Chapter 1
The Acoustics of Speech

The acoustics of speech includes in a broad sense both the theory of speech as wave motion and how speech waves are produced and heard. This is a field of study which has intrigued researchers of various specialities during the last centuries and it has ancient traditions.

Classical phonetics has been and is still articulatory phonetics dealing with an inventory of speech sounds defined from their production within the vocal tract. The speech research of communication engineers is more concerned with the speech wave which we will define by the sound pressure variations at a point in front of the speaker.

With modern sound recording and analysis techniques it is possible to undertake rather complete specifications of the speech wave. However, a maximally detailed description is unmanageably complex and the great problem is to find useful approximation. The physiology of the speaking mechanism on the other hand cannot be studied and described with the same exactness. When it comes to hearing, there are even less possibilities to make complete specifications. The neurophysiology of speaking and hearing are the least accessible links of the complete communication system but they carry the key to many interesting problems.

The following presentation concentrates on the structure of speech waves and the theory of speech production.

SPECTROGRAPHIC ANALYSIS

The "Visible Speech" spectrographic techniques [1], introduced by the Bell Telephone Laboratories some fifteen years ago, are still our most important means of studying the characteristics of speech waves. The most useful records are the well-known spectrograms with time in horizontal direction, frequency in vertical direction, and intensity of time-frequency bounded areas displayed by the relative blackness or brightness of the picture marking.

The spectrograms of Fig. 1 were obtained with the Sona-Graph-analyzer which is a commercial development of the original Bell Telephone Laboratories speech spectrograph. This is a heterodyne analyzer with a fixed filter of alternative 45 c/s or 300 c/s bandwidth. A piece of speech maximally 2.4 sec long is analyzed by repetitive analysis with frequency increments of 15 c/s between successive closed loop repetitions of a stored piece of speech. A doubling of the broad or narrow bandwidths can be accomplished by the trick of replaying the speech material from a tape-recorder to the Sona-Graph storage loop at half the normal speed as is exemplified in Fig. 1. Adjustments have been made in the frequency scale in order to retain the same frequency scale (expanded) as in the normal speech processing.

This article originally appeared in *Proceedings of the Third International Congress on Acoustics, Stuttgart, 1959*, edited by L. Cremer (Amsterdam: Elsevier Publishing Company, 1961). Reprinted with permission.

The overall intensity as a function of time has to be recorded by means of supplementary instrumentation to the spectrograph, in the form of an amplitude display curve on the same sheet as the spectrogram or as a separate display on an oscillograph[2].

The spectral distribution of intensity or energy within a specific short time interval of the speech wave is defined by an intensity (db amplitude) *vs.* frequency curve. A

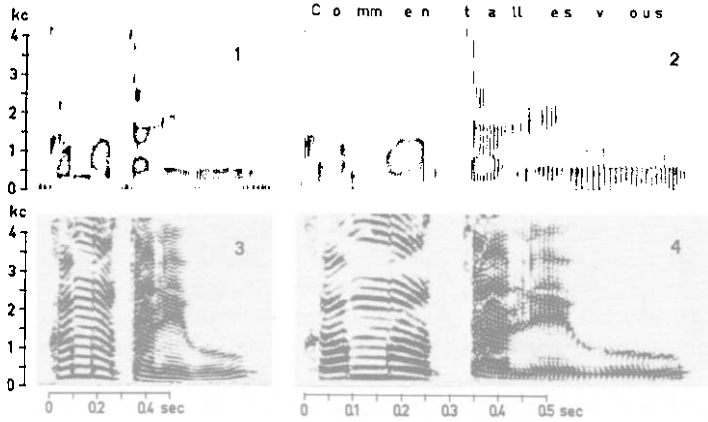

Fig. 1. Time-frequency-intensity spectrograms illustrating the effects of various analysis bandwidths 300 c/s in 1, 600 c/s in 2, 45 c/s in 3, 90 c/s in 4.

spectrum section of this type may be produced on a spectrograph by synchronous sampling of the separate frequency channels. In case the sound to be analyzed is produced in a sustained form it may be convenient to utilize a sweep-frequency method of analysis. The spectra of Fig. 2 pertain to synthetic and human vowels each of 3 sec duration analyzed by means of a filter of 32 c/s width moving at constant speed of 1.3 kc/s through the frequency range of 0–4000 c/s.

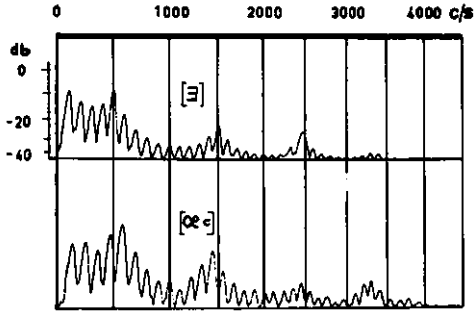

Fig. 2. Harmonic spectra obtained from narrow bandwidth sweep frequency analysis of sustained sound œ, and of the synthetic reference sound [ɜ].

Vowels and other voiced sounds possess periodic or rather quasi-periodic wave forms and accordingly display harmonic spectra. This fine structure originates from

the opening and closing movements of the vocal cords periodically modulating the volume of the exhaled air during phonation at a rate of F_0 c/s, which is the voice fundamental frequency[3, 4]. In narrow-band spectrograms F_0 is the harmonic spacing and in broad-band spectrograms $1/F_0$ is the time interval between successive striations each reflecting a single voice cycle. The time variation of F_0 is the physical basis of intonation.

The train of successive airpulses emerging from the vibrating glottis is the primary source of voiced sounds. The air cavities within the vocal tract act as a multiresonant filter on the transmitted sound and impress upon it a corresponding formant structure superimposed on the harmonic fine structure. This can be clearly seen in Fig. 2. The frequencies of the three lowest formants, F_1, F_2, F_3, are the main determinants of the phonetic quality of a vowel.

The resonance frequencies of the vocal tract F_1, F_2, F_3, F_4, conceptually contained in the term *F*-pattern, vary more or less continuously across the often sharply time localized breaks in the spectrographic time-frequency-intensity picture. Such breaks may for instance indicate shifts from voice to noise source or vice versa. Each position of the articulatory organs has its specific *F*-pattern. Some ambiguities do exist due to compensatory forms of articulation but these are not very important in normal speech. The time-variation of the *F*-pattern across one or several adjacent sound segments, which may be referred to as the *F*-formant transitions, are often important auditory cues for the identification of a consonant supplementing the cues inherent in the composition of the sound segments traditionally assigned to the consonant.

In general, the continuous elements of speech are due to the continuity of the position of the articulators. The discrete breaks are mainly due to a shift in manner of production, that is a change in type of source (fine structure), or a radical change in the active resonator system through which the sound is filtered (open/closed mouth passage with and without a lateral or a nasal by-pass of the sound). A sudden shift in the *F*-pattern and in the overall intensity following the step from a closed to an open mouth passage may thus be regarded as a discontinuity.

Spectrographic pictures convey an overflow of data which are non-essential for descriptive purposes. This redundancy is in part a matter of interrelations, repetitions, and continuities within the signal structure, in part the presence of a fine structure the details of which carry very little or no information. Any description of the speech wave, for speech typewriter coding purposes or for speech bandwidth compression applications or merely for the study of acoustic correlates to phonetic categories, must be based on approximations. Binary coded pattern aspects as well as quantized parameter data belong to the inventory of such specifications.

When processing the spectrographic data on connected speech the first object is to identify the boundaries of successive sound segments. A sound segment generally carries information on more than one phoneme of a sequence. Conversely, each phoneme may be physically encoded to a smaller or greater extent in the pattern aspect of several adjacent sound segments. The number of successive sound segments of a piece of connected speech is generally larger than the number of phonemes. Stop sounds, for instance, can be considered to be made up of at least two typical sound segments, the occlusion and the burst, and the latter phase may in some instances be split up into three successive and partly overlapping phases, the explosion transient, a short fricative, and an h-sound. The description of a sound segment for the purpose

of identification may be based on the following parameters, previously mentioned and summarized below.

1. Duration
2. Intensity
3. Energy (Area under the intensity-time curve)
4. Voice fundamental frequency, F_0
5. The F-pattern ($=F_1, F_2, F_3, F_4$, etc.)
6. The formant structure (Frequency-intensity distribution)
7. The fine structure; referring to speech production, the source (Voiced, unvoiced, mixed, or silence)

In addition there enter the dynamical aspects of speech patterns[5] in terms of the time variation of each of the variables 2-7. The identification of a phoneme from the physical data contained in successive sound segments involves first a phonetical categorization, essentially with regard to "manner of production", and then within each category, a choice related to "position of articulators", for instance the choice of one of [b], [d], [g] when the phoneme has been identified as a voiced stop.

The techniques of automatic speech recognition are still in an initial phase of development. Instrumental problems are severe and specificational theory is not fully established. The main difficulty in any speech writing coding scheme[6] is the variability of human speech. However, this area of research is developing rapidly.

THEORY OF SPEECH PRODUCTION

Acoustic theory of speech production[7,8] in its present form is largely based on equivalent circuit concepts. As visualized in Fig. 3 any speech sound is regarded as the filtered output of a network in which a sound source is inserted. The characteristics of any

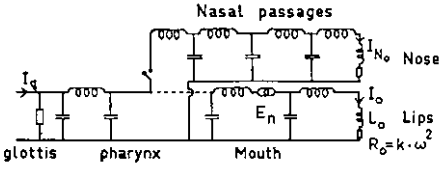

Fig. 3. Equivalent circuit representation of human and synthetic speech production applied to voiced sounds. The coils and condensers of the circuit should be regarded as distributed elements rather than lumped elements pertaining to specific cavities.

quasi-stationary sound segment thus contains the characteristics of the source and those of the network, the latter referred to as the vocal tract transfer function or filter function. In terms of Laplace transforms

$$P(s) = S(s)T(s) \tag{1}$$

where $P(s)$ pertains to the radiated sound. $S(s)$ to the source, $T(s)$ to the vocal tract transfer function, and $s = \sigma + j\omega$ to the complex frequency variable.

The transfer function $T(s)$ of voiced sounds is defined as the ratio of the Laplace transforms of the sound pressure at a distance l cm from the speaker to the volume velocity of the pulsating airflow passing the vocal cords. If the coupling to the nasal cavities is negligible this function has no other zero than that at the origin of the complex frequency plane. This differentiation approximates the transfer from volume

velocity at the lips to the sound pressure in the radiated wave. The ideal transfer function of voiced sounds,

$$T(s,l) = \frac{s}{4\pi l} \cdot \frac{1}{\prod_{n=1}^{\infty} (1 - s/\hat{s}_n)(1 - s/\hat{s}_n^*)} \tag{2}$$

is thus essentially an infinite pole product, where

$$\hat{s}_n = \sigma_n + j\omega_n \quad \text{and} \quad \hat{s}_n^* = \sigma_n - j\omega_n$$

are conjugate complex poles. For synthesis applications the infinite product is substituted for a finite (3, 4, or 5) number of poles and a "higher pole correction"[8].

The air-filled cavities within the vocal tract constitute a continuously inhomogeneous transmission line with low losses, and the equivalent network may thus be described in terms of the distributed series inductance and parallel capacitance per length unit along the vocal tract. Series and parallel resistances representing finite losses enter a complete representation. A lumped element representation of a series inductance for a constriction and a capacitance for the volume of a specific cavity is not permissible, except for very low frequencies.

Each resonance of the vocal cavities may be described in terms of its frequency F_n and bandwidth B_n which are related to the conjugate complex poles of $T(s)$ as follows

$$\left.\begin{aligned} F_n &= \omega_n/2\pi \\ B_n &= -\sigma_n/\pi \end{aligned}\right\} \tag{3}$$

The average spacing within the frequency scale of these resonances is of the order of 1000 c/s or more specifically $c/2l_v$ were l_v is the effective length of the vocal tract and c the velocity of sound. This inverse dependency of formant frequencies on vocal cavity length dimensions explains the higher formant frequencies of females compared to males, and of children compared to adults.

The two constituents of a pole, the frequency and the bandwidth, may be studied by various means of exciting the vocal cavities. One is merely to thump the outside of the throat with a finger and measure the damped exponential, the decay characteristics of which provide a measure of the bandwidth according to (3). The vocal tract response to any transient excitation must contain as a component a damped oscillation

$$p_n(t) = A_n\, e^{-\pi B_n t} \cdot \cos(2\pi F_n t + \varphi_n) \tag{4}$$

which is the inverse transform of a formant number n.

The same frequencies and bandwidths may be obtained from the sine-wave response of the vocal tract as determined experimentally from driving the vocal tract with a larynx microphone utilized as a sound source and a pickup microphone close to the lips. This is exemplified by Fig. 4. Typical values of resonance bandwidths are shown in Fig. 5. They are of the order of 50 c/s in the frequency region occupied by the first and second formant. Formant bandwidths are slightly greater than resonance bandwidths due to additional losses through the glottis slit.

The equivalent circuit theory of speech production suggests a convenient method[9] of deriving the properties of the vocal source[4] without bringing any probes into the vocal cavities. This is the inverse filtering technique of passing the speech wave

through anti-resonance circuits, one for each formant. The first step is to integrate the speech wave thus removing the radiation zero of $T(s)$.

Some results of this technique are illustrated in Fig. 6*. It may be observed that integration alone provides a first approximation to the voice flow. The apparent starting point of the damped oscillations appears to coincide with the offset of the flow, *i.e.* the closing phase. These curves display the well-known facts that increased

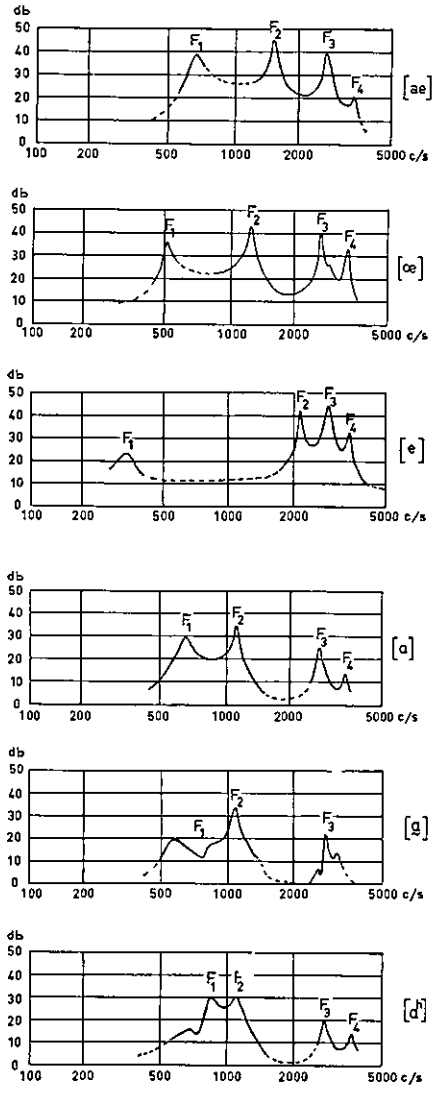

Fig. 4. Sine-wave response curves of the vocal tract driven externally from the pharynx and measured 2 cm in front of the lips. The effect of lowering the soft palate as in a nasalized vowel and of opening the vocal cords as in h-sounds is illustrated for the vowel [a].

* These illustrations of inverse filtering originate from a thesis work by C. CEDERLUND of the Speech Transmission Laboratory, Royal Institute of Technology, Stockholm (Sweden).

voice efforts sharpen the wave shape of the vocal airpulses. At low voice intensities the closure phase is relatively short and the wave form is rounded. A tendency of a double peaked voice flow period has been found for one of the subjects.

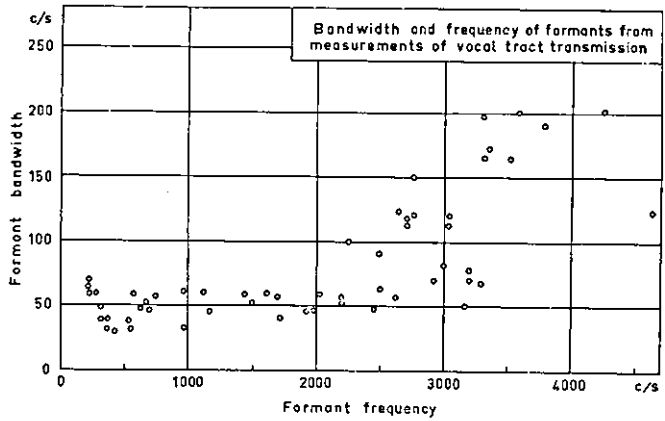

Fig. 5. Frequency dependency of the bandwidth of vocal resonances under conditions of closed glottis.

Another aspect of the Laplace transform representation is the frequency domain decomposition of vowels into elementary resonance curves. This is illustrated in Fig. 7, which pertains to idealized vowels. A shift of F_1 one octave up in frequency is apparently followed by an increase in the spectrum envelope level of 12 db at all frequencies well above F_1. When any two relatively close lying formants approach in frequency there occurs an increase in intensity of each which is 6 db per halving of their distance. These and other rules relating spectrum shape and spectrum levels to formant frequencies, *i.e.* to the *F*-pattern may be observed from Fig. 8, which illustrates the effects of changing F_1 and F_2 and also F_3 within the spectra of synthetic

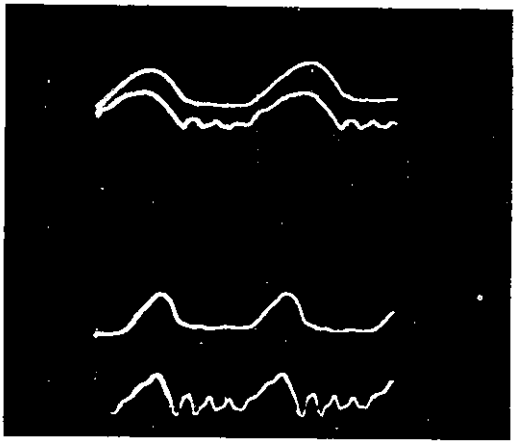

Fig. 6. Wave forms of the regenerated voice flow and, for comparison, the merely integrated speech wave. The upper pair of curves pertains to the vowel [æ] produced with a low voice effort and the bottom pair pertains to the same vowel produced with a high voice effort. The first four formants were filtered out in the top curve of each pair, but appear as damped oscillations in the merely integrated wave.

vowels. Most of these are close to Swedish vowels, the articulatory positions of which are shown in Fig. 9. Here as well as in Fig. 8 the vowels are arranged in terms of increasing F_1 to the right and increasing F_2 upwards in the diagram.

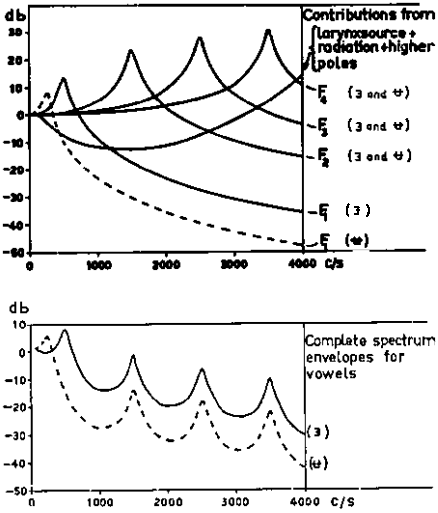

Fig. 7. Spectrum decomposition of ideal vowels in terms of elementary resonance curves, one for each formant plus additional constant characteristics. The latter include a voice source spectrum sloping —12 db/octave and a high frequency emphasis representing the residual contribution from formants higher than the fourth. The effect of shifting F_1 down one octave is indicated by the broken line. Each elementary resonance curve is analogous to a low-pass filter of 12 db/octave attenuation above its cutoff frequency.

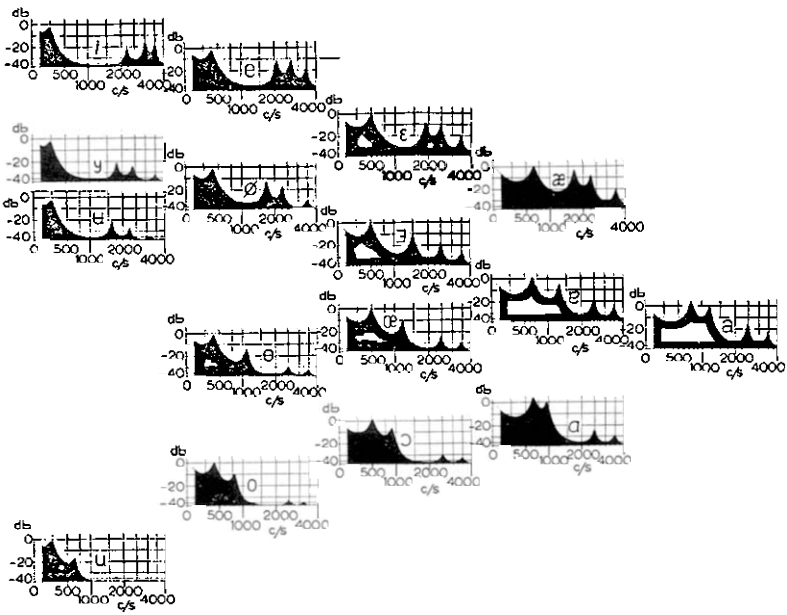

Fig. 8. Spectra on an approximate mel scale of synthetic vowels ordered according to the particular F_1 and F_2. The changes in spectrum shape and in formant levels following a shift in one or more of the formant frequencies should be observed.

The main articulatory variables are

1. the location,
2. the degree of constriction of the main narrowing between the tongue and the opposite wall of the vocal cavities, and
3. the degree of constriction and lengthening of the lip passage. The generalized relation suggested in older phonetics literature, that F_1 is due to the cavity behind the tongue constriction and F_2 to the cavity in front of the constriction is an impermissible oversimplification, sometimes contradicting actual relations. All parts of the vocal cavities have some influence on all formants and each formant is dependent on the entire shape of the complete system[7, 10]. The general rules are that a tongue constriction located in the middle of the mouth cavity is optimal for a high F_2 and that a maximally high F_1 requires the main constriction to be located just above the larynx and the mouth cavity to be wide-open. A constriction location slightly advanced from that of maximum F_2 provides maximal F_3.

A decrease of the lip-opening area or increase of the length of the lip passage

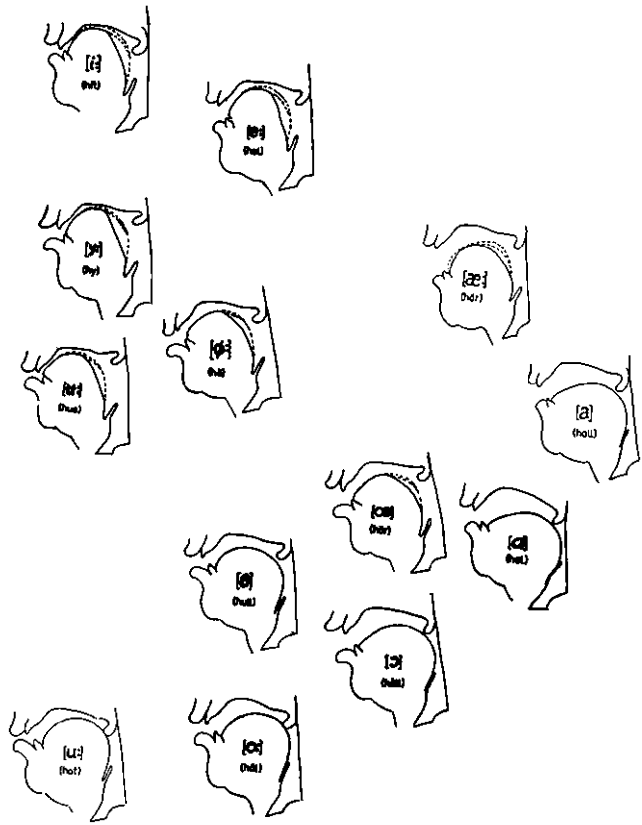

Fig. 9. X-ray tracings of Swedish vowels arranged as in Fig. 8 according to increasing F_1, right, and increasing F_2, up.

causes a lowering of the frequencies of all formants. F_1 is maximally low when the mouth cavity is constricted and F_2 is maximally low when the tongue constriction is in the upper part of the pharynx.

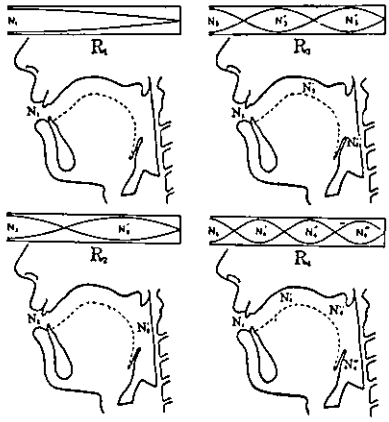

Fig. 10. Distribution of volume velocity at the frequencies of each of the first four resonances of an ideal neutral articulation in which the vocal tract simulates a tube of constant cross-sectional area. (After CHIBA and KAJIYAMA[12].)

All these relations observed when correlating articulatory and spectrographic data and corroborated model experiments[7, 11] may be inferred from a simple consideration of the distribution of pressure of volume velocity inside a neutral state idealized model of the vocal tract defined by a tube of constant cross-sectional area open at the lip end and closed at the larynx end. As shown in Fig. 10 there is a volume velocity maximum at the lips and a minimum at the glottis independent of the particular resonance frequency. The second resonance has an additional volume velocity maximum at $^1/_3$ of the vocal tract length above the glottis and a volume velocity minimum at a place $^1/_3$ of the total tube length from the lip end. The homogeneous tube has resonance frequencies at 500, 1500, 2500, 3500 c/s, etc.

If this neutral tube is constricted at a volume velocity maximum, *i.e.* at a pressure minimum, there results a shift down in the frequency of the particular resonance[12, 13]. This is to be expected since the minimum pressure at the constriction implies that the distributed capacitance in this region is small compared to the inductance and that an area change thus is effectively an inductance change[7]. The simple rule may thus be stated that a constriction of the vocal cavities at the place of a volume velocity maximum causes a shift down in the particular resonance frequency and that a constriction at a volume velocity minimum causes a shift up of the particular resonance frequency.

CONSONANT SPECTRA AND THEIR DETERMINANTS IN SPEECH PRODUCTION

Some of the sounds referred to as consonants, *e.g.* [j], [w], [v], [r], and [l] are often produced as voiced continuants with little or insignificant noise added. Sound segments of the speech wave belonging to this category, except [l] in a strict sense, may be analytically treated as vowels, *i.e.* formant intensities are only dependent on the *F*-pattern of resonance frequencies (poles) and on the particular source spectrum. In all other categories of sound segments, *i.e.* nasal consonants, unvoiced stops, fricatives and

affricates, and the unvoiced parts of the corresponding "voiced" sound segments, there enters in addition a 0-pattern of anti-resonances (zeroes) as an additional determinant of formant levels. This is also true of [l] but to a lesser extent.

The common denominator of all non-nasal speech sounds, voiced or not, is the *F*-pattern defining the frequencies where formants may be found. In some instances the bandwidths of the higher resonances are so broad that adjacent formants merge into a single formant area. Thus, typically for the [s]-sound, the formants *F*1, *F*2, *F*3, *F*4 are very weak and the main spectral energy is contained in *F*5, *F*6, *F*7, *F*8, *F*9 generally seen in the spectrogram as a single or two formant areas. In the sound segment of nasal consonants there are more resonances (poles) than those of the *F*-pattern in which case the *F*-pattern is defined to comprise those frequencies which show the greatest continuity with the *F*-positions of orally open, adjacent sound segments. The *F*-pattern of nasalized vowels is similarly defined from a continuity to non-nasalized sound segments disregarding those resonances introduced by the nasal coupling.

The zero function of a vocal tract transfer function is due to the cavity system behind the source (typical for all noise sounds) or to the presence of a cavity system in front of the source shunting the main path of wave propagation (typical for nasal consonants, nasalized vowels, and laterals).

When discussing the spectral effects of poles and zeroes, it is convenient to distinguish between free poles and zeroes and bound poles and zeroes, the latter comprising pairs of a pole and a close lying zero, providing small combined spectral contributions only. In a frequency region of low coupling between a front cavity and a back cavity or between a shunting cavity and a main outlet, all the poles of the back cavities or of the shunt are bound. This is the reason for the low intensity of formants lower than *F*5 in the spectrum of the [s]-sound and the low intensity of *F*2

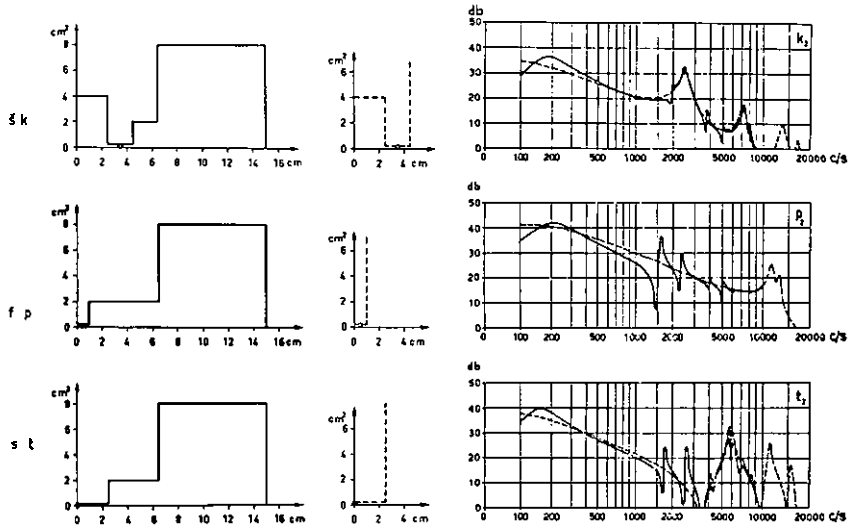

Fig. 11. Simple idealized vocal tract models and corresponding spectra of the sounds [k], [p] and [t]. Coarticulation with a front vowel is suggested by the relatively narrow mouth cavity. Broken line figures and curves pertain to models involving front resonators only and the corresponding spectra.

in the spectrum of most nasal consonants. The free poles are generally due to the cavities in front of the source and have a considerable association with the back cavities only when the coupling is great. When approximating a sound spectrum in terms of poles and zeroes it is evidently possible to discard the bound poles and zeroes. Those bound poles entering the *F*-pattern are, however, of specificational importance as the starting points of formant transitions towards adjacent sound segments.

Pole-zero patterns of palatal, labial, and dental sounds, derived from idealized cavity configurations, are shown in Fig. 11. The spectrum of the stop [k] is dominated by a free pole which is the first resonance of the front cavity. The labial consonant [p] has no free poles and no free zeroes unless the tongue is in a high palatal position, and the dental consonant [t] has a free zero in the region of 3000 c/s and a free pole at 6000 c/s originating from the narrow tongue passage which contribute to the relative emphasis of the spectrum above 4000 c/s. In natural speech there is an additional free pole originating from the resonance of the cavity in front of the teeth.

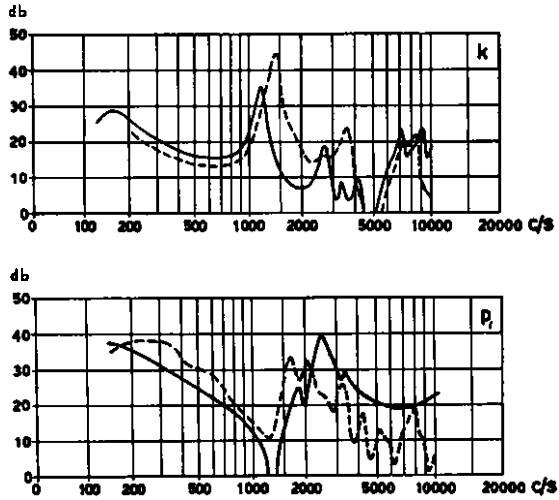

Fig. 12. Calculated spectra (solid lines) derived from X-ray studies of vocal tract dimensions, and spectra of the same sound sampled from the subject's connected speech (broken lines). The tongue articulation of [k] was postvelar and for [p] palatal. The sampling pertains to the first 10 msec of the explosion.

The accuracy which can be achieved in predicting the spectrum of a stop consonant of a subject's connected speech from X-ray pictures of his stationary articulation of the "same" sound is illustrated by Fig. 12 which pertains to the explosion phase of a [k]-sound of velar articulation and a palatalized [p]-sound. The predictability is good considering the apparent difficulties.

SPEECH SYNTHESIS*

Depending on how well the synthesis instrumentation preserves the general properties of speech, various levels of naturalness may be reached from very machine-like qualities to a rather natural sounding speech. Synthesis is made either to simulate a human model or to generate an impersonal speech by rule. Systematic synthesis experiments

* A review of synthesis methods may be found in ref. 2 (chapter 3.3).

are generally directed towards the evaluation of the relative importance of various pattern aspects. An important contribution to our understanding of the distinctive sound cues stems from the investigations at the Haskins Laboratories[14]. Their classical investigations were based on a constant pitch harmonic synthesizer, which provides a high degree of approximation of the speech wave. The results obtained from their studies should be checked by similar experiments with formant coded synthesizers capable of producing a more natural speech. Such studies are under way or are planned now in several laboratories in U.S.A., England, and in Sweden. Much of this work is directed to the realization of analysis-synthesis telephone systems enabling bandwidth reductions greater than those of a channel vocoder[15]. The signals extracted at the transmitting end and controlling the synthesis at the receiving end are of a parametric nature and have a low information rate. The formant coding implies an extensive use of Laplace transforms for the parametric decomposition. One example of such a synthesis scheme is shown in Fig. 13.

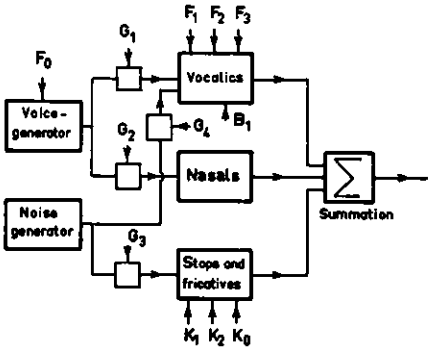

Fig. 13. Block diagram of the Swedish formant coded speech synthesizer OVE II.

Formant coded speech synthesis may be adopted as a supplement to analysis for expressing speech spectra in terms of pole-zero patterns providing a synthetic copy approximating the natural sample. This procedure has been called analysis by synthesis and is a promising approach to descriptive problems.

REFERENCES

1. R. K. Potter, A. G. Kopp and H. C. Green, *Visible Speech*, Van Nostrand, New York, 1947.
2. G. Fant, *Acta Polytechnica Scand.*, 246 (1958) 1.
3. J. W. van den Berg, *J. Speech and Hearing Research*, 1 (1958) 227.
4. J. L. Flanagan, *J. Speech and Hearing Research*, 1 (1958) 99.
5. H. M. Truby, *Acta Radiol.*, (1959) Suppl. 182.
6. D. B. Fry and P. Denes, *Language and Speech*, 1 (1958) 35.
7. G. Fant, *Acoustic Theory of Speech Production*, RIT Div. of Telegraphy-Telephony, Report No. 10, 1958; to be published by Mouton & Co., 's-Gravenhage, 1960.
8. G. Fant, *Ericsson Technics*, 15 (1959) 3.
9. R. Miller, *J. Acoust. Soc. Am.*, 31 (1959) 667.
10. H. K. Dunn, *J. Acoust. Soc. Am.*, 22 (1950) 740.

11. K. N. Stevens and A. S. House, *J. Acoust. Soc. Am.*, 27 (1955) 484, and 28 (1956) 578.
12. T. Chiba and M. Kajiyama, *The Vowel, its Nature and Structure*, Kaiseikan Publ. Co., Tokyo, 1941.
13. G. Ungeheuer, *Z. Phonetik allgem. Sprachwiss.*, 11 (1958) 35.
14. A. M. Liberman, P. Delattre and F. S. Cooper, *J. Acoust. Soc. Am.*, 29 (1957) 117.
15. M. R. Schroeder, *Recent Progress in Speech Coding at the Bell Telephone Laboratories*, this book, p. 201.