THE RELATIONS BETWEEN AREA FUNCTIONS AND THE ACOUSTICAL SIGNAL

Gunnar Fant, Department of Speech Communication, Royal Institute of Technology, S-10044 Stockholm, Sweden

Chairpersons: Wiktor Jassem and Kenneth N. Stevens

Introduction

The topic of this paper is to discuss how configurations, shapes, and detailed outlines of the vocal tract cavity system influence the acoustic signal and the reverse, how to predict vocal tract resonator dimensions from speech wave data. As far as the direct transform is concerned, this is a re-visit to my old field of acoustic theory of speech production.

What progress have we had in vocal tract modeling and associated acoustic theory of speech production during the last 20 years? My impression is that the large activity emanating from groups engaged in speech production theory and in signal processing has not been paralleled by a corresponding effort at the articulatory phonetics end. Very little original data on area functions have accumulated. The Fant (1960) Russian vowels have almost been overexploited. Our consonant models are still rather primitive and we lack reliable data on details of the vocal tract as well as of essential differences between males and females and of the development of the vocal tract with age.

The slow pace in articulatory studies is of course related to the hesitance in exposing subjects to X-ray radiation. Much hope was directed to the transformational mathematics for deriving area functions from speech wave data. These techniques have as yet failed to provide us with a new reference material. The so-called inverse transform generates "pseudo-area functions" that can be translated back to high quality synthetic speech but which remain fictional in the sense that they do not necessarily resemble natural area functions. Their validity is restricted to non-nasal, non-constricted articulations and even so, they at the best retain some major aspects of the area function shape rather than its exact dimensions. However, some improvements could be made if more representative acoustic models than LPC analysis are considered.

Once a vocal tract model has been set up it can be used, not only for studying articulation-to-speech wave transformations, but also for a reverse mapping of articulations and area functions to fit specific speech wave data. These analysis-by-synthesis remapping techniques, as well as perturbation theory for the study of the consequences of incremental changes in area functions or of the inverse process, are useful for gaining insight in the functional aspect of a model. However, without access to fresh articulatory data the investigator easily gets preoccupied with his basic model and the constraints he has chosen.

The slow advance we have had in developing high quality synthesis from articulatory models is in part related to our lack of reliable physiological data, especially with respect to consonants, in part to the difficulty involved in modeling all relevant factors in the acoustic production process. The most successful attempt to construct a complete system is that of Flanagan et al. (1975) at Bell Laboratories. A variety of studies at KTH in Stockholm and at other places have contributed to our insight in special aspects of the production process such as the influence of cavity wall impedance, glottal and subglottal impedance, nasal cavity system, source filter interaction, and formant damping.

From area function to the acoustic signal

The acoustic signal or, in other words, the speech wave is the product of a source and a filtering process. The most common approach is to disregard the source and relate a vocal tract area function to a corresponding formant pattern only, i.e. a set of formant frequencies F_1 F_2 F_3 F_4, etc. This correspondence is illustrated by Fig. 1. I shall not go into the mathematics of the wave equations and the equivalent circuit theory. Instead I will attempt to develop a perspective around some basic models and current problems.

To derive an area function from X-ray data on vocal tract dimensions is by no means a straightforward procedure, see Fant (1960; 1965) and Lindblom and Sundberg (1969).

The estimation of cross-sectional shapes and dimensions in planes perpendicular to the central pathway of propagation through the vocal tract has to rely on crude conventions and involves uncertainties, e.g. with respect to variations with articulation and for different types of subjects. The lack of basic data is especially apparent for female and child speech and for consonants, e.g. laterals and nasals. In spite of the accessibility of the speech wave to quantitative analysis there is a similar lack of reference data concerning the acoustic correlates. Most studies have been concerned with male speech and vowels.

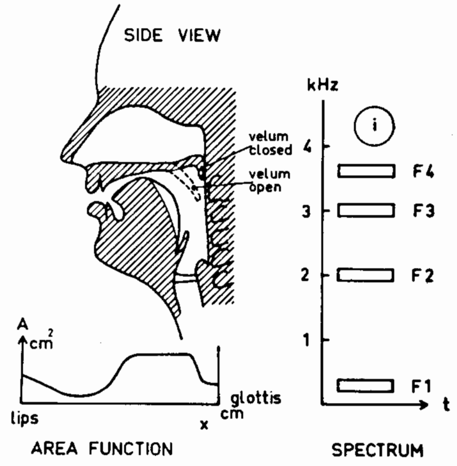

Figure 1. Principle illustration of vocal tract sagittal view with area function and corresponding resonance frequency pattern.

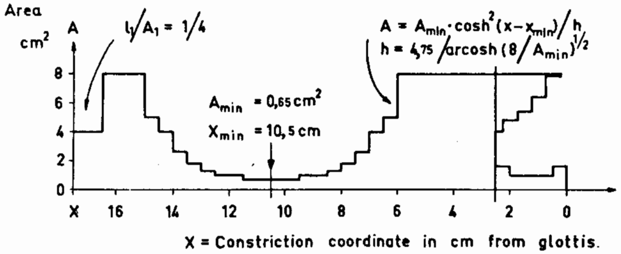

Figure 2. Three-parameter voçal tract model (Fant, 1960).

A specification of an area function as a more or less continuous graph of cross-sectional area from the glottis to the lips allows detailed calculations of the acoustic response but is not practical for systematic descriptions. A data reduction in terms of parametric models brings out the acoustically relevant aspects. The three-parameter models of Stevens and House (1955) and Fant (1960) differ somewhat in the details but have the same set of parameters, the place of minimum cross-sectional area of the tongue section, the area at this coordinate, and the length over area ratio l_0/A_0 of the lip section.

My model is shown in Fig. 2. The shunting sinus piriformis cavity around the outlet of the larynx tube was a constant feature in my model. A weakness is that it is not reduced in volume for back vowels which does not allow F_1 to reach a sufficiently high value for [a]. Fig. 3 shows the variation of the F-pattern with the place of tongue constriction. This is a well established graph which retains basic patterns such as the rise of F_2 with advance of the tongue constriction from back to front up to an optimal place at a midpalatal location after which F_2 drops again. A limitation of the parameter range to a region bounded by [ɑ], [u], and [i] as proposed in several articulatory models, e.g. Lindblom and Sundberg (1969), would exclude the standard Swedish pronunciation of the vowel [ʉ] which, contrary to traditional classifications, has a constriction somewhat anterior to that of [i] (Fant, 1973).

The constriction coordinate is an acoustically more relevant classifier than the "highest point of the tongue" of classical phonetics. Most stressed vowels have a definite "place of articulation" as evidenced by a region of minimum cross-sectional area which we may exemplify by [i], [u], [o], [ɑ] ending with a variant of [æ] with major narrowing just above the glottis (Fant, 1960). On the other hand, it may be argued that the traditional classification in terms of tongue locations and related parameters belongs to a production stage one step higher up than area functions and could be directly related to formant patterns.

The [a] and [i] vowels are polar opposites, the [i] vowel requiring a wide pharynx and narrowed mouth, whilst the opposite is true of [a] type vowels. A production of a vowel [u] requires a double resonator configuration with a narrow lip opening to ensure

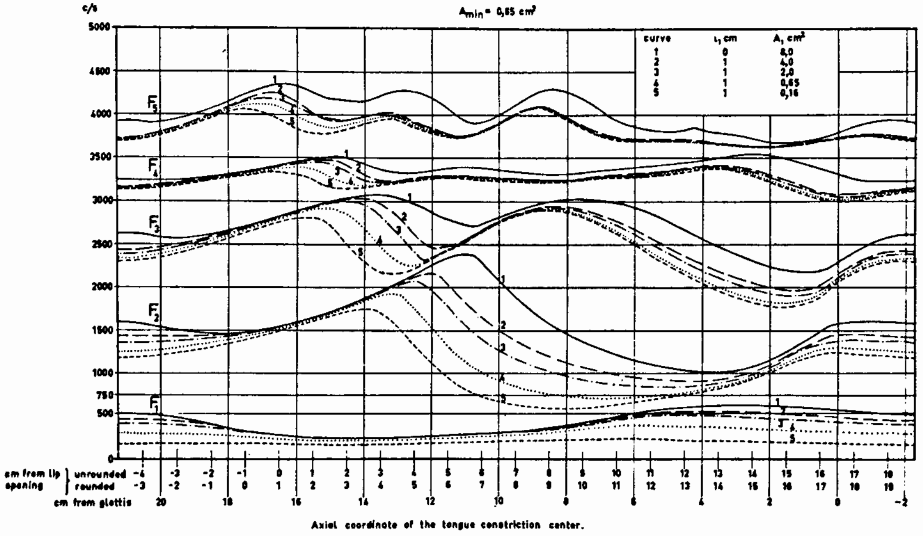

Figure 3. F-pattern variation with constriction coordinate x_C at different sets of lip parameter l_1/A_1 at constant constriction area A_{min}. The constriction coordinate is zero at the glottis.

a low F_1 and a narrow constriction between the two major cavities as a correlate of a low F_2. These shape aspects are brought out in the stylized area functions of Fig. 4. A basic issue in acoustic phonetics is that it is not possible to produce these vowels without retaining the major shape aspects of the area functions. To this extent area functions are predictable from the acoustic signal as will be discussed in greater detail in a later section. Peter Ladefoged would back me up here with his competence of transforming phonetic qualities to equivalent resonator configurations.

Another basic issue is that the vocal tract filtering is determined by the location of formants only and that the spectrum envelope between peaks cannot contain any other irregularities than those originating from the source function. Minor irregularities in the outline of the area function may have some influence on formant locations but will not give rise to irregularities in the spectrum envelope. This is not evident without an insight in the mathematical constraints imposed by acoustic theory. It is related to the one-dimensional wave propagation, wavelengths generally being short compared to vocal tract cross dimensions. Systematic perturbations of vocal tract area functions will be discussed in a later section.

Highly simplified area functions of fricatives (or corresponding stops) and their filtering functions are shown in Fig. 5. As discussed by Fant (1960), the "compact" sibilant [ʃ] or the stop [k] has a definite cavity in front of the major constrictions which accounts for a central dominance of the spectrum, usually a single formant, if the cavity is abruptly terminated by the constriction. The [s] or [t] has a narrow channel of a few centimeters length behind the source which may combine with a small front cavity to produce resonances above 4000 Hz which build up a high-pass filtering. The [f] or [p] has no significant resonance in its closed state.

In general, the cavities behind the source do not influence the spectrum much, provided that the consonantal constriction is effective. Resonances of the back cavities may appear if the constriction tapers off gradually as in palatals or if a palatal tongue articulation builds up a supporting constriction behind the lips. Back cavity resonances combine with and are cancelled by spectral zeroes at complete closure but move away from their

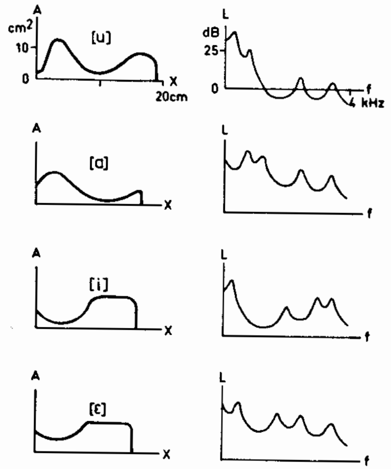

Figure 4. Stylized area functions and corresponding spectrum envelopes of [u] [ɑ] [i] and [ɛ]. The constriction coordinate is zero at the lips.

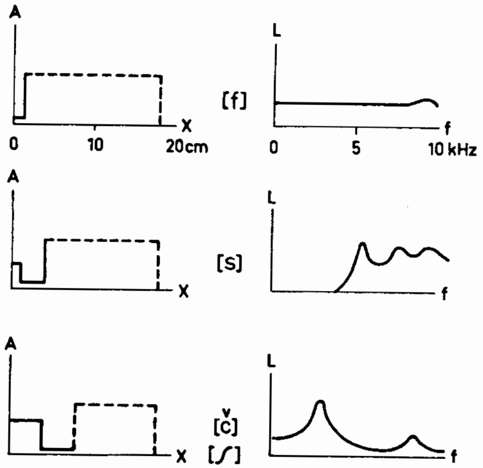

Figure 5. Stylized area functions and corresponding spectra of three basic consonant categories. The constriction coordinate is zero at the lips.

zero mates during release and are then more or less free to appear. In Fig. 6 we can study measured and calculated spectra of [k] and a palatalized [p'] (Fant, 1960). The labial burst spectrum contains peaks at around 2-3 kHz but has a free spectral minimum at 1400 Hz. In contrast, the [k] spectrum has a single formant peak around 1400 Hz. It is interesting to note that the calculations from the area function data back up the measured spectra. We need more studies of this type.

Vocal tract boundary constraints and dynamics

The simplified static models relating a single area function without parallel branches to a set of formant frequencies have obvious limitations. On a higher level of ambition we must include proper boundary conditions such as radiation load and a finite coupling to the subglottal and nasal systems. In order to predict formant bandwidths we must consider the energy loss during an oscillatory cycle of a formant associated with "loss elements" on the surface of the vocal tract resonator system and other dissipative elements (Fant and Pauli, 1975). Source functions must be defined with respect to place of insertion in the vocal tract, their spectrum or waveform, and the degree of coupling to other parts of the system (Stevens, 1971). In addition, these properties are highly time variable within a voice fundamental period (Fant, 1979) and within intervals of transition from various states of the glottis or of other terminations of the vocal tract. Rapid opening and closing gestures pose specific problems in relating area functions to acoustic data. In a proper analysis of connected speech we need two sets of acoustic variables: the continuous variations of the F-pattern as a correlate of the continuous movements of the articulators and the often abruptly varying patterns of spectral energy distributions associated with discrete events of production.

The acoustic production model of Fig. 7 may serve as a starting point for a brief discussion of these problems. First of all, we should note an important element in converting area functions to a filter function. The walls of the vocal tract are not rigid. They may expand during a voiced occlusion as represented by the element C_w in the equivalent circuit of a small slice of the area function, Fig. 8, and they have a finite mass L_w which adds to the tuning of vocal resonances and which dominates the impedance of

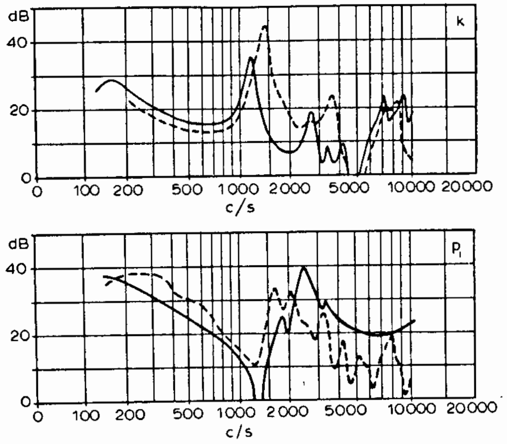

Figure 6. Calculated (solid line) and measured (broken line) stop release spectra of a velar [k] and a palatalized [p']. The minimum in [p'] at 1400 Hz is a free zero in the sub-lip impedance whilst the main formant of [k] is a mouth cavity formant. After Fant (1960).

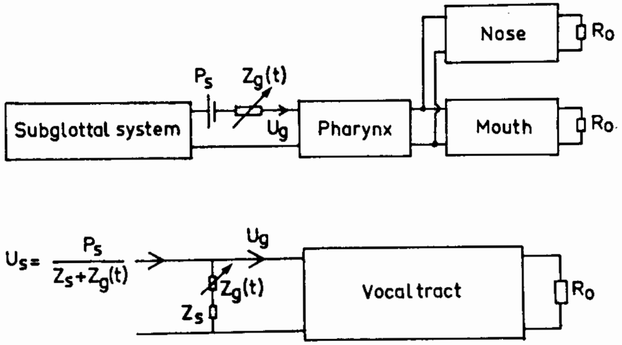

Figure 7. Block diagram of the production of voiced sounds illustrating the principle difference between an ideal source U_s and a glottal flow U_g.

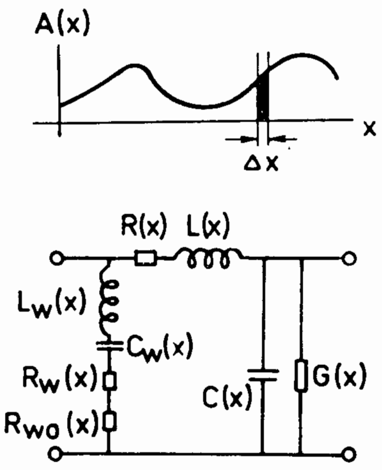

Figure 8. Lumped constant approximation of a small slice of the area function.

the shunting branch at frequencies above 40 Hz. A small fraction of sound is radiated externally from the outside of the head through R_{wo}. It is negligible except as a constituent of the voice bar of a voiced occlusion.

Disregarding the cavity wall mass element L_w, calculations would provide $F_1 = 0$ for an area function starting and ending with complete closure. The finite F_1 of around 150-250 Hz found in the spectrogram of the voiced occlusion is determined by the resonance of the entire air volume compliance in the tract with the total lumped cavity wall mass shunt. This resonance can easily be measured acoustically (Fant et al., 1976) and amounts to $F_{1w} = 190$ Hz with a bandwidth of $B_{1w} = 75$ Hz, typically for a male voice, and around 20% higher for females. The wall mass element L_w is thus an important constituent in calculating F_1 from the area function. The procedure is to start out with a derivation of an ideal F_{1i} without mass shunt and add a correction factor

$$F_1 = F_{1i}(1 + F_{1w}^2/F_{1i}^2)^{1/2} \qquad (1)$$

The distribution of the wall impedance along the vocal tract and its dependence on particular articulations are not known. The experiments of Fant et al. (1976) suggest that regions around the larynx and the lips are especially important. Experiments by Ishizaka et al. (1975) provide data of the same order of magnitude but have not revealed conclusive distribution patterns.

The resistive component R_w in the cavity wall branch determines a major part of the bandwidth B_1 of low F_1 formants. The resistive part of the radiation load which is proportional to frequency squared is the essential bandwidth determinant of resonances above 1000 Hz originating from an open front resonator. Internal surface losses from friction and heat conduction enter through the elements R and G in Fig. 8. They are proportional to the half power of frequency and to the inverse of the cross-sectional area. A detailed analysis of formant bandwidths and their origin appears in Fant (1972), Fant and Pauli (1975), and Wakita and Fant (1978).

The time variable glottal impedance accounts for variations of formant frequencies and bandwidths within a voice fundamental period (Flanagan, 1965). A more detailed analysis of glottal damping requires a reconsideration of the process of voice generation (Fant, 1979) and adoption of perceptual criteria for deriving equivalent mean values (Fant and Liljencrants, 1979). The main excitation of the vocal tract occurs at the instant of interruption of glottal flow by glottal closure. At this instance, damped oscillations are evoked and subjected to the damping from supraglottal loss elements.

When the glottis opens for the next flow pulse the vocal tract becomes loaded by the time variable glottal plus subglottal impedance. Providing a resonance mode is much dependent on the part of the area function immediately above the glottis, the glottal damping becomes severe. This is especially apparent if the lower pharynx is narrowed thus facilitating an impedance match between the cavity system and the glottal resistance. A complete extinction of the formant oscillation in the glottal open interval may result. This is typical of F1 of the vowel [a] produced at low or moderate voice effort by a male subject.

In general most of the energy excited during a voice fundamental period is lost during the timespan of the following period. Since glottal resistance decreases with lowered transglottal pressure the damping effect is especially apparent at weak voice levels. The mean glottal bandwidth in normal voice production is of the order of 0-100 Hz with 20 Hz as a typical value for male medium intensity phonation.

It is apparent that any model of voice production which adopts the actual flow through the glottis as the primary source will create problems. With this convention, which happens to apply to inverse filtering techniques, the source attains components of formant oscillations and becomes dependent of the vocal tract area function (Mrayati and Guérin, 1976). Their approach is intended to define a proper source for a formant synthesizer.

A different approach more suited for production models is to incorporate the combined glottal and subglottal impedance as a termination paralleling the input end of the tract and to define the source as the flow through the glottis which would have occurred with the input to the vocal tract short circuited. This representation adopted by Fant (1960) preserves a realistic definition of the vocal tract transfer function but fails to take into account source modifications due to aerodynamic losses in supraglottal constrictions. In the transition from a vowel to a voiced consonant there is generally some loss of transglottal pressure which reduces the excitation strength of the voice source.

The interplay of glottal and supraglottal sources associated with articulatory narrowing and release becomes an important part of a dynamically oriented theory of predicting acoustic signals from area functions (Stevens, 1971).

What about the subglottal system? How does it influence speech? In normal voice production the influence appears to be small. As long as the glottal opening is small and the flow velocity high, the glottis impedance becomes high compared to the subglottal impedance. Unless there is a constant leakage bypassing the vibrating part of the glottis, the subglottal system should have a minor influence only.

This reasoning is concerned with the modification of the supraglottal formants only. At the instance of flow interruption when the glottis closes there is a simultaneous excitation of resonances in the trachea and other parts of the subglottal system. Potential frequencies are 600, 1250, and 2150 Hz for a male voice (Fant et al., 1972). The transmission losses associated with the penetration of these components through the walls of the trachea and the chest to externally radiated sound appear to be sufficiently high to rule out any significance, but this remains to be proved.

As shown by Fant et al. (1972), subglottal formants may occasionally be seen in spectra from aspirated sound segments, e.g. in the release phase of unvoiced stops. "F1-cutback" in the first part of the voiced interval after release, which appears as a relative delay in onset of F1 compared to F2 and higher formants, may be explained as an instance of excessive F1 damping through an incompletely closing glottis. The upper formants are less dependent on the glottal termination and thus less affected. This relative weakening of F1 is a filtering effect, whilst the relative weakness of F1 in a preceding unvoiced, aspirated segment is also a matter of low source energy in the F1 region. The F1 intensity reduction is also seen in the terminating periods of a vowel before the occlusion of an unvoiced stop (pre-occlusion aspiration).

Nasalization and aspiration have similar effects on F1. In nasalized sounds the F1 intensity is typically reduced by a spectral zero (Fant, 1960; Fujimura and Lindqvist, 1971). The nasal model of Fant (1960) produces too high values of the lowest nasal pole. The possible occurrence of several low frequency pole-zero pairs is made plausible by the study of Lindqvist and Sundberg (1972). More anatomical and acoustic data are needed.

In connection with the voice source studies of Fant (1979) it has been noted that the spectral maximum often seen below F_1 in vowels is a voice source characteristic, which becomes especially enhanced in contrast to a weak F1 in nasalized or aspirated, voiced segments. This is especially apparent in a time domain study. Another way of expressing this finding is to say that nasal sounds retain more source characteristics than non-nasal sounds.

If an area function is subjected to a substantial change in a very short time, one may expect some deviations from the linear stationary behavior. Point-by-point calculations of resonance frequencies are still valid but additional bandwidth terms enter which may be positive or negative. A rapid opening of a constriction is accordingly associated with a negative bandwidth component and a rapid closure with a positive bandwidth component. The analysis is simple. Consider a flow U(t) through an acoustic inductance $L(t) = \rho l/A(t)$. The pressure drop is:

$$P(t) = \frac{d}{dt}[L(t)U(t)] = L'U + LU' \qquad (2)$$

L' = dL/dt apparently has the dimension of a resistance R_d

$$R_d = \frac{dL}{dt} = \frac{-A'(t)\rho l}{A^2(t)} \qquad (3)$$

In a single resonator system the bandwidth component associated with a resistance R in series with an inductance L is simply R/2πL.

Accordingly, the bandwidth associated with R_d is

$$B_d = \frac{-A'(t)}{2\pi A(t)} \qquad (4)$$

which implies a bandwidth component of opposite sign to that of the rate of change of the area. Fig. 9 illustrates the temporal course of the bandwidth when a resonator of volume 100 cm^3 is coupled to a neck of length 4 cm and a cross-sectional area A(t) varying exponentially from closure to complete opening of 2 cm^2 with a time constant of 10 milliseconds.

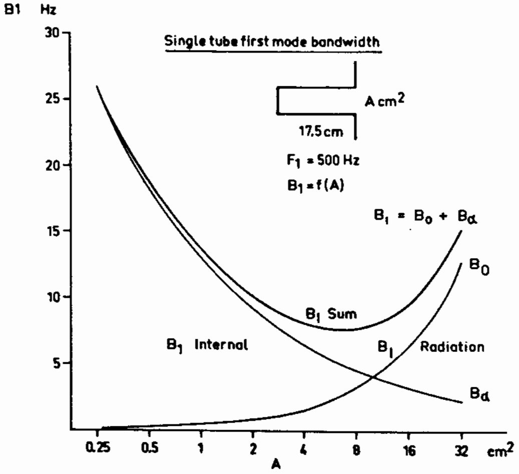

Figure 9. Resonator outlet area A, resonance frequency F, and total bandwidth B as a function of time during an exponential release with a time constant of 10 milliseconds. B_d is the negative dynamic component of the bandwidth.

The time varying negative bandwidth overrides the frictional bandwidth components up to 8 milliseconds after release which could tend to increase the amplitude of the oscillation during that period. However - in the speech case there enter additional positive bandwidth components related to flow dependent resistance and to cavity wall losses and possibly also glottal losses which tend to reduce the importance of the negative terms. In a detailed analysis of the glottis resistance the dynamics calls for some decrease of glottal resistance in the rising branch of the glottal pulse and an increase in the falling branch, as noted by Guérin et al. (1975). Except for the analysis above, a proper evaluation of the practical significance has to my knowledge not been performed. The most detailed thesis on the theoretical aspects is that of Jospa (1975). I feel that dynamic effects are of academic rather than practical significance. Of greater importance is probably the mere fact that a rapid transition of a formant creates a special perceptual "chirp" effect.

Perturbation theory and vocal tract scaling

Perturbation theory describes how each resonance frequency, F_1 F_2 F_3, etc., varies with an incremental change of the area function A(x) at a coordinate x and allows for a linear summation of shifts from perturbations over the entire area function. The relative frequency shift δF/F caused by a perturbation δA(x)/A(x) is referred to as a "sensitivity function". We may also define a perturbation δΔx/Δx of the minimal length unit Δx of the area function which will produce local expansions and contractions of the resonator system. It has been shown by Fant (1975b), Fant and Pauli (1975) that the sensitivity function for area perturbations of any A(x) is equal to the distribution with respect to x of the difference $E_{kx}-E_{px}$ between the kinetic energy $E_{kx} = \frac{1}{2}L(x)U^2(x)$ and the potential energy $E_{px} = \frac{1}{2}C(x)P^2(x)$ normalized by the totally stored energy in the system.

Fig. 10 from Schroeder (1967) illustrates perturbations of a single tube resonator by changes in the area function derived from sinusoidal functions. These have been chosen to influence F_1 only (a), none of the formants (b), and F_2 only (c). The middle case is of special interest. There exists an infinite number of small perturbations applied symmetrically with respect to the midpoint of the single tube, which will have almost no influence

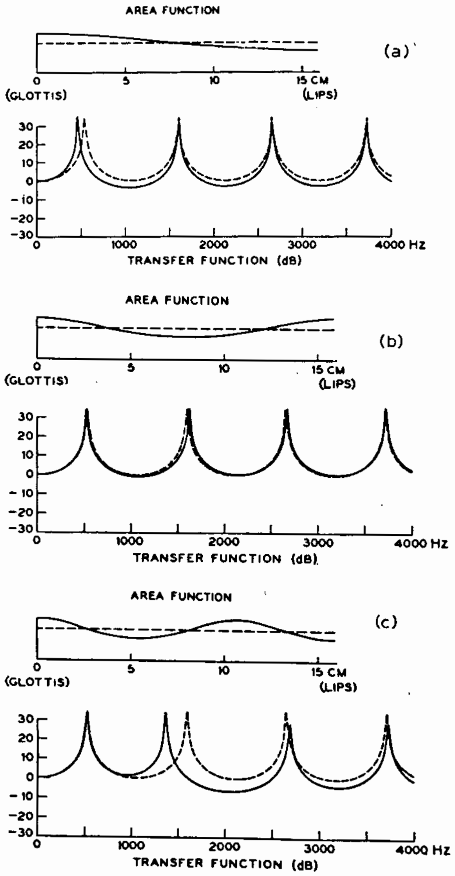

Figure 10. Perturbations of the single tube area function affecting F_1 only (a), almost no influence (b), and F_2 only (c) (after Schroeder, 1967).

on the formant pattern. In the general case of an arbitrary area function the rule of symmetry is upset (Heinz, 1967) but there still exists a tendency of compensatory interaction between front and back parts (Öhman and Zetterlund, 1975).

Sensitivity function for area perturbations of my six Russian vowels are shown in Fig. 11. This chart is useful as a reference for general use. Given the relative amount of area change, the corresponding relative frequency shift $\delta F_n/F_n$ is proportional to the product of $\frac{\delta A(x)}{A(x)}$ and the amplitude of the sensitivity function, $E_{kx}-E_{px}$. As an example we may note that F_1 of the vowel [u] rises with increasing area at the lips, i.e. decreases with increasing degree of narrowing and that narrowing the tongue constriction of [u] causes F_2 to fall and F_3 to rise. A narrowing of the outlet of the larynx tube will apparently have the effect of tuning F_4 to a lower frequency.

With the area function sampled at intervals of Δx, e.g. $\Delta x = 0.5$ centimeter for practical use, we may ask what happens if we increase Δx at the coordinate x by the amount $\delta\Delta x$. The local expansion thus introduced causes a frequency shift $\delta F_n/F_n$, which is proportional to $-\delta(x)/1+\delta(x)$ and to $(E_{kx}+E_{px})$ of resonance n.

The distribution of $(E_{kx}+E_{px})$ is uniform for a single tube resonator. The effect of a length increase is obviously the same irrespective of where along the x-axis the tube is lengthened. An overall increase of the length by, say $\delta(x) = 0.2$, causes a shift of all resonance by a factor $-0.2/(1+0.2) = -0.17$. The same calculation performed directly from the resonance formula $(2n-1)c/4l_t$, where l_t is the total length and c=35300 cm/s is the velocity of sound, would provide the same answer, i.e. a frequency ratio of $1/(1+0.2)=0.83$.

The distribution $E_{kx}+E_{px}$ along the vocal tract is also a measure of the relative dependence of the particular resonance mode on various parts of the area function. This is the best definition we have of "formant-cavity" affiliations. From Fig. 12 we may thus conclude that most of the energy of the second formant of [i] is stored in the pharynx, whilst the third formant of [i] "belongs to" the front part of the system. F3 of the back vowels [u] [o] and [ɑ] are associated with a central part of the tract, and F4 of all vowels has a substantial peak of energy located in

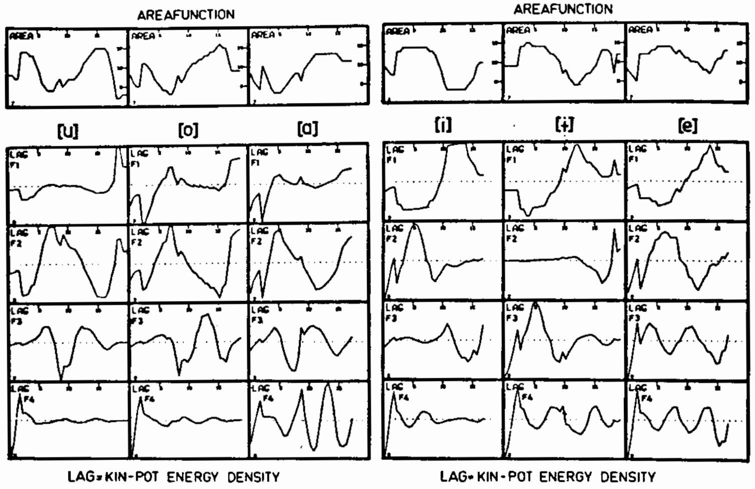

Figure 11. Sensitivity functions for area perturbations of the six Russian vowels (Fant, 1960). From Fant (1975b). The constriction coordinate is zero at the glottis.

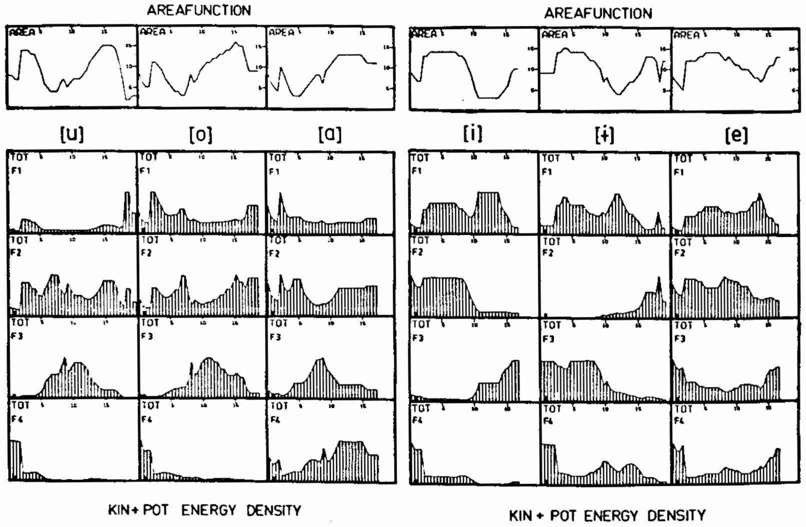

Figure 12. Sensitivity functions for length perturbations of the six Russian vowels (Fant, 1960). From Fant (1975b). The constriction coordinate is zero at the glottis.

the larynx tube. Expanding the length of the pharynx will have a large effect on F_2 of [i] and a small effect only on F_3 and vice versa for a length expansion of the mouth cavity. This analysis would apply to the relatively short pharynx of females compared to males.

If a perturbation of the entire area function is expressed as a function of as many parameters as there are formants, it is possible to calculate the change in area function from one F-pattern to another (Fant and Pauli, 1975). This indirect technique has been used by Mrayati et al. (1976) for deriving plausible area functions for French vowels on the basis of their deviation from my reference Russian vowels. This procedure must be administered in steps of incremental size with a recalculation of the sensitivity function after each major step. It may involve length as well as area perturbations.

In practice, when aiming at direct transforms only, it may be easier to resort to a direct calculation of the response of the perturbed area functions than to derive it from the energy distributions. The perturbation formulas and especially their energy based derivations are more useful for principal problems of vocal tract scaling or for gaining an approximate answer to a problem without consulting a computer program.

The area functions of male and female articulations of the Swedish vowels [i] and [u] and corresponding computed resonance mode pattern in Fig. 13 may serve to illustrate some findings and problems. The data are derived from tomographic studies in Stockholm many years ago in connection with the study of Fant (1965; 1966) and were published in Fant (1975a; 1976). It is seen that in spite of the larger average spacing of formants in the female F-pattern related to the shorter overall vocal tract length, the female F_1 and F_2 of [u] and the F_3 of [i] are close to those of the male. This is an average trend earlier reported by Fant (1975a), see Fig. 14. Differences in perceptually important formants may thus be minimized by compensations in terms of place of articulation and in the extent of the area function narrowing. Such compensations are not possible for all formants and cannot be achieved in more open articulations. The great difference in F_2 of [i] is in part conditioned by the relatively short female pharynx but can in part be ascribed to the retracted place of

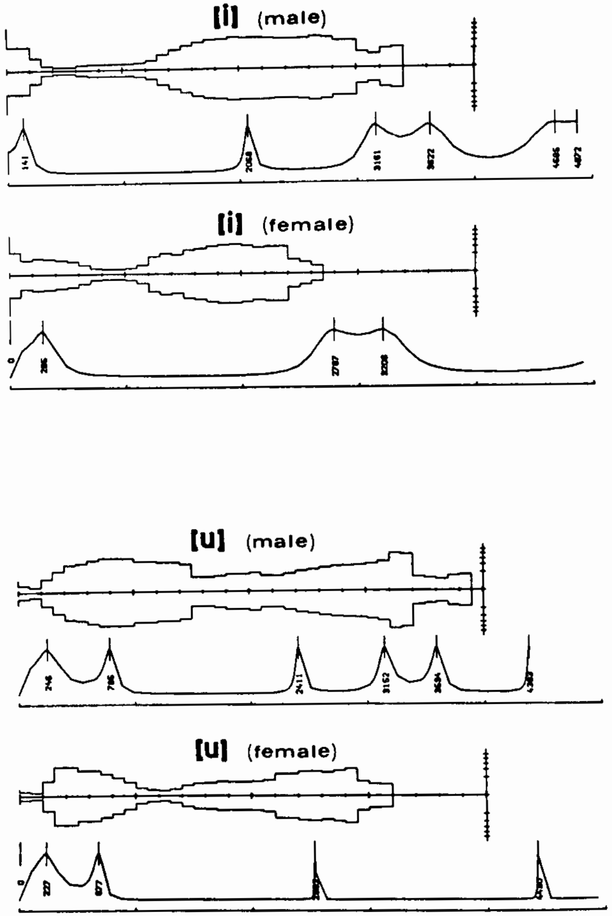

Figure 13. Male and female vocal tracts (equivalent tube representation) and corresponding F-patterns from the tomographic studies of Fant (1965).

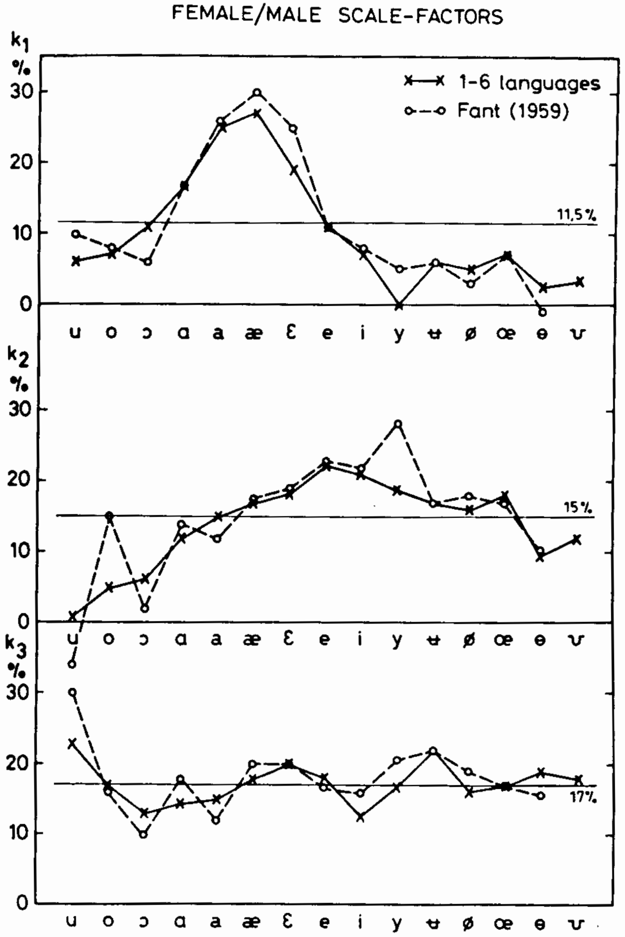

Figure 14. Female/male scale factor variation with vowel and the particular formant (Fant, 1975a).

articulation. It is also disputable whether this particular female articulation serves to ensure an acceptable [i] or whether there is a dialectal trend towards [ɩ]. Also, it is to be noted that X-ray tomography may impede the naturalness of articulations because of the abnormal head position required.

Much remains to be studied concerning how the vocal tract area functions of males, females, and children are scaled in actual speech and what kind of compensation occurs for minimizing perceptual differences or maybe the reverse, to mark contrasts between age and sex groups.

The lack of reference data on area functions is severe and the attempts to overcome this lack by means of area function scaling performed by Nordström (1975) were not conclusive except to support the general issue that the vowel and formant specific female-male differences, documented by Fant (1975a), Fig. 14, do not always come out as a result of the particular scaling assumed. The agreement was good for F_3 and fair for F_2 and rather bad for F_1. The predictability of F_3 is expected in view of the high dependency of F_3 on length dimensions.

A weakness in the Nordström study is that his [æ] and [ɛ] vowel area functions were interpolated from the Russian [ɑ] and [e] vowel and accordingly attain a centralized quality not representative of the [a] and [æ] category vowels which normally display a very large female-to-male F_1 ratio, see Fig. 14.

It is interesting to note that the non-uniform differences between females and males are paralleled by similar patterns comparing tenor and bass male singers. These vowel and formant specific trends are not only the automatic consequence of different anatomical scalings but also reveal compensations according to criteria that are not very well understood yet. A promising project on vocal tract modeling from anatomical data, now carried out at MIT (Goldstein, 1979), should provide us with fresh insight in female, male, and child differences.

From Goldstein's still unpublished graphs of vocal tract outlines I have noted that the length of the pharynx measured from the glottis to the roof of the soft palate grows from 3.3 cm in the newborn child to 7.6 cm for the female aged 21 and 10 cm for the male aged 21. The length of the mouth measured from the back wall of the upper pharynx to the front teeth (alveolar ridge for the newborn infant) grows from 5.5 cm for the newborn infant to 8 cm for the female of 21 and 8.5 cm for the male of 21. The tendency of relatively small variations of mouth cavity length with sex and age is more apparent than anticipated from earlier studies and would tend to minimize the range of "mouth cavity formant frequencies". The radical variations in relative pharynx length suggest that the relative role of front and back parts of the vocal tract could be reversed for a small child, i.e. that F_2 of the vowel [i] would be a front cavity formant, whilst F_3 is more dependent on the shorter back cavity. When front and back cavities are of more equal length, the dependency is divided and the F_3/F_2 ratio smaller than for males, which is typical of females or children of an intermediate age.

The inverse transform

As noted already in the introduction, there has been a substantial amount of theoretical work directed towards the derivation of area functions from speech wave data. In practice, however, these techniques are limited to non-nasal, non-obstructed vocal productions and the accuracy has not been great enough to warrant their use in speech research as a substitute for cineradiographic techniques. In the following section I shall attempt to comment on some of the main issues and problems. The usual technique, e.g. Wakita (1973), is to start out with a linear prediction (LPC) analysis of the speech wave to derive the reflection coefficients which describe the analog complex resonator. The success of this method is dependent on how well the losses in the vocal tract are taken into account. Till now the assumptions concerning losses have been either incomplete or unrealistic. Also the processing requires that the source function be eliminated in a preprocessing by a suitable deemphasis or by limiting the analysis to the glottal closed period. In spite of these difficulties the area functions derived by Wakita (1973; 1979) preserve gross features.

In general, a set of formant frequencies can be produced from an infinite number of different resonators of different length. We know of many compensatory transformations, such as a symmetrical perturbation of the single-tube resonator. However, if we measure the input impedance at the lips (Schroeder, 1967) or calculate formant bandwidths, we may avoid the ambiguities. A technique for handling tubes with side branches has been proposed by Ishizaki (1975).

According to Wakita (1979), the linear prediction method is capable of deriving an area function quantized into successive sections of equal and predetermined length providing the LPC analysis secures an analysis equivalent to M formants specified in terms of frequency and bandwidth.

An estimation of the total length and of the area scale factor require additional analysis data. An incorrect length estimate automatically generates compensatory changes in the area function which may be appreciable.

LPC analysis is a simple and powerful method of analysis but it fails in naturalness of representing the production process and as such is a poor substitute for a lossy transmission line representation. With the fresh eyes of a non-expert on the inverse transform, I would attempt to make the following suggestions. One is that M formants with associated bandwidths could have a greater predictive power than noted by Wakita. The area scale factor could be included in addition to the 2M relative areas of his model. In general, with reservation for possible uniqueness problems, 2M formant parameters - including bandwidths but not necessarily as many bandwidths as frequencies - would suffice for predicting 2M independent area function parameters.

Thus, adding one more formant frequency to the M pairs of frequencies and bandwidths would suffice for estimating the total length of the 2M system. Alternatively, from the 2M formant measures, we could derive a model quantized into M equivalent tubes each specified by cross-sectional area and specific length, thus also predicting the total length, Fig. 15. The rationale for this reasoning is that all losses in the transmission line analogs are unique functions of the area and length dimensions. One could also design a three-parameter model of the vocal tract as in Fig. 15 with a constant larynx tube. The four parameters (lip parameter A_1/l_1, x_c and A_c, and the total length) would hopefully be predictable from a specification of F_1, F_2, and F_3 and a bandwidth, say B_3, which appears to be more discriminating than B_1 and B_2. If we omit the total length and sacrifice the bandwidth, we have approached the articulatory modeling of Ladefoged et al. (1978)

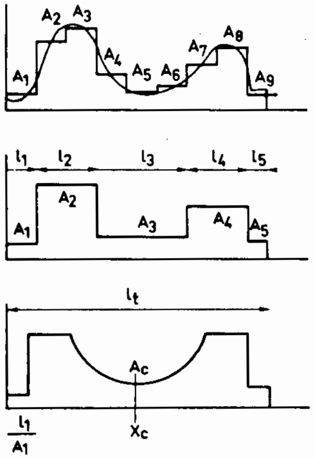

Figure 15. Continuous area function approximated by a constant larynx tube and 8 sections of equal length (top), by 4 sections of variable length and area (middle), and by a three-parameter model extended to include the total length (bottom). The constriction coordinate is zero at the lips.

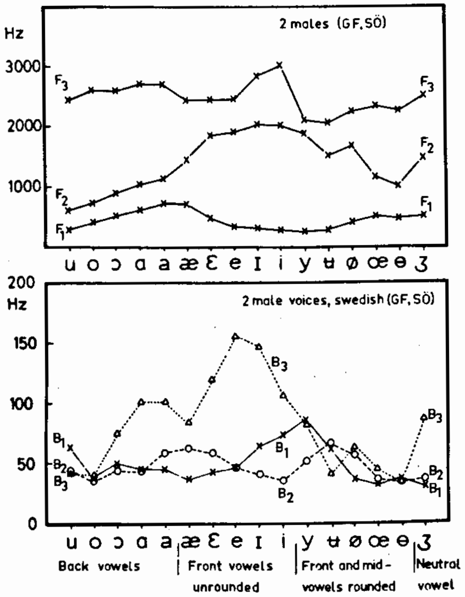

Figure 16. Frequency and bandwidth patterns of Swedish vowels (Fant, 1972).

which is based on correlational methods for deriving three articulatory parameters from F_1, F_2, and F_3.

In general, bandwidths have less predictive power than frequencies. They are to some extent predictable from formant frequencies (Fant, 1972), Fig. 16. Furthermore, bandwidths vary with speaker, voice effort, and laryngeal articulations and are inherently difficult to measure.

Still, I do not want to rule out the use of bandwidths. The following examples may serve to illustrate their predictive power and limitations. First, a test of the uniqueness in predicting 2M area function parameters from 2M formant data. Take the simple case of M=1 which implies a single tube resonator. What are the length and cross-sectional area of a tube with a specified first resonance frequency and bandwidth? The length is immediately given by $F_1=c/4l$. As shown in Fig. 17 the area is a single-valued function of bandwidth providing only one loss element is postulated (as in LPC analysis). If we include both the internal surface losses of a hard-walled tube and the radiation resistance, the bandwidth versus area attains a minimum at 10 cm^2 and there are two alternative areas that fit the same bandwidth. The higher value could possibly be ruled out as being outside the possible range of human articulation. Similar ambiguities could also be expected in a more complex lossy transmission line model, as pointed out by Atal et al. (1978). However, one should note that their treatment of the invariance problem is not quite fair. They introduce more articulatory parameters than acoustic descriptors which obviously exaggerate the ambiguities. Next consider a two-tube approximation of the vocal tract, Fig. 18 (A), with a back tube of length 8 cm and area 8 cm^2 and a front tube of length 6 cm and cross-sectional area 1 cm^2. The formant frequency pattern of F_1=275 Hz, F_2=2132 Hz, F_3=2998 Hz, F_4=4412 Hz and all higher formants is exactly the same as that of a two-tube system with the same areas but the lengths reversed, i.e. a front tube of length 8 cm and a back tube of length 6 cm (Fig. 18 B). This length ambiguity rule is apparent from the expression for resonance conditions

$$\frac{A_2}{A_1}\,\mathrm{tg}\,\frac{\omega l_1}{c} \times \mathrm{tg}\,\frac{\omega l_2}{c} = 1 \qquad (5)$$

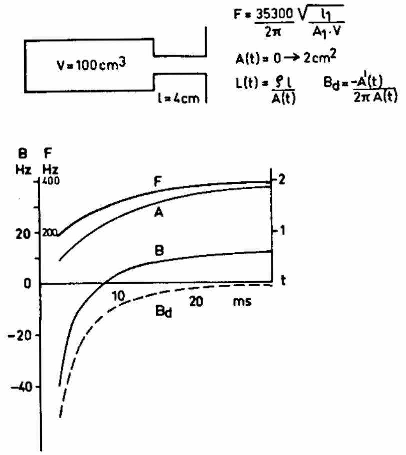

Figure 17. Bandwidth versus area of a single tube resonator taking into account internal losses and radiation load losses.

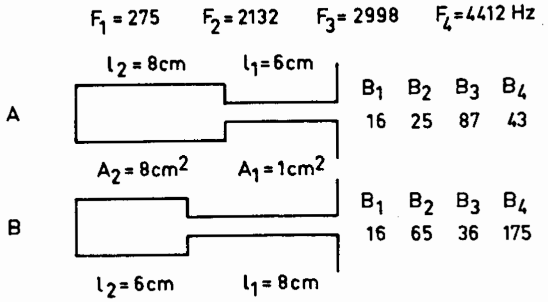

Figure 18. Two twin-tube resonators that provide the same F-pattern appropriate for the vowel [i], differing in terms of bandwidths.

If bandwidths are calculated taking into account both the interior surface losses and the radiation resistance by formulas given by Fant (1960), we find that B_2 and B_4 of Fig. 18 (A) are relatively low compared to B_3. In Fig. 18 (B), B_2 and B_4 are large compared to B_3. The different bandwidth patterns resolve the ambiguity. The physical explanation is that F2 and F4 of the first model are essentially determined by the back cavity and by the front cavity in the second model. The high damping associated with the surface losses in the narrow tube and the radiation resistance affect B_3 of (A) and B_2 and B_4 of (B).

The two models do not differ in terms of B_1. Theoretically it would be possible to choose the correct l_1, l_2, A_1, A_2 of the two-tube model from a specification of F_1, F_2, F_3 and either B_2 or B_3 or the ratio B_2/B_3 or B_4 or some combination of B_4 and other bandwidths, e.g. $(B_2+B_4)/B_3$. In a real speech case the situation might be different if the glottal losses are large and execute high damping of the back tube resonances.

In practice it may take a ventriloquist to produce something similar to these two models. Possibly the one with a shorter back tube would fit into the vocal tract anatomy of a very small child, as suggested in the previous section.

In conclusion - to improve techniques for inferring vocal tract characteristics from speech wave data we need a better insight in vocal tract anatomy, area function constraints, and a continued experience of confronting models with reality - a balanced mixture of academic sophistications and pragmatic modeling.

References

Atal, B.S., J.J. Chang, M.V. Mathews, and J.W. Tukey (1978): "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique", JASA 63, 1535-1555.

Fant, G. (1960): Acoustic theory of speech production, The Hague: Mouton (2nd edition 1970).

Fant, G. (1965): "Formants and cavities", Proc.Phon.5, 120-141, Basel: Karger.

Fant, G. (1966): "A note on vocal tract size factors and non-uniform F-pattern scalings", STL-QPSR 4, 22-30.

Fant, G. (1972): "Vocal tract wall effects, losses, and resonance bandwidths", STL-QPSR 2-3, 28-52.

Fant, G. (1973): Speech sounds and features, Cambridge, Mass.: MIT Press.

Fant, G. (1975a): "Non-uniform vowel normalization", STL-QPSR 2-3, 1-19.

Fant, G. (1975b): "Vocal-tract area and length perturbations", STL-QPSR 4, 1-14.

Fant, G. (1976): "Vocal tract energy functions and non-uniform scaling", J.Acoust.Soc.Japan 11, 1-18.

Fant, G. (1979): "Glottal source and excitation analysis", STL-QPSR 1, 85-107.

Fant, G. and S. Pauli (1975): "Spatial characteristics of vocal tract resonance modes", in Proc. Speech Comm. Sem. 74: Speech Communication, Vol. 2, G. Fant (ed.), 121-132, Stockholm: Almqvist and Wiksell.

Fant, G., K. Ishizaka, J. Lindqvist, and J. Sundberg (1972): "Subglottal formants", STL-QPSR 1, 1-12.

Fant, G., L. Nord, and P. Branderud (1976): "A note on the vocal tract wall impedance", STL-QPSR 4, 13-20.

Fant, G. and J. Liljencrants (1979): "Perception of vowels with truncated intraperiod decay envelopes", STL-QPSR 1, 79-84.

Flanagan, J.L. (1965): Speech analysis synthesis and perception, Berlin: Springer (2nd expanded ed. 1972).

Flanagan, J.L., K. Ishizaka, and K. Shipley (1975): "Synthesis of speech from a dynamic model of the vocal cords and vocal tract", Bell System Techn. J. 54, 485-506.

Fujimura, O. and J. Lindqvist (1971): "Sweep-tone measurements of vocal-tract characteristics", JASA 49, 541-558.

Goldstein, U. (1979): "Modeling children's vocal tracts", JASA 65, S25(A).

Guérin, B., M. Mrayati, and R. Carré (1975): "A voice source taking into account of coupling with the supraglottal cavities", Rep. from Lab. de la Communication Parlée, ENSERG, Grenoble.

Heinz, J.M. (1967): "Perturbation functions for the determination of vocal-tract area functions from vocal-tract eigenvalues", STL-QPSR 1, 1-14.

Ishizaka, K., J.C. French, and J.L. Flanagan (1975): "Direct determination of vocal tract wall impedance", IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP-23, 370-373.

Ishizaki, S. (1975): "Analysis of speech based on stochastic process model", Bull. Electrotechn. Lab. 39, 881-902.

Jospa, P. (1975): "Effets de la dynamique du conduit vocal sur les modes de résonances", Rep. de l'institut de phonétique, Université Libre de Bruxelles, 51-74.

Ladefoged, P., R. Harshman, L. Goldstein, and L. Rice (1978): "Generating vocal tract shapes from formant frequencies", JASA 64, 1027-1035.

Lindblom, B. and J. Sundberg (1969): "A quantitative model of vowel production and the distinctive features of Swedish vowels", STL-QPSR 1, 14-32.

Lindqvist, J. and J. Sundberg (1972): "Acoustic properties of the nasal tract", STL-QPSR 1, 13-17.

Mrayati, M. and B. Guérin (1976): "Etude des caractēristiques acoustiques des voyelles orales françaises par simulation du conduit vocal avec pertes", Revue d'Acoustique 36, 18-32.

Mrayati, M., B. Guérin, and L.J. Boë (1976): "Etude de l'impédance du conduit vocal - Couplage source-conduit vocal", Acustica 35, 330-340.

Nordström, P.-E. (1975): "Attempts to simulate female and infant vocal tracts from male area functions", STL-QPSR 2-3, 20-33.

Öhman, S.E.G. and S. Zetterlund (1975): "On symmetry in the vocal tract", in Proc. Speech Comm. Sem. 74: Speech Communication, Vol. 2, G. Fant (ed.), 133-138, Stockholm: Almqvist and Wiksell.

Schroeder, M.R. (1967): "Determination of the geometry of the human vocal tract by acoustic measurements", JASA 41, 1002-1010.

Sidell, R.S. and J.J. Fredberg (1978): "Noninvasive inference of airway network geometry from broadband long reflection data", J. of Biomedical Eng. 100, 131-138.

Sondhi, M.M. and B. Gopinath (1971): "Determination of vocal tract shape from impulse response at the lips", JASA 49, 1867-1873.

Stevens, K.N. (1971): "Airflow and turbulence noise for fricative and stop consonants, static considerations", JASA 50, 1180-1192.

Stevens, K.N. and A.S. House (1955): "Development of a quantitative description of vowel articulation", JASA 27, 484-493.

Wakita, H. (1973): "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms", IEEE Trans. Audio and Electroacoustics, AU-21, 417-427.

Wakita, H. (1979): "Estimation of vocal tract shapes from acoustical analysis of the speech wave: the state of the art", IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-27, 281-285.

Wakita, H. and G. Fant (1978): "Toward a better vocal tract model", STL-QPSR 1, 9-29.

DISCUSSION

Hisashi Wakita, Raymond Descout and Peter Ladefoged opened the discussion.

Hisashi Wakita: In determining the interrelationship between speech articulation and acoustics, we are particularly interested in the inverse problem, i.e. the estimate of vocal tract shapes from the acoustic waveform. There are various uncertain factors in deriving vocal tract area functions from the waveform, but it is an attractive method, because it is both the safest and easiest. (The problem with recent articulatory models for vocal tract shaping is that we do not yet know the exact parameters that control vocal tract shapes, in terms of articulators, and we do not have sufficient methods to obtain the data.) One of the most promising methods is the linear prediction (LPC) method, to estimate area functions from acoustic data. We do not know to what extent we can describe the details of the vocal tract shape, but by combining the LPC method with physiological data, we hope to improve this method.

One problem is the non-uniqueness, i.e. we can generate an infinite number of shapes having exactly the same frequency spectrum within a limited frequency band. To solve the uniqueness problem we have to impose constraints, physiologically determined constraints, or constraints determined by the higher harmonic structure. So far, the LPC method has been using formant frequencies and bandwidths, and in fact the final area function is sometimes quite sensitive to bandwidth. But we would like to get rid of bandwidth in the calculations: From the first three formant frequencies we can obtain the midsagittal view of the vocal tract, like in the Peter Ladefoged model, and to get at the unique shape of this midsagittal area function we may employ physiological constraints.

Another problem with LPC analysis is the vocal tract excitation and the losses, both within the vocal tract and at its boundaries, and these problems have to be solved in order to get more accurate vocal tract shapes. In fact, with the LPC method we can detect the closed glottis portion, where the interaction between sub- and supraglottal cavities is minimized, which makes for more accurate area functions. A further draw-back of LPC is that we have to start from the very simple assumptions of a simple loss at the glottis and a lossless acoustic tube. On the other hand, you can make a production model as complex as you wish, - you can add any realistic losses along the vocal tract or at the glottis that you like, but as analysis model there is a strong limitation in incorporating losses and other factors. So at this moment, the imminent problem is how to attack the loss problems and the source uncertainties.

Raymond Descout: Very little original data has accumulated on area functions, because collecting it is difficult, from a technical point of view. On the other hand, deriving vocal tract area functions from acoustic data has some disadvantages: with LPC techniques we only get pseudo area functions, and with acoustic measurements, which I previously worked on, there is a great problem in dynamic measurements, especially. Further, interest has largely centered on the midsagittal view of the tract, but we need information about the frontal view as well, which may be obtained with the new techniques of computerized tomography. We need this information in order to turn the midsagittal view into a three-dimensional area function, and to determine the shape factors that are necessary for the introduction of losses in our models.

All the articulatory models proposed are based upon vowel configurations, and when we try to make dynamic simulations on the articulatory model, everything that we do not know about the consonants is put into a special coarticulation and transition rule. We need more information on the consonants.

The acoustic model of the vocal tract is derived from the propagation equations, based on assumptions of symmetrical, equal length sections, - but to do an inverse transform you really need a very appropriate model which includes the shape factors that are necessary for the loss calculations, because the mathematical technique involved in the transformation is stupid in the sense that the result will be adjusted according to mathematical criteria, but this may not result in a realistic vocal tract. Therefore, I think that doing inverse vocal tract transforms is premature: we must work first of all on the proposition of the best production model, including shape factors and losses, before trying to do inverse vocal tract transforms.

Due to the progress made in articulatory modelling and to the limitations of LPC-techniques, we have witnessed a come-back of studies on vocal tract and vocal source simulations. To refine the articulatory model, we need further physiological data.

In conclusion: I do not think that LPC will give us a better understanding of speech production (it is, however, excellent for synthesis purposes). We need more studies on the relationship between articulatory parameters / area functions / vocal tract shapes.

Peter Ladefoged: Gunnar Fant showed us many years ago that what is important in characterizing speech are the first three formant frequencies, and you can even get a great deal of a speaker's personal quality with just three formant frequencies. But with the inverse transform, to get as far as eight tubes (which is only a coarse model of the vocal tract), you need at least four formant frequencies and their bandwidths, and with eighteen tubes you need nine formant frequencies and bandwidths, etc. Now something is wrong here: any phonetician can draw, more or less accurately, the midsagittal view of a given speaker's vowels, and we ought to be able to develop an algorithm that will go from the acoustics to the tract shape. There are of course problems - we do not actually observe the tract shape, only the midsagittal dimensions, and there are only very limited sets of data that tell us how to derive the tract shape from the sagittal dimension.

The work of Lindblom and others has shown that you can produce an [i:] with your jaw in a more or less open position, i.e. one has the ability to control tract shapes using different articulatory procedures, and it is of great interest to us to know how we exert that control and less interesting what the muscles do. Eventually, we have got to be able to go from acoustic structures, finding out what the tract shape is, and then deducing from that what the underlying control signals must have been.

Gunnar Fant: I agree with the main points of the discussants. Inverse transforms cannot make up for our great lack of physiological reference data.

My suggestions for improving inverse transform techniques in part supported by the previous discussions are: (1) we should model the vocal tract in terms of lossy transmission line sections instead of the simplified LPC model, (2) we should not expect to generate a larger number of independent production parameters than we have independent and well specified speech wave descriptors relating to the vocal tract transfer function. Overspecified area functions are necessarily non-unique, whereas a balanced specification can be, but need not be, unique. With proper model and parameter constraints, a 32-section area function model may be generated from a set of 3-6 articulatory parameters and controlled by the same number of acoustic parameters. It remains to be seen if we can extract more than four independent acoustic parameters. (3) The vocal tract total length should be derivable from one extra independent acoustic parameter.

Our discussion concerning bandwidths is still rather academic and we appear to share a doubt concerning the specificational value of bandwidths. Theoretically the set F_1 F_2 B_1 B_2 could suffice to specify a three-parameter model extended with a fourth parameter, e.g. the total length. This might hold for a resonator model only but not for a true vocal tract with less predictable bandwidth sources and the limited accuracy in bandwidth measurements. A more efficient set of acoustic parameters would be F_1 F_2 F_3 and B_3. From my Fig. 16 illustrating bandwidths of Swedish vowels it is seen that B_3 is a good correlate of degree of lip opening and also mouth opening. However, vowel bandwidths including B_3 are to a high degree predictable from formant frequencies. The role of bandwidths in an LPC model is not the same as that of a true vocal tract model. This is an important distinction. The LPC bandwidths, e.g. B_3, may come out quite different from those of real speech or from simulations by an improved model. The bandwidths we need for the inverse LPC based transforms are the bandwidths of a production model which has losses at the glottis only and locks the cavity wall shunt. From the true formant frequencies and bandwidths we thus have to make a best guess of what bandwidths the LPC model would generate. This is in the line of the recent work of Hisashi Wakita (1979).

Kenneth Stevens: With regard to what a male speaker does in order to compensate relative to the [u:] of a female: if we define narrow vowels as having so narrow a constriction that turbulence is just not generated, is it conceivable then that males, who generate a greater air flow than women, cannot round the vowels as much as can women, and therefore the formants are not lower than those of women?

Gunnar Fant: It could be, but in Swedish the vowel [u:] as well as [i:], [y:], and [ʉ:] are generally produced, by males and females alike, with a diphthongal glide passing through a relatively constricted phase in which some turbulence may be generated. I would rather expect different male and female articulations to be aimed at some criterion of perceptual invariance of which we do not know too much yet.

Antti Sovijärvi asked Gunnar Fant what his concept is about nasalized vowels.

Gunnar Fant: An essential characteristic of nasalization independent of the specific resonances added is the reduced F1 amplitude which is especially apparent in an oscillographic analysis. What appears to be a sub-F1 nasal formant is often a voice source feature which is relatively re-inforced because of the F1 reduction.

Hisashi Wakita: As long as the calculations are based on the first few formant frequencies, the problems in inverse transformation are rather equivalent with different methods. To uniquely determine a six tube vocal tract shape, LPC uses the first three bandwidths. If you want a smooth area function, you have to specify one of the higher frequency characteristics, and to do that you have to impose some kind of constraint, which is what Dr. Ladefoged does. And whatever the method, if you do not want to use bandwidth, you have to use some other kind of information to uniquely determine the spectra, and any information will do as long as you are able to reconstruct the original spectrum with its original bandwidths - so bandwidth is in fact a very important parameter.

Gunnar Fant: It would be interesting to see how far you would get if you started out with F1, F2, and F3 and then predicted B1, B2, and B3 from the formulas that I have.

Peter Ladefoged: I have tried using Hisashi Wakita's formulae with Gunnar Fant's type of predicted bandwidths (and other bandwidths from the literature), and it did not work, - I got absolutely impossible vocal tract shapes. Regarding Atal's vocal tract shapes that produce identical formant frequencies: some of them are quite impossible, the tongue just cannot produce some of those shapes.

John Holmes: I wish to emphasize the difficulty of mathematically deriving the vocal tract from the speech waveform, because we know too little about the glottal source. Gunnar Fant emphasized that the closed glottis portion is better suited than the open glottis portion to work out the supraglottal characteristics, but (as can be seen on the Farnsworth vocal chord movie of about 1940 and from Tom Baer's work), even when the vocal chords are closed there is sufficient ripple and surface movement for there to be an effective volume velocity input into the vocal tract, which means that your resultant waveform is never a force-free response, - and this is one of the things that makes bandwidths so difficult to estimate, because it is quite possible that ripple in vocal chord surface could actually be causing the formant amplitude to be still building up even, in exceptional cases, during the closed glottis period. I think this supports the view that we have to work from much more basic information and use articulatory constraints rather than to derive vocal tracts by purely mathematical techniques from some artificial and unrealistic production model.

Gunnar Fant: I can only agree with your statements. It is necessary to learn more about the human voice source in order to improve our methods of inverse transforms.

Osamu Fujimura: We can obtain cross-sectional vocal tract shapes with the regular computerized tomography, but only at great costs, because the X-ray dosage is tremendously high, a requirement of brain diagnoses that demand a very good density solution.

But I think the machine can be adjusted and the X-ray dosage reduced for our purposes, where we are really only interested in the distinction between matter and air.

Mohan Sondhi at the Bell Laboratories has proposed an acoustic impedance measurement using an impulse-like excitation at the lips, which can give us complete information about the area function of the vocal tract, because we obtain two sets of infinite series, i.e. the poles and the zeroes of the impedance function that together uniquely determine the vocal tract shape, without having to assume or measure losses. I think that there is one major difficulty with this technique: the subject articulates silently, i.e. he has no auditory feed-back, and we cannot be sure about the actual gestures. That problem can be overcome if we simultaneously monitor the vocal tract with e.g. the X-ray micro-beam method.

Gunnar Fant: The micro-beam system will certainly provide us with excellent data about speech articulation, but will it provide us with all the details that we want about the vocal tract, like the exact dimensions of the pharynx and larynx cavities?

Osamu Fujimura: We can obtain data on cross-sectional shapes, because we can place pellets also outside the midsagittal plane, - the only constraint being that we cannot use too many pellets at the same time, which will increase the X-ray dosage, but it is not easy to place pellets on the pharyngeal walls, which is a limitation of the method. However, we have a new stereo-fiberscope which can be used for three-dimensional optical observations of the pharynx, and I hope in the future to be able to develop a technique that will supplement the X-ray technique with this kind of optical information.

Raymond Descout: I am presently working with a prototype CT (computerized tomography) scanner, which scans in five seconds, and we are trying to lower the X-ray dosage to ten percent the normal dosage, because all we need is to see the difference between air and flesh. There is still a problem with the CT technique, though, and that is determining exactly the position of the slice relative to the skin and the rest of the person.