *vers.* voltage at the input of analyzer, is also shown in Fig. 4.

The experimental procedures are illustrated schematically in Fig. 5, where, by means of audio-oscillator and VU meter, the pitch and level of sustained vowel are monitored by the subjects themselves and supervised also by the experimenter. The subjects are instructed to utter the vowels naturally, as in conversation and not in

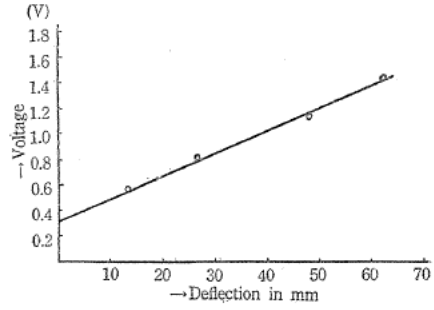


FIG. 4. Calibration curve of deflection of the vibrater *vers.* voltage at the input point of the analyzer.

singing and to keep them sustaining as steady as they can for about ten seconds until the analysis from 0 to 10,000 c.p.s. is accomplished. The level differences between vowels of each subject are to be compensated by the control of amplifier gain. The prescribed level for matching is −2 db re. 1 mw/600 ohms at the input of analyzer.

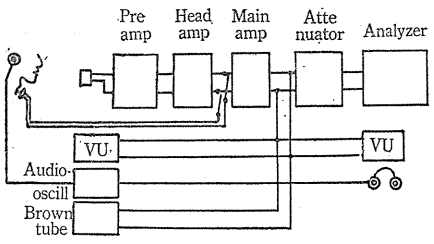


FIG. 5. Scheme of experimental procedure.

Number of oscillograms for spectre analysis obtained in this experiment is given in the table.

Sub.	Voiced vowel	Voiceless vowel	Throat vibration	Total number
T. F.	151	31	214	396
S. M.	127	24	35	186
H. H.	221	27	46	294
Y. H.	165	25	40	230
H. O.	134	22	25	181
Total	798	129	360	1287

Preliminary data on unstability of voices

The ordinary voiced vowels in sustained utterances without any vocal techniques such as *vibrato, tremolo* or pitch warble, being composed of an amount of components which usually stand in a beautiful harmonic relation, we are inclined to confine our study to deal only with the timbres in harmonic construction under the name of pitched timbre. We can find there the vital points of static study of timbre structure. And no inconsiderable part of timbre problem may be solved. As a consequence, the principal aim of this study consists naturally in a closer re-examination about the timbre problem in harmonic configuration as its internal structure, catched in such a static state as their components remain always stable as possible. It is an ideal condition to be desired. But the human voices are liable to vary. Even when we try to utter them in order to keep them as constant as possible, more or less deviations are inevitable. We stand in need of ten and more seconds in one analysis. With a view of keeping stable utterance, we prepare not only the pitch matching but also the level supervising. By these ways, the pitch deviation during one whole utterance almost gets extinguished, but the level deviation during one breath still remains to certain extent. The stability itself in timbre would not necessarily be guaranteed even under the ideal condition with constant pitch and level, because the timbre is another thing than pitch and volume. If there are still some level fluctuations, the timbre will receive more deviation

accelerated by them. Therefore, the first work to be done in static analysis is to meet such an inevitable deviation. Let us begin with collecting the data about the unstableness of voice.

Deviation in level

Level fluctuation during one breath of utterance depends considerably on pronouncing subject, and in the case of the same talker it also depends on pitch and to some extent on the kind of vowel. Fig. 6 shows the case of Sub. Y. H., where she utters five vowels with 5 or 6 pitches. This corresponds to the example of unexperienced subject. In Fig. 7 we can see the mean value of level fluctuation of 5 vowels in which 3 subjects uttered. In this figure we show only the cases of 2 pitches, that is the lowest and highest among all pitches employed. Sub. T. F. (♂) is an expert on voicing; Sub. H. H. (♀) is unexperienced but skilful in level control. There is relatively large deviation in lower pitches in the case of T. F., and we find a reversed case in the example of H. H. We cannot find any conspicuous trend as to the vowels, as far as our experience is concerned.

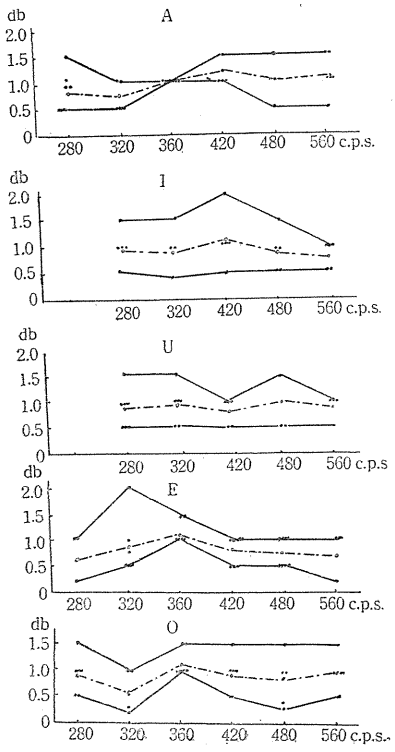


FIG. 6 (*left*). Level deviation of five vowels of Sub. Y. H. at several pitches.

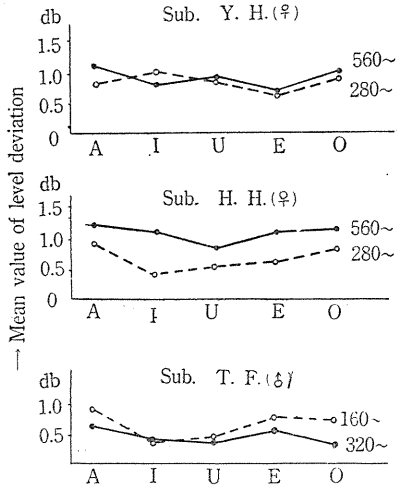


FIG. 7 (*right*). Mean deviation in level of five vowels of three subjects.

Fluctuation in timbre structure

Deviation of levels of the components which constitute the timbre construction is of the most importance. We investigate it about 5 subjects, a part of which is shown here. The most evident is the difference between the experienced and un-

experienced subjects. Fig. 8 shows the example of relatively slight fluctuation in timbre, where the amplitude of components is measured in linear scale. In the case of unexperienced talker, the pronounciation at certain pitches in certain vowels seems to be unfixed, resulting in the relatively large fluctuation. Fig. 9 shows it. It seems that the mode of vibration thereby receives some particular unstable change.

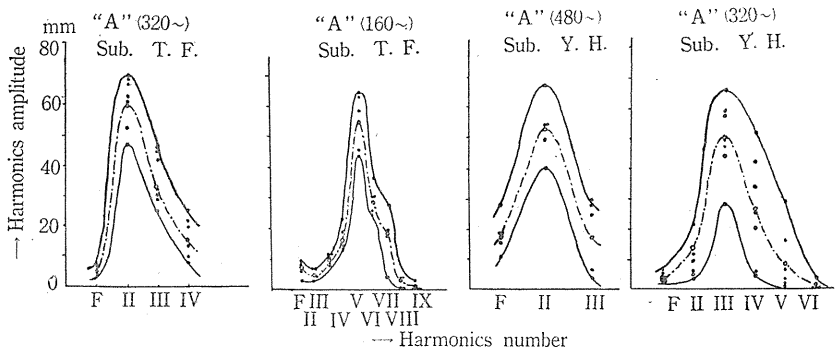


FIG. 8. Fluctuation in timbre structure.

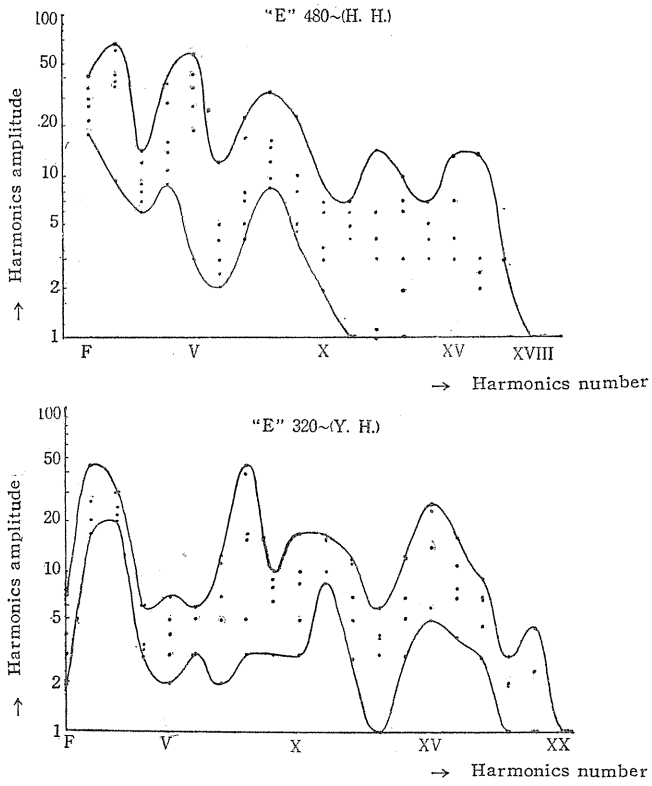


FIG. 9. Fluctuation in timbre structure in "E" vowel.

Choice of uttering subjects and vowels to be uttered

As uttering subjects, we have chosen two young men and two young women and one child (a boy still before change of voice). They have no defect in uttering and no dialect in pronounciation. They have undergone the otorhinolaryngologic diagnosis, sitting for Röntgengraph. They are naturally Japanese, all born at Tokai-district in the central part of Japan. Except two male subjects, they are almost unfamiliar with foreign languages. Also they all are not specialist in vocal music. The vowels employed are those which are usually found in ordinary conversation, but under the sustained condition, with a little more strain (nearly *mezzo forte* in strength) than in speech. Oral vowels are only used, excluding as well nasals as neutral vowels. With their utterance, it is requested not to use such a *vibrato* or *tremolo* as in vocal technique. Needless to say that these amateurs in vocal music have no finesse to use consciously the so-called *gedeckte Stimme*. Their vocalisations are so natural and plain as they do in ordinary conversation. Moreover they all being very young and normal, we cannot find any trace of *son rauque* in their voices.

Choice of pitches

In order to obtain precisely and completely the formant structure of vowels, it is very important to determine the necessary range of pitch and more in detail the kind of pitches actually used. In accordance with our present purpose that is to study plain vowels in colloquial expression, it will be proper to decide the pitch range nearly one octave, which is fit for the conversation of the subject of interest. To determine concretely the actual form of formant structure, the choice of the kind and number of pitches with which the vowels are to be uttered proved to be important. We know that it is thereby very needful to bear evidence not only the existence of formant but even the non-existence of it. It is because, the complicated combination of resonance and absorption determine the complete form of formant structure as a whole, and if we desire to call it in question minutely down to its fine structure, it will be necessary to obtain even the utmost delicate combination of feeble resonance and imperfect absorption. In some occasion, therefore, we are obliged to verify that there is none of such and such a formant in such and such a position. As we will see later, this proves to be important in deciding the real position of vocal glen or gorge in the vocal alps.

In choosing the pitches we take a constant difference method, that is,

$$p_n - p_{n-1} = p_{n-1} - p_{n-2} = \ldots p_3 - p_2 = p_2 - p_1$$

and

$$p_n/p_1 \fallingdotseq 2$$

For instance, for male subject, 320~, 280~, 240~, 200~, 160~ pitches are used.

The manner of choosing the pitches can bring influences upon the determination of formant structure. We show the actual example in Fig. 10, where are shown the 6-pitch process, and upper 3-pitch process, and lower 3-pitch process respectively. We know that there exist a slight difference with the positions of formant peak, and moreover an evident difference with the number of formant peak according to the choice of pitches.

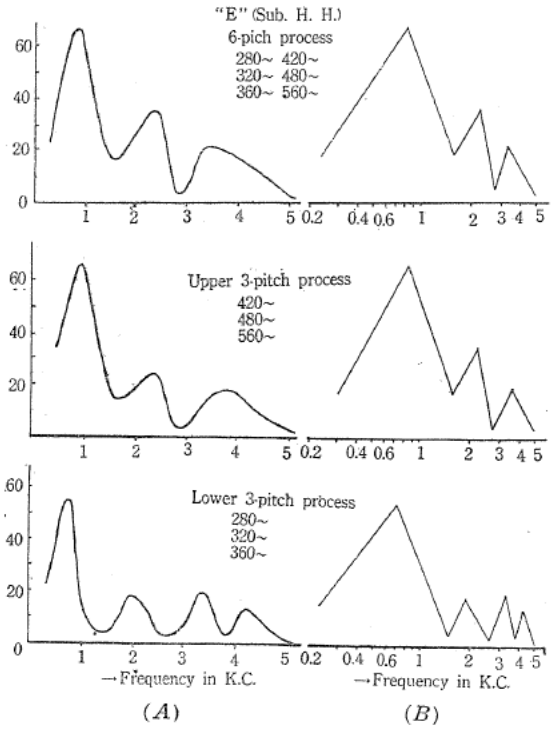


FIG. 10. Effect of pitch choice upon the determination of formant structure.

Mean structure of timbre

Even if we are successful in controlling of level or pitch of uttering subjects, we cannot assert the accomplishment of timbre control. Because, timbre is another thing. In sober truth, even if we can specify the kind of vowel that is to be uttered, the vocalic voice cannot thereby be specified without any ambiguity. Simply speaking, we can utter the same vowel with various voices, even at the same pitch and on the same level.

But all that we can do in timbre study in the present stage, is to minimize any sort of deviation caused by as well timbre fluctuation in itself as level fluctuation. Especially in our method where every pitched vowel is overlapped only once in every process, it is none the less useful to seize hold of any pitched vowel in such a state as steady as possible. It is for this reason, that we repeat the analysis of the same vowel under the same condition. If it be stable, we take ordinarily only 5 or 6 samples of photogram, if it be unstable, we are in need of from 7 to 10 samples of them, from which we obtain the mean value of every component. Thus we come to attain the mean structure of vocalic timbre. We think, the process of mean structure is indispensable in realizing the conception of static sturcture.

Juxtaposition method

Juxtaposition method is employed in the determination of formant structure, as the best method of obtaining the contour of formant structure, on the assump-

tion that the complete whole of every static structure of every pitched vowel comes to compose the formant structure that marks the vowel in question. The mechanism of our recognizance of vowel as such is thoroughly unknown. But if the formant itself comes into play in such recognizance of our perception, a multiple exposure process in photograph is likely to be useful in their interpretation. Instead of direct summation method, the juxtaposition method is thus adopted here. In this method, the matching and balancing in level among pitched vowels is essentially important. There are two methods; component enveloping method, and resonance-curve enveloping method, Fig. 11 (*A*) (*B*) (*C*) shows their actual examples.

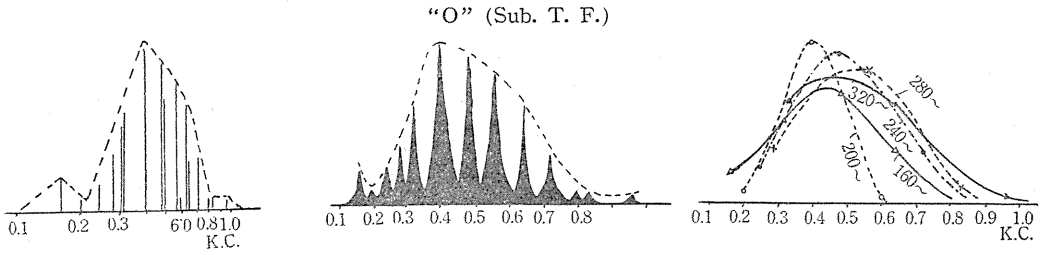
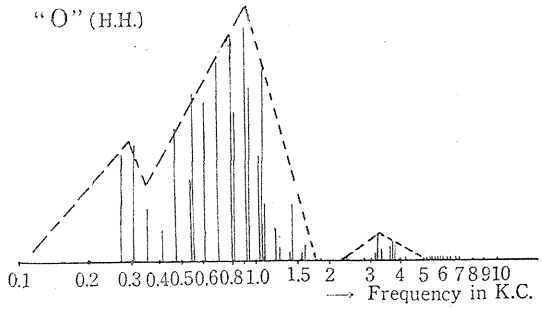


FIG. 11 (*A*). Formant structures obtained by component- and resonance-enveloping methods.



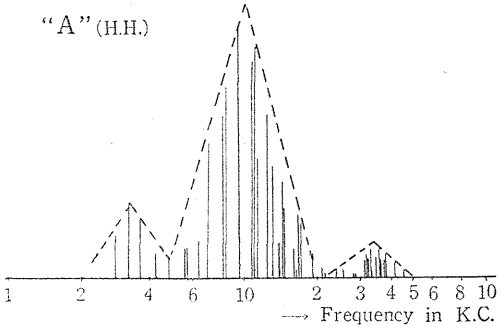


FIG. 11 (*B*). Component-enveloping method.

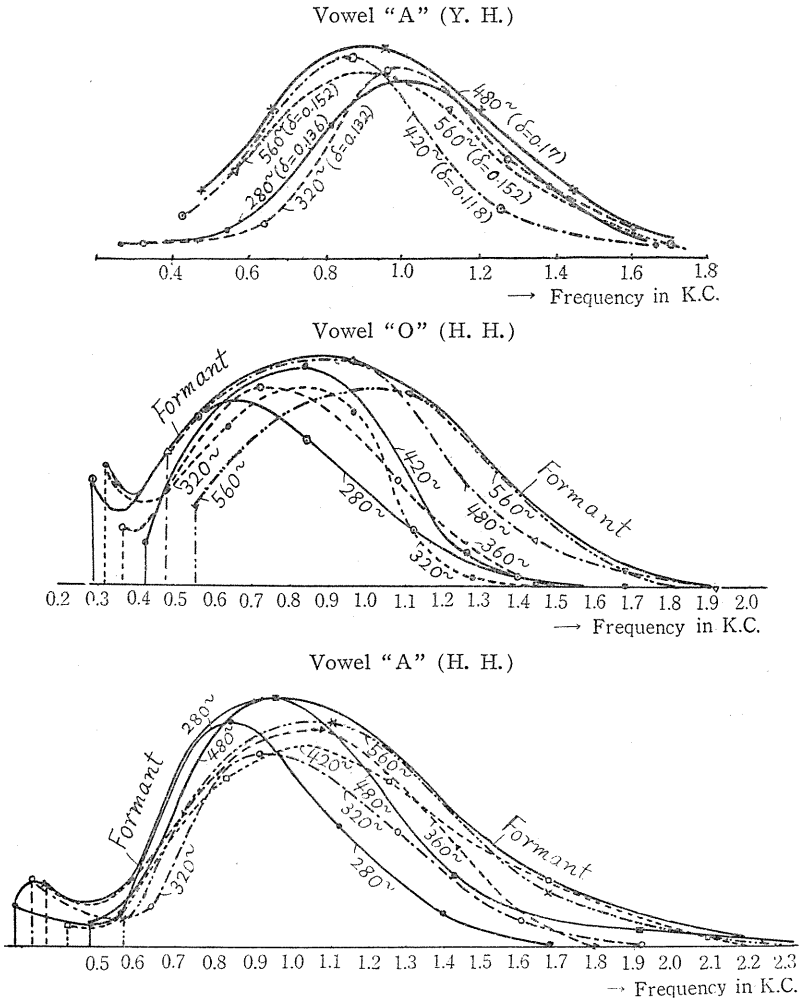


FIG. 11 (*C*). Resonance-curve enveloping method.

Superposition method

For obtaing the vocal pattern, we adopted the superposition method, where each component is superposed according to its harmonic number. In accordance with our purpose of searching the invariant structure of timbre notwithstanding the kind of vowels, the *devocalizing* process is adopted, where all pitched vowels are superposed, as they are uttered at the same pitch. In this case, the level matching among vowels is very important. Needless to say here that the devocalization by superposition of vowels is not always perfect. It will be more available to use the throat pick-up method by taking the so-called white voice.

Representation of Timbre

We have had enough vowel analysis. It is not the description about the vowel

components, thus analyzed, however in full it may be done, but the representation or expression of it as timbre, that we have here indeed need of. Suppose that we have chosen for instance the formant structure as a representation of vocalic timbre. Our present question is to see how the timbre can be represented by a structure of this kind. To know quite precisely the peak position of formants is naturally most important. But that alone will not be enough. We want to know the minute form of the contour of formant structure as a sort of timbre pattern. In order to call in question the subtle point of personal nuance of vocalic timbre, we must be capable of obtaining sensibly fine structure of it. We must not anyhow take a *partial* representation of formant, we must adopt a representation of timbre as a *complete whole.*

Formant structure of vowels

Formant structure of 5 vowels about 5 subjects are shown in Figs. 12, 13. As we are in need of formant structure as a sort of representation of timbre pattern, each formant structure of each vowel is shown in percentage scale of amplitude and also in logarithmic scale of frequency. We have endeavoured to express its fine structure as faithfully as possible. We have tried also to express about all subjects its white-vowel pattern (W) such as $W = T_{V_1} + T_{V_2} + T_{V_3} + T_{V_4} + T_{V_5} + \cdots = \sum_V T_V$, which is very convenient to interpret the vowel quality. We will come back again on this point. In these representations, high fidelity can be expected in the frequency range below about 6,000 c.p.s.

Two kinds of formants

If now the formant contour comes into question as a whole, instead of the partial representation, we can see, there are strong formants and weak ones in it. There are still more interesting things, namely, when we try to run through the group of vowel formants, we will find the variant formant and invariant formant; the former is sensibly movable depending on the kind of vowel, the latter is almost fixed in position, almost independent of the kind of vowel. Invariant formant will be more evidently repre-

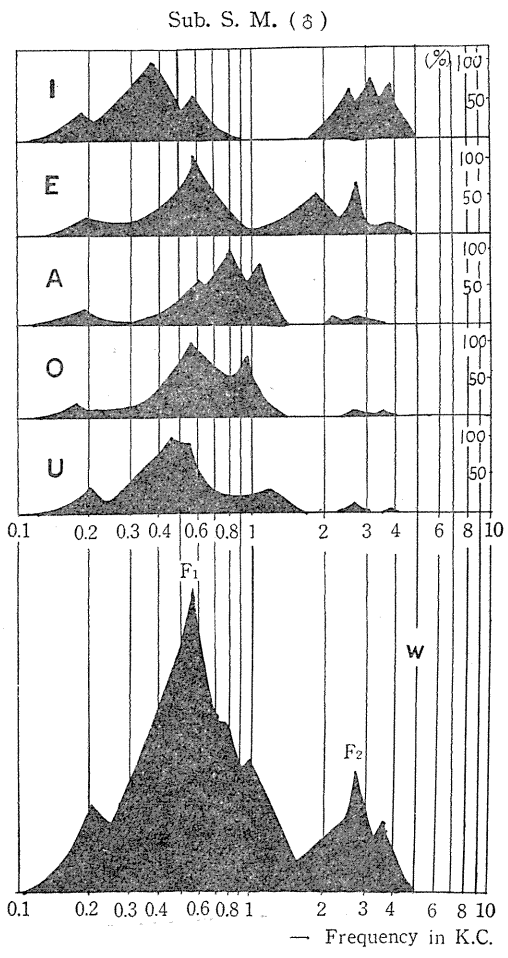


FIG. 12. Formant structure and white phone pattern of Sub. S. M.

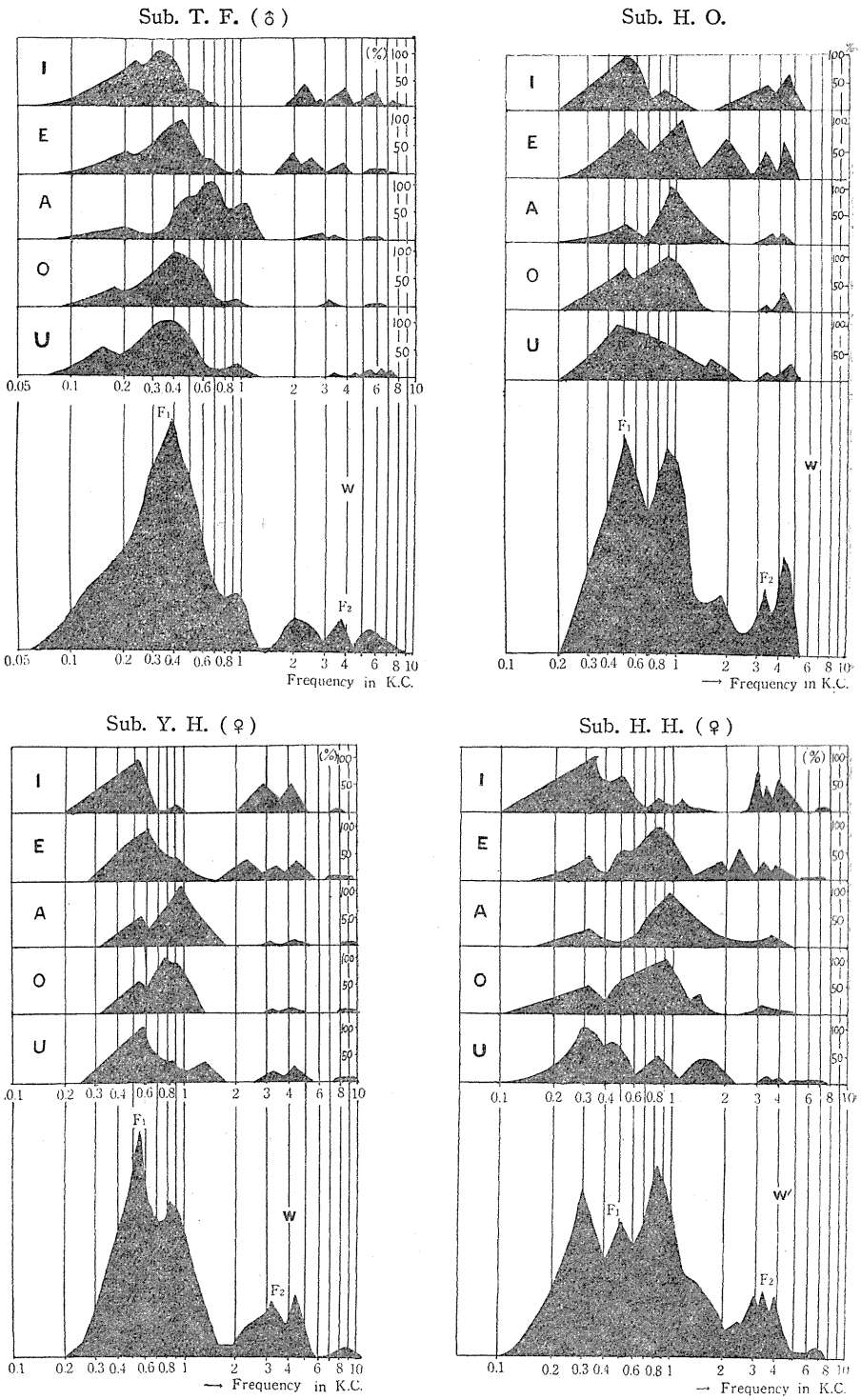


FIG. 13. Formant structure and white phone patterns of Subjects T. F., H. O., Y. H. and H. H.

sented in white-vowel expression. The signs F_1, F_2 denote their positions. In the case of Subs. H. O., Y. H., the existence of invariant formant is most conspicuous. In the case of Sub. T. F. it is slightly unconspicuous. As a good example, we will take the case of Sub. H. O., showing it in another expression. Fig. 14 shows it. This is obtained from the so-called changing-car phenomena which are expressed in Fig. 15.

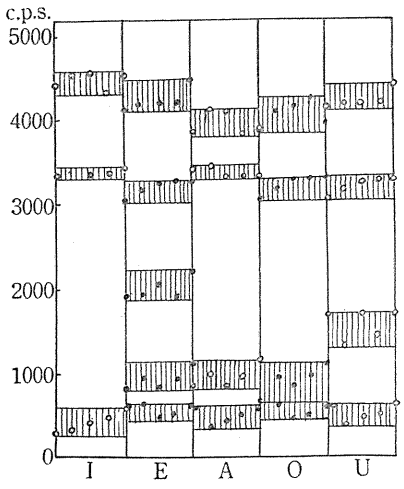


FIG. 14. Formant-peak position diagram obtained from changing-car phenomena.

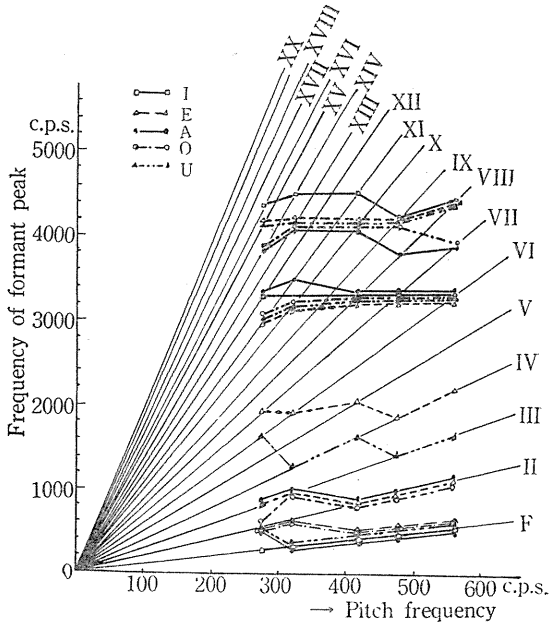


FIG. 15. Changing-car phenomena observed in the voice of Sub. H. O.

White vowel or white phone

By summing up the phonal structure of oral vowels, we come to attain the white vowel or white phone of which the outline is something like a mountain silhouette. We should like to call it, therefore, VOCAL ALPS.* This vocal alps, as we can see in it, has its own peaks and its own glens. There is a main peak and an auxiliary peak, and also there is a main glen and an auxiliary glen. By running through the silhouette in detail, we are able to find, in some occasions, a tiny gorge in the chain of peak. We see the main peak chain composed, in general, of single summit, or double summit or triple summit, so long as our subject is concerned. We must furthermore keep in mind the appearance of vocal glen which is very important in such a sense that it divides a main peak from an auxiliary one. The deepest glen functioning as a sort of demarcation between two formant groups (that is, two chains of peak) proved to be sensibly significant in our quality theory. Because, it exactly corresponds to the balancing-point frequency of band-articulation quality. This is, so to speak, a qualificative meaning

* More precisely, vocalic alps.

of VOCAL GLEN in vocal alps. A qualificative meaning of vocal peak will be evident in itself. We can find out some invariant formant among them. But the highest peak does not always correspond to the invariant formant. Variant formants being unfixed in their position, but appearing usually very powerfully one by one, are possible to take the position of highest peak. The variant formant is usually intermingled so confusedly with the invariant one, that it is not seldom uneasy to differenciate one from another. Before quitting this subject, it is fit to take some notice about the meaning of white vowel. In quality theory, we need very often the conception of vowel as a whole. White vowel justly corresponds to this conception.

Invariant formant

We have now some very interesting points to study. Turn our attention to the invariant formant. Because a thorough consideration of this part of the question will be found best introduction to the study of naturalness. Fig. 16 shows the position diagram of invariant formant of 5 vowels with regard to 5 subjects. Sub. T. F. has a little scattered type, but in comparison with its real variant formant, the invariant looks quite different in the order of its dispersion. In the case of H. H., a part of lower invariant is missed in the complexity of its fine structure.

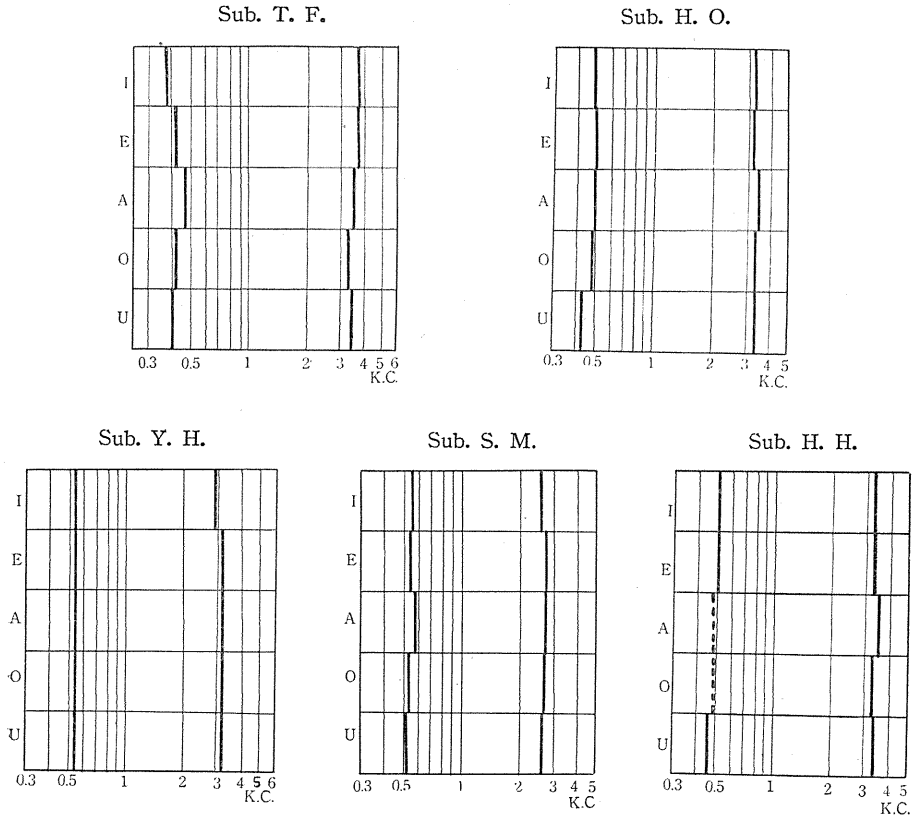
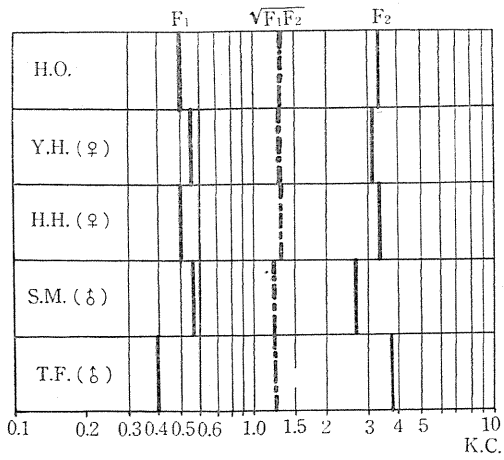


FIG. 16. Position of invariant formant of five subjects.

But its higher invariants stand in good order. The mean position of invariant is obtained in respect to 5 subjects. It is shown in Fig. 17. It is very interesting to see how the invariant interval differs according as the subject differs. But there is some inclination about the invariant, that is, invariant-frequency product remains alway nearly constant notwithstanding the subject. In the diagram, we mark the mid-frequency position ($\sqrt{F_1F_2}$) with dotted line.



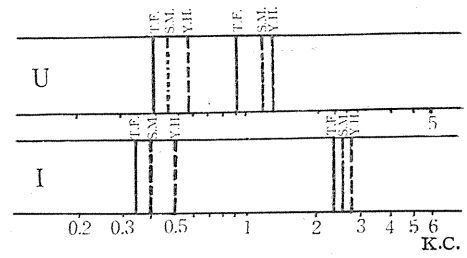


FIG. 17 (*left*). Invariant formants in outsider positions and mid-frequencies of five subjects.

FIG. 18 (*right*). Position of variant formant of two vowels about three subjects.

Variant formant

We need scarcely say about variant formant. The formant study was, in most cases, nothing but the study of variants. Therefore we have not so much interest in it. But it is worthy of mention that the interval of two main variants remains almost constant independently of the subject, under the good control of pronunciation. In Fig. 18 we can show its example where 2 vowels of "U" and "I" are examined about 3 subjects.

Vocal pattern of voices

Vocal pattern of 3 subjects obtained by superposition of 5 vowels at the pitch of 320~ is represented in Fig. 19, by which we are possible to see how the quality of voice differs according as the subject differs. Strange to say, the vocal pattern of male voice of T. F. is sensibly drawn in, meanwhile his phonal pattern considerably stretches along. The female voice of H. H. is rich in harmonics, of which the intensities are not so strong; the female voice of Y. H. (younger girl) is constituted by strong harmonics, the number of which is not so many. These vocal patterns in general are likely to be composed of 2 or 3 groups. The first group (containing the fundamental pitch) differs most conspicuously as the subject differs. The impression of various voices is; the voice of T. F. deep and soft, that of Y. H. a little hard and thin, that of H. H. clear but a little ringing. Fig. 20 shows the trend of vocal pattern with the pitch change, taking the example of Sub. S. M., which is of more complicated type than Sub. T. F.

Fig. 21 shows the comparison of the vocal pattern taken by superposition method with that taken by throat pick-up method in respect to the voice of Sub.

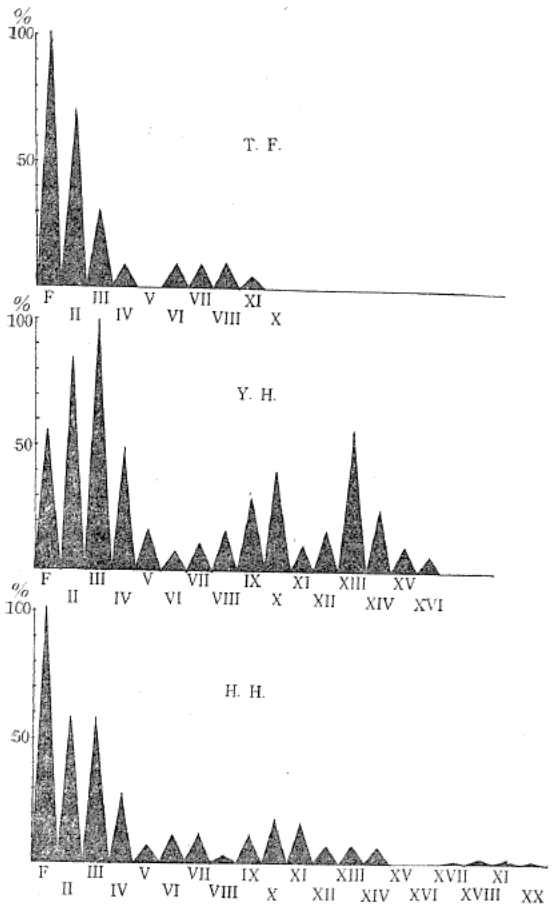


FIG. 19. Vocal pattern of three subjects at 320～ pitch.

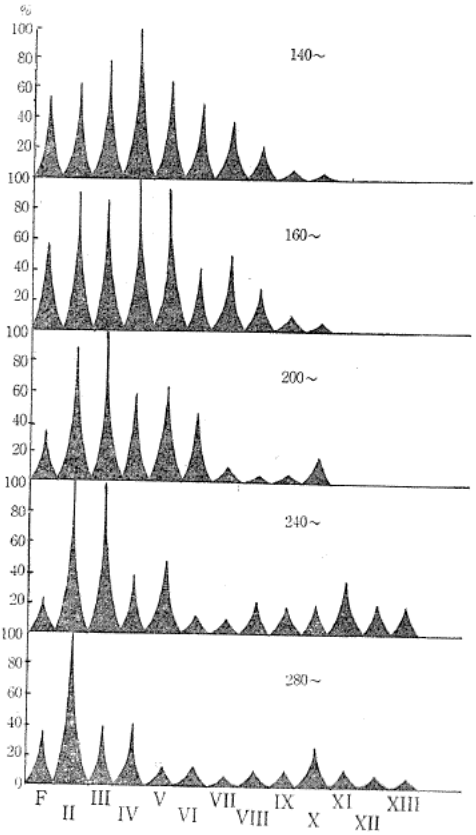
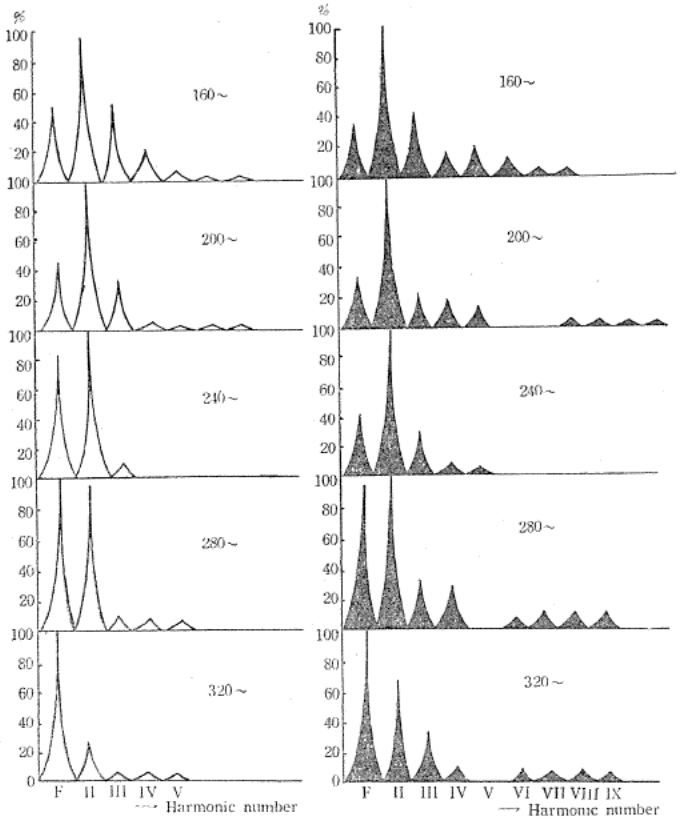


FIG. 20. Vocal pattern of Sub. S. M.

T. F. These patterns are both represented in percentage expression for the same reason mentioned above. Both are much alike. Truth to say, the former pattern is still under the remaining influence of variant formant, but the latter is almost free from it. Moreover, the exact data with regard to the attenuations of vibration in tissue or on body-surface remains to be studied. We are also destitute of the precise response characteristics of pick-up equipment. Nevertheless, the two patterns bear a striking likeness. To evince the tact for obtaining such good similarities, it is absolutely essential to balance the uttering level in both cases.

White voice

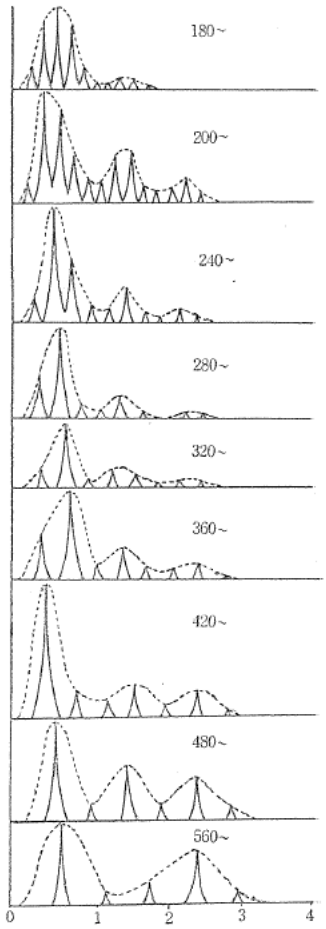
We have noticed that the white voice produced by the superposition of oral vowels cannot always be free from the influence of oral cavities. The compensation of phonal qualities of vowel is thereby perhaps imperfect. In other words, the effect of variant formants more or less still remains. The pure detection of influence of invariant formant will be performed by the throat pick-up method.



(*A*) By throat pick-up method.

(*B*) By vowel superposition.

FIG. 21 (*left*). Vocal pattern of Sub. T. F.



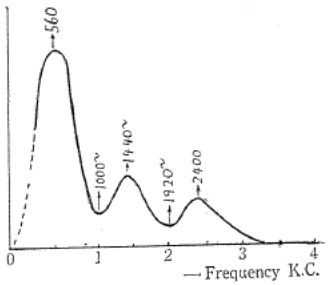


FIG. 22 (*right*). Invariant formants detected by throat pick-up method in the case of Sub. H. H.

The analysis example obtained in the keenest way is shown in Fig. 22. This is the voice of Sub. H. H. where nine pitches are used. The totalized pattern of them is shown in the last. There is 3 invariant formants. The middle formant was not found in the example of superposition method. We can see here how the invariant formants fix their positions independently of the pitch. With the pitch change, the most predominant component must change their harmonic number. We call this phenomena changing-car phenomena. These vocal patterns were

obtained in another occasion than that where the formant structure was obtained. The uttering level was not the same in both cases. The position of outsider invariants differs each from other. But their product is almost equal in two cases. The fact that the vocal pattern changes its pattern form depending on level is shown in Fig. 23 which is taken from the example of Sub. T. F. Here we see the patterns of 2 pitches on 3 different levels. The changing-car phenomena happen between *piano* and *mezzo* levels, of which the difference in VU level is 6 db. Between *mezzo* and *forte*, this phenomenon can not take place evidently but the pattern form undergoes thereby considerable change.

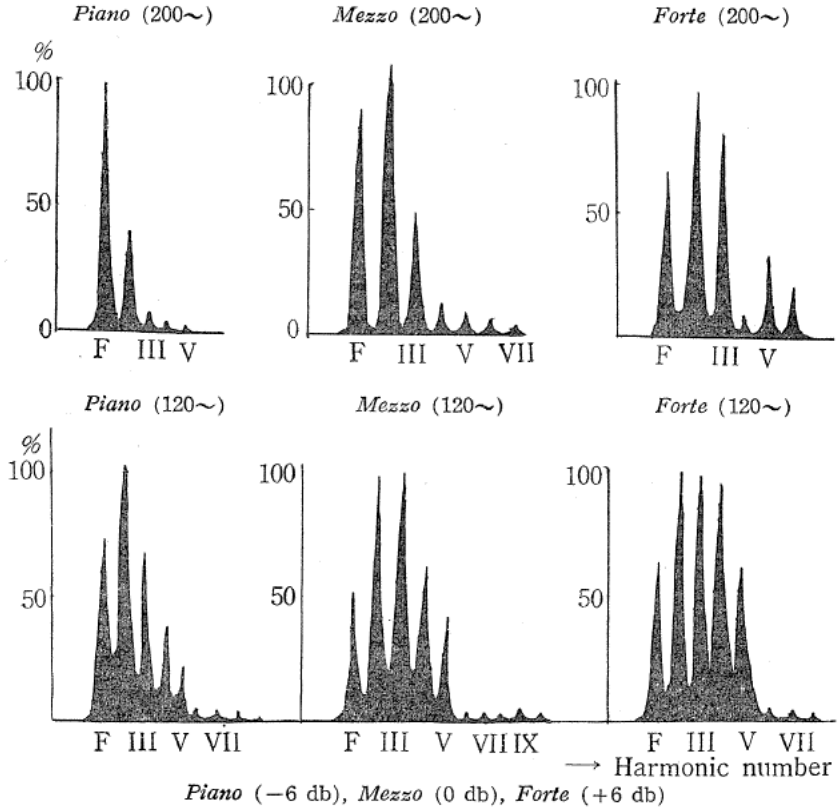


FIG. 23. Change of vocal patterns effected by uttering level and detected by throat pick-up method in the case of Sub. T. F. The uppers are $Z(200\sim)$, the lowers $Z(120\sim)$.

Voiceless vowels

For reference we have investigated also the voiceless vowels. Wave form of voiceless vowel is sensibly unstable, because the sustained utterance of this vowel is not so easy for a certain subject. Each vowel has 5 repetitions which is averaged as mean structure of its contour. Fig. 24 shows the two cases of Sub. Y. H. and H. H. The variant and invariant formants being so evidently separated each from other, we can trace here the two kinds of formants without such difficulty that was experienced in the voiced vowels. In the diagrams, the white bar denotes

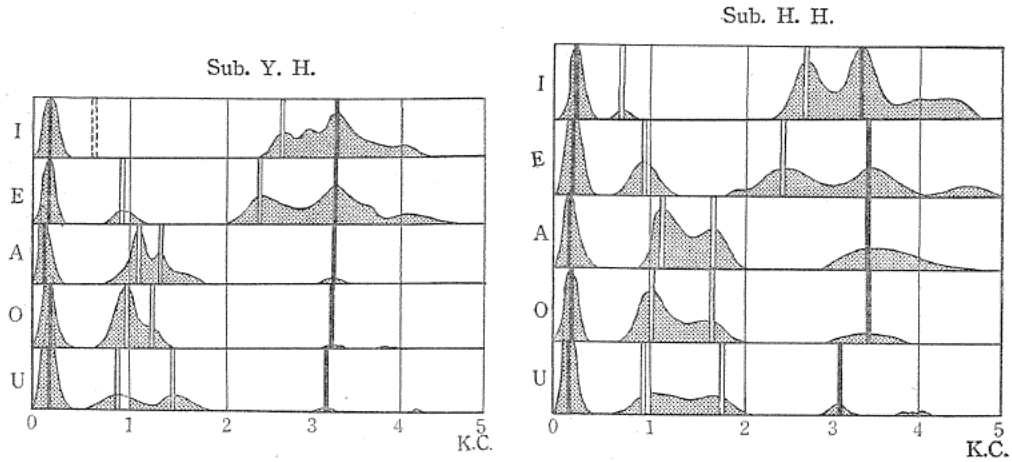


Fig. 24. Invariant and variant formants of five voiceless vowels of two subjects.

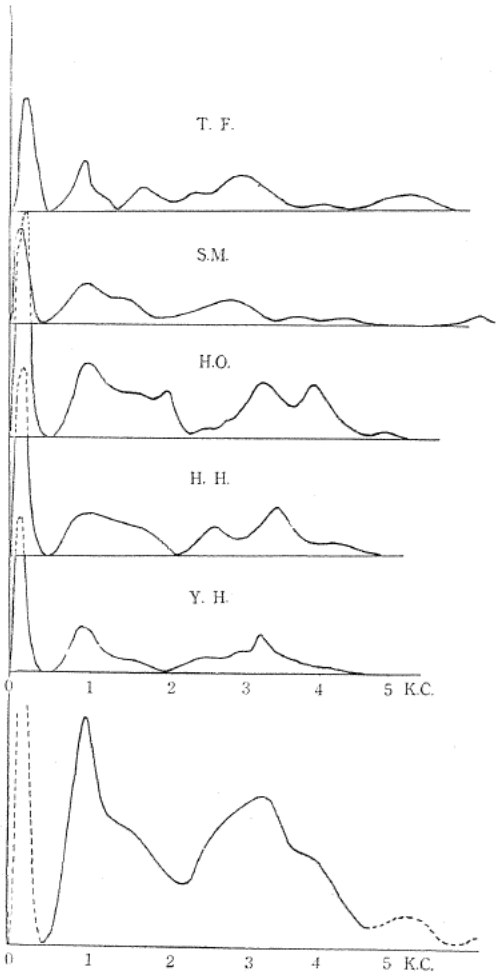


FIG. 25 (*left*). White phone in voiceless condition of five subjects and their averaged mean structure.

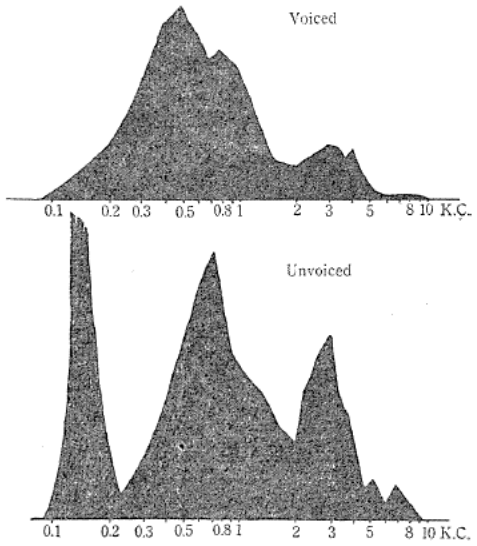


FIG. 26 (*right*). Averaged mean structure of white vowel of five subjects in voiced and voiceless utterances.

the variant, and the black bar the invariant. We can give also the white phone in whispered condition in Fig. 25 where 5 subjects are examined and lastly the mean pattern of 5 subjects is given there. We can see everywhere the evidently appeared invariants, of which the position is, however, almost fixed independently of subject. It will be meant that by and by the voiceless voices come to lose their individuality which they continued to preserve persistently in voiced condition. In voiceless condition, the higher invariant looks to keep nearly the same position as in voiced case, but the lower invariant gets lower position, displacing itself down considerably. Fig. 26 shows the white phone in both utterances, which is represented in totalizing the patterns of 5 subjects. These patterns can show their marked contrast.

Conditioning of level

The secret tact of timbre study of voices consists in the control of the mode of utterance, which is to be expected on the side of uttering subject. The con-

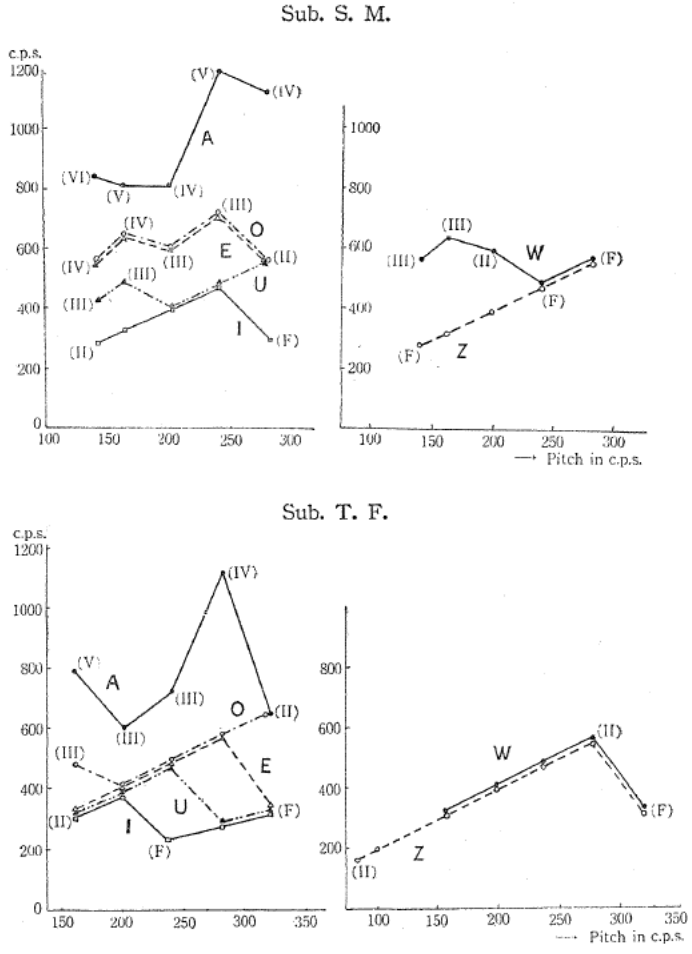


FIG. 27. Changing-car phenomena in lower formant region.

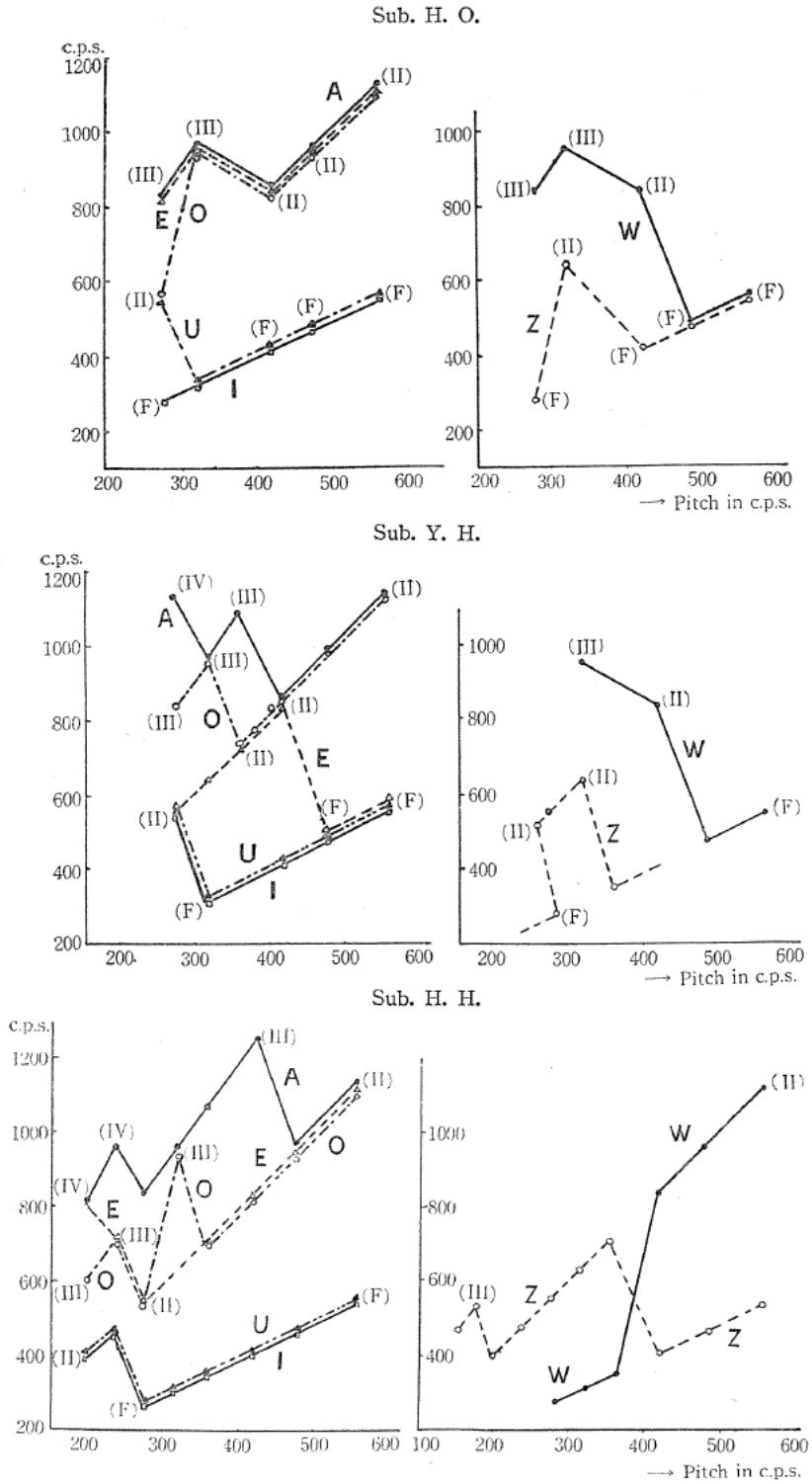


FIG. 28. Changing-car phenomena in lower formant region.

venient method of supervising this mode of utterance could not be found. All that we can do to meet our immediate requirement in this respect is to condition as much as possible as well the uttering level of subject as the pick-up level for analyzing on the one hand, and on the other hand to make a check-up, investigating, for instance, the changing-car phenomena somewhere in the domain of lower harmonics. Figs. 27, 28 show this example. These figures are obtained about five vowels as regards five subjects. The changing-car phenomena are going, in most cases, in good order. But we can see some of them thrown in disorder on rare occasions. For example, it happens rarely the disorder in changing harmonics about a certain pitch. This means indirectly the disorder of the uttering mode which influences significantly on the timbre analysis, resulting in the disturbance of timbre pattern. In these diagrams, the level conditioning of white voices is also checked. Such a check-up can be used as a material of judgement for the perfectness of the timbre pattern thus obtained.

Band interval and band position

Judging from the result of the timbre analysis of this kind, there will be needful at least two conceptions of frequency band in respect to the voice transmission; band interval and band position. For the transmission of articulation of vowels, we are in need of a conception of frequency band interval, that is, f_2/f_1. For the transmission of naturalness of voices there will be needful a conception of the position of frequency band, that is, $F_1 \cdot F_2$. The frequencies that determine the frequency ratio are those which come into play with the variant formants of vowels. The frequencies that determine the frequency product are those which come into contact with the invariant formants of vowels.

If the articulation center of vowels is not far from the frequency point of vocal glen, and moreover if the naturalness center of voices comes also to the point of vocal glen nearby, the system design for adequate transmission of voice and vowel will not be so difficult as to be considered at first glimps. We show in Fig. 29 the mid-frequency of main variants of "I" vowel of all subjects and also the mid-frequency of outsider invariants of all subjects. The mid-frequency of "I" variants shifts according as the physical constitution differs; a tall male has a lower position, a short female a higher position. This perhaps has some concern with the spatial capacity of oral cavities.

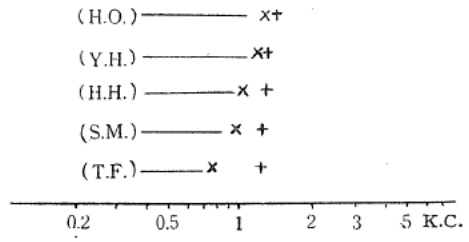


FIG. 29. + mid-frequency of invariant formants. × mid-frequency of variants of "I" vowel.

The mid-frequency of invariants however keeps almost the fixed position, notwithstanding the physical constitution of subjects. We cannot know what is meant by this. But this phenomenon can not hastily pretend to its absolute correctness. Because our subjects are only 5 and they are too young and too short (except one)

to be considered as the typical representatives of Japanese. We will be necessary to have more examples. With reserve we must speak of our data here. This is just the thesis on which nearly everything remains to be studied.

All that we know in inferring from these phenomena is; where the constant frequency interval is of meaning in its transmission quality, there the conception of the position of frequency band can not always hold good. Meanwhile, where the meaning of constant band interval does not hold good, there seems to be set up the conception of the band position. This is, however, a matter of relativeness. Because, this is nothing but the phenomena in quality, and not the physical phenomena. If the center of naturalness is consistent with the center of articulation, our design of transmission system will become easy. It is true. But there is no positive reason that these two centers must necessarily coincide with each other.

Discussion

As it is rather difficult to get a conception of naturalness with the whispered vowels, so it is also a little difficult to test the existence of two qualities with the sound phenomena of consonant, although it does not mean that the aspects of naturalness and articulation are to be denied utterly in the qualificative interpretation of consonant sounds.

The real substance of telephone articulation consists naturally in the consonant transmission. The reason why we call in question here merely the naturalness and articulation restricting ourselves only to the problem of voiced vowel in harmonic construction, is nothing but that we try to find the relation of two qualities under the most severe condition, that means, to trace the most secret clues to timbre study by groping about in the internal sturcture of harmonic configuration. For establishing the conception of naturalness, it is useful to grasp this idea as such that forms a fine contrast to the idea of articulation. It will be meaningless to place these two ideas on the same side of things. From the monismic conception of physical fidelity, we cannot lead these two qualities.

As phenomenological effect, it might be without so much difficulty to find the existence of the so-called invariant formants. The real difficulties lie rather in giving the meaning to them. To give the meaning correctly, we must place the basis of quality theory in communication in the most right position. Without this foundation of quality theory, we cannot succeed in giving the meaning, and, more in some occasion, we will not achieve in catching the existence of invariant formants even in the face of them.

That we try to classify the formants into two groups, it means that we prepare one stage to mount up for looking out over the complicated landscape of timbre and for treating it at home. In other words, such a classification of formant will be helpful as necessary lines of approach to the impregnable fortress of timbre. We shall not fail therefrom to jump into the problem at once, ceasing to go round and round about in vain.

We are destitute now of giving physical meaning to the results of timbre analysis. It will be none the less useful to do so. For example, the physical meaning of vocal glen and that of invariant formant are both yet unknown. We are trying to go along furthermore.

We should like to lay stress on the fact that it is at first needful to give a

qualificative meaning to the result of timbre analysis and to give it the physical meaning is but the second step to be done. The full description of conceptional study will be set down in another case.*

Conclusion

We have picked up the problem of human voice as a signal sound in communication. If we succeed in solving the voice problem, we can say we are half way already in the course of sound study. Timbre study of vocalic voice is reasonably the vital question of timbre of all sounds. We try to interpret qualificatively the result of the timbre analysis and also try to represent the timbre itself. We have undertaken fundamentally the conceptional study of timbre, and starting therefrom, we have tried to study the internal static structure of human voices as an example of measuring the timbre construction. We have tried also to know how the timbre is to be represented, for the purpose of good understanding as well of the transmission quality as of the timbre sensation.

This study will be a little serviceable also for phonology, phonetics or science of vocal music. Because, that sort of relation between voice and phoneme comes here into question.

There are many things of which the full meaning cannot be realised until subjective experience brings them home. It is quite so with the timbre problem. No solid results have been yet attained. But we are content only with proposing the steps of thinking in such a manner, and with suggesting the timbre interpretation in such a way.

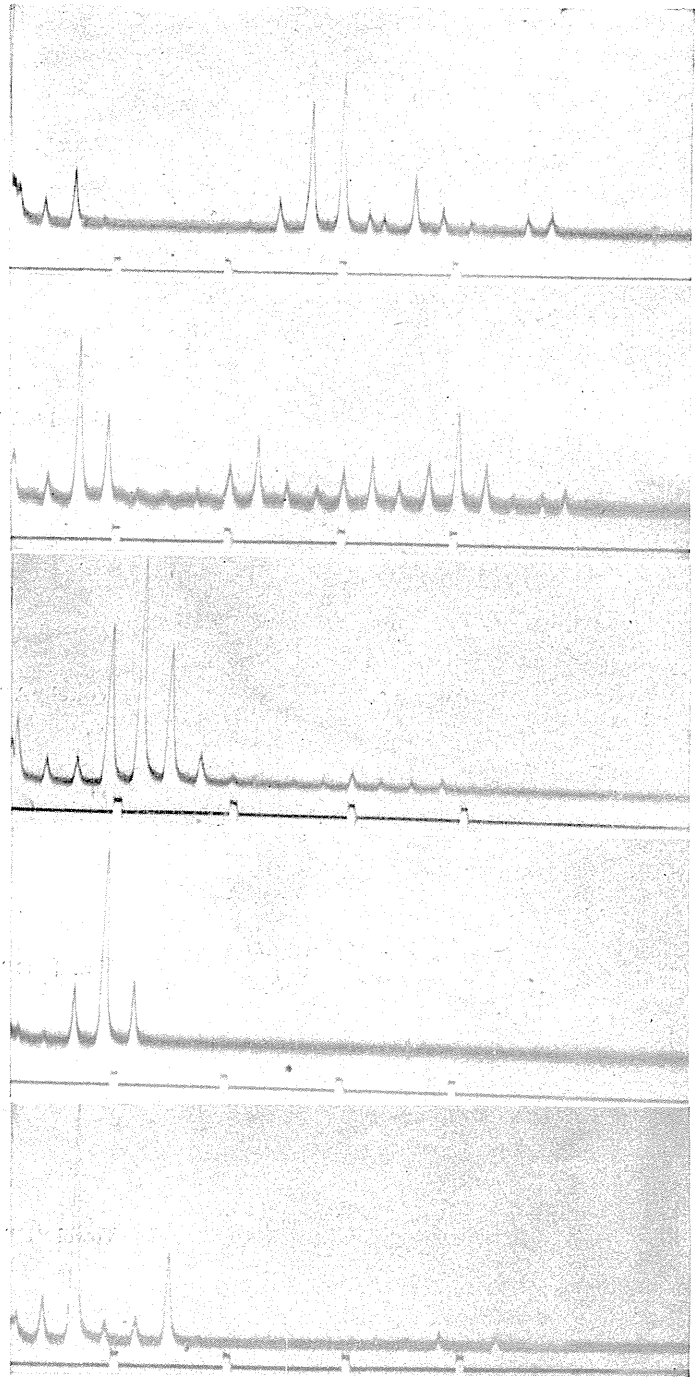
A part of this research is indebted to the grant of Developmental Scientific Research of the Educational Department of this country. Grateful acknowledgment must by payed to the Telecommunication Board of Tokai-district, that offered us several conveniences as regards the equipment and the personal assistance. The assistance and help of Mr. Kenzo Hara are also profoundly appreciated.

References

1) Y. Ochiai, T. Yamashita and T. Fukumura: Denso-hinshitsu ni tsuite (Über den Grundbegriff der Übertragungsgüte). Research Report of the Faculty of Engineering (R.R.F.E.). Nagoya University **3**, 1 (1950).
2) Y. Ochiai and T. Yamashita: Chujitsusei no ronri (Logik der Treue). R.R.F.E. Nagoya Univ. **3**, 2 (1950).
3) Y. Ochiai: Shizensei ni tsuite (Sur la naturelesse). R.R.F.E. Nagoya Univ. **4**, 1 (1951).
4) Y. Ochiai: Mémoire sur les sons des voix humaines. Memoirs of the Faculty of Engineering (M.F.E.). Nagoya Univ. **4**, 2 (1952).
5) Y. Ochiai, T. Yamashita and T. Fukumura: Boin-onshoku no kenkyu(A la recherche du timbre des voyelles; Quelques considérations basées sur expériences sur la structure de timbre sonore des voyelles soutenues). R.R.F.E. Nagoya Univ. **4**, 2 (1952).
6) Y. Ochiai et H. Kato: Sur la netteté et la naturalité de la voix humaine, réflechie du point de vue de la qualité de transmission. M.F.E. Nagoya Univ. **1**, 2, Oct., (1949).
7) Y. Ochiai and T. Fukumura: Studies on qualities of speech and voice by timbre distortion. M.F.E. Nagoya Univ. **4**, 2, Nov., (1952).
8) T. Yamashita: Some data on voices observed in the wave modes of body-surface vibration. M.F.E. Nagoya Univ. **4**, **1**, July, (1952).
9) Y. Ochiai et S. Saito: Sur la représentation de la structure phonétique des langages. M.F.E. Nagoya Univ. **4**, 2 (1952).

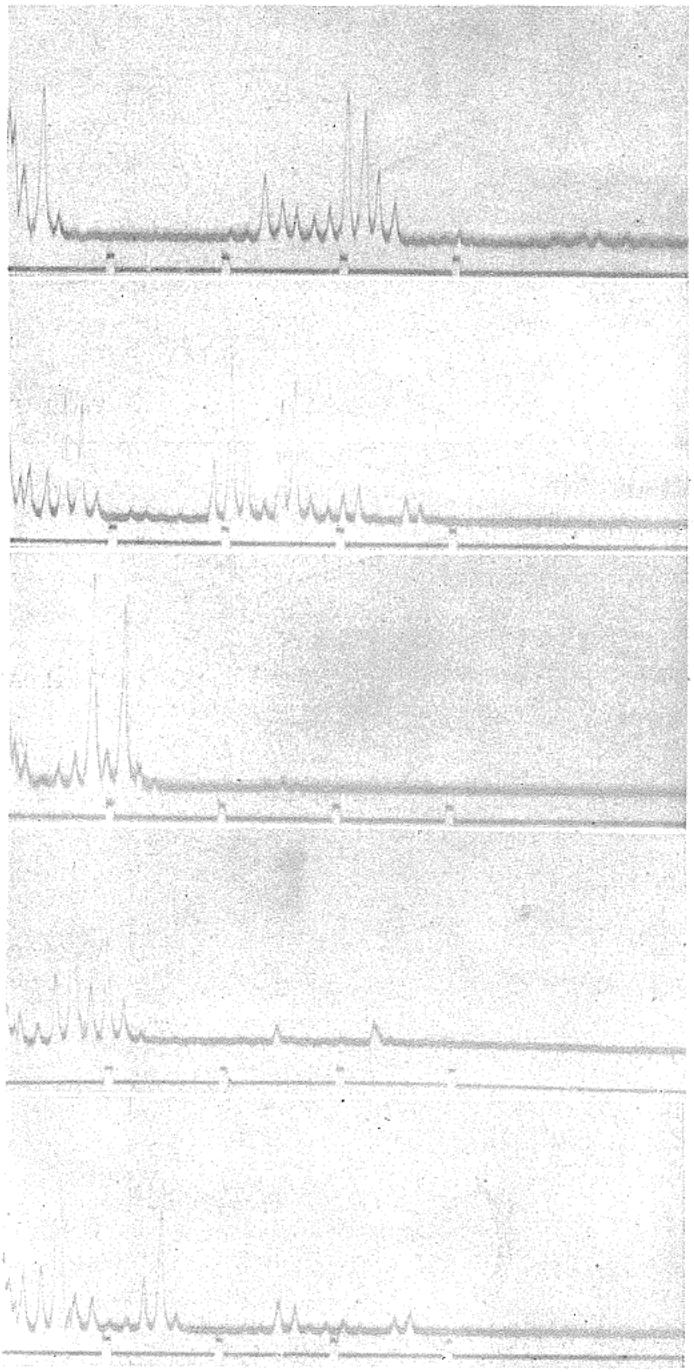
* Y. Ochiai: Reflection on the basis of transmission in communication. (unpublished).

PLATE 1.

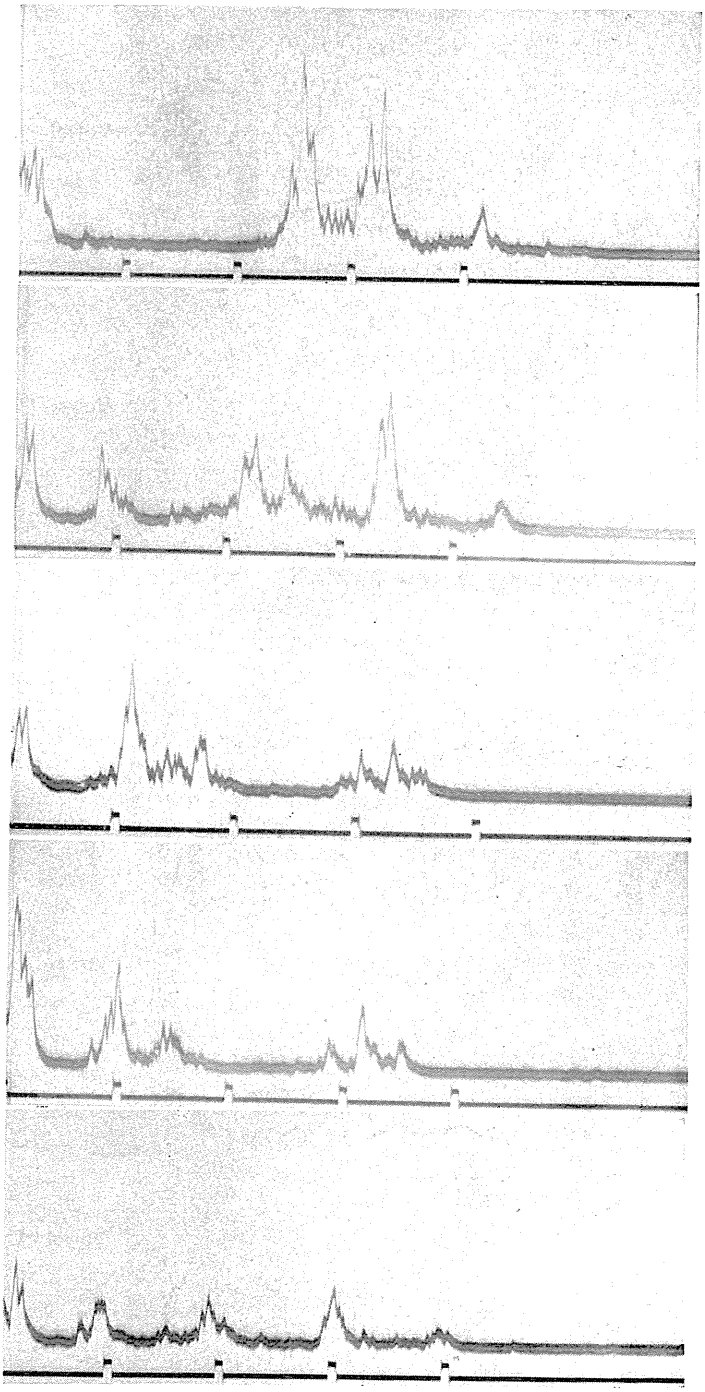


Five vowels of Sub. Y. H., voiced at the pitch of 280~.

PLATE 2.

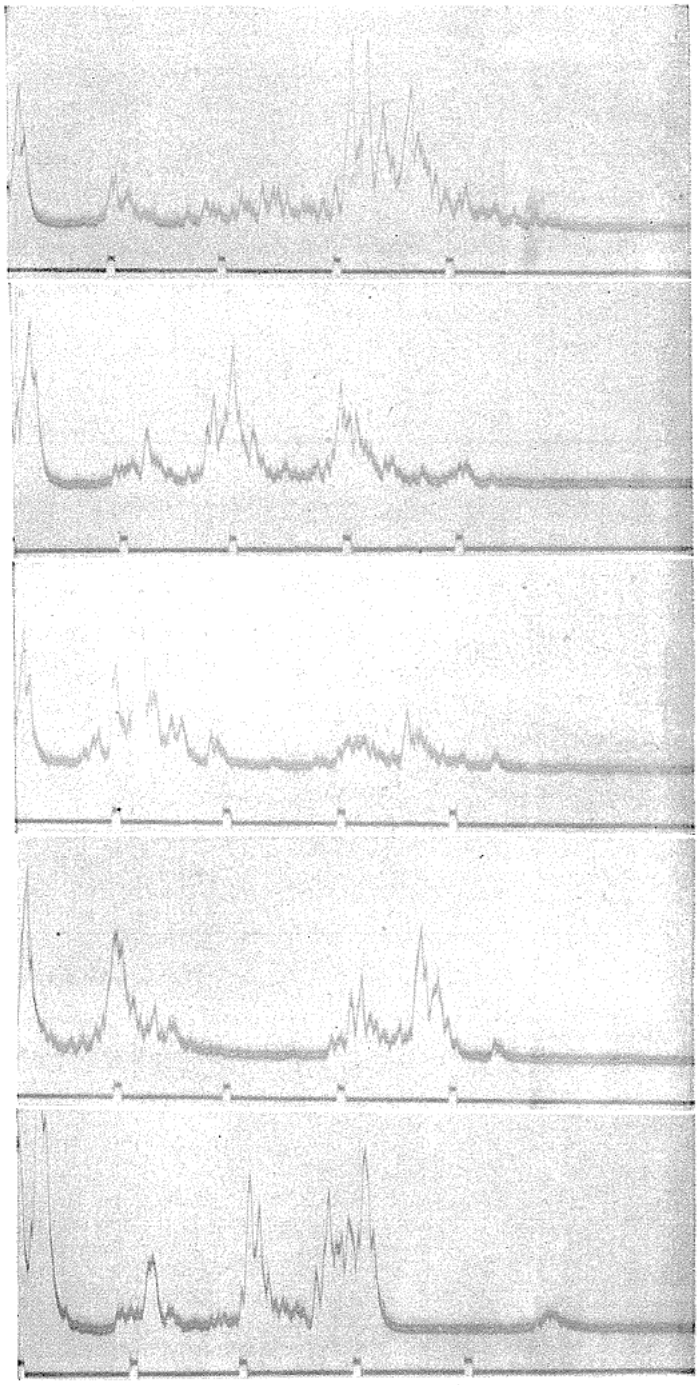


Five vowels of Sub. S. M., voiced at the pitch of 160~.

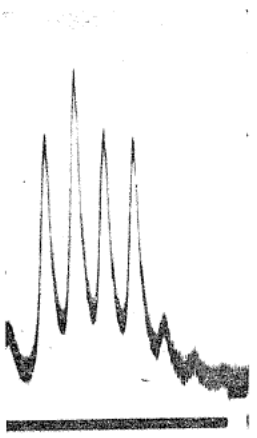


Five voiceless vowels of Sub. H. H.

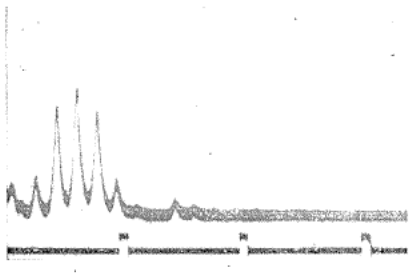
PLATE 4.



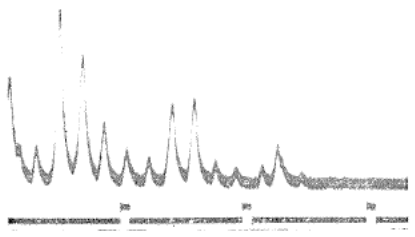
Five voiceless vowels of Sub. H. O.



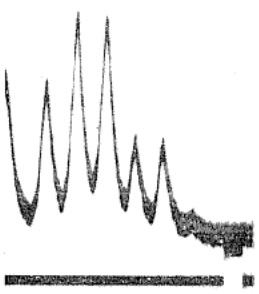
Vowel "I"



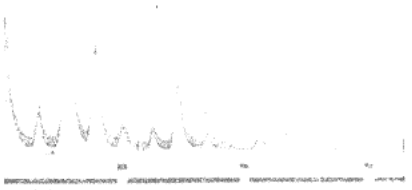
Pitch 180~



Pitch 200~

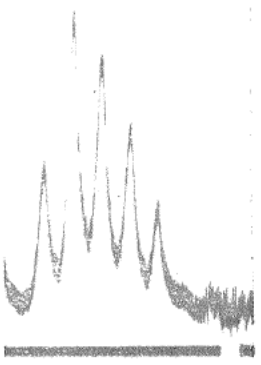


Vowel "E"

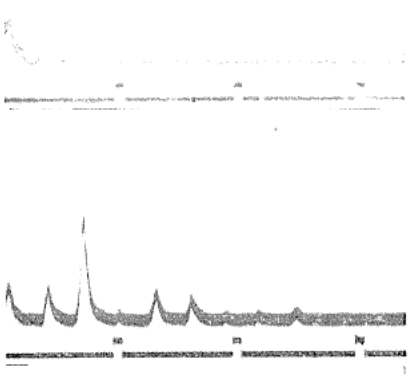


Pitch 240~

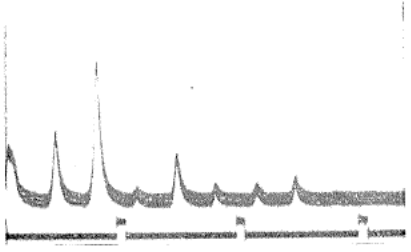
Pitch 280~



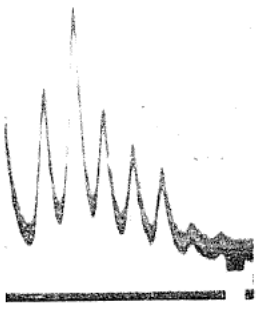
Vowel "A"



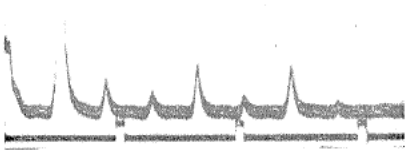
Pitch 320~



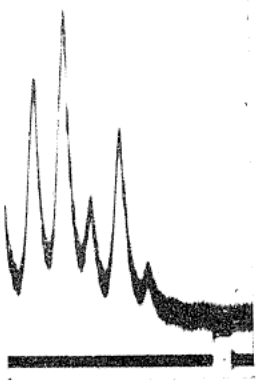
Pitch 360~



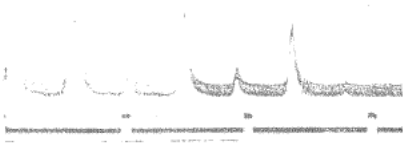
Vowel "O"



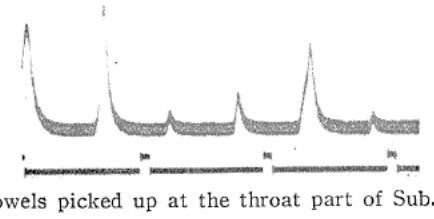
Pitch 420~



Vowel "U"



Pitch 480~



Pitch 560~

(*left*) Vibration figure of five vowels picked up at the throat part of Sub. T. F., when the vowels are voiced at the pitch of 120~.

(*right*) Vibration figure of "A" vowels picked up at the throat part of Sub. H. H., when the vowels are voiced at several pitches.