Received 7 July 1970 9.3, 9.9

Sweep-Tone Measurements of Vocal-Tract Characteristics *

OSAMU FUJIMURA

University of Tokyo, Tokyo, Japan

JAN LINDQVIST

Royal Institute of Technology (KTH), Stockholm, Sweden

The vocal tract was excited transcutaneously at a point just above the glottis by an external sweep-tone signal, in order to measure its transfer characteristics acoustically as continuous frequency functions. An analysis-by-synthesis procedure derived reliable data of vowels, in particular of the formant bandwidths, for three male and three female normal subjects. It has been shown for the closed glottis condition that the first formant bandwidths are higher for close vowels (typically 70 Hz for male subjects) than for semi-open vowels (typically 35 Hz for male subjects). Stationary consonantal articulations including stops, nasals, and nasalized vowels also have been studied, as well as the effect of opening the glottis on the vocal-tract transfer characteristics. The stop articulations give rise to a first-formant frequency slightly below 200 Hz. This fact and the high dissipation of the first formant is explained by assuming nonrigidness of the surrounding wall. Characteristics of nasalized vowels and nasal murmurs are also discussed based on the data obtained in this experiment.

INTRODUCTION

Direct measurements of the vocal-tract transmission characteristics by use of sinusoidal waves as the excitation have been reported by van den Berg, who applied a sweeping pure-tone signal at the larynx of a hemilaryngectomized subject.[1] Fant adopted a similar method to a normal subject by exciting his vocal tract through the skin near the larynx. His sound source was a small moving-iron-type transducer, and by applying the vibrating diaphragm directly on the outer surface of the throat somewhat above the larynx, he recorded frequency response curves of the vocal tract for different vowels and some nasalized vowels. By visual inspections of the continuous response curves near formant peaks, he estimated the bandwidths of the formants under a closed-glottis condition, and derived a graph of the bandwidth values plotted against frequency over a frequency range up to above 4000 Hz.[2]

These sweep-tone techniques as a method for exploration of the acoustic characteristics of the vocal tract are characterized by two major advantages. The continuous-frequency response curves that can be obtained by these methods can reveal all details of the transfer characteristics, whereas the harmonic structure of natural voiced speech samples obscures these spectral details. More important is that these measurements exclude the unknown factor of the source spectrum, which we usually, in the case of analyses of natural utterances, cannot separate from the transfer function. The source characteristics change depending not only upon individual subjects but also upon individual utterances or even different portions within the same stretch of voicing, in partial correlation with articulatory changes.[3] Particularly in the low-frequency range, we have lacked precise knowledge both of the voice characteristics and of the vocal-tract characteristics because they could not be measured separately. Several specific points concerning the formant bandwidths of vowels, transfer characteristics of nasal consonants, and effects of nasalization and of subglottal coupling through an incomplete closure of the glottis, etc., are discussed in this paper.

The sweep-tone methods cited above, however, also suffer from serious disadvantages. When we apply the vibration signal transcutaneously as in Fant's experiment, we are not sure if the data are free from crucial errors caused by the nonflatness of the transfer characteristics through the tissue and the cartilages, even if we could estimate the characteristics of the vibrator. In the case of van den Berg's experiment, we do not have this unknown factor, but we scarcely have a choice of

subjects or sufficient control of the experimental environment, and we cannot expect sufficient data for drawing quantitative conclusions. Also, pathological cases may tend to be abnormal even when the physiological anomaly does not seem directly relevant to pertinent articulatory actions. In fact, the formant data obtained in his experiment show apparent inaccuracy even with respect to the number of formant peaks in a certain frequency range.

Fortunately, these difficulties can be largely circumvented by the experimental technique proposed here. In the case of the external excitation through the laryngeal wall, it is actually not necessary for us to assume a flat transfer characteristic of the body wall, in order to obtain accurate measurements of the frequency-response curves for various articulations. If we adopt as our starting hypothesis certain conclusions of the acoustic theory of speech production[4] and take a heuristic approach which is familiar to us in the analysis-by-synthesis experiments,[5] all we need here is a much weaker assumption that this transcutaneous transmission characteristic remains constant during a comparatively short period of an experimental session, for the set of different articulatory poses to be examined in the session. Furthermore, this assumption can to a large extent be tested in the experiments, as will be shown in this study.

I. EXPERIMENTAL PROCEDURES

A. Recording of the Vocal-Tract Response

Figure 1 illustrates the experimental setup for the recording of the response curves. The subject sits in an anechoic chamber and applies a vibrator to his throat. A high-quality moving-coil-type electromagnetic transducer (manufactured by the Goodman Corporation, England, type V-47) is employed as the vibrator, and a special brass case with an internal lining of lead is used for the transducer in order to minimize the direct sound radiation to the free field. A conically shaped piece of plastic is attached to the transducer as the mechanical output terminal. This plastic piece has a flat circular area of about 1 cm in diameter at the top. The circular top of the moving part is surrounded by the neck of the container leaving a small gap, and the former outstands the latter by an adjustable difference in level, so that both fit to skin with appropriate pressures.

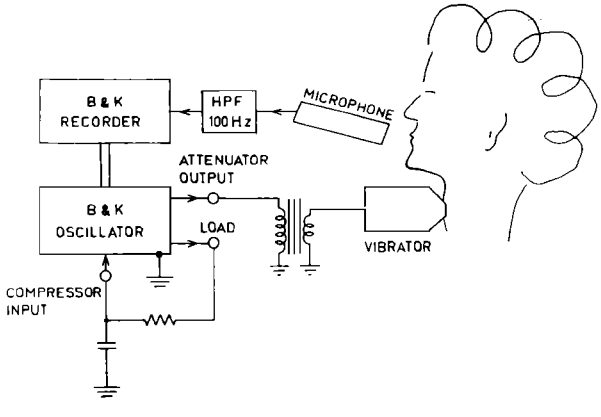

FIG. 1. Schematic diagram of the data-recording procedure.

A plaster model of the anterior part of the neck was cast for each subject in order to form a special clay adapter which was exactly fitted to the subject's neck. The clay thus filled in the space between the skin and the container of the vibrator over a comparatively large area. For some of the subjects this adapter was not made and, instead, a simple ring of model clay filled in the space between the neck and the container. This temporary setup was also used for selecting the optimal location of the vibrator before the clay adapter was made. The use of an adapter generally proved helpful in both eliminating the sound leakage and stabilizing the location of the vibrator on the throat.

In the sessions of data acquisition a condenser microphone picked up the sound signal at the mouth opening, normally making a distance of about 1 cm from the lips to the closest edge of the microphone. The tip of the microphone, about 23 mm in diameter, was held vertically so that the effect of reflection of sound back to the vocal tract was minimized. A manual switch was provided in order to let the subject choose either a buzz signal from a pulsetrain generator or a sinusoidal output of a beat-frequency oscillator as the input signal to the transducer. By using the buzz source for an artificial "voice" signal, the subject could listen to his "articulation," and when he judged he was ready for a run, he gave a cue to the experimenter by switching from the buzz to the sinusoidal tone. The experimenter outside the chamber monitored the microphone signal by an oscilloscope and, on observation of the switching, he started sweeping from 100 Hz. The subject held the intended articulation as constant as possible, normally with his glottis completely closed (see *infra*), until the tone had gone up to a high enough frequency. A sweeping from 100 to 5000 Hz took about 8.5 sec; this sweeping rate was selected by considering the stability of the articulatory pose and the sharpness of the vocal-tract resonances. Some preliminary tests were made on the stability and reproducibility of articulations under this condition, and this sweeping time proved appropriate both for the subjects and for the frequency resolution.

The signal level from the microphone was recorded by a high-speed pen recorder (Brüel & Kjær, type 3304) as a function of frequency. The recording paper was driven in a mechanical link with the frequency dial of the oscillator (Brüel & Kjær in combination with the recorder above) at a paper speed of 10 mm/sec. The "writing speed" of the pen recorder was selected at 500 dB/sec, using a 10-cm-wide paper roll for the full swing of 50 dB.

The accuracy of recording the frequency response with the equipment employed and under the sweeping

condition above was empirically tested by simulating various resonant characteristics of the vocal tract by use of passive electrical circuits. For recording of the vowel curves, and in fact for any other articulatory conditions, as we shall see, the sweeping rate employed proved slow enough for recording the rapid changes in the frequency response curves with very rare exceptions (see *infra*). For the most difficult cases of high-frequency sharp resonances, a simple resonance with a half-power bandwidth of 40 Hz at 2000 Hz was recorded without error in the peak height, and for a bandwidth of 15 Hz at the same frequency the recorded peak was erroneous by slightly over 1 dB. Even in the latter case the bandwidth could be estimated accurately with appropriate correction charts (under the assumption that the peak represented a simple resonance). For the first-formant region, the correction was not necessary for the bandwidth down to 20 Hz.

A high-frequency emphasis was given to the oscillator output (see Appendix A) in order to optimize the dynamic range over the entire frequency range of interest in consideration of both nonlinear distortions of the vibrator output at a high level (in low frequencies) and the electric and acoustic noise caused by the microphone and other sources. The frequency response of the vibrator adopted was reasonably smooth and good for the frequency range from 100 to 5000 Hz. The exact frequency characteristics varied considerably, depending on the loading condition. The mechanical impedance looking into the body wall from outside is not well known. Therefore an accurate estimate has not been made of the characteristics of the vibrator under the actual loading conditions. In our study of the vocal-tract response, an assessment of the effective over-all characteristics including the vibrator and the body wall suffices our purpose, and this has been derived for individual recording sessions of each subject (see *infra*). Some supplemental measurements were also made with simulated loading conditions, in addition to free-field measurements. From these measurements with models, it has been concluded that under the real loading condition the vibrator had a reasonable frequency response and the lowest resonance would be located slightly above 100 Hz and would be highly damped.

The transmission characteristics through the body wall depend heavily on the choice of the location of the vibrator in application. In some cases, we obtain very irregular curves as the frequency response, but it has been possible for our subjects to improve and obtain reasonably smooth curves after some trials. Some typical examples of the recorded curves are given in Fig. 2 together with the high-frequency preemphasis characteristic employed in the data-recording procedure. The curves pertain to different vowel articulations. Clearly defined formants are observed up to the fourth formant. An apparent antiresonance occurs at some frequency near 4000 Hz (cf., e.g., the curve for [ϕ]). The location of this antiresonance is dependent on the position of the

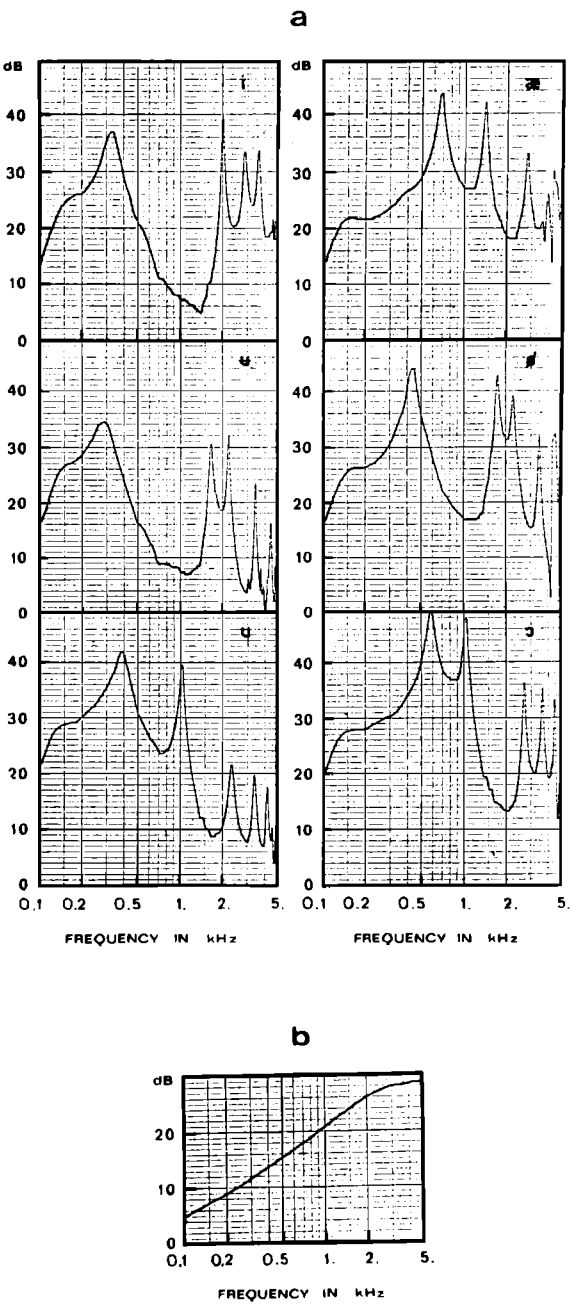

FIG. 2. (a) Response curves for different vowels (female subject). (b) The preemphasis characteristic employed in the data-recording process.

vibrator. It seems that, at least in typical cases, this reflects a resonance of the part of the vocal system that stretches from the glottis up to the location of the effective sound source in the vocal tract. A side cavity like the sinus morgagni also may be involved in this subsystem. It has been found that generally a high position of the vibrator gives rise to an antiresonance at an undesirably low frequency like 3000 Hz or even lower. When the vibrator is located too low and the effective source is created below the glottis, this mispositioning

can be discovered by an appreciable attenuation of the sound output that is observed when the subject closes his glottis.

B. Spectral Matching for Vowels and Derivation of Consonantal Characteristics

The data curves for vowels obtained in this manner have been matched against similarly recorded frequency-response curves of a vowel-synthesis circuit. The formant parameters of the circuit (see *infra*) were adjusted in a cut-and-try fashion until the circuit appeared to simulate the natural vocal-tract characteristics satisfactorily well.

The purpose of this data processing for vowels, as in any analysis-by-synthesis study, is twofold.[6] One is to corroborate our theory and assumptions on which the design of the simulation system is based, and the other is to collect the most accurately estimated descriptions of the samples of the vocal-tract transfer characteristics in terms of the parameters that are used in the simulating synthesizer. Bandwidth values for vowel articulations, under the closed-glottis condition, thus have been collected for three male and three female subjects. Having obtained an empirical corroboration about the validity of our assumptions for vowels, we now can expand the range of application of the same technique, in order to obtain detailed data about consonantal articulations, which are less understood compared with vowels.

For deriving the transfer characteristics of the vocal tract from the recorded data we assume the following two conditions: (1) Relevant acoustic properties of the body wall at the throat—or, more exactly, we assume that the transmission characteristics from the electrical signal at the input of the vibrator to the effective sound (volume velocity) source in the vocal tract remain constant regardless of the articulatory differences within the series of samples among which comparisons are made; (2) the vocal-tract transmission characteristics for vowels, when the glottis is closed, are effectively given by the framework of the acoustic theory of speech production.[4] Specifically, the frequency-response curve of the vocal tract is assumed to be completely simulated by a product of simple resonances, representing several formants, and the so-called higher-pole correction.

Thus we express the frequency function that we obtain in the form of recorded data as:

$$D(f)\approx W(f)\cdot T(f)\cdot R(f)\cdot P(f), \tag{1}$$

where $T(f)$ represents the amplitude of the transfer function of the vocal tract (for pure imaginary frequency $s=if$); $W(f)$ represents the *transfer characteristic* of the system consisting of the transducer and the body wall; $P(f)$ is the preemphasis function used in the recording session; and $R(f)$ is the radiation characteristic that transfers the volume velocity at the vocal-tract outlet to the sound pressure at the microphone. According to the acoustic theory of vowel production, $T(f)$ for vowels can be expressed as:

$$T(f)=H(f)\prod_{i=1}^{4}F_i(f), \tag{2}$$

where

$$F_i(f)=\left|\frac{f_i^2+(B_i/2)^2}{f_i^2-[f+j(B_i/2)]^2}\right|. \tag{3}$$

Here f_i and B_i stand for the frequency and the half-power bandwidth of the ith formant, respectively.

The function $R(f)$ can be assumed to be fixed apart from a numerical constant, and it is approximately proportional to f. We can absorb this function together with $P(f)$ into a constant correction function and rewrite Eq. 1 as

$$\begin{aligned} D(f)&=C(f)\cdot T(f),\\ C(f)&=W(f)\cdot R(f), \end{aligned} \tag{4}$$

where $T(f)$ is given by Eq. 2. The function $H(f)$ in Eq. 2 is an approximate function that represents the effective influence of the higher formants ($i=5, 6, \cdots$ in Eq. 3) upon the vocal-tract transfer characteristic within the frequency range of interest. We limit this pertinent frequency range to that from 100 to about 3000 Hz for conclusive data and for this we need to match the curves by adjusting f_i's at least up to the fourth (or fifth) formant, and B_i's up to the third formant.

If the subject is successful in maintaining a stationary articulation during the sweeping period, and if our assumptions made above are valid, then we should be able to match the recorded curves $D(f)$ for different vowel articulations by use of one fixed frequency function $C(f)$ in Eq. 4. By adjusting the values of parameters for $T(f)$, namely f_i's and B_i's and also a subsidiary parameter of $H(f)$, we here try to obtain good matches for all vowel samples within one session by using the same $C(f)$. The curve $C(f)$ is unknown and is derived from these matchings by a method of successive approximation. Usually, one set of preliminary matchings of sampled data for a new recording session is sufficient for deriving an appropriate function for $C(f)$.

For example, the curve for the vowel [ɔ] in Fig. 2 was simulated as shown in Fig. 3. First, the recorded curve is visually inspected and the formant frequencies a e approximately determined. By adjusting the resonant frequencies and bandwidths of an electric circuit which is represented by Eq. 2 (see Appendix A), then we try to simulate the resonant characteristics in the vicinities of the peaks for the lowest four formants. The frequency response curve for this circuit is recorded by the same pen recorder with the same sweeping conditions as in the acoustic data recording session (cf. Curve a in Fig. 3). The difference (in decibels) between the two responses is plotted as a function of frequency. Similar preliminary matchings are attempted for other recorded

curves in the same session. The difference functions thus derived are supposed to represent the function $C(f)$ in Eq. 4 plus the constant frequency function for the high-frequency preemphasis represented by Curve b in Fig. 2. Thus, if our assumptions for vowel samples are valid, we should be able to derive the same frequency function for the difference curves. Taking the average of the obtained difference curves, we derive a first approximation to $C(f)$. By addition (in decibels) to this function to $T(f)$, we can readjust the formant frequencies and bandwidths for each vowel sample in order to obtain

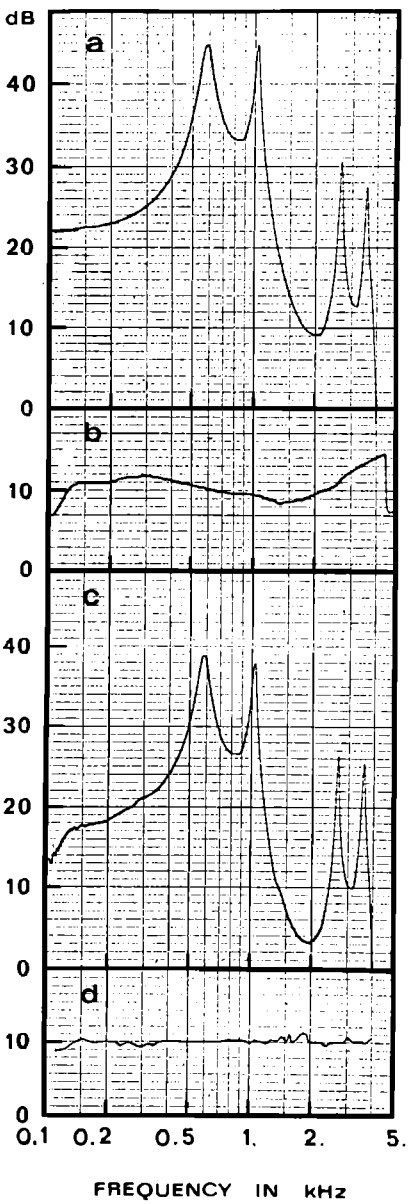

FIG. 3. Decomposition of the simulating frequency response curve for the vowel sample [ɔ] in Fig. 2. (a) Estimated vocal-tract response curve $T(f)$ as generated by the simulating circuit. (b) The correction curve $C(f)$. (c) The sum of Curves a and b, viz. the simulated response curve $C(f)\cdot T(f)$. (d) The difference curve $D(f)-C(f)\cdot T(f)$, showing the amount of mismatch.

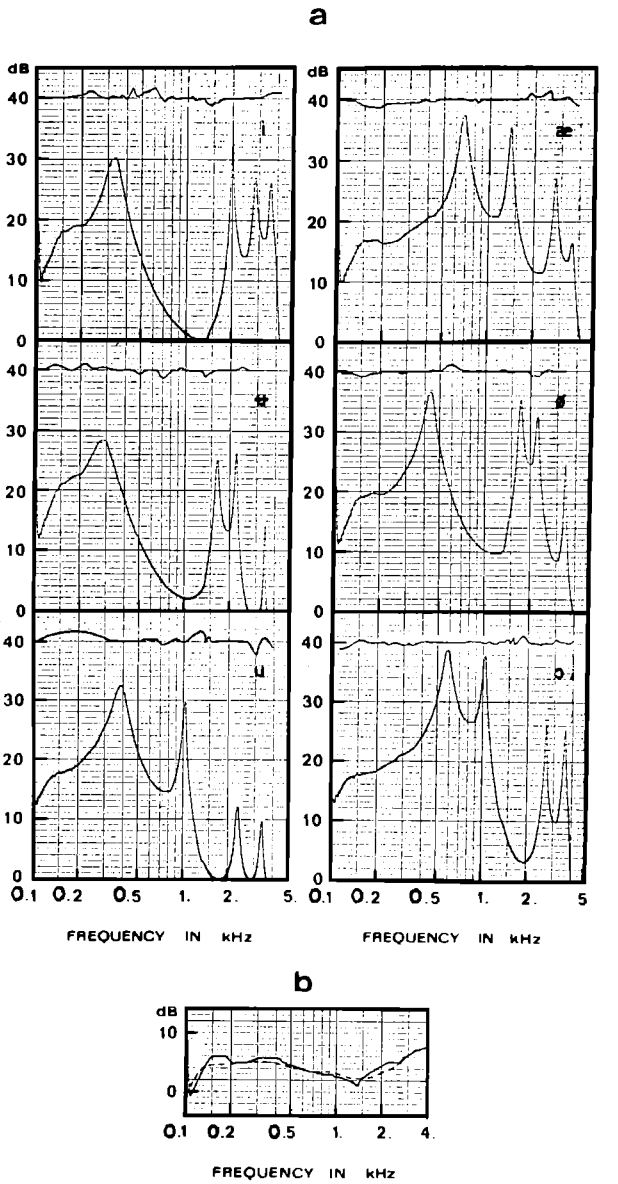

FIG. 4. (a) Synthesized curves for the vowel samples shown in Fig. 2 and the amount of mismatch (upper curves). (b) The "constant characteristic" used for matching of this series of samples.

better matches. Taking the average of the difference curves for these matchings, we can obtain a second approximation to $C(f)$. The resulting frequency function for $C(f)$ is shown in Fig. 3(b). By use of this curve as $C(f)$, we record the sum (in decibels) of Curves a and b (i.e., Curve c), by use of a special selfadjusting circuit and the same recorder (cf. Appendix A). Overlaying this synthesized function $C(f)-T(f)$ onto the recorded data curve $D(f)$, we can estimate the discrepancy and draw a difference curve for this sample, as shown in Fig. 3(d), Figure 4 shows samples of such simulated curves for several different vowel articulations by a female Swedish subject. The difference function is given in the upper portion of each curve. Slightly

different curves for the correction function $C(f)$ are used for front vowels and back vowels [cf., the solid and broken curves in Fig. 4 (b)]. The response data for these two groups of vowels were recorded in two subsessions of the same subject, separated by a short pause. The vibrator was reapplied after the pause with use of the same clay adapter.

The electrical circuit for simulating $T(f)$ consisted of a series connection of four simple-tuned inductance–resistance–capacitance (LRC) circuits, with buffer amplifiers in between, plus (also in series) an effective fifth-formant (another well-damped LRC) circuit with a variable resonant frequency, and the higher pole correction circuit with a few selective parameter values.[4,7]

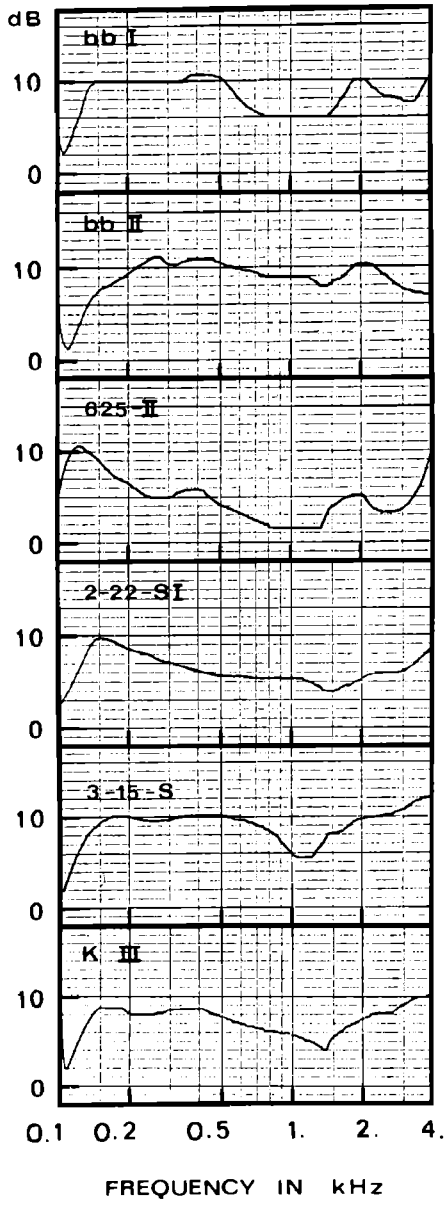

FIG. 5. "Correction functions $C(f)$'s" for different sessions of different subjects.

The selections of the last two subsidiary functions were meant for a good match in the high-frequency region, in particular the third-formant region, and had little influence on the frequency region up to 2000 Hz.

The function $C(f)$ is an empirical semivariable curve, which, in general, varies from one data-collection session to another. In order to generate such *ad hoc* curves in simulation, one would consider the use of an interactive digital computer. In some practical considerations for this work at the time, we devised instead a special analog-type function generator which is suitable for use in combination with our oscillator–recorder apparatus. A description of this device is given in the Appendix A. In short, the device senses the width of the unpainted area on an optical program disk and lets the output level of the oscillator adjust itself instantaneously and servomechanically in accordance to this specification of the program pattern.

The results of matchings in our experiment were acceptably good for selected subjects (see *infra*). For very close vowels, however, the signal level was sometimes too low to outstand the noise level in the valleys of the curves (see also the results for stop consonants, *infra*). Another cause of difficulty was that a comparison of drastically different conditions of articulation, such as close vowels versus extremely open vowels or front vowels versus back vowels, made it difficult to maintain the same condition near the larynx, and the function $C(f)$ did not remain constant as perfectly as in comparison between relatively similar articulations.

In deriving the transfer functions of consonantal articulations, we compare the data curve with a not too different vowel articulation selected as the reference. A safe choice for a reference vowel is the neutral vowel [ɜ], unless a vowel context is specified for the consonantal articulation. From matchings of the vowel samples distributed among consonantal articulations within a session, we derive the curve $C(f)$ and at the same time check the stability and reproducibility of the articulations. Then their correction function $C(f)$ is used to derive the unknown $T(f)$ from the recorded $D(f)$ for the consonant in question.

In principle, the correction function $C(f)$ may be irregular and complex without invalidating the purpose of our study, as long as it remains unchanged for different samples within a session. Our equipment is in fact capable of handling these cases. In practice, however, it is obviously preferable to have curves for $C(f)$ that are as smooth as possible, since it simplifies the process of successive approximation and avoids any possible ambiguity. Actual curves for $C(f)$ obtained for our subjects are exemplified in Fig. 5. The sessions are identified by the mark at the upper left corner of each curve: bb-I and bb-II refer to two sessions of a female subject, 625-II represents one of many sessions of a male subject, 2-22-S-I and 3-15-S refer to two sessions of another male subject, and finally K-III refers to a session of a different female subject. Each of these

curves was recorded by the Brüel & Kjær pen recorder with the program disk while the synthesis circuit for $T(f)$ was bypassed. The scale for the ordinate is given an arbitrary zero point for each curve, and a constant multiplication factor, i.e., any vertical translation of the curve, is disregarded in matching the data curves unless otherwise stated in this paper (see *infra*).

II. EXPERIMENTAL RESULTS

A. Vowels

Approximately 250 articulatory samples have been analyzed by the process described above for three male and three female subjects. Consistent results have been obtained, and the assumptions described above for vowel articulations are largely supported by the present data. With appropriate selections of the formant frequencies and bandwidths, the differences between the observed and simulated curves were made for the most part to be less than ± 1 dB by the use of a fixed correction function for typically about 10 samples within each recording session. Occasional deviations have been observed, particularly when the signal level was very low in frequency regions between formants. The dynamic range in better cases was found to be more than 35 dB (cf. Fig. 4, for example).

Accurate estimations of formant frequencies and bandwidths have been obtained in this way. For formant frequencies, this study does not attempt to collect data of standard articulations of any language, but articulations of various phonetic values have been observed. For example, an interesting case of an extremely low second formant at 427 Hz for [u] was found for a tall Swedish male subject. The first formant for this vowel articulation was located at 290 Hz. The highest value for the second formant frequency measured for the same subject in this experiment was 2100 Hz for an articulation of [i]. The frequency range of the second formant for this subject is thus well above 2 oct.

Another interesting case, which was discovered by this accurate estimation of the vocal-tract transfer function, is illustrated in Fig. 6. These data pertain to a tense and strongly rounded vowel [y] by the same subject. As seen in the curve, the second and the third formants are located very close to each other and this extreme proximity of the formants results in an apparent single peak even in this continuous response curve. The analysis-by-synthesis readily reveals that this peak actually consists of two formants, but an accurate estimation of the frequencies and bandwidths in such an extraordinary case is rather difficult partly because of the limitation in the frequency–time resolution of our recording system. It is highly probable, however, that at least one of the two formants is comparatively heavily damped.

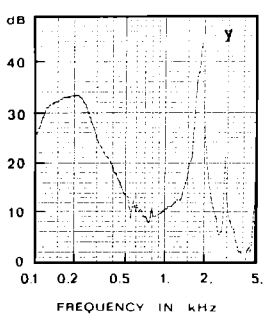

FIG. 6. An example of the recorded response curve for a front rounded vowel. Note the merging of the second and the third formants into one apparent peak.

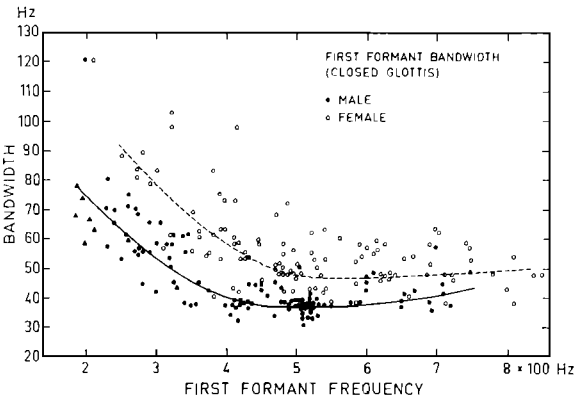

FIG. 7. Bandwidth values for the first formant plotted against the formant frequency. Each closed circle represents a vowel sample of one of three male subjects, and an open circle represents a sample of one of three female subjects. Representative values are estimated by visual inspection of the plots, and curves are drawn for male and female subjects separately. Bandwidth values for articulations with bilabial closures by a male subject are also added in this graph (closed triangles).

B. Formant Bandwidths

The data of formant bandwidth values obtained in this experiment are of particular interest because accurate measurements are difficult in analyses of vowel samples produced by speech utterances. There have been reports on the formant bandwidths by several investigators using different measuring techniques, including a direct acoustic method by the use of an impulse excitation, but some appreciable disagreements between authors have been noted.[8,9] For cases of lower first-formant frequencies, the estimation of the bandwidth value by spectral examination is particularly difficult. The impulse-excitation method, as reported by House and Stevens, also is not applicable for the close vowels. Apparently, however, it has been generally accepted that the formant bandwidth value can be approximated by either a monotonically increasing function of the formant frequency or by a constant regardless of frequency in the first formant region,[10] but this experiment has clearly shown that this is far from the truth in the lowest frequency range.

The data obtained in this study are compiled in Fig. 7 for the first-formant bandwidth values of all the vowel samples analyzed in this study. Data for three male subjects are plotted with closed circles, each circle representing a vowel articulation. Similar plots are

given for three female subjects by open circles. An apparent difference is seen between the male and female subjects in distributions of both the formant frequencies and the bandwidths. Even though there is appreciable scattering of data, it is clear that the bandwidth values are generally high for low formant frequencies, i.e., for close vowels such as [i], [y], [u′], and [u]. (Here, and below, [u′] represents IPA barred u.) Occasional scattering of data in the higher range of bandwidths may have been due to failure in completely closing either the glottis or the velum, or due to difference in some unknown conditions near the larynx. Slight movements of the articulators also would result in apparent broadening of the peak. Average curves for male and female subjects are given in the figure by visual inspection of the data, ignoring apparent sporadic deviations toward higher values. It must be noted that these data pertain to the closed-glottis conditions and that the bandwidth values in the usual voicing conditions for vowel articulations are expected to be slightly higher, owing to the dissipation of sound energy in the subglottal system.[11]

The high dissipation of the low first formant is explained by considering an appreciable participation of the surrounding soft tissues in the vibration of the acoustic system. In the low-frequency region, the wall surrounding the cavity acts as a highly dissipative mass that can be considered as connected in parallel to the radiation load at the mouth outlet. The contribution of this wall vibration to the bandwidth of the formant is greater when the impedance looking toward the outside through the mouth opening is higher, which is the case for closer vowels. The result is that we observe higher values of bandwidth for lower frequencies of our first formant. In these cases of comparatively low-frequency sounds, other sources of dissipation, e.g., the heat-conduction loss and the frictional loss of the wall, are estimated not to be appreciable compared with the above-mentioned cavity wall, and the radiation loss at the orifice is also negligible. When the first formant is high in frequency, the radiation loss is expected to contribute several hertz to the bandwidth.[12] This would probably explain a slight rise of the bandwidth toward the right of the figure.

The marked difference between the data for male subjects and female subjects indicates that the mass of the surrounding wall is significantly smaller for females than for males. The portion of the wall of the vocal tract that effectively yields to the low-frequency acoustic pressure may be a particular area that has a comparatively small mass and a high compliance, most probably the glottal and supraglottal laryngeal areas. The actual system, in terms of its electrical analog (direct analog), has distributed branches of inductance with high dissipation connected to each other through compliances, and these would constitute a distributed parallel shunt circuit for the vocal-tract transmission network. In the lower-frequency range, the effect of this shunting network can be represented by a single lumped-constant inductance (and a large series capacitance) with a high damping factor, whereas the vocal tract with its radiation load can be approximated as a Helmholtz resonator.

This model was proposed by Fant and Sonesson[13] in their study of degradation of vowel qualities in high ambient pressures. They observed that formant frequencies, particularly of the first formant, became considerably higher when vowel samples were pronounced in a high-pressure tank. Their data pertained to an atmospheric pressure as high as six. Consequently the effect of sound transmission through the tissue structure surrounding the vocal tract was much more pronounced, owing to less abrupt change in characteristic impedance at the boundary between air and the flesh.[14]

In this connection, it is of particular interest to observe the lowest resonance of the vocal tract when the mouth outlet is completely closed. Owing to the shunting element, the first-formant frequency, with decrease of the mouth opening towards a complete closure, does not tend to zero but tends rather to a certain finite value. As discussed in a later section, we can measure the resonant frequency and also approximately estimate the bandwidth for this closure condition. The values for bilabial stops are represented in Fig. 7 by closed triangles (for male subjects). As expected from the shunting model, the data for vowels can be continuously extended to these stop conditions without showing any singularity.

Similar data have been obtained for the second-formant bandwidth values (see Fig. 8). Scattering of data is observed to a larger extent than in the case of the first formant, and the dependence of the bandwidth on the formant frequency is less apparent. A minimum value of the second-formant bandwidth appears to be independent from the formant frequency, and it is about 35 Hz for male subjects and 40 Hz for female subjects. The difference of the bandwidth values between male and female also seems to be smaller in the case of the second formant than for the first formant.

There are some correlations of the bandwidth values with some particular articulatory features. The second-formant bandwidth appears to be generally higher for more-open vowels than for closer vowels. This is clear in the case of front vowels. Thus, if we compare [ɛ], [æ], and [a] (open circles in Fig. 8) with [i], [e], [ɸ], etc. (filled circles), we can see that the former are clearly more distributed towards upper portions of the figure, compared with the plots for the latter at the comparative frequencies. It is also seen that on the average, and particularly for more-open vowels, the bandwidth value tends to rise as the formant frequency increases. Thus, the open back vowels [ɔ] and [ɑ] have a typical value of 40–45 Hz for male and 50–60 Hz for female subjects, and open or semiopen front vowels [ɛ], [æ], and [a] have values scattered mainly in the range 50–70 Hz for male and 50–80 Hz for female subjects. The second-formant bandwidth values for close or semi-

TABLE I. Typical bandwidths (in hertz) for the third formant. Values within parentheses are not reliable.

Vowel	i	y	u'	e	ø	ɛ	a	ɑ	o	u	ɜ
Male	110	75	40	115	45	135	120	90	(65)	(50)	60
Female	95	110	70	150	85	125	105	120	55	125	75

close vowels, on the other hand, are seen to be more concentrated to lower values, typically 35–40 Hz for males and 40–60 Hz for females.[15] When there is scattering of data for the vowels of similar phonetic values, the lower values may be considered more reliable as data, because, as noted above, any deviation from a well-defined vowel condition, e.g., slight relaxation of either the glottal or velar closure or momentary fluctuations of the articulatory gesture, would cause an apparent increase of the bandwidth values. Any error in the other direction is much more unlikely.

Apparently higher values are observed for the Swedish vowel [u'], which has a close front tongue position and a marked labial constriction without protrusion. This point, which remains inexplicable at present, is particularly true in male samples and is true for both of the two Swedish subjects who contributed to the data plots.

The generally higher values of bandwidth for more-open vowels and their increase with the formant frequency can be explained by appreciable contribution of the radiation loss. According to a calculation by Fant,[16] the contribution of the radiation loss to the formant bandwidth for a neutral vowel (uniform tube) is 3.4, 43.6, 133, and 225 Hz for the first-, second-, third-, and fourth-formants, respectively. For a twin-tube model which simulates the vowel [I], the contributions are given as 0.3 Hz for F_1 (255 Hz), 5.5 Hz for F_2 (2045 Hz), 40.1 Hz for F_3 (2664 Hz), and 16.0 Hz for F_4 (4290 Hz). Thus, even though the radiation loss contributed very little to the first-formant bandwidth, its contribution can be quite significant for higher formants depending on the phonetic characteristics. Qualitatively, compact vowels with open lips have gradual widening of the vocal tract toward the free field, and the good acoustic impedance matching of the tract to the open field causes high dissipation of acoustic energy.

It may be noted that all vowels covering a wide range of articulatory conditions have shown most of representative data of the second-formant bandwidth values within a range up to 70 Hz for males and 80 Hz for females. The smaller difference between male and female data in the case of the second formant can be explained by consideration of more appreciable contributions from the radiation loss and the air friction loss, both of which are less dependent on the physiological properties of the human body.

Estimates of the bandwidths were also obtained for the third formant by the matching but the accuracy was not as high as the estimates for the lowest two formants, partly because of the effects of the higher formants and also because of a possible influence of a high-frequency zero. Typical values for various vowels are listed in Table I.

C. Open-Glottis Conditions

The vowel data above pertain to the closed-glottis condition, which was assumed by the subjects.[17] Bandwidths for natural speech data would be expected to be somewhat larger because of the partially open glottis during the phonation. It is hard to measure the dependence of the bandwidth on various degrees of opening of the glottis. Wide-open conditions for the glottis have been tried by some of our subjects and examples are given in Fig. 9. Curve a was obtained under the regular conditions of the glottis and the source location. The articulation is for the neutral vowel [ɜ]. In Curve b, the glottis was kept wide open. A significantly higher damping of the first formant is observed. The coupling of the subglottal system apparently causes some complexity of the curve. Thus, for example, a zero is observed at about 700 Hz accompanied by a pole just above it. Curve c was obtained also for the open-glottis condition. In this case, however, the vibrator was placed in a lower position. The effective location of the source

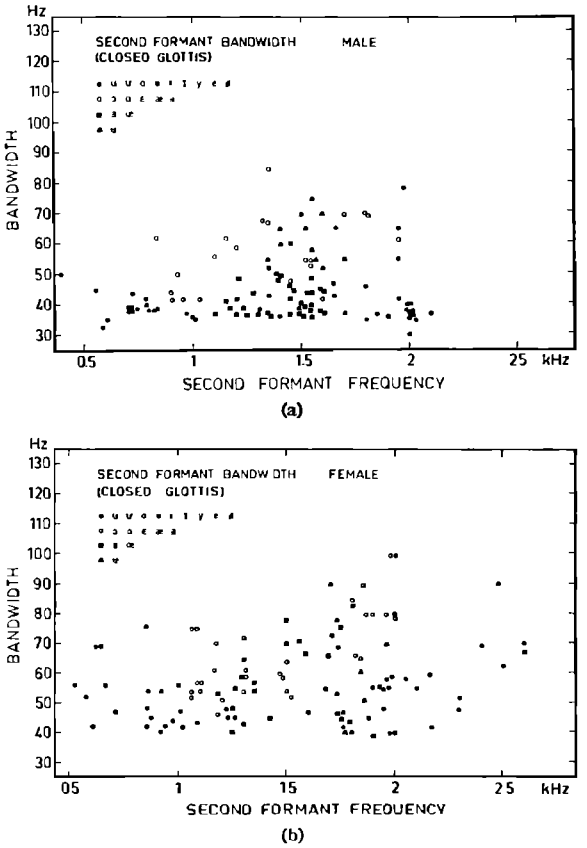

FIG. 8. Bandwidth values for the second formant for three male subjects (a) and three female subjects (b). Articulations of different vowels are categorized and represented by different marks, as shown in the figure.

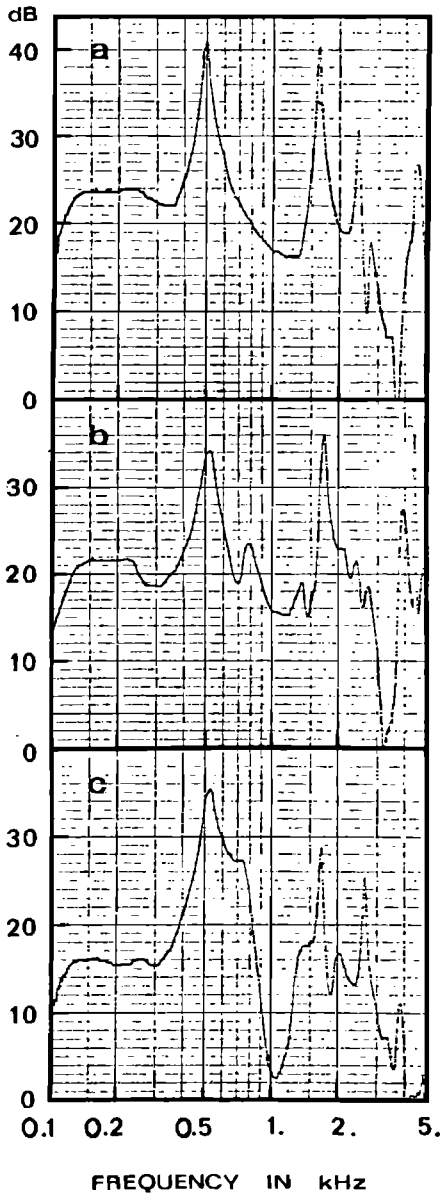

FIG. 9. Frequency response curves for a neutral vowel recorded under different conditions: (a) Regular response curve with the closed glottis (supraglottal source). (b) Open glottis, supraglottal source. (c) Open glottis, subglottal source.

was just below the glottis. This was ascertained by observing that the sound output was markedly reduced by closing the glottis. In order to see the reproducibility of the glottal condition, which was not felt obviously for the subject, the same trial was repeated and very similar curves were obtained for several trials.[18]

The subglottal system is coupled to the vocal tract through the orifice of the glottis in both of Cases b and c. The poles should consequently be the same. The zeros are different, and a very effective zero can be observed near 1000 Hz in the case of subglottal excitation. Evidently, this is caused by a tuning of the subglottal system (so-called antiresonance). Some more data are discussed later in connection with nasalization of vowels.

By a graphical method similar to that one discussed in connection with nasalization of vowels, we can semi-quantitatively predict from the data above that the impedance singularity and zero will be found near 800 and 1000 Hz, respectively, if we measure the impedance looking into the trachea from the glottis (without the glottal constriction). This compares very well with preliminary results of direct acoustic measurement with a laryngectomized subject.[19]

D. Stops

As an extreme case of close vowels, we can apply our method to a study of stop consonants where we have a stationary and complete closure of the vocal tract at the point of articulation. In the case of natural utterances, it is known that the spectrogram of the outcoming sound shows only a weak low-frequency component, which often is called a buzz bar.[20] The structures in higher-frequency regions are usually not observed in spectrograms. It should be kept in mind, however, that we cannot assume the source characteristic to be the same as that for vowels, because in this particular case of voicing the loading condition for the vocal vibration is extremely different. Some investigators apparently assume for voiced consonants even a motor control that is substantially different from that for vowels.[21] With our external source, there is no problem about the source characteristics. The present study has revealed that the peak level difference between the first and the second formants is not necessarily grossly greater in comparison

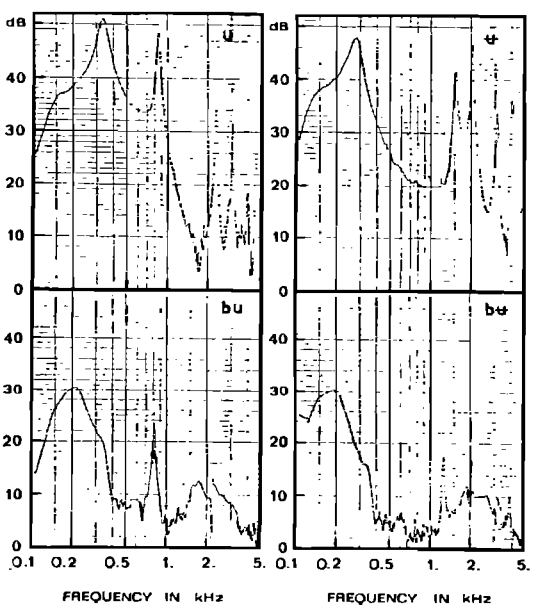

FIG. 10. Comparisons between vowels with a very small labial opening and a bilabial stop consonant with approximately the same tongue articulation. The correction function for this set of data is given in Fig. 5 (2-22-SI).

TABLE II. The first-formant frequency (in hertz) and bandwidth (in hertz) by a male subject (average of three samples).

	u	u′	b_u	$b_{u'}$
F_1	295	266	204	189
B_1	56	57	62	73

to that for vowels, the high formant being sometimes quite sharp.

Figure 10 shows two samples of recorded curves of the bilabial stop in comparison with vowels that have approximately the same articulations as the corresponding bilabial stops except for the small labial opening for the latter. All the four data here were recorded in the same recording session under the same conditions including the amplifier gains. Three similar samples have been obtained for each of the bilabial stops, and the first-formant bandwidths values for the six samples were estimated by using the correction function that was obtained from matchings of the vowel curves. These data were included in Fig. 7 (triangles). The Swedish vowels [u] and [u′] have extremely narrow lip openings but only the former has a marked labial protrusion. By closing the lips completely to make [b_u], the second formant shifts downward very little in frequency. In the case of [$b_{u'}$] compared with [u′], the third formant remains at the same position whereas the second formant shifts downward by about 250 Hz. The third formant of [b_u] is obscured by some interference effects in the low signal level. The apparent heavy damping of the stops (apart from the visual artifact due to the logarithmic frequency scale) is partially due to the contribution of the damped resonance of the vibrator to the recorded curves (in the form of the correction function). The representative first-formant frequencies and bandwidths are tabulated in Table II. The first-formant level is higher for the vowels than for the stops, as seen in Fig. 10. If we assumed a vowel production model for the stop articulations by using the formant data in Table II, we would predict by simple calculation that the first-formant levels of the stops in comparison with the corresponding vowels (of Fig. 10) would be on average about 12 dB higher than the actual values. This is, of course, not surprising because the formant frequency for a stop is not determined by an actual labial opening through which the sound is emitted.

Similar level differences are observed for the second formant. The level difference between the vowel and the stop, as observed in Fig. 10, is on average about 25 dB. From this fact we may conclude that the higher formants would not be observable in the spectra of natural voiced stops even if the buzz source waveform were maintained as in the case of vowels.

Figure 11 compares different palatal–velar consonants, i.e., [g]'s with different vowel coarticulations. Occasionally, particularly when there is an appreciable lip rounding, we can see a peak that apparently shows the resonance of the mouth cavity in front of the tongue closure (compare [u] with [w], for example). The low-level signals in these stop articulations suffer from noise, and some peculiarities appear due to interferences of the sound transmitted through different portions of the subject's body and through leakage near the larynx.

For estimations of formant data of the bilabial stop consonant, it is possible to obtain clearer curves by inserting the tip of a thin probe microphone inside the lip closure.[18]

E. Nasalization

The effects of nasalization of vowels have been discussed both experimentally and theoretically by several authors.[4,22–26] According to a model proposed by one of the present authors, the transfer functions of nasalized

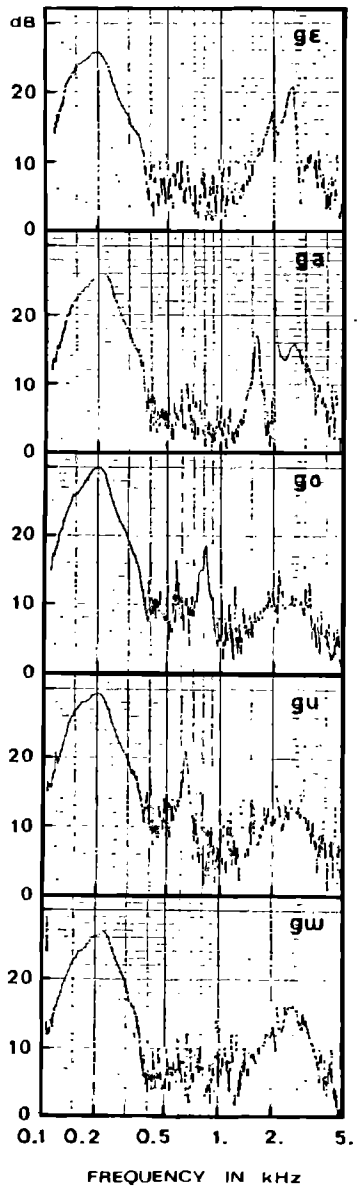

FIG. 11. Comparisons between [g]'s with different vowel coarticulations.

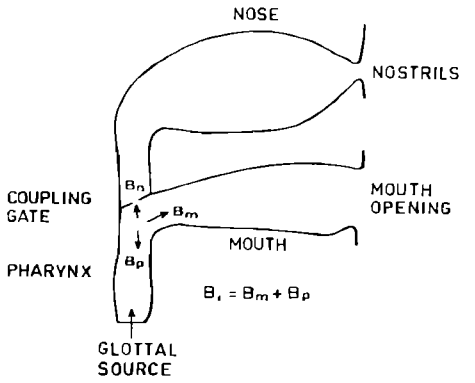

FIG. 12. Simplified model of the acoustic system for production of nasalized vowels. The degree of nasalization is varied by changing the opening area of the coupling gate.

vowels will be matched in the lower frequency region by the present analysis-by-synthesis method with a set of two formants and one antiformant in place of the first formant of the oral vowel.

This model of nasalization assumes that a comparatively small opening at the velum couples the nasal tract to the main vocal tract through a variable gate without changing the oral cavity area function. By considering the driving point admittances B_p, B_n, and B_m at the coupling point looking into the pharyngeal, the nasal, and the oral passages, respectively (see Fig. 12), and assuming a lossless system for consideration of the formant–antiformant configuration for the oral output, a somewhat detailed qualitative acoustic characteristic of slightly nasalized vowels can be theoretically predicted.[24]

When we decrease the degree of coupling continuously from a finite coupling area to zero, each antiformant approaches one of the formants of the combined vocal-tract system and the pair will finally be annihilated. Since we assume that the change in the nasal-tract geometry takes place only in its innermost portion, it can be shown that as the coupling is decreased, each pole of B_n, that is, the antiformant of the nasalized vowel, moves downwards until it reaches a zero of B_n. This set of pole and zero of B_n can be considered to be associated with each other. The zeros remain at the same positions as the degree of nasalization is varied. This fact and the knowledge that both B_n and $B_i = B_p + B_m$ are monotonically increasing functions of frequency lead to several general rules concerning the behavior of the poles and zeros of the transfer function by use of a graphical method (see Fig. 13).[4,24,26] Some of the principal conclusions which are relevant here are:

(1) Nasalized vowels, in general, have two kinds of formants which we may call "nasal formants" and "shifted-oral formants." Each nasal formant is paired with an antiformant, i.e., an antiresonance observed at the mouth opening.[27] The pair of nasal formant and antiformant approach each other, and finally an annihilation results when the coupling of the nasal tract to the vocal tract reaches zero. When a vowel is heavily nasalized, however, the antiformant can be closer to one of the other formants than to its own mate.

(2) The lowest formant of the coupled system can be either a nasal formant or a shifted oral formant. If the first formant of the nonnasalized vocal tract is of higher frequency than a certain critical frequency, the lowest formant is a nasal formant; and if the first formant is of low frequency, the lowest formant is an oral formant. Physically, the critical frequency is the lowest resonant frequency of the nasal-tract proper when it is closed at the coupling end.[28]

(3) All formants of a nasalized vowel shift monotonically upwards as the degree of coupling increases.

(4) A nasal formant always originates from the characteristic frequency of the nasal tract, and a shifted oral formant from one of the formants of the nonnasalized vowel with the same vocal-tract configuration.

(5) When the degree of nasalization is increased, the formants will shift upwards in the frequency domain in various ways, but no formant can meet or overtake another.

(6) This invariance of order does not hold for the formant–antiformant set. There can, therefore, be an annihilation of the antiformant with a nonconjugate formant for a certain finite degree of nasalization.[29]

(7) The location of the lowest formant of a nasalized vowel is always in between the lowest characteristic frequency of the nasal tract and the first-formant frequency of the nonnasalized vowel. Consequently, the range of frequencies in which this lowest formant of nasalized vowels can be located is much more limited compared with the range of the first formant for different nonnasalized vowels.

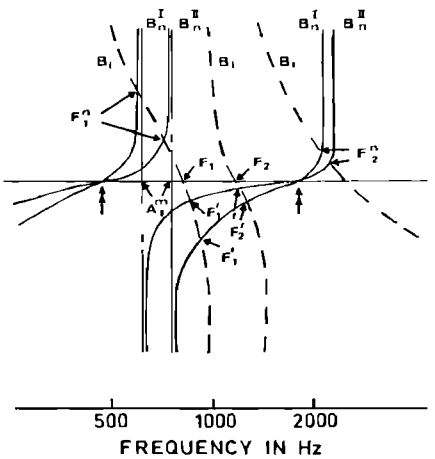

FIG. 13. Sketch of the driving-point susceptance of the nasal tract B_n with two degrees of nasalization B_n^{I} and B_n^{II} and the internal susceptance of the vocal tract B_i (inverted). The vocal-tract configuration is appropriate for the vowel [a]. The zeros and poles of these functions in this example were estimated from unpublished data of analog experiments by House and Stevens. F_1, F_2 represent first and second oral formants; F_1', F_2', shifted oral formants; F_1^n, F_2^n, nasal formants; and A_1^m, the antiformant to be observed in the output spectrum at the mouth. Double arrows indicate the characteristic frequencies of the nasal tract.

In one recording session a subject articulated alternately nonnasalized and nasalized articulations of a neutral vowel [ɜ]. The session contained seven samples of the oral vowel and six samples of the nasalized vowel. The microphone was placed, as in the case of vowels, quite close to the lower lip.

A typical pair of consecutive samples are shown in Figs. 14(a) and 14(b). Curve a represents a regular oral articulation. By an attempt at nasalizing the same vowel without changing the tongue articulation, Curve b (solid line) was obtained. The two curves, as well as other samples in the same session, were matched with a correction function derived from the matchings of the nonnasalized vowel samples, and satisfactory matches were obtained for the pair of samples with the following locations of the poles and the zero (in hertz, bandwidths in parentheses) (see Table III).

Curve a was matched without any appreciable deviation. Curve b was matched perfectly well except in the frequency range from 700 to 1700 Hz. The mismatch is indicated by the overlaid broken line which is identical with part of Curve c. Curve c shows the synthesized curve with the settings of the parameters given above. The correction function for this session is shown in Fig. 5 (Curve 625-II).

When we subtract the correction function from Curve c we obtain Curve d. This can be considered as the transfer function of the vocal tract without including the radiation transfer characteristic.

A perfectly good match, as seen in Curve b, assures us that the formant–antiformant–formant (Fn–AF–F1′, see *supra*) structure is satisfactory for describing the nasalized vowel in the lower-frequency region. Nasalization of different vowels has been studied similarly, and comparable results have been obtained.[18] The match in the second-formant region in Fig. 14(b) seems not bad when the damping factor is appropriately chosen, but this is not necessarily true for other tongue articulations. Theoretically, also, it is clear that a spectral complexity similar to that in the first-formant region must exist for the higher-frequency regions in the case of nasalized vowels in general.

Figure 15 illustrates comparisons of recorded curves for another subject, showing effects of nasalization of a series of back vowels. The set of graphs also compares the effects of slightly opening the glottis. The details of spectral characteristics as shown here are hard to obtain from analyses of regular voiced samples.

In the case of the close vowel [u], we find that the first peak of the nasalized vowel is slightly lower than the original first formant and above it there is a pair of a valley and a peak. According to the theoretical consideration given above, we interpret the first peak as appearance of the nasal formant, the valley as the antiformant reflecting the resonance of the nasal shunt, and the second peak (which is influenced appreciably in its shape by the consecutive antiformant) as reflection of the shifted first formant, which is located slightly above its original (nonnasalized) frequency. A similar complexity is observed in the third-formant region. The difference in frequency between the original first formant F_1 and the nasal formant $F_1{}''$ is very small. In consideration of the theoretical conclusion above that

TABLE III. Pole and zero frequencies for non-nasalized vowel [ɜ] and a nasalized vowel with approximately the same tongue articulation [ɜ̃].

	pole 1	pole 2	pole 3	extra pole	zero
[ɜ]	470 (33)	1450 (42)	2300 (62)	...	...
[ɜ̃]	517 (56)	1230 (55)	1830 (154)	270 (46)	312 (60)

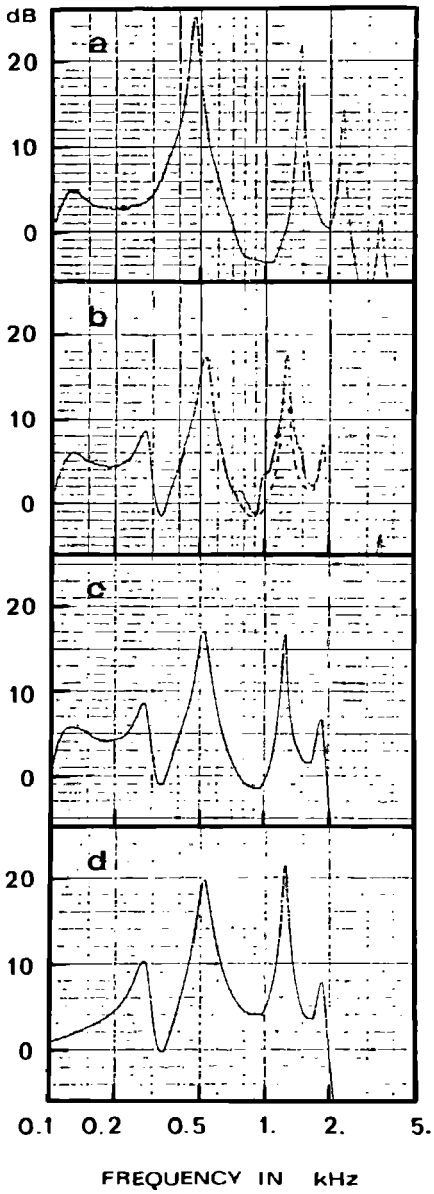

FIG. 14. Data of matching articulations of a neutral vowel with and without nasalization. (a) Recorded response curve for [ɜ]. (b) Same for [ɜ̃] (the broken line is an overlaid tracing of Curve c. (c) Synthesized curve for [ɜ̃]. (4) Same without the correction function (the correction function is given in Fig. 5, Curve 625-II).

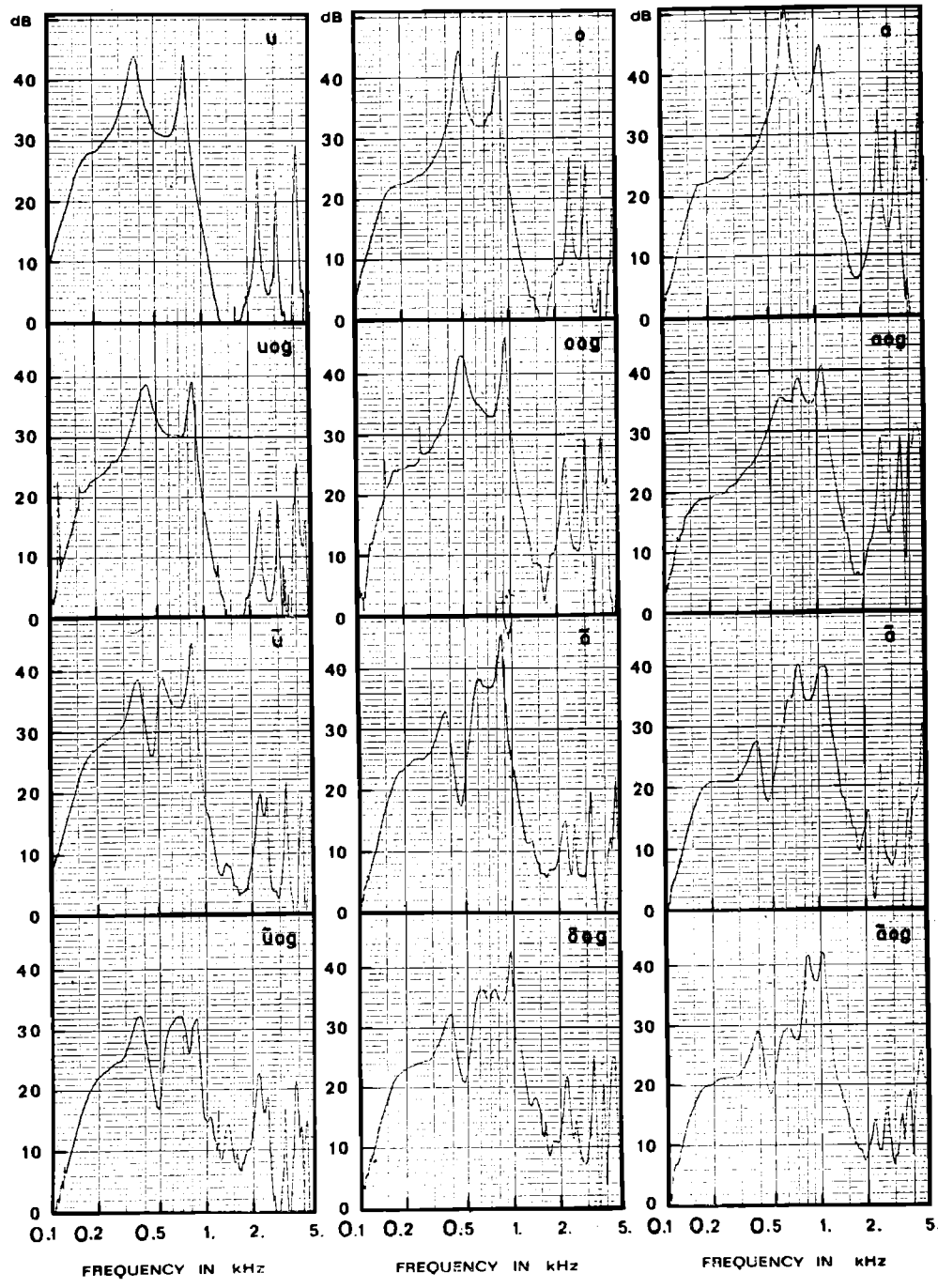

FIG. 15. Comparisons of nonnasalized, open glottis, and nasalized conditions, where "og" stands for open glottis.

$F_1{}^n$ is always located between the characteristic frequency and F_1, we may assume that the characteristic frequency of the nasal tract of this Swedish male subject is located close to this frequency, somewhere between 300 Hz and 400 Hz. When the first formant is higher, as in the case of the vowel [o], the antiformant occurs at the same frequency as in [u], showing the same degree of velum opening. This is true also for the nasalized open vowel [ɑ̃], and it indicates reproducibility of the subject's articulatory gestures. The nasal formant for [õ] is slightly higher than for [ũ], and for [ɑ̃] it is still higher. The correlation between the location of the nasal formant and the first-formant frequency can be readily predicted from the graphical considerations. The shifted first formant is also, as predicted, always higher than the original formant position.

The gross acoustic effects of nasalization of these back vowels are to complicate the spectral shape in the first-

formant region and to damp the second formant at the same time, thus causing an appreciable deviation from the vocalic F pattern and a spread of acoustic energy into the lowest frequency range. In this respect, the slight opening of the glottis is very similar in the effects to nasalization, except that the spread of energy into the lowest frequency range is not observed. The trachea system is much larger than the nasal tract, and the lowest resonance of the combined system is too low to be considered. The effect is probably more like introducing a free-field radiation through a small hole at the glottis, which substantially damps the formants. The combination of nasalization and glottis-opening results in a suppression of any singling out peak, as demonstrated at the bottom of this figure for the three vowels.

F. Nasal Consonants

Nasalization of a stop consonant results in an articulation of a nasal consonant (nasal murmur). Comparisons between the nasal murmurs and the nonnasalized stops with identical tongue articulations have been made by this sweep-tone method. Figure 16 illustrates some examples.

It is noted that quite often we can relate a peak in the nasal curve to the second or the third formant of the corresponding stop. When this is true, it may be said that the nasal passages leading back to the velum can be regarded as a kind of well-damped probe tube for the frequency region.[30] It is apparent, however, that the vocal-tract transfer functions for the nasal murmurs are substantially different from those of vowels. Even in the lowest frequency region up to 500 Hz, the envelopes obtained here for nasals with different coarticulations cannot be represented by a single pole. The middle-frequency range also entirely lacks characteristics of the F patterns for vowels. It is also noted that the transfer characteristics of nasal consonants obtained in this experiment varied greatly from subject to subject. Thus, the common gross features of nasals as a class in opposition to stops are found to be: (1) A marked but not necessarily simple low-frequency boost around 200–300 Hz; (2) a gross deviation from the vowel F pattern as an over-all pattern; and (3) a higher total energy both in the low-frequency boost and in the rest of the frequency domain compared with stops. These results compare with previous conclusions of analyses of natural utterances and also of synthesis experiments that have been reported by several different authors.[4,31]

The present data show that the transfer functions in details are, in fact, very complex and variable depending on both the subject and the articulatory characteristics, and there is not very much more to say for general accurate descriptions of the transfer characteristics. When the regular voice source is used for excitation, however, the spectral characteristics of the outcoming sounds in any case cannot be determined exactly, and details in spectral information are usually not of any

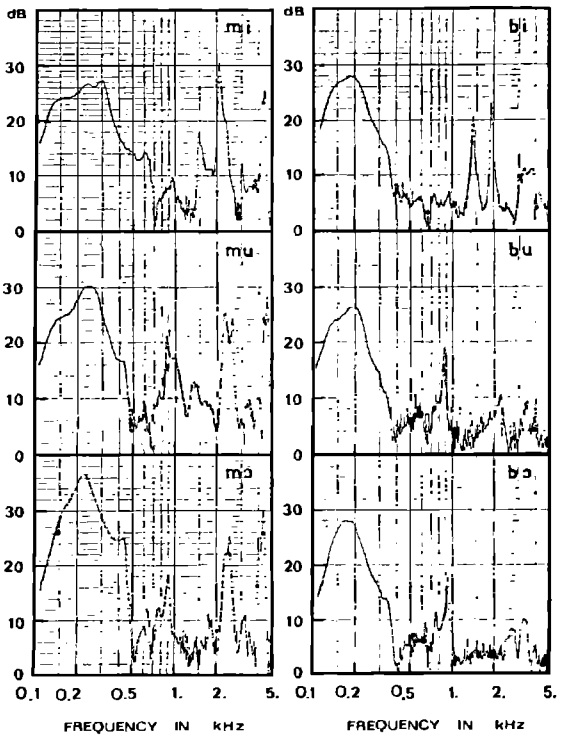

FIG. 16. Comparisons between nasals and stops with approximately the same articulation except for lowering of the velum.

immediate concern. In gross terms for practical considerations, particularly in comparison with vowels, an approximation by a damped-vowel F pattern (probably preferably with a slightly higher formant density) works well perceptually when appropriate formant transitions can give strong enough cues in regard to the place of articulation.[32] For a better approximation of spectral characteristics of nasal murmurs, however, introduction of a pole–zero pair in the middle-frequency range is required to account for the gross spectrum envelopes characterizing different nasal consonants.[33] The location of the antiformant constitutes a minimally simplified acoustic determinant of the place of articulation as far as the stationary murmur is concerned, but whether it is obviously identified in the isolated spectrum as a local characteristic depends on the over-all complexity of the spectrum.[34] For nasalization of vowels, too, a similar spectral complexity will be expected when the degree of nasalization is comparatively high.

III. CONCLUDING REMARKS

A new technique of estimating the vocal-tract transfer characteristics by direct acoustic measurements has been proposed, and it has been experimentally used for obtaining data that led to some new findings about the acoustic characteristics of the articulatory system. The method to some extent depended on the articulatory skill of the subjects. In contrast to the revealing curves for the vowel F patterns elicited by many subjects as

reported here, mostly without any special practice for this experiment, some subjects did not provide any comparative resonant curves. One common difficulty seemed to be in the control of the glottal conditions. Sometimes, the separation of the glottal control from the velar control seemed to be particularly difficult. It is suspected that in some cases subjects might have tended to relax the esophagus opening in trying to close the glottis tightly. No attempt was made at the time of this experiment to monitor the laryngeal condition but by listening to the acoustic effect by use of the external buzz excitation.

Some data for Swedish fricative consonants have been collected from one of the male subjects in this experiment, but the data have been omitted from our analyses in this study. A combination of this method with some other articulatory measurements, in particular radiographic observations, is also one of the themes for future studies. Observations of dynamic changes in articulation perhaps could be achieved by the same technique if we replace the sweep tone by a pulsetrain, and the level recording by a computational analysis of the output time function,[35] if the signal level could be high enough.

ACKNOWLEDGMENTS

This experimental study was performed at the Royal Institute of Technology (KTH), Stockholm, mainly during the period 1964–1965. Professor Gunnar Fant as the director of the Department of Speech Communication at the KTH, participated in the study in many aspects, and his direct and indirect contributions to the content of this paper cannot be fully listed. We simply express our gratitude and appreciation. Also thanks are due to many members of the laboratory and particularly to Si Felicetti for her efficient secretarial assistance. We also acknowledge our appreciation of participations of Sven Öhman, Kerstin Engström, Bibi Bergendahl, and Birgitta Bartning, as the most successful subjects.

* The experiment described here was carried out at the Speech Transmission Laboratory, Department of Speech Communication, Royal Institute of Technology, Stockholm, during the period of 1964–1965. Parts of the content of this paper have been reported in several issues of the Speech Transmission Laboratory, Quarterly Progress and Status Report.

[1] J. van den Berg, "Transmission of the Vocal Cavities," J. Acoust. Soc. Amer. 27, 161–168 (1955).

[2] G. Fant, "The Acoustics of Speech," *Proceedings of the Third International Congress on Acoustics Stuttgart 1959* (Elsevier, Amsterdam, 1961), Vol. I, p. 188. Also, G. Fant, "Formant Bandwidth Data," Speech Transmission Lab. Quart. Progr. Status Rep. No. 1, Royal Inst. Technol., Stockholm (1962), pp. 1,2.

[3] J. L. Flanagan and L. Landgraf, "Self-Oscillating Source for Vocal-Tract Synthesizers," IEEE Trans. Audio Electroacoust. AU-16, 54–62 (1968).

[4] G. Fant, *Acoustic Theory of Speech Production* (Mouton, The Hague, 1960).

[5] For the first proposal, see K. N. Stevens, "Toward a Model for Speech Recognition," J. Acoust. Soc. Amer. 32, 47–55 (1960). For an example of actual data processing of speech material by use of this principle, see, for example, C. G. Bell, H. Fujisaki, J. M. Heinz, K. N. Stevens, and A. S. House, "Reduction of Speech Spectra by Analysis-by-Synthesis Techniques," J. Acoust. Soc. Amer. 33, 1725–1736 (1961).

[6] K. N. Stevens, Ref. 5

[7] See C. G. Bell *et al.*, Ref. 5.

[8] For the direct acoustic measurement, see A. S. House and K. N. Stevens, "Estimation of Formant Band Widths from Measurements of Transient Response of the Vocal Tract," J. Speech Hearing Res. 1, 309–315 (1958).

[9] For comparisons of different methods and discussions on accuracy of the estimations, see H. K. Dunn, "Methods of Measuring Speech Formant Bandwidths," J. Acoust. Soc. Amer. 33, 1737–1746 (1961); see also G. Fant, "Formant Bandwidth Data," Ref. 2.

[10] See references cited above. It may be mentioned that several data points in the graph given by Fant ("Formant Bandwidth Data," Ref. 2) actually show that the close vowels have wider bandwidths for the first formant, even though the fact is not noted as a significant result.

[11] According to a calculation by House and Stevens (Ref. 8), the contribution of the subglottal coupling due to voicing in normal conditions is estimated to be approximately 20 Hz at lower frequencies and somewhat less in higher frequencies. It may be argued that the frequency dependence shown in Fig. 7 is roughly the same or perhaps even more pronounced when this contribution is taken into consideration.

[12] According to House and Stevens (Ref. 8), the contribution of the radiation loss and that of the wall vibration to the bandwidth value are predicted to be about 2 and 70 Hz, respectively, for the neutral vowel. See also Fant (Ref. 4) and J. L. Flanagan, *Speech Analysis, Synthesis and Perception* (Springer, Berlin, 1965) for calculations of different contributions to the bandwidth values of various vowel articulations.

[13] G. Fant and B. Sonesson, "Speech at High Ambient Air-Pressures," Speech Transmission Lab. Quart. Progr. Status Rep. No. 2, Royal Inst. Technol., Stockholm (1964), pp. 9–21.

[14] According to our finding above that the female subjects show higher dissipation, we would expect the formant shift due to high pressure would be more pronounced for female subjects than for male.

[15] The bandwidth data obtained here are generally comparable to those reported by House and Stevens (Ref. 8) for the closed glottis condition, both for the first formant and for the second formant, except that their values are significantly higher than ours for nonclose front vowels.

[16] Ref. 4, pp. 309, 310.

[17] All of our subjects had experience in more or less professional phonetic training.

[18] O. Fujimura and J. Lindqvist, "Experiments on Vocal Tract Transfer," Speech Transmission Lab. Quart. Progr. Status Rep. No. 3, Royal Inst. Technol., Stockholm (1964), pp. 1–7.

[19] K. Ishizaka (personal communication).

[20] R. K. Potter, G. A. Kopp, and H. Green Kopp, *Visible Speech* (Dover, New York, 1966).

[21] N. Chomsky and M. Halle, *The Sound Pattern of English* (Harper and Row, New York, 1968), Chap. VII. Also see, M. Halle and K. N. Stevens, "On the Mechanism of Glottal Vibration for Vowels and Consonants," Res. Lab. Electron., MIT, Quart. Progr. Rep. No. 85 (1967), pp. 267–271.

[22] A. S. House and K. N. Stevens, "Analog Studies of the Nasalization of Vowels," J. Speech Hearing Disorders 21, 218–232 (1956).

[23] S. Hattori, K. Yamamoto, and O. Fujimura, "Nasalization of Vowels in Relation to Nasals," J. Acoust. Soc. Amer. 30, 267–274 (1958).

[24] O. Fujimura, "Spectra of Nasalized Vowels," Res. Lab. Electron., MIT, Quart. Progr. Rep. No. 58 (1960), pp. 214–218.

[25] M. H. L. Hecker, "Studies of Nasal Consonants with an Articulatory Speech Synthesizer," J. Acoust. Soc. Amer. 34, 179–188 (1962).

[26] J. L. Flanagan, Ref. 12.

[27] S. Hattori, K. Yamamoto, and O. Fujimura, "Nasalization of Vowels in Relation to Nasals," J. Acoust. Soc. Amer. 30, 267–274 (1958).

[28] It can be seen, therefore, that the lowest formant is always a nasal formant, regardless of the vowel configuration, if the nostrils are closed.

[29] It is quite possible, for example, that the shifted first formant of [ɑ] may disappear, and the lower nasal formant may appear

conspicuously as if it were the first formant shifted downward. If the degree of nasalization is continuously decreased, however, the origin of the formant should be traced correctly.

[30] Care has been taken also in these recordings in regard to the gain of the recording system that the relative levels of the curves can be compared.

[31] O. Fujimura, "Analysis of Nasal Consonants," J. Acoust. Soc. Amer. **34**, 1865–1875 (1962); S. Hattori, K. Yamamoto, and O. Fujimura, "Nasalization of Vowels in Relation to Nasals," J. Acoust. Soc. Amer. **30**, 267–274 (1958); A. S. House, "Analog Studies of Nasal Consonants," J. Speech Hearing Disorders **22**, 190 204 (1957); and K. Nakata, "Synthesis and Perception of Nasal Consonants," J. Acoust. Soc. Amer. **31**, 661–666 (1959).

[32] For a terminal analog simulation with a highly damped first formant, see K. Nakata, Ref. 31.

[33] See O. Fujimura, Ref. 31. Some perceptual experiments show that nasals must be identified in some phonetic environments by characteristics of the nasal murmur. See, e.g., A. Malécot, "Nasal Syllabics in American English," J. Speech Hearing Res. **3**, 268–274 (1960).

[34] Note that in the analysis-by-synthesis work cited above (Fujimura, Ref. 31), the locations of the additional pole–zero pair were estimated by evaluating their influences in higher-frequency regions, too.

[35] This idea was suggested by M. R. Schroeder (personal communication).

Appendix A. An Instrumentation for Spectrum-Matching Experiments

In connection to the pole–zero analysis of the vocal-tract response curves, an optical programming device (function generator) has been developed and has proved useful for the pertinent purposes. The system is described here in some detail together with supplementary remarks on other relevant techniques of the experiment.[A1]

The Brüel & Kjær oscillator–recorder (type 3304) is employed for recording both the measured response curves and the synthesized response curves. The built-in compressor circuit is utilized in order to add (in the logarithmic scale) a constant frequency characteristic to the response curve. In the case of data recording, as shown in Fig. 1, the output voltage of the oscillator is fed into the compressor–input terminal through a simple resistance–capacitance network, with the result that the signal level applied to the vibrator is a monotonically rising function of frequency [see Fig. 2(b)]. This suppression of low frequencies is introduced in order to obtain a maximum signal power without distortion, and at the same time a grossly flat average signal level at the microphone over the frequency range of 100–5000 Hz. A more exact correction for the constant characteristics, including the transmission characteristics through the body wall, is attempted in the synthesizing (matching) stage, rather than in the data recording, because these characteristics vary from subject to subject and also from session to session, depending on the condition at the throat of the subject.

In the case of synthesizing frequency-response curves for matching, it is therefore necessary to generate complex semivariable frequency characteristics. For this purpose, an optical control system has been designed as an independent attachment to the Brüel & Kjær oscillator–recorder. A frequency function is specified in the form of a painted plexiglas disk, and the output level of the sinusoidal signal produced by the Brüel & Kjær oscillator is automatically adjusted exactly in accordance with this specification. A schematic diagram for this system is shown in Fig. A-1.

The light flux that originates from a special light source transmits through the transparent portion of the program disk and is then measured by a special light sensor. The control disk is prepared by painting on the disk; any desirable curve can be given as the outer borderline of the unpainted transparent portion. This disk is then attached to the dial shaft of the oscillator through a toothed rubber band. In the mechanical link, the rotation angle is amplified by a factor of 3, so that 360° rotation of the disk covers a frequency range from 100 to 4500 Hz. The range of the radius available on the disk for programming is from 75 to 100 mm. Three small incandescent lamps are embedded in a light guide of plexiglas to form the light source. One edge of this plexiglas light guide transmits light with a nonlinear intensity distribution along the radius of the program disk. The shape of the plexiglas piece and the positions of the lamps as well as the voltages given to them are adjusted in such a way that the position of the boundary between the painted and transparent portions to the program disk gives directly a decibel scale for the measured total flux. After some *ad hoc* trials, a dynamic range of 13 dB was attained, as shown in the calibration curve (see Fig. A-2). The width of the edge is 2 mm corresponding to about 2% of the frequency value within the dynamic range.

The light intensity is received by the sensor unit that also consists of a plexiglas light guide and photosensitive diodes embedded therein and the composite output of the diodes gives the reference voltage for the oscillator

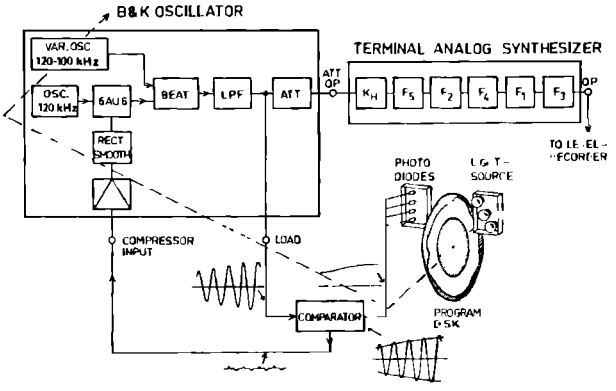

FIG. A-1. Synthesis scheme for matching the response curves by the use of an optical program disk for the semivariable correction function.

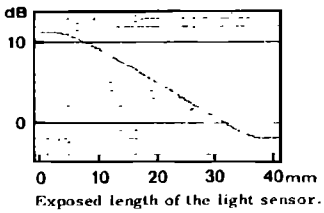

FIG. A-2. Calibration curve for the optical control system.

output level. A reasonably accurate and quite reproducible tracking is achieved by means of an electrical servosystem consisting of the built-in compressor circuit and an external transistorized comparator circuit. In the latter, the sinusoidal signal from the oscillator output terminal is compared with and sliced by the reference voltage. This is done in a balanced form, mainly with the intention to eliminate the transient d.c. component (thump). Essentially, the central portion of the waveform is deleted and the remaining parts made abutted upon each other. The signal is amplified and fed into the compressor terminal, so that the output from the oscillator automatically follows, in the envelope, the specified value at each frequency. A calibration can be given on the absolute output levels for reference points in the program. A simple adjustment of the lamp voltage in reference to the absolute output level, in the beginning of a series of matchings, seems to guarantee a reproducibility with no observable variation in the relative frequency curve, both within a session and among sessions on different days. Particular care has been taken in the transistor circuit-design in order to achieve this stability. The temperature dependence of transistors are largely taken care of by the circuit itself, and a possible slow change in the characteristics of the ac amplifiers after the comparison slicing is not significant because the circuit is put in a servo loop.

The control disk is painted with a water-soluble black paint (casein color). Fine adjustments of the edge of the painted portion, which determines the frequency function, can be made by scraping part of the dry paint or by repainting, but the best is to use a thin adhesive black tape in order to define the boundary. A special radial ruler with a calibrated decibel scale has been made to facilitate drawing the frequency response curve on the disk. The result of the calibration is illustrated in Fig. A-2, where the amount of the signal reduction is plotted against the width of the transparent portion on the control disk. Note that, for the dynamic range to be covered for correction functions (as exemplified in Fig. 5), the input–output relation is almost perfectly linear (the wiggle is due to the stepwise function of the servosystem in the Brüel & Kjær recorder).

[A1] O. Fujimura and J. Lindqvist, "An Instrumentation for Spectrum-Matching Experiments," Speech Transmission Lab. Quart. Progr. Status Rep. No. 2, Royal Inst. Technol., Stockholm (1964), pp. 6–8.