TIMBRE STUDY OF VOCALIC VOICES VIEWED FROM SUBJECTIVE PHONAL ASPECT

PART III. GENERALIZED TREATMENT OF TIMBRE CONFUSION SECTION A—PHONEMIC CONFUSION

Yoshiyuki OCHIAI and Teruo FUKUMURA

Department of Elctrical Engineering

(Received October 31, 1956)

Here we deal with an exact and detailed method of treating confusion phenomena in quality aspect, citing an example of confusion in band-eliminating distortion (BED). We treat these phenomena from three aspects: First, from that of timbre quality; second, from that of distortion of band-eliminating nature, and, third, from that of confusion-matrix. TIMBRE-QUALITY consists of two qualities, phonemic and vocal, which take priority over the others. In DISTORTION DIRECTION, there are two, low-cut and high-cut, both of which are necessary for our co-ordinate of observation. CONFUSION PHASE, of which there are two in confusion-matrix display, incoming and outgoing, are both at our disposal. In addition, we here present confusibility data as to phonemic and vocal confusion derived from our original experiment of 1953. In our study we emphasize three points of importance: (1) Confusion study gives the most effective means for timbre-quality study because confusion is essentially timbre phenomena in a differential sense; (2) Outgoing confusibility defined as uncertainty of correspondence viewed from signal toward quality in matrix representation, and incoming confusibility defined as uncertainty of correspondence viewed from quality toward signal; (3) The intricate nature of the confusion problem can be clarified by the introduction of both incoming and outgoing confusion concepts. As the technique for introducing both confusions, we adopt, for the moment, the difference form of these two confusibilities, $|C_{in}-C_{out}|$, both useful for characterizing essential confusion trend for which we here give actual data in phonemic confusion and in vocal confusion. Thus from the joined inspection of both incoming and outgoing confusions, we can propose *quality-formative* and *quality-ruinous* effects of band-cutting process, which, viewed from still another angle, is the quality formative and quality-ruinous nature of band region in frequency domain.

CONTENTS

Introduction

Already having finished a study[1] on how to obtain the timbre patterns of our calling subjects aiming not only at phonemic but at vocal patterns, as our next step we attained the quality measurement of these timbre signals. In Part I[2] of this series report we discussed the two general quality characteristics, *i.e.*, general articulation characteristics AC (G) and general naturalness characteristics NC (G), and further discussed individual quality characteristics, *i.e.*, the four kinds of characteristics, AC/V_c, NC/V_c, AC/V_1, NC/V_1. In Part II[3] we explained the meaning and significance of our study on the confusion problem and in addition gave a brief outline of the confusion phenomena, using the data of our original experiment in 1953. Clear as those phenomenological descriptions may be and naïve as were the descriptions without any accompanying dexterious technique of thinking, it may yet be well to describe the phenomena in a somewhat different way by resorting to mathematical expression.

Thus we are at the stage in our study where, in a generalized and systematized procedure, we must give precise form to a problem so complicated and so important. If we can find the best method of treatment to accomodate this confusion phenomenon, then surely a considerable part of the quality theory will be established and consequently a highway leading to practical application will be opened up. Our treatment of confusion presented here, however, is but a rough schema in an attempt to interpret the core problem of quality phenomena by using the confusion data in our preliminary experiment.

Outline of Fundamental Schema of Treatment of Confusion

In our present treatment of confusion phenomena, we must keep to the same line as that taken in the previous studies (Parts I and II) which is that our timbre study is to be looked on as a study of correspondence between timbre signals and timbre qualities. It must be remembered that in Part I we presented a comparison study between timbre patterns (vocal and phonemic) and quality characteristics AC(G), NC(G), discussing therein the outstanding points in quality distribution and that in Part II we attempted to compare such detailed confusion characteristic with each timbre pattern, this time discussing the type of quality distribution in frequency domain more directly, conforming to the frame-work of our quality theory.

In the present study, Part III, we must advance our fundamental basis along this fruitful line of inquiry. We must stand on that basis and look at timbre confusion as a mathematical problem of correspondence between a set of timbre signals and a set of timbre qualities. By "timbre signal" we mean input of timbre information given to or impressed upon a recipient and by "timbre quality" we mean the timbre response which is evoked by signal or the timbre information understood by and carried through a recipient. The problem of correspondence between any two sets is, of course, made accessible by mathematical treatment. Thus the most subtle one of all timbre-quality subjects is made accessible by the display of matrix in ensemble.

Here we want to call attention to one point. What we adopt is neither the signal-to-signal matrix (which we ourselves studied[4] previous to the attempt by G. A. Miller who made an exhaustive consonant study[5] based on the information theory) nor the quality-to-quality matrix which we shall discuss another time. What

we present here is, in fact, our signal-to-quality matrix and the reason we have adopted this matrix above all others will be explained later under the heading of Discussion.

Common Basis of Generalized Description of Timbre Confusions

We have already given a brief outline of confusion phenomena, depicting only the most outstanding features and discussing some fundamental quality principles. Now we must pay attention to a more exact and precise formulation of the phenomena by searching for the details of confusion phases. First we begin our study by giving some precise form to the ensembles in question.

Designate a set of timbre-signals, *i.e.*, signal ensemble, by $(X_1, X_2, \ldots, X_n)$, and in the same manner a set of timbre-qualities, *i.e.*, quality ensemble, by $(x_1, x_2, \ldots, x_n)$, then the correspondence-relation of two ensembles is schematized by either

	x_1	x_2	$\cdots$	x_n
X_1	α_{11}	α_{12}	$\cdots$	α_{1n}
X_2	α_{21}	α_{22}	$\cdots$	α_{2n}
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
X_n	α_{n1}	α_{n2}	$\cdots$	α_{nn}

or

	X_1	X_2	$\cdots$	X_n
x_1	β_{11}	β_{12}	$\cdots$	β_{1n}
x_2	β_{21}	β_{22}	$\cdots$	β_{2n}
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
x_n	β_{n1}	β_{n2}	$\cdots$	β_{nn}

where α_{ij} is the confusion coefficient which expresses some measure of correlation between signal X_i and quality x_j and where β_{ij} is the same between quality x_i and signal X_j, and both can be measured by any one of the quantities such as Frequency of Occurrence, Transition Probability and Digram Probability. By α_{ii}, β_{ii} we can measure correct identifications and by α_{ij}, β_{ij} $(i \neq j)$ incorrect identifications, *i.e.*, the confusions between signals and qualities. Thus the formulation given by the ensembles of coefficients is available for complete display of every detail of confusion phenomena.

As that may be, we must guard ourselves against being misled by the following point. In an ensemble of sufficiently large number, the all-inclusive treatment of confusion necessitates the measurements and computations of all possible coefficients α_{ij}, β_{ij} the labor of doing which is quite troublesome; in some cases the researcher becomes dazzled by an excess of detail in such a way as to cause him to lose sight of the essential meaning of confusion phenomenon itself. For that reason, as a guard we take into consideration the following three parameters when we make a compendious and time-saving study of confusion: (1) The uncertainty of correspondence viewed from any individual signal toward quality ensemble; (2) the uncertainty of correspondence viewed from any individual quality toward signal ensemble; (3) the difference between these two prescribed uncertainties when examined from the viewpoint of the reciprocity of confusion, *i.e.*, viewed from the symmetricity of matrix. For every parameter we must prepare numerical representations to show the detailed features of confusion characteristic.

Suppose now that α_{ij}, β_{ij} mean transition probabilities, *i.e.*, conditional probabilities $P_{x_i}(x_j)$, $P_{x_i}(X_j)$, then we can have two quantities given by

$$C_{out}(X_i) = P(X_i)\sum_{j=1}^{n} P_{X_i}(x_j) \log P_{X_i}(x_j)$$

$$C_{in}(x_i) = P(x_i)\sum_{j=1}^{n} P_{x_i}(X_j) \log P_{x_i}(X_j) \qquad (i = 1, 2, \ldots, n)$$

where

$$P(X_i) = \sum_{j=1}^{n} P(X_i, x_j)$$

$$P(x_i) = \sum_{j=1}^{n} P(X_j, x_i).$$

By the former, $C_{out}(X_i)$, we define *elemental outgoing confusibility from signal* X_i, that is, the uncertainty of correspondence between a signal X_i and ensemble of qualities, and by the latter, $C_{in}(x_i)$, we define *elemental incoming confusibility to quality* x_i, that is, the uncertainty of correspondence between a quality x_i and ensemble of signals. We will explain further by citing an example or two of simple nature. For example, equally large values of both confusibilities $C_{in}(x_i)$, $C_{out}(X_i)$ indicate the fact that from quality viewpoint the constitution of ensemble is so characterless that there are no considerable dissimilarities between any two elements in ensemble; no definite quality discriminations of x_i from others in ensemble result. Both outgoing and incoming confusibilities are alike high. As another example we consider the case where outgoing confusibility is larger than incoming, *i.e.*, $C_{out}(X_i) > C_{in}(x_i)$. This means that the signal X_i in question is more often confused with qualities other than x_i, and at the same time signals other than X_i are less frequently confused with the quality x_i. This case may be called *negative confusion trend.* The inverse relation $C_{in}(x_i) > C_{out}(X_i)$ may be called *positive confusion trend* which means that the signal X_i has a greater incoming to the quality x_i from other signals, at the same time having a smaller outgoing from this signal X_i to qualities other than its own quality x_i.

By developing confusibility coefficients, we can acquire in addition to the elemental, a more collective meaning of confusion. For instance, by summing up elemental outgoing confusibilities for all signals, and incoming confusibilities for all qualities, we define *ensemble outgoing confusibility* and *ensemble incoming confusibility* respectively, by expressions

$$C_{out} = \sum_{i=1}^{n} C_{out}(X_i)$$

$$C_{in} = \sum_{i=1}^{n} C_{in}(x_i)$$

each of which represents the average mean for each ensemble. Since these quantities are equivalent to Shannon's *equivocations*,[6] we are led to the following relation

$$H(X) - C_{in} = H(x) - C_{out}$$

from which this

$$H(X) - H(x) = C_{in} - C_{out}$$

can be derived, where

$$H(X) = \sum_{i=1}^{n} P(X_i) \log P(X_i)$$

$$H(x) = \sum_{i=1}^{n} P(x_i) \log P(x_i).$$

We obtain the most distinctive general trend

$$C_{in} \gg C_{out}$$

by using the logatome in which all elements are of equal occurrence in order to compose the signal ensemble in such a way that $H(X)$ takes its maximum value, and by adopting the careful process of experiment by which any kind of elision phenomena except confusion is kept out as far as possible.*

It is noteworthy that while there can be positive and negative trends in *elemental confusibility* there cannot be a negative trend with respect to *ensemble confusibility*. Here there are only two cases, one of equality $C_{in} = C_{out}$ and one of inequality $C_{in} > C_{out}$. Equally large values of both ensemble confusibilities indicate that in the sense of ensemble-mean the correspondence between signal and quality ensembles is high in its uncertainty. The relation of inequality, that is, the superiority of ensemble-incoming over ensemble-outgoing, indicates that in the sense of ensemble-mean the signal ensemble also is susceptible to an eccentrically distributed high response toward some particular qualities.

Phonemic Confusion

Following the steps of computation prescribed, the confusibilities are obtained in every condition of both low frequency cut-off (LCD) and high frequency cut-off (HCD) for five vowel-phonemes. The designations of phonemic timbre are: "I", "E", "A", "O", "U" for phoneme-signals, and "i", "e", "a", "o", "u" for phoneme-qualities. In Fig. 1(a), the abscissae represent the cut-off frequencies (kc) in logarithmic scale; the ordinates represent the confusibilities in bit; the solid- and dotted-line curves represent respectively the outgoing and incoming confusibilities which we shall hereafter refer to as Elemental Confusibility Characteristics. We see in Fig. 1(a) the individual phonemic confusibility characteristics under the distortion BED. For comparison of these characteristic curves with the phonemic patterns of the five vowels averaged for five voices they are shown side-by-side in Fig. 1(b).

We must now lay stress on the joint study of the relative movement of the two characteristics $C_{in}(x_i)$ and $C_{out}(X_i)$. This means that it is more important for us to examine the general tendency of combined mutual relationship between characteristics $C_{in}(x_i)$ and $C_{out}(X_i)$ than to scrutinize separately every detailed feature of individual characteristics. By examining the relative movement of two characteristics, we become familiar with the characteristic points, such as the crossover point and the maximum point of their discrepancies. The crossover point which shows the equilibrium state of two confusibilities, *i.e.*, the reciprocity of confusion, means that through this point the superiority of one over the other in some determined band region becomes lost and takes a turn toward inferiority in some region which follows. Next we must look at the direction of $C_{in}-C_{out}$, calling into question the so-called positive and negative trends of confusibility characteristics, because the trend has an important meaning in timbre interpretation. In phonemic confusion, by "positive trend" we mean *quality-formative* or *quality-gaining*, and by "negative trend" we mean *quality-ruinous* or *quality-losing*,

* Level-recovering process under cutting condition adopted for further experiments is used for this purpose.

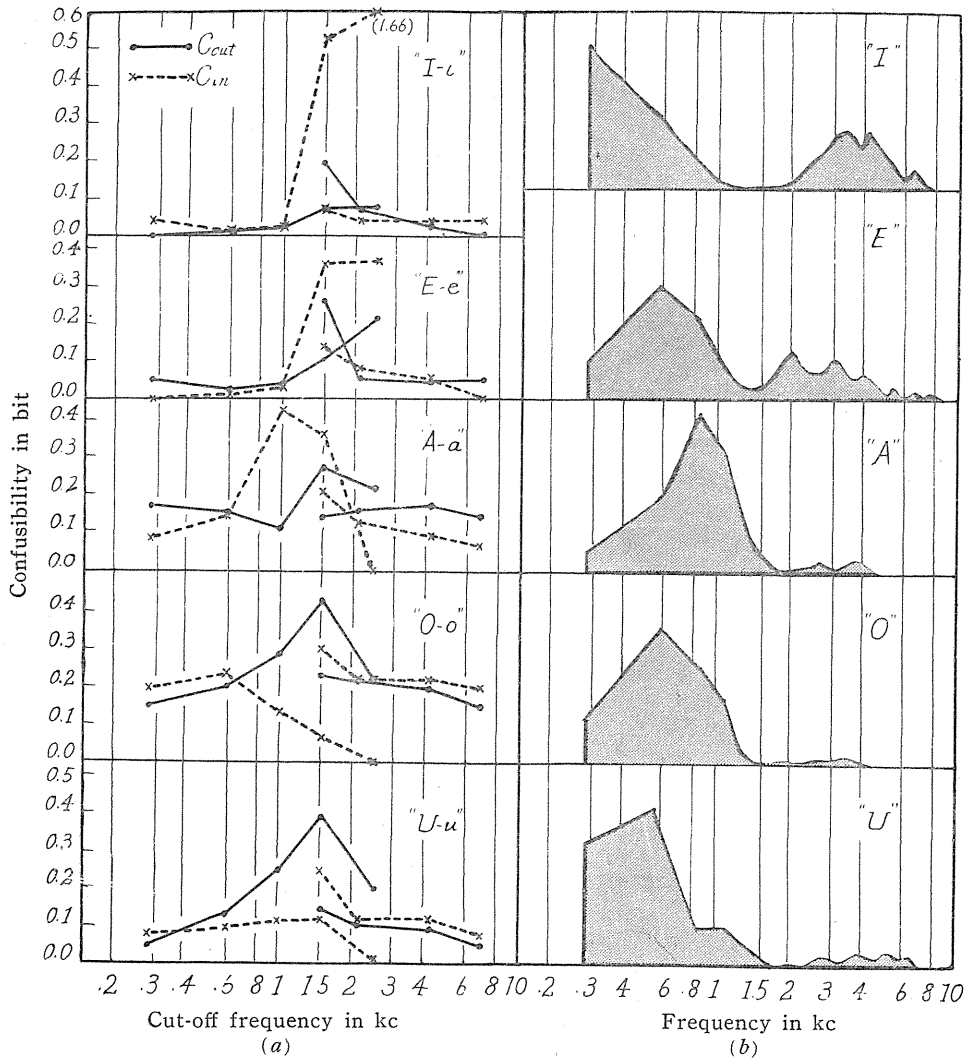

FIG. 1. Confusibility characteristics (*a*) and phonemic patterns (*b*) for individual phonemes.

both of which are affected by the band-cutting process under consideration.

Characteristics $C_{out}(I)$ *and* $C_{in}(i)$

HCD characteristics—The elemental outgoing confusibility characteristic $C_{out}(I)$ gradually increases with slightly increasing distortion. On passing downward below the frequency 2.0 kc, characteristic $C_{out}(I)$ increases abruptly, attaining considerable value at $f_c = 1.5$ kc. On the contrary, the incoming confusibility characteristic $C_{in}(i)$ does not change noticeably when the distortion varies.

LCD characteristics—Outgoing $C_{out}(I)$ does not change greatly, keeping relatively low in value even at extreme distortion of $f_c = 2.5$ kc, while incoming $C_{in}(i)$ remaining nearly at zero in the distortion range below 1 kc, undergoes an abrupt and severe increase when distortion passes upward beyond 1 kc and reaches the value of 1.66 bits at $f_c = 2.5$ kc which corresponds to about 72% of possible maxi-

mum value. Serious positive discrepancies between $C_{in}(i)$ and $C_{out}(I)$ in the band region of 1.5–2.5 kc reveal that by cutting down the region 1.5–2.5 kc in the LC direction, phonemes other than "I" show almost equally vigorous tendencies toward unique confusion with "I", directly and immediately suggesting there is a sudden violent increase of incoming to quality "i". In other words, for phoneme "I" this region 1.5–2.5 kc especially the band region surrounding frequency point of 2.5 kc* is most important as an indispensable band of *quality-formative* nature.

Characteristics $C_{out}(E)$ and $C_{in}(e)$

HCD characteristics—Both outgoing $C_{out}(E)$ and incoming $C_{in}(e)$ begin to increase only for the band downward below 2 kc and the discrepancies $C_{out}(E)$ and $C_{in}(e)$ become noticeable when the distortion reaches $f_c=1.5$ kc.

LCD characteristics—Outgoing $C_{out}(E)$ and incoming $C_{in}(e)$ run low in the region below 1 kc and on extending upward and surpassing this 1 kc point $C_{out}(E)$ increases linearly, and $C_{in}(e)$ increases more suddenly to a greater degree but is soon saturated, their discrepancies $C_{in}(e)$–$C_{out}(E)$ forming a maximum at $f_c=1.5$ kc. For phoneme "E" the band region surrounding frequency point 1.5 kc is most indispensable as quality-formative band.

Characteristics $C_{out}(A)$ and $C_{in}(a)$

HCD characteristics—Because of the poor pronunciation by some particular calling subject, both $C_{out}(A)$ and $C_{in}(a)$ have considerable values even at the referential condition of no-distortion. The curve of $C_{out}(A)$ does not show any increase during the whole course of this distortion, rather tending to drop as f_c approaches the middle band. On the contrary, the curve of $C_{in}(a)$ rises gradually accordingly as the distortion increases.

LCD characteristics—In this LC direction, a pair of two confusibility curves shows a distinctive mark with respect to its characteristic form, suggesting the peculiarity of this phoneme; the superiority of $C_{out}(A)$ over $C_{in}(a)$ in referential condition decreases with the increase of distortion, and on passing through the balancing state, *i.e.*, $C_{in}(a)=C_{out}(A)$ at 0.6 kc, this relation is reversed, and thus, after reaching the maximum discrepancy at the cut-off frequency 1 kc, this discrepancy begins to decrease, and only through the second changing of the mutual relation by crossing at 1.7 kc, is the superiority of $C_{out}(A)$ recovered. The quality-formative band for phoneme "A" is to be found in the frequency band at 1.0 kc.

Characteristics $C_{out}(O)$ and $C_{in}(o)$

HCD characteristics—The high starting points of both confusibility curves in referential condition must also be attributed to some mispronunciation of this phoneme by some caller. Generally speaking, confusions which occur most frequently and noticeably in the referential condition are restricted to the two phonemes "A" and "O". In outgoing and incoming confusibility curves $C_{out}(O)$ and $C_{in}(o)$ there are no distinctive features which deserve mentioning except the gradual growth of $C_{in}(o)$ and $C_{out}(O)$ and their departure which begins at 2 kc and seems to continue in the region below 1 kc.** To fully reveal the quality feature

* In further experiment, 2–3 kc band region is to be considered most indispensable for "I".

** That this tendency kept increasing even below 1 kc and that the departure of $C_{in}(o)$ from $C_{out}(O)$ is not maximum at the cut-off 1.5 kc, was clearly ascertained in further experiments which followed.

of "o", a further advancing of cutting conditions in this direction is necessary.

LCD characteristics—Two confusibility curves in this direction make a distinctive feature in striking contrast to those in HC direction. After $C_{in}(o)$ reaches a weak peak-point at $f_c=0.6$ kc, it gradually decreases as the band elimination is augmented. Meanwhile $C_{out}(O)$ increases one-sidedly thereby, finally experiencing its prominent maximum value at $f_c=1.5$ kc. The slight dominancy of incoming confusibility at 0.6 kc seems to suggest the meaningful position of phoneme quality and the existence of a comparable amount of outgoing confusibility in the same position, at the same time seeming to reveal some effect of mispronunciation of some caller.

Characteristics $C_{out}(U)$ and $C_{in}(u)$

HCD characteristic—Neither confusibility curve seems to have any distinctive feature. Inasmuch as the distortion range is restricted to the band experimented there is no essential difference in characteristics between phonemes "U" and "O". The incoming confusions at 1.5 kc point most frequently come from phoneme "I": $C_{in}(u)$ is furnished for the most part by $C_{out}(I)$. In the same manner, the major part of $C_{in}(a)$ at 1.5 kc is fed by $C_{out}(U)$.

LCD characteristics—Upward to the cutting point of 1.5 kc, the curve of $C_{in}(u)$ continues without any noticeable increase, and after passing through this point it descends to zero at $f_c=2.5$ kc, while the curve of $C_{out}(U)$ climbs steeply, reaching its maximum point at $f_c=1.5$ kc. In this experiment, during almost the whole course of the characteristics in LC direction, the tendency of negative trend, *viz.*, $C_{out}(U)>C_{in}(u)$, is observed. But this apparent tendency is a fallacy. A later experiment shows clearly that there is a positive trend at 1.5 kc.* This latter is rather reasonable.

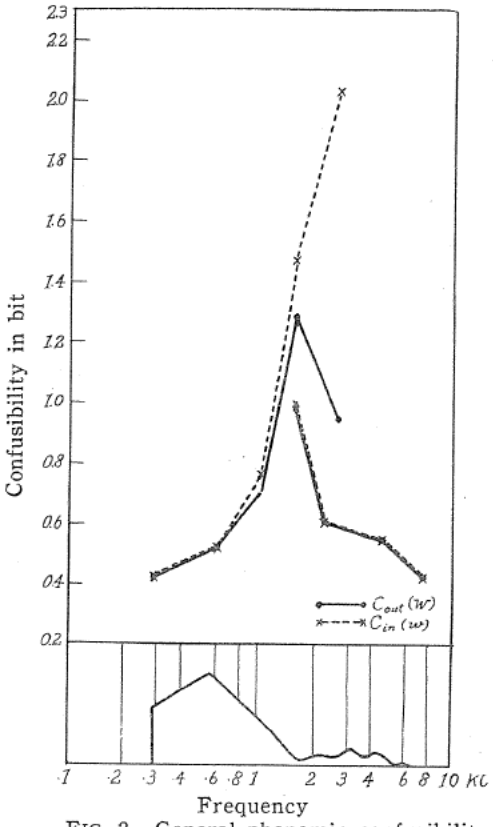

FIG. 2. General phonemic confusibility characteristics.

General Phonemic Confusibility Characteristics

Two kinds of confusibility characteristic are obtained by summing up respectively the elemental confusibilities for five vowels and those for five qualities. We designate $C_{in}(w)$ for ensemble incoming confusibility and $C_{out}(W)$ for ensemble outgoing confusibility, and we can acquire General Confusibility Characteristics for

* As was verified in this succeeding experiment, the positive trend is maximum at 1.5 kc. This is the main difference in confusion of "U" and "O" in which latter the discrepancy $C_{in}(o)-C_{out}(O)$ increases still more in the distortion downward below 1 kc.

the white vowel from the ensemble confusibility *vs.* distortion BED, as shown in Fig. 2. By these curves, we obtain an understanding of the most common confusion features of the vowels which in general never fail to reveal very important information on the localization of phoneme quality. Usually there is no noticeable discrepancy between two confusibility curves in HCD; on the contrary, in LCD there is a conspicuous departure of $C_{in}(w)$ from $C_{out}(W)$, *i.e.*, the incoming confusibility tends to exceed the outgoing confusibility accordingly as the distortion increases. It is noteworthy that the difference between two curves becomes astonishingly large at the point $f_c=2.5$ kc. The sudden falling of $C_{out}(W)$ and the simultaneous increasing tendency of $C_{in}(w)$ in this extreme region contribute the horn-shaped characteristics. This is an unmistakable feature of confusion in LCD. The fact that there is no noticeable inverse trend of confusion in HCD is due partly to the yet insufficient cutting-step and partly because of an inherent trend in HCD characteristics in contrast to LCD characteristics.

Some Remarks on Phonemic Confusion

Judged simply by the shape of white-phone and white-voice patterns, Japanese vocalic sounds uttered not in singing voice but in speaking voice are always characterized by timbre patterns formed in two parts, the lower part with a very high amplitude, and the upper part with a quite slight amplitude. When we consider the subject from the quality-response characteristics aspect, we obtain the understanding that usually the lower part of highest amplitude does not correspond to the quality of most dense concentration and also that the upper part of slight amplitude does not correspond to the quality of most dilute concentration. These facts must be brought into evidence more positively and more directly by resorting to the confusibility characteristics as shown in Fig. 2. The portion of the characteristics resembling the mouth of a horn, $C_{in}-C_{out}$ in the band region of 2–2.5 kc in LCD where the existence of the most violent confusion-tendency is caused by the biased incomings to quality "e" and especially to quality "i", corresponds exactly to the upper part of white-phone structure which is mainly determined by the upper part of formant structures of the phonemes "E" and "I". From this tendency observed in characteristics LCD, we can infer the following fact which holds good in the case of HCD: The lower part of white-phone structure which is chiefly composed of the lower formants of the vowels "U", "O" and "A" might be detectable from the phase of confusibility-characteristics in HCD. Solid evidence of this, however, is lacking now only because of insufficient cutting of HCD.*

It is probably worthy of note that the quality most clearly detected by confusibility characteristics in LCD only is based exclusively on the upper structure of white-pattern, and the quality most probably detectable by confusibility characteristics in HCD only is based on the lower structure. Reversely, the quality which is scarcely revealed by LCD characteristics is found chiefly in the lower structure of the pattern, and the quality which is scarcely revealed by HCD characteristics is found mainly in the upper structure, in spite of conditions of distortion appearing priorily in both cases. In short, what the confusion phenomena duly reveal by utilizing the one-directional distortion of band-eliminating nature, is that only the last remaining qualities are sharp enough to be detected in the last coming distortions.

* We were able to prove this in further experiments twice repeated.