CONTRIBUTION OF THE TIME PARAMETER TO THE PERCEPTION OF SPEECH

A. COHEN, J. F. SCHOUTEN, J. 't HART

1. INTRODUCTION

When looking at visual registrations of speech, be it in the shape of oscillograms, spectrograms or X-ray pictures, the phonetician is confronted with the notorious problem of dividing what presents itself as a continuum into relevant segments, individual speech sounds, i.e. the consonants and vowels of the language under consideration. Normal practice is for the linguist to superimpose a linguistic pattern, derived from a phonemic analysis, onto this speech continuum and try, as best he can, to establish a one-to-one relation with the patterns to be found in the visual registration.

Of late a number of acoustic studies have appeared to the effect that such a procedure is doomed to failure since, taking into consideration what is actually perceived by a listener, it turns out that he is often unable to identify unmistakably the acoustic representation of a single consonant or even vowel phoneme. Thus cues for the perception of plosive sounds have very convincingly been shown to subsist largely in the vowel portion, notably the bending of the vowel formants, the so-called transitions, of the syllable containing the plosive.[1]

We have deliberately chosen a different line of approach, away from scrutinizing visual patterns and consisting in carrying out an analysis by ear directly. The burden of deciding on the elements into which the speech continuum is to be divided is now clearly on the human listener, whose only qualification for this task is that he should be a native speaker of the language investigated.

First of all in the following section a device is described which can be used to obtain a segmentation in time. The perceptual segments thus obtained form the basis of the approach in synthesizing speech to be described in section 3, synthesis.

2. ANALYSIS

By means of an electronic gating circuit, a word spoken on a closed loop of tape can be made audible in gradually increasing portions, starting at the very beginning

[1] An excellent survey is to be found in P. Delattre, "Les indices acoustiques de la parole, Premier rapport," *Phonetica*, 2 (1959), pp. 108–118, 226–251.

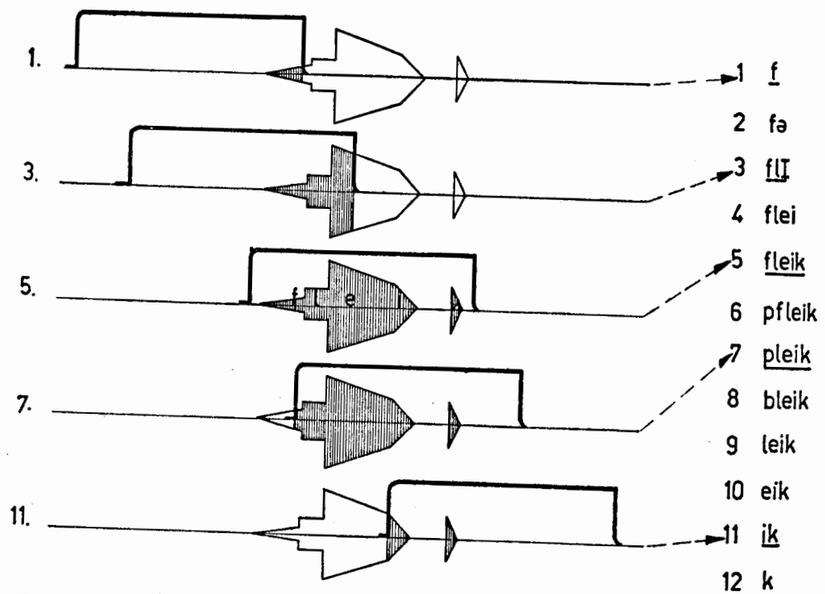

Fig. 1. Five phases of the phonetic analysis of the word *flake* by means of a moving gate, corresponding to No's 1, 3, 5, 7, 11 of the various stages to be heard.

of the first sound and building up in successive presentations until the complete word is being passed by the gate.[2]

Continuing this process, the very first part is cut off and subsequently larger and larger portions are suppressed until only the very last part of the word is left over.

Fig. 1 illustrates how this effect can be brought about, starting from a tape on which the English word *flake* was spoken, which is to be analysed in the way just described.

A number of stages can be distinguished, five of which (nrs. 1, 3, 5, 7, 11) are portrayed as stylized oscillograms. This method of analysis implies a segmentation in time which gives rise to some interesting results:

1. it takes a certain time for the first consonant /f/ to be perceived as such, the first few presentations constituting merely an undefinable hiss;
2. the first presentation of /l/ has a vowellike nature;
3. the first few presentations of the diphthong /ei/ are heard as [I], until, at sufficient length, a new element appears, contributing to an /ei/ perception; when the first part is suppressed it turns out that this new element consists of [i], in other words: no intermediate glides can be perceived;
4. when sufficiently shortened the /f/ loses its character and is heard as [p] and even [b].

[2] In principle this device resembles the one adopted by P. Menzerath, *Der Diphthong* (Bonn, 1939).

In other words, variations of the time parameter give rise to different perceptual judgements – shortened [f] e.g. is not just short [f] but [p]. This seems to warrant the hypothesis that two sounds, having the same colour, i.e. the same spectral components, may be perceived as radically different as long as the time cues differ.

3. SYNTHESIS

To prove the hypothesis that speech can be considered as a concatenation of perceptual segments, some of which may share spectral components and differ only in the time parameter, a number of electronic gating circuits were developed enabling the experimenter to vary at will and independently of one another the rise, duration and decay of an acoustic signal.

Fig. 2 shows the definition of the three time parameters t_1, t_2 and t_3, expressed in milliseconds (a), and four different settings of these parameters for the segments /p/, /f/, /I/ and /i/ (b-e). Synthetic sounds were produced by means of two sources: a periodic one for generating vowellike sounds and a noise source for the production of consonantlike sounds.

In order to obtain the segments to be used in building up a synthetic word, some spectral modification has to be conferred upon the sources: actually, two frequency regions for the vowel sounds and one frequency region for the consonant sounds are selected to provide sufficient colour. When presented in this shape they do not sound very much like speech sounds. It is possible, however, to make these rough sounding components play the part of speech sounds by giving them the required t_1, t_2, t_3-values, in other words their respective amplitude envelopes.

Building up a word means sewing these segments together by triggering the respective gates at the appropriate moments. It will be clear from this procedure that no formant bendings are applied, nor are additional glides of whatever kind introduced.

Fig. 3 (top) represents the overall time pattern required for producing the word *phonetics*. No attempt was made to indicate the relative amplitudes of the individual segments. The dotted line represents the place and the shape of the intonation contour.

When the decay of the /ɛ/ gate is lengthened, as indicated on the second line of Fig. 3 and with slight attenuation of the volume of the /t/ segment, the resulting time pattern gives rise to the perception of the American version [fənɛ:ʤIks].

The next line portrays the time and intonation pattern of the French word *phonétique*. For this purpose a slight change of the spectral components of the vowel sounds [ɛ] and [I] into [e] and [i] was desirable.

The bottom line renders the situation for German *Phonetik*, which differs from the previous pattern mainly in the length of the /e/ segment and the position of the intonation peak.

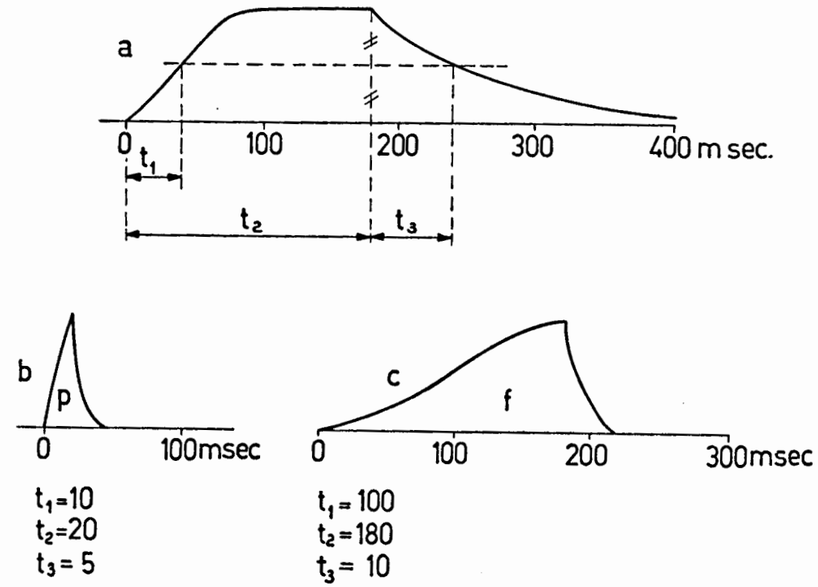

Fig. 2. *a*: Definition of the three time parameters t_1, t_2, t_3, shaping the envelopes of segments used in synthetic speech. *b–e*: Different values of t_1, t_2, t_3 used to synthesize a plosive sound (b), a fricative (c), a checked vowel (d) and a free vowel (e).

The various stages of the process of synthesis were demonstrated by means of a tape recording: two sources, spectral modifications by means of filtering, the application of the individual gate settings, the correct ordering in time of the segments necessary for building up the word *phonetics* as well as the three modifications of this word into American English, French and German.

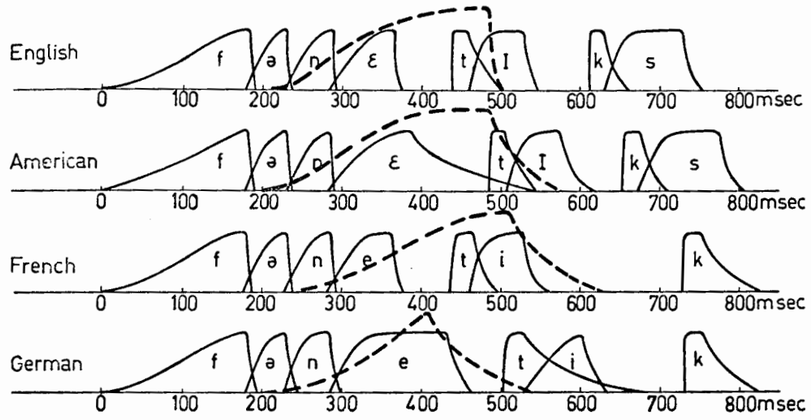

Fig. 3. Time patterns required for building 4 synthetic words: *phonetics* (Eng. and Am.), *phonétique* (French), *Phonetik* (German). No attempt was made to show actual amplitudes; the dotted line represents the intonation pattern used, i.e. frequency of periodic source.

4. CONCLUSION

1. It turns out to be possible to synthesize speech by starting from perceptual segments, which in number would roughly correspond with that of the phonemes of the language; some of these segments may differ only in the time parameter.

2. In experimenting with this type of synthetic speech a number of interesting problems are touched on. The hypothesis seems warranted that frequency transitions need not play such an integral part in speech perception as has sometimes been claimed since in our synthesis only steady-state portions have been employed apparently without harm to recognition or even naturalness. Furthermore, the adjustment of the proper time pattern provides a useful means of dealing with the problem of syllable division and of testing experimentally the relative contribution of pitch, intensity and length to syllable prominence.

3. The gating device enables one to monitor directly the outcome of the parameter settings. It seems therefore to be more flexible than the older devices employing painted spectrographic displays and more in line with the technique developed by Haskins laboratories in Octopus.[3] In this way analysis through synthesis can be carried out literally on the spot, providing data on how well a human listener is able to detect variations of the time parameter, which, in our opinion, plays a basic part in speech perception.

[3] J. M. Borst, *Journal of the Audio Engineering Society*, 4 (1956), pp. 19ff.

We would not be surprised to find that, in speech recognition, the perceiving mechanism is more vitally concerned with information carried by the dynamic pattern than that inherent in the spectral data.

Instituut voor Perceptie Onderzoek
Eindhoven