Special Lectures

THE RELATIONS BETWEEN AREA FUNCTIONS AND THE ACOUSTICAL SIGNAL

Gunnar Fant, Dept. of Speech Communication, Royal Institute of Technology (KTH), S-100 44 Stockholm 70, Sweden

Vocal-tract modeling

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, even with respect to the possibility to track a side branch of the vocal tract.

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 has 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. These will be dealt with in a separate paper.

An example

The area functions of male and female articulations of the Swedish vowels [i] and [u] and corresponding computed resonance mode pattern in Fig. 1 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 by Fant (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). 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 articulation. It is also disputable whether this particular female articulation serves to ensure an acceptable [i] or whether there is a dialectal trend towards [ɪ]. Also, it is to be noted that X-ray tomography may impede the naturalness of articulations because of the abnormal head position required.

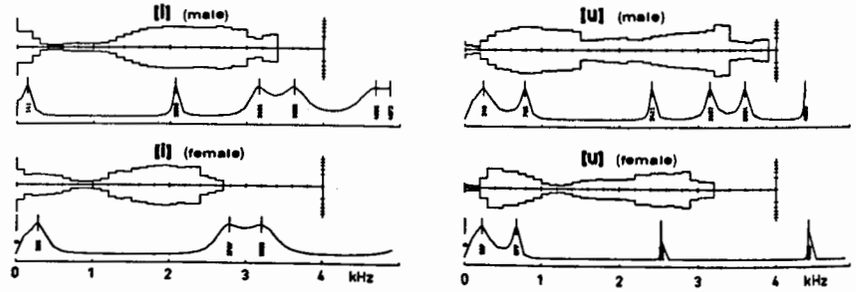

Figure 1

Multicylinder representation of VT area functions of male and female vowels [i] and [u] together with corresponding F-pattern. The shunting effect of sinus piriformis and of cavity walls is not included.

Perturbation theory and the inverse transform

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 $\delta F/F$ caused by a perturbation $\delta A(x)/A(x)$ is referred to as a "sensitivity function". We may also define a perturbation $\delta\Delta x/\Delta x$ 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 (1974), 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.

The distribution of the sum of the kinetic and potential energies describes the sensitivity to length scale perturbations and provides furthermore a realistic quantitative measure of the dependency of the resonance mode on various parts of the area function. Length perturbation has been applied to the problem of scaling the pharynx and the mouth differently, comparing male and female articulations, Fant (1975b).

If the perturbation 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 (1974). This technique has been used by Mrayati and Guérin (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 step.

I shall not go into details of the mathematics of the inverse transform. 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) 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).

The following very general discussion of the inverse transform is based on a lossy transmission line representation of each section of the area function. The approach is similar to that of Atal et al. (1978).

It can be shown that a number of m formants, specified by their frequencies and bandwidths potentially define a unique area function with 2 m degrees of freedom providing that the resistive elements that determine the bandwidths are unique functions of frequency and of the resonator configuration. It follows that given any total length of an area function, it can be quantizied in 2 m sections of equal length and there could exist a unique solution for the 2 m area values. The non-uniqueness of the overall length may be overcome by adding one more formant to the specification.

Another solution which is unique with respect to vocal-tract length is a configuration of a cascade of m cylindrical tubes, each specified by area and length derived from the m formant frequencies and bandwidths. We can exemplify this model by the single-tube resonator. Its length determines the lowest resonance frequency and the area is determined from the bandwidth measure. An F-pattern with F_1 = 260, F_2 = 1990, and F_3 = 3050 Hz appropriate for the

vowel [i] would be generated by a two-tube system in which the back tube has an area of 8 cm^2 and a length 8.7 cm, and the front tube an area of 1 cm^2 and effective length 5.8 cm. The compensatory articulation with the same areas but exchange of lengths has exactly the same pattern of all formant frequencies, Fant (1960), but a different bandwidth pattern. A minimum of two frequencies and two bandwidths would theoretically suffice for a unique derivation of either configuration. In practice it may take a ventriloquist to produce both variants. Possibly, the variant with short back cavity would fit the shape of a child's vocal tract. Other aspects of front-back compensations have been treated by Öhman and Zetterlund (1974).

On the whole, we are free to choose any parametric specification to fit a continuous area function providing the number of parameters is twice the number of formants specified in both frequency and bandwidth. Unless the total length is a unique function of the parameters we need one more formant to be specified. We could thus construct a four-tube model with or without smoothing between sections to be uniquely defined by four frequencies and four bandwidths. A combination of this technique with specific constraints, such as a fixed larynx tube, may be introduced to concentrate the predictive capacity to other parts of the system.

This simple reasoning has potentialities to be exploited more than has been done. In practice, however, as pointed out by Atal et al. (1978), we might find that bandwidths may come out the same in two alternative configurations or that their difference may turn out to be smaller than what we can accurately measure. Some additional redundancy could be introduced to overcome such difficulties.

A lack of bandwidth measures can generally not be compensated for by introducing more formant frequencies. On the other hand, if we resort to an articulatory model with natural constraints on possible area functions we may base the prediction on formant frequencies alone, Lindblom and Sundberg (1969), Ladefoged et al. (1978). However, the same pattern of, say, F1 F2 and F3 could generate somewhat different area functions in other models with other constraints, e.g. in terms of a different overall length. A combination of formant frequencies, bandwidths and articulatory constraints should be optimal.

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, 's-Gravenhage: Mouton (2nd edition 1970).

Fant, G. (1965): "Formants and cavities", in Proc. 5th Int. Congr. of Phonetic Sciences, Münster, E. Zwirner (ed.), Basel: Karger.

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

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

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

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

Fant, G. and S. Pauli (1974): "Spatial characteristics of vocal tract resonance modes", in Speech Communication, Vol. 2, G. Fant (ed.), 121-132, Stockholm: Almqvist & Wiksell 1975 (Proc. SCS-74, Stockholm).

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.

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

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 (KTH, Stockholm).

Mrayati, M. and B. Guérin (1976): "Etude des caractéristiques acoustiques des voyelles orales françaises par simulation du conduit vocal avec pertes", Revue d'Acoustique No. 36, 18-32.

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

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.

Öhman, S.E.G. and S. Zetterlund (1974): "On symmetry in the vocal tract", in Speech Communication, Vol. 2, G. Fant (ed.), 133-138, Stockholm: Almqvist & Wiksell 1975 (Proc. SCS-74, Stockholm).