AURAL STIMULI AND THEIR INTERPRETATION

H. MOL

Already the title of this paper draws the attention to the fact that in speech aural stimuli call for *interpretation.* The comparatively recent development of electroacoustics has provided the investigators with the technical means to prove that the aural stimuli do not directly label the phonemes of the words understood by the listener. The mechanism of recognizing spoken words and sentences is no doubt voice-operated but it may not be regarded as the acoustic counterpart of an electrical teleprinter which prints letters that are unambiguously labelled by the electric signals it receives from a transmission line. When, in a teleprinter system, a key on the keyboard of the sending typewriter is pressed a normalized electrical signal corresponding to that key is generated and sent along a line to the receiving machine to print one special letter, to wit the letter corresponding to the signal and the pressed key on the sending machine. The receiving teleprinter has no choice, it acts as a slave and no interpretation is expected.

Engineers who try to construct so-called voice-operated typewriters and speech recognition systems experience that the process of speech recognition is not analogous to the teleprinter system. In spite of the hopeful announcements in the popular press a voice-operated typewriter can perhaps be made to react reasonably to its masters voice but it fails to cope with the variety of voices with which a human listener has no difficulties. The machine needs a brain.

The following examples, selected from recent publications, illustrate the absence of the much hoped for analogy with the teleprinter.

In the almost classic measurements of Peterson and Barney[1] on vowels 76 talkers pronounced 1520 isolated words containing 10 American vowel types. The first two formants of every vowel were measured and plotted in the wellknown way of plotting F 2 horizontally and F 1 vertically. They found 10 targets around which the realisations spread but although the talkers pronounced isolated words there was considerable overlap between the targets. When the 1520 words, recorded on magnetic tape, were presented to a jury of 70 listeners only about one half of the words were unanimously correctly recognized.

By carefully selecting a new jury this fraction could be raised to about two thirds.

[1] G. E. Peterson and H. L. Barney, *JASA*, 24 (1952), 175.

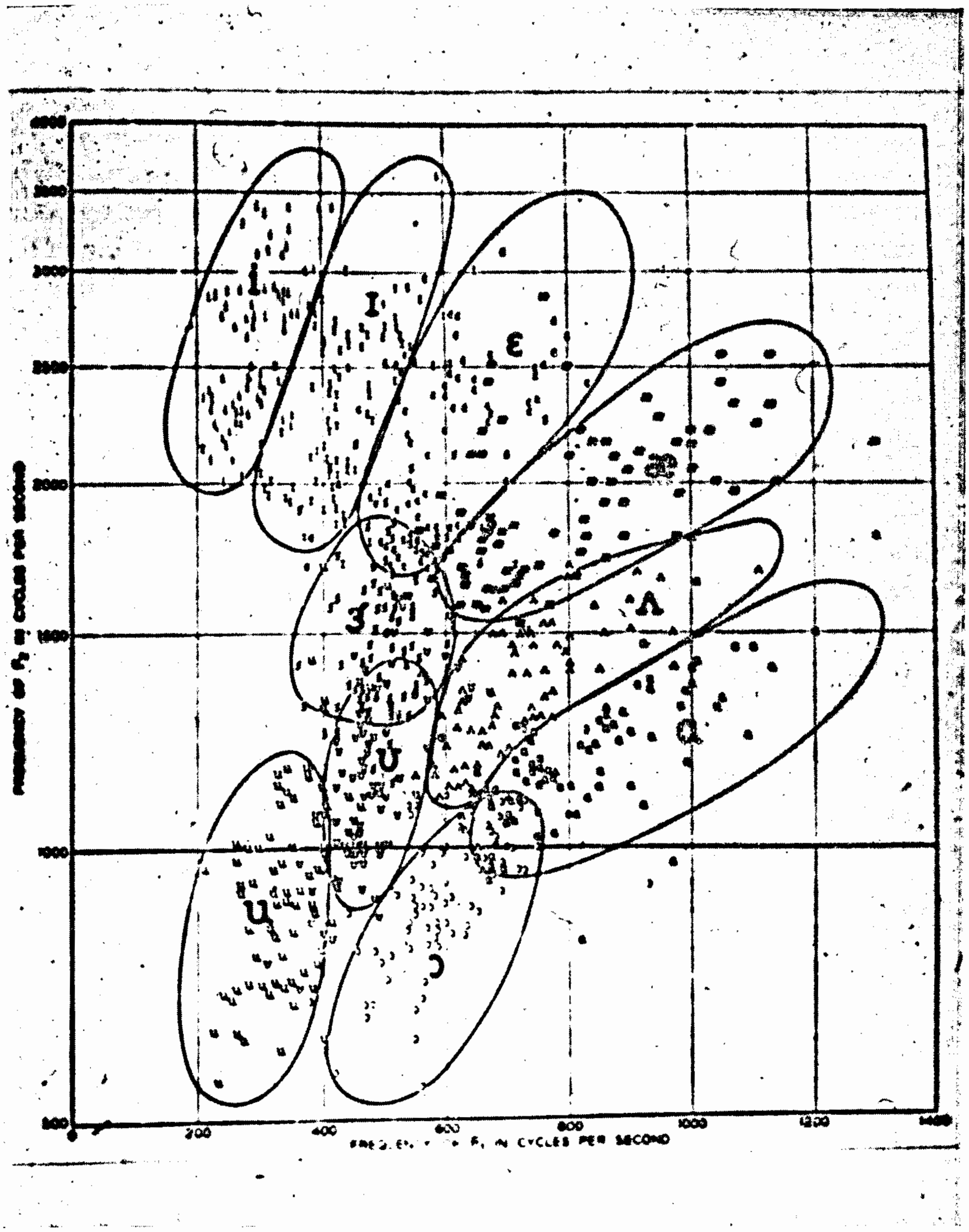

Fig. 1. Frequency of second formant versus frequency of first formant for 10 vowels by 76 speakers (Peterson and Barney).

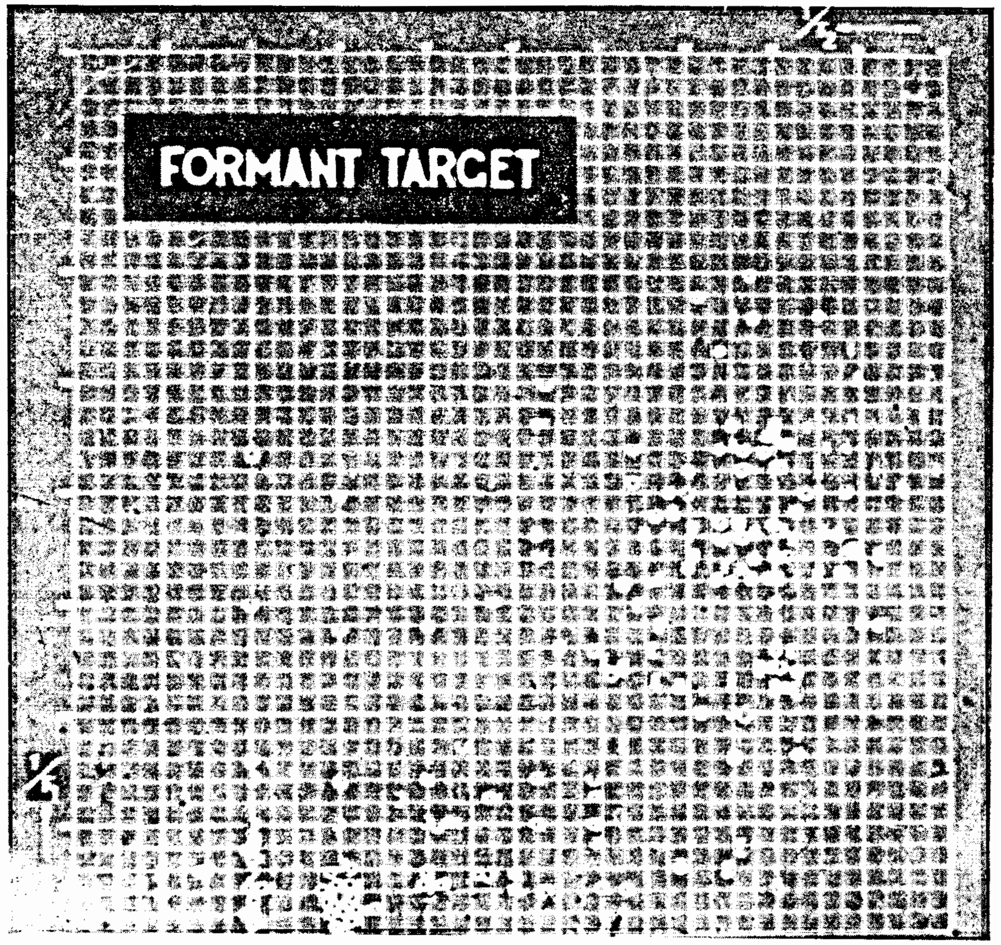

Fig. 2. Period of second formant versus frequency of first formant for 12 Dutch vowels appearing in a simple, freely pronounced sentence. The data of this pilot investigation refer to 15 talkers taken at random from a group of 100 talkers. The numbers on the axes represent periods, in harmony with the fact that the formants were indeed measured as periods and not as frequencies.

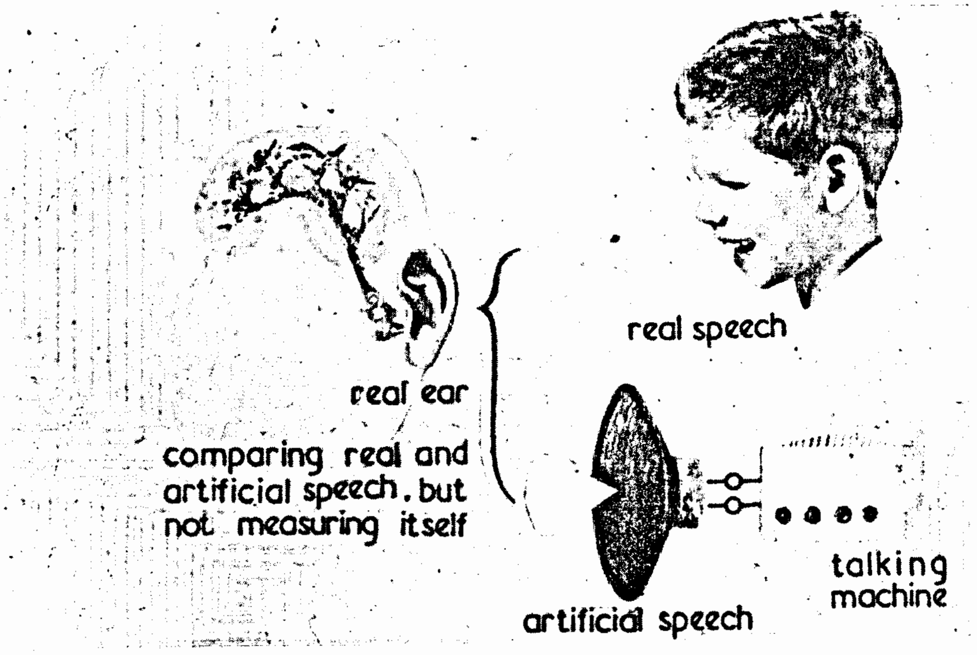

Fig. 3. Schematic representation of the process of setting the controls of a talking machine.

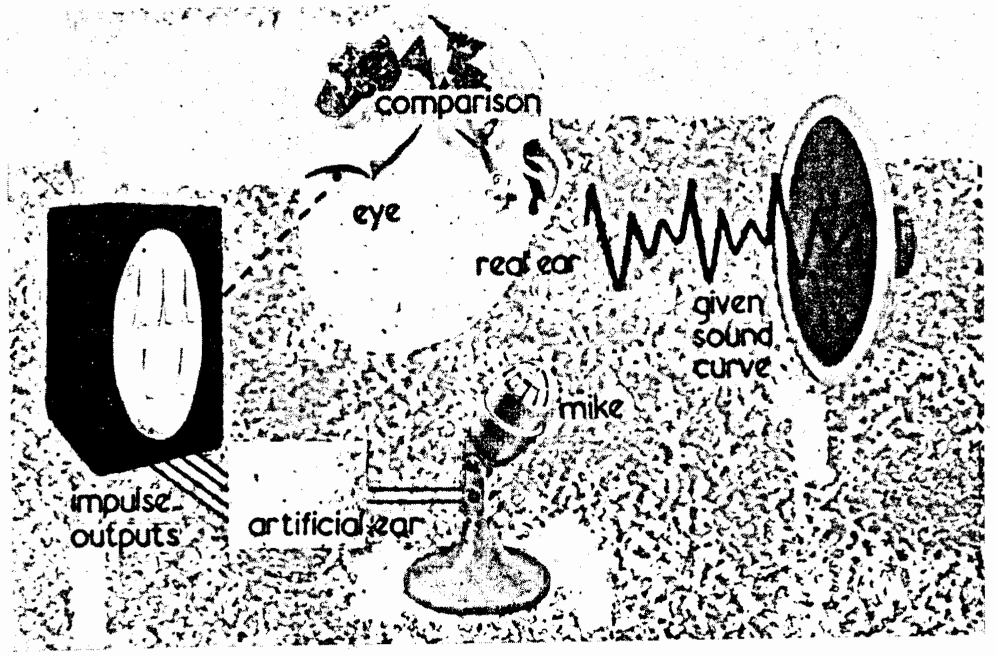

Fig. 4. A method for studying the mode of action of the ear by using a second sense organ. It must be emphasized here that the action of the brain is *highly* schematized in this figure, but that does not reduce the value of the method.

These experiments show that even for isolated monosyllables there was uniformity neither among the talkers nor among the listeners.

Ladefoged and Broadbent,[2] in 1956, showed that the recognition of a synthesized word could be influenced by the phonetic character of a likewise synthesized, introductory sentence. This experiment confronts us with the amazing ability of a listener to adapt himself to the articulatory habits of a speaker in a very short time.

In 1957, Blom and Mol plotted the formants of 12 Dutch vowel types appearing in a simple sentence freely pronounced by a large number of talkers.[3] For this current speech they did not find 12 targets but instead only 2 targets, showing that the talkers, when left to themselves, reduced distinction to a bare minimum. One target contains the vowels [i], [I], [ü], [ə], [ö], [e] and [ɛ]. The other target embraces the vowels [a], [ɑ], [o], [ɔ], and [u].

The lesson of all those investigations is that the phonic data the ear extracts from the sound waves do not form the only source on which the listener bases his identifications.

We can also derive a warning from the just mentioned and similar investigations that it is dangerous to draw conclusions on the mode of action *of the ear itself* from hearing tests in which speech is used as a stimulus. There is always the problem of separating the passive action of the ear from the interpretative activity of the brain. Let us try to start from scratch in order to get a clear picture of the problems involved.

The sound waves produced by the articulators and, in fact, all sound waves, are transformed by the ear into nervous quantities suitable for processing and interpretation by the nervous system. This transformation is performed in the organ of Corti by a very big number, more than say 23000, of surprisingly simple sensory elements that can only say yes or no when driven by the excursions of the partition containing them. Galambos,[4] in his excellent review of 1954, also drew the attention to the simplicity of the sensory elements. Traditional theory attributes an important selective power to the mechanical and hydrodynamical action of the cochlear partition and perilymphe. It speaks of a spatial spread along that 35 mm long partition. But the parrots and other speaking birds, many of which display a remarkable capability of producing fully recognizable speech, have only a rudimentary cochlea with a partition of merely 3 mm. Furthermore they do not have a clear distinction between external and internal haircells like the mammals display. The fact that some birds can speak throws doubt on the necessity of a long cochlea for the processing of speech and indicate that the simple sensory elements themselves can perform some type of analysis.

[2] P. Ladefoged and D. E. Broadbent, "Information conveyed by vowels", *JASA*, 29 (1957), pp 98–104.

[3] H. Mol, *Belief and superstition in phonetics* (in Dutch). Inaugural address Amsterdam 1960 (The Hague, 1960).

[4] Robert Galambos, "Neural Mechanisms of Audition", *Physiological Reviews*, Vol. 34, no. 3 (July 1954), pp. 497–528.

Since the outstanding work of Tasaki[5] we know that a sensory element can *at best* induce a nerve impulse in a single fibre of the auditory nerve every time the partition containing the organ of Corti passes its position of equilibrium in the direction of the oval window. In fact this is the wellknown technique of indicating zero-crossings in a sound wave.

In *Lingua*, Uhlenbeck and Mol[6] described how this simple mechanism allows formant extraction indeed. We did not, however, present a complete and definite description of the mode of action of the ear but merely showed the tremendous achievements of simple mechanisms in order to pave the way to a better understanding of the mechanism of hearing.

In a recent issue of *Nature*[7] mr Broadbent, discussing the perception of pitch says: "one may argue that the pitch of a sound will depend not so much on the precise receptors stimulated (as traditional theory holds) as on some features of the message travelling up the nerve fibres". "A likely candidate, mr Broadbent suggests, is the frequency of impulses in the auditory nerve".

It is exactly this principle which is incorporated in the mechanisms described by Uhlenbeck and Mol.

The only safe way to discover the role of the organ of hearing plays in the phonemic classification of sound waves is to perform experiments on real ears with the aim of measuring the nervous activity in the fibres of the auditory nerve for a given sound wave.

A given sound wave is defined here as a sound wave of which the time-function is exactly known.

Up to now most experiments have been performed with pure tones which are exactly known time-functions indeed. As the production of a nerve impulse by a sensory element is a highly non-linear affair, however, the reaction of an element to a superposition of pure tones is not the sum of the reactions of the element to single tones. In other words: Fourier analysis fails here, the system being non-linear.

Therefore, in order to study the nervous activity evoked by vowel-like sounds we must stimulate the ear by means of a repeated damped oscillation the time-course of which is exactly known. We are not allowed, at this stage, to call this sound a vowel, because that would already be an interpretation.

It is evident that many investigators try to avoid the just-sketched procedure because it requires an advanced operation technique, the making and positioning of micro-electrodes, the means of making simultaneously visible the given sound wave and the impulses it evokes in the nerve fibres and perhaps other intervening phenomena.

Some investigators concentrate themselves on experiments with talking machines. Though talking machines may considerably differ in principle and construction they are all alike in that they possess *controls* that can be adjusted automatically by a programme or by hand.

The programme may be presented to the machine on punched cards, painted slides rotating belts etc. By trial and error, sometimes guided by the results of some form of analysis, one arrives at a "best" setting of the controls. The best setting is the setting for which a group of listeners displays the highest degree of agreement in classifying a synthesized sound as a given phoneme.

Very much, however, depends on the instruction and previous training given to the jury and on the number of phonemes from which the listeners are allowed to choose. When, for instance, the choice is one out of three even a deaf listener has a chance of 33⅓% for being correct. The maker of the talking machine should never be in a jury because, like all parents, he understands his own child better than anyone else does.

Now, properly speaking, with the ear as a *zero* instrument, we learn the properties of the talking machine rather than the properties of the ear. Though we are able to synthesize, in some way or another, a sound that a jury is willing to classify as one of a given number of phonemes, we have not thereby prooved that the ear has a direct access to the properties the controls master, not do we know whether the ear in actual speech uses other criteria not available on the machine. It is dangerous, therefore, to call the controls "the parameters of speech".

For studying the mode of action of the ear without performing operation we need a different set-up. We need an artificial ear, in the ideal case consisting of a sound input and many nerve outputs. By listening to a given sound wave at the same time seeing the impulses of the artificial ear on the screen of a cathode-ray oscilloscope we use the brain as an instrument for comparing the output of our acoustic nerve to the impulses of the artificial nerve fibres. When we hear a change in the given sound wave we can see on the screen what change in the pattern of impulses was at the root of that change.

The construction of the artificial ear should be based on what we really know about the physiology of the organ of Corti and not on wishful thinking. We should not coquet with our supposed ignorance on that subject and we must not take refuge in mathematical ways of describing speech sounds like, for instance, Fourier-analysis, just to do something. We must not misuse the spectrograph.

Let us first enumerate items in favour of the spectrograph.

A good spectrograph is a valuable technical achievement. Though it does not depict phase it is still very convenient for those who study the vocal tract by investigating how the latter reacts to sinusoidal waves, often called pure tones. The spectrograph is very helpful for comparing the output of an artificial vocal tract to that of the real mouth, especially when the artificial vocal tract is based on spectrographic principles.

We must not, however, apply the spectrograph in the study of the organ of hearing.

In this respect it is interesting to read what G. Fant states on page 160 of his excellent

[5] I. Tasaki, "Nerve impulses in individual auditory nerve fibers of guinea pig", *J. Neurophysiol.* 1954, 17, 97–122.

[6] H. Mol and E. M. Uhlenbeck, "Hearing and the concept of the phoneme", *Lingua*, vol. VIII, 2, (May 1959), pp. 161–185.

[7] D. E. Broadbent, "The perception of speech", *Nature*, vol. 189 (Febr. 18, 1961), no 4764, pp. 528–529.

book *Acoustic Theory of Speech Production*: "It has become the normal technique in speech analysis to illustrate acoustic sound quality attributes by spectral curves. However, oscillograms produced at a high paper speed may sometimes provide a comparable or even clearer insight into specific details of the signal structure".

After a critical study of what has been written about aural stimuli and their interpretation I came to the conclusion that progress in this field is retarded by superstitions.

For instance, more often than not one reads in papers and textbooks that frequency, intensity and time *are* the physical components or parameters of speech sounds, as if the Fourier components were physical realities indeed.

The majority of people saying loose things about the ear assert that it is seeking for those alleged physical components.

Now the only physical reality of a speech sound is the time function depicting the barometric pressure as a function of time. In its graphical form it is often called sound curve or oscillogram.

Physical realities are, for instance, the zero-crossings and peaks in a sound curve because they are events that really happen in the vibrating air at the entrance of the ear. The nerve endings, however, have no direct access to an airborne sound curve which has to penetrate into the ear until it reaches the organ of Corti where it is transformed into an electrical phenomenon that stimulates the nerve endings. No doubt during these adventures the shape of the curve changes but in a way that can be both predicted and measured nowadays. The nerve endings react to the zero-crossings in the ultimate curve by which they are stimulated. Also the performance of groups of nerve endings must be seen in this light. In other words, the whole pattern of nervous activity in the acoustic nerve is governed by the ability of the nerve endings to indicate zero crossings.

When, in the future, the art of predicting or actually measuring the nervous activity evoked by speech sounds has developed sufficiently, investigators will be in a better position. They will then be able to study the purely interpretative faculties of the listener apart from the actual nervous activity at his disposal.

It will also be possible then to settle the following question. As far as we can overlook at the moment the overlap between consonants is less outspoken than between the vowels because the articulatory freedom in vowels is greater. We have a hunch that in running speech the consonants form the cues for identification of the words rather than the vowels which, as it were, are filled in by the listener on basis of interpretation.

Summarizing, we can underline the following points.

The aural stimuli are those physical events in the air that can be transformed into patterns of nervous activity by the ear.

The nature of this transformation must be studied further by means of experiments on real ears involving the measurement of the nervous activity evoked by known

sound waves. In current speech the nervous activity does not yield a one-to-one description of the vowel phonemes intended by the speaker.

A human listener is superior to a machine in that he can interpret using his brain. Determination of the degree in which he has to use this faculty in practice is a problem that can only be solved by a close cooperation between linguists, psychologists, physiologists, and phoneticians.

Amsterdam
Leiden