AC/EP
No.62/48
28 mei 1962
MS33.

INSTITUUT VOOR PERCEPTIE ONDERZOEK

EINDHOVEN

Speech communication Seminar Stockholm.

Speech synthesis of steady-state segments.

Phonetic studies have always been largely concerned with the problem of adequately segmenting the speech continuum, even before the introduction of modern synthetic speech machines. In a paper read before the 4th I.C.A. in Copenhagen a method of perceptual analysis was described, which can be carried out by means of an electronic gating device. It enables the experimentor to perform listening tests of successively presented increasing or decreasing portions of a word, spoken on tape, and to have subjects determine where one segment ends and the next one begins.

The number of the so-called perceptual segments thus obtained roughly equals that of the phonemes of the language investigated. One of the salient points of this type of analysis is that it tends to indicate that the glides, observed in studying articulatory movements and detectable in spectrographic recordings, need not always be present as perceptual cues. In fact, transitions were not heard to be gradual at all and clear turning points could generally be observed.

The outcome of this perceptual analysis lends itself to be subsequently tested by means of experiments with synthetic speech segments. During the analysing stages, the duration of the segments heard seemed to play an important part in identifying individual speech sounds. Assuming e.g. the spectral composition of fricative sounds like /f/ and /s/ to be fairly homogeneous, as can be seen from spectrographic recordings, short presentations of these sounds give rise to unmistakable /p/ and /t/ judgments respectively.

Speech synthesis of segments of roughly phoneme size can therefore serve two different purposes:

1) to test hypotheses arising from the perceptual analysis approach;
2) to test the contribution of the time parameter to speech perception.

An added advantage is that the number of segments required is comparatively small. (For a detailed exposition of the various systems of speech synthesis by means of the building block approach, see E.Sivertsen [1])

---

[1]) E.Sivertsen, Segment inventories for speech synthesis, Language and Speech, 4; 1961, 27-89.

Synthesis.

As an investigation of the influence of the time parameter seemed a profitable approach, we made this the pivotal point in our synthesis programme. We therefore required conditions of great flexibility of varying time patterns for experimental purposes which had to be met by electronic means. This resulted in a device which might be called a variable function gating circuit, whose main features are independently variable settings for rise and decay times as well as the total duration of acoustic signals. These requirements can be fulfille by means of an adjustable monostable multivibrator whose slopes can be modified by RC circuits.

This device enables us, in fact, to vary the time envelope of the individual segments to be synthesized. The definition of the operational values of rise, duration and decay time are illustrated in fig.

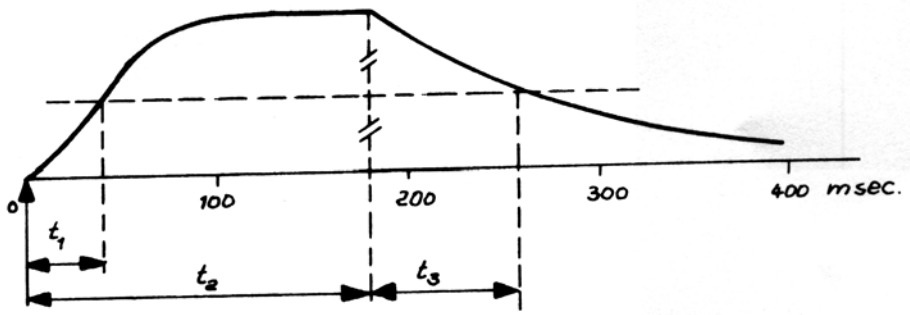

Fig.1 Definition of t_1, t_2, t_3 as independently variable parameters for rise, duration and decay time.

Two different sources are used for generating periodic and noise like sounds, viz. a multivibrator and a noise generator.

Vowel-like sounds.

In order to find the spectral components necessary for synthesizing vowel-like segments, a number of perceptual tests were carried out, based on the following assumptions:

a) spectral data concerning two-formant regions would be sufficient, a supposition that can be backed by spectrographic findings on spoken vowels in the literature;
b) as time cues in themselves seemed to be capable of causing qualitative differences in perceptual analysis (involving similar spectral composition), the influence of the time parameter should be taken into account.

By means of 50 fixed single-tuned RCL filters periodic signals generated by a multivibrator could be modified by choosing any two filter values as the characteristic formants of the synthetic vowel sounds to be tested perceptually.

In all, 350 such two-formant synthetic vowels were presented as a first scanning of the total perceptual field. The time envelopes are the same for all vowels (t_1= 40, t_2 = 180, t_3 = 70 msec.). The results averaged over two subjects are plotted in fig.2.

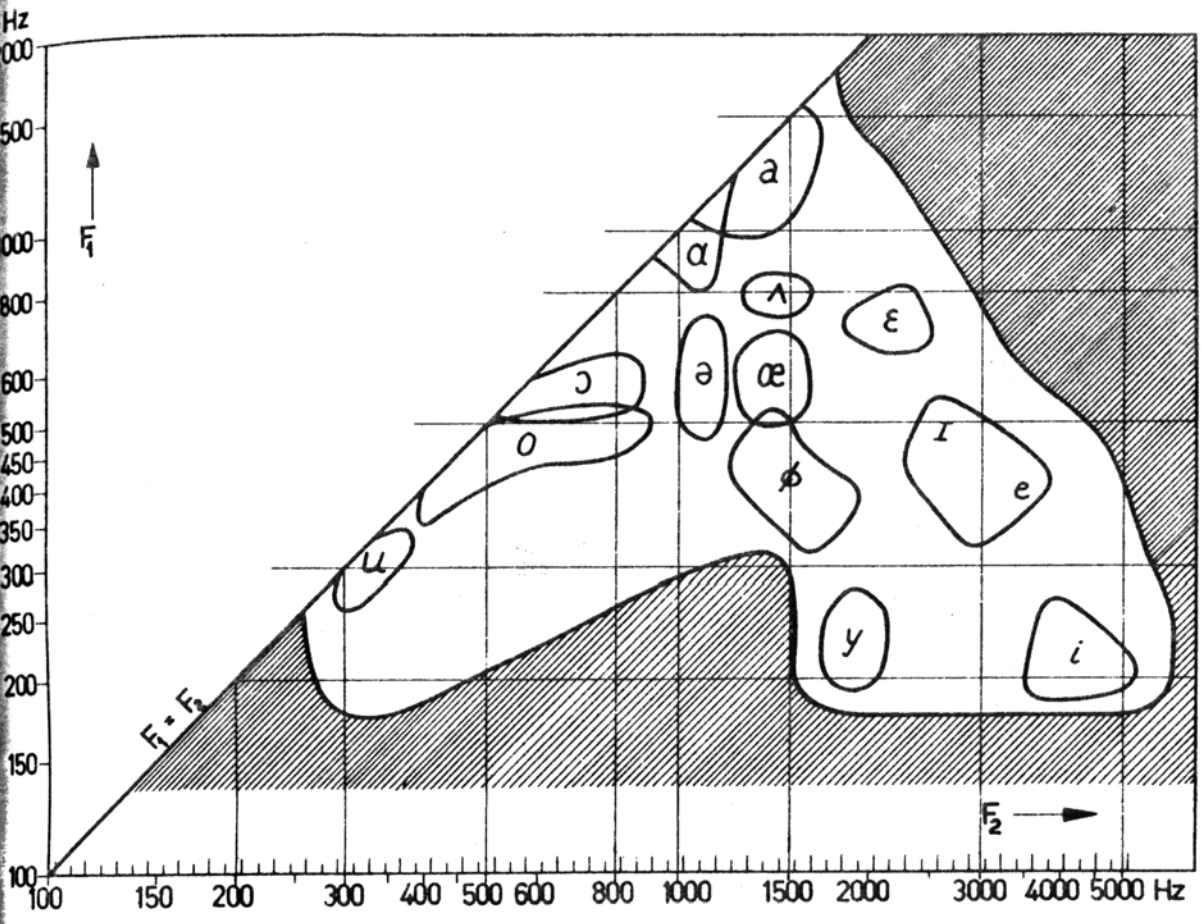

Fig.2 Perceptual judgments of 350 synthetic F1-F2 combinations of constant length (one subject).

A second series of perceptual tests was run to obtain information about the expected contribution of the time parameter. This experiment was based on the assumption that listeners, as native speakers of the language, disposed of a built-in time pattern, enabling them to judge whether a synthetic vowel had to appropriate length, or else was either too short or too long. A suitable combination of F1-F2 values was selected for each vowel, based on the findings illustrated in fig. 2. The total material was divided into two classes, long and short vowels, differing mainly with respect to the decay times ("long" t_3 = 75 msec vs. "short" t_3 = 15 msec). Durations (t_2) were varied in ten discrete steps ranging from 125 to 300 msec. The individual vowels within each group showed a remarkable consistency in the distribution of "good" judgments: the optimal values, i.e. the peaks of the distribution curves, show as little as 5% variation. The results, averaged over each class and over two subjects, are plotted in fig.3.

The difference in t_3 values was compensated for so as to express the perceptual lengths of the stimuli in terms of t_2 only.

As a third series once more variable F1-F2 combinations were presented as in the first experiment, this time with the optimal time envelopes. Comparing the results of the long and short presentations (similar t_1 = 20, t_2 = 180, different t_3; viz. 75 vs. 15 msec) the most interesting outcome is to be found in cases where the same spectral composition gives rise to qualitatively

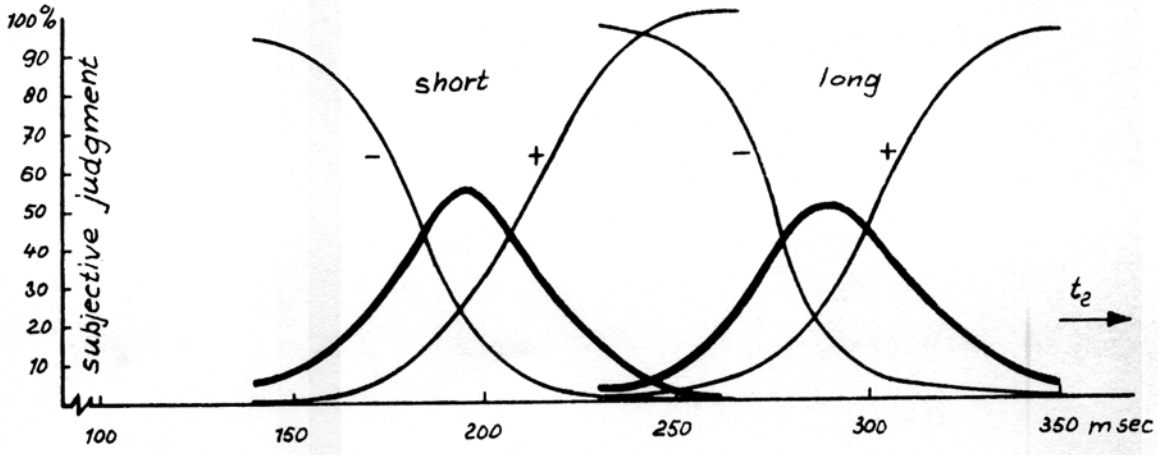

Fig.3 Subjective judgments in %, averaged over 2 subjects, in determining perceptual lengths by means of 2 classes of synthetic vowels, "short" and "long". "Good" judgments are marked with bold lines. The thin lines represent "too short" (-) or "too long" (+) judgments.

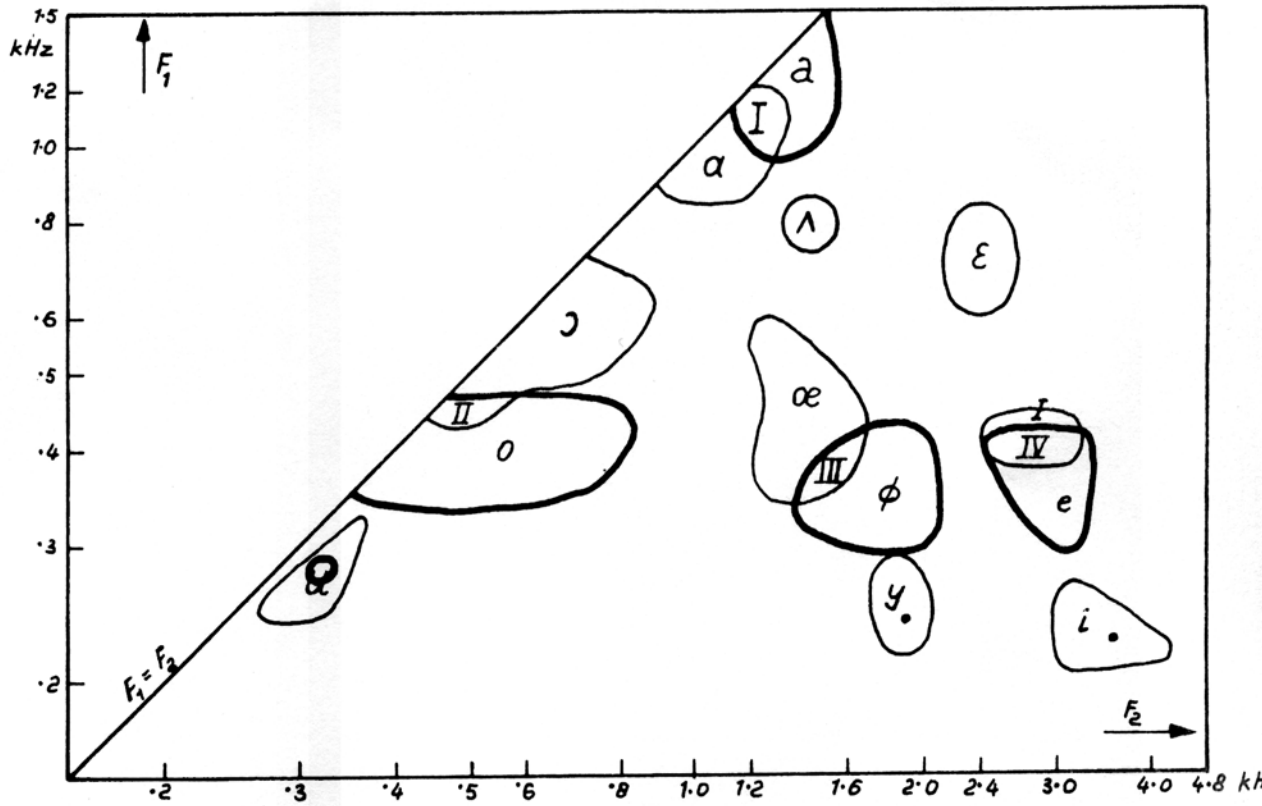

Fig.4 Perceptual judgments of vowel quality of synthetic F1-F2 combinations. Bold lines indicate acceptable quality when long; thin lines idem at short presentation. Overlapping areas I,II,III,IV include different vowel quality dependent on either long or short presentation.

on the length of the stimulus, a/ɑ, ɔ/o, œ/ø, ɪ/e.

A further point of interest is to see that inherently short vowels such as /u,y,i/, which have no similarly coloured long counterparts, show a narrower region of acceptability when presented with long tʒ. Moreover, /ɑ,ʌ,ɛ,ɔ,œ,ɪ/ when long were rejected.

A by no means unimportant factor about the manner of presentation in these tests has not yet been mentioned. This has to do with the fact that in order to make the synthetic vowels sound more or less natural, a certain pitch inflection was found to be helpful. For this purpose a special facility inherent in the over-all electronic set-up was employed. This consists in the possibility of applying the variable direct voltage, necessary for controlling the gate, to the multivibrator, the source of the periodic sounds. This results in a frequency deviation of the multivibrator, linearly proportional to the amplitude envelope of the gate output. In other words we have here a means of introducing an intonation pattern correlated with intensity variations of the acoustic signal, for which link there is some corroboration to be found in the literature on the physiological mechanism [2]). The name micro-intonation is suggested for this feature since it is made to operate only on the level of individual segments. The term macro-intonation will then be available for use on the level of prosody, involving intonation features as a relevant factor in bringing about degrees of prominence among various segments.

Noise-like sounds.

As results of analysis indicated that, for perceptual purposes, stops could be regarded as abbreviated fricatives, a start was made to synthesize fricatives first /f,s,ʃ/. In determining the source for synthesizing these sounds, our main preoccupation has been to select the character of the noise in such a way that the potential energy maxima of /s/ and /ʃ/ should be present, whereas for /f/, which does not seem to have a characteristic energy maximum, we should dispose of a diffuse distribution. This complex requirement can be fulfilled by choosing a prefiltered noise signal as the basic source from which /f,s,ʃ/ can be obtained individually by applying a single filter. The spectral distribution of the common noise source is given in fig. 5.

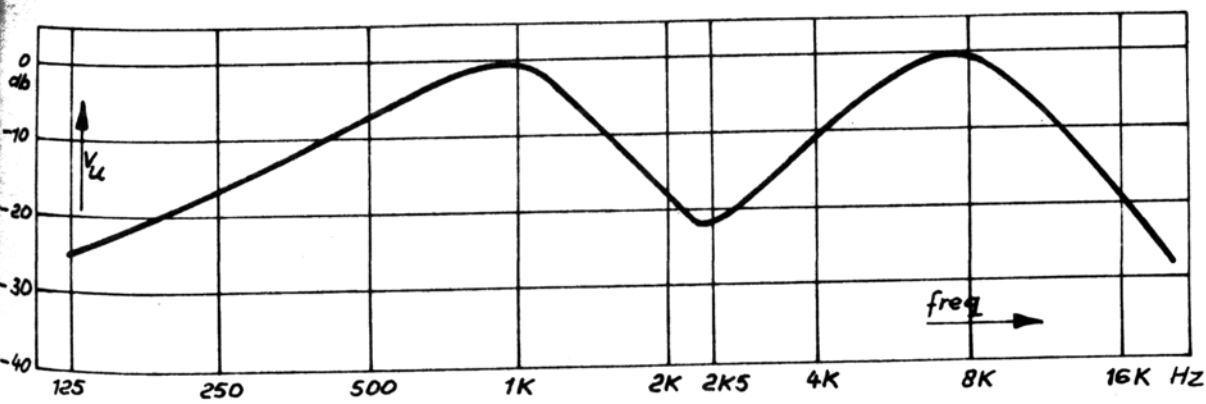

Fig.5 Spectral distribution of noise, used as a source of stops and fricatives.

---

[2]) Most recently, P.Ladefoged,"Subglottal activity during speech", Proceedings 4th. Int.Congr.of Phon.Sciences, Helsinki 1961

As the character of the fricatives to be generated was expected to depend on the nature of the following vowel, the ultimate choice of filter values for the various fricatives has been determined by means of perceptual tests for establishing the optimal consonant identification before various vowels.

Each of the synthetic fricatives with variable spectral composition was combined with the synthetic vowels /i,a,u/. These constitute the vertices of the acoustic vowel triangle, and as such represent extreme values to the extent that /i/ is the vowel with highest F2, /a/ with highest F1, and /u/ with lowest F2.

It turns out that in a subjective judgment scale the peaks, representing optimal values for the corresponding spectral colour, are correlated with the F2 value of the following vowel for /s/ and /x/. No appreciable variation of judgments could be brought about by changing the filter value for /f/ stimuli. A survey of the results of these perceptual experiments for 2 subjects is given in fig.6. The specific filter value for the individual consonants is designated consonant formant.

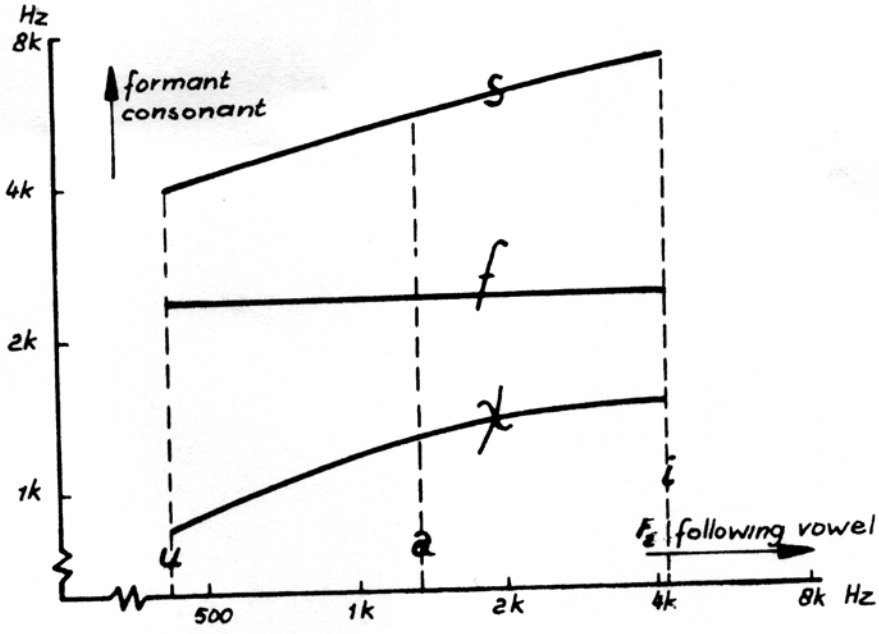

Fig.6 Representation of optimal judgments of synthetic /f,s,x/ plotted in terms of preferred formant values as a function of F2 of following vowel /u,a,i/.

Apart from differences in spectral composition the need was felt to vary the time patterns of the three fricatives notably with respect to the onset: for /f/ t_1 should be 100 msec, for /x/ about 60 msec, for /s/ about 80 msec in prevocalic position.

To generate the corresponding voiceless stops /p,t,k/ it suffices to apply the same spectral composition as had been found for the fricatives /f,s,x/. The perceptual difference between these two groups is mainly brought about by a difference in time pattern; t_1 for stops should have an upper limit of 10 msec;

t_2 is generally around 25 msec, but may be raised to as much as 40 msec with /k/ sounds. t_3 ; which is dependent on position, should not exceed 10 msec prevocalically and be at least 25 msec postvocalically, notably in word final position.

Mixed consonants.

To generate the voiced counterparts of these fricatives and stops, mere application of the so-called voice bar, a periodic signal with a single formant in the neighbourhood of the fundamental frequency of the periodic signals used, does not lead to acceptable results. For perceptual purposes the distinctive cue should perhaps rather be looked for in characteristic time and intensity differences, which are apparent from spectrographic recordings.

Diphthongs.

A special case is presented by the synthesis of diphthongs. Perceptual analysis shows that they can be regarded to consist of only two perceptual segments, which seem to overlap slightly. This means that in synthesizing diphthongs we are forced to introduce this slight overlap to guarantee acceptable quality.

An interesting feature of this procedure is that again no frequency glide or formant bendings seem necessary for evoking the right response in the listener.

As a matter of fact, diphthongs are produced with straight, i.e. unvarying, formants throughout each segment, whereas the frequency jump that is physically present is not perceived.

Application of the experimental findings on perceptual cues for synthetic speech segments makes it possible to synthesize words by presenting the required segments one after the other, each with its specific time and frequency values. The appropriate moment of triggering each segment can be derived from an electronic ring counter, providing pulses at intervals of 10 msec. In fact, the gates themselves can also be arranged in cascade to the effect that each gate triggers the following one.

Fig. 7 is given as an instance of the spectrographic result of a synthetic word phonetics, compared with the spoken version.

Conclusion.

The ultimate aim of speech synthesis is to find out whether the hypotheses derived from analytic procedures as to the contribution of the physical parameters inherent in speech can be corroborated. For this purpose listeners' judgments have to be incorporated. It therefore seems to be the most adequate approach to introduce such judgments already at the analytic stages of the investigation.

The resulting segmentation, tested in synthesis experiments, seems to indicate that the intermediate stage of spectrographic re-

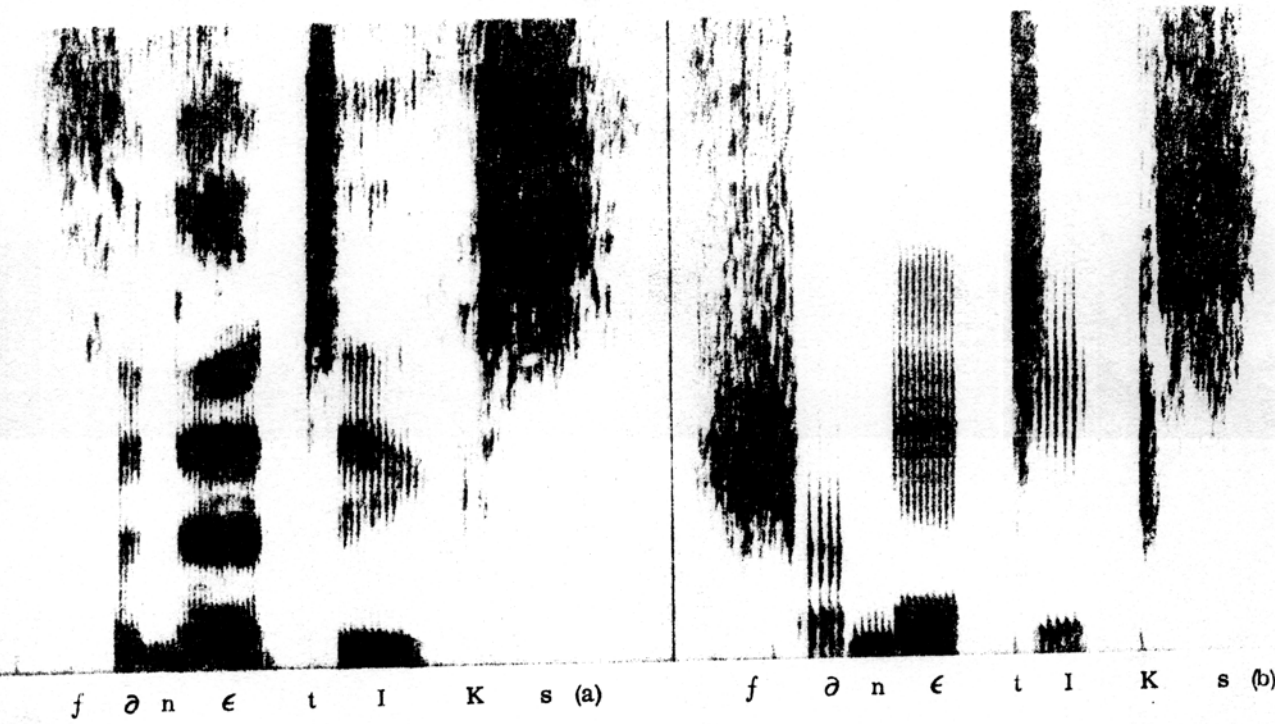

Fig.7 Spectrograms of the word phonetics in real speech (a) and synthetic (b).

cording with its characteristic frequency bendings can largely be bypassed.

The speech synthesis technique described above employs a limited number of segments, of phoneme size, and it has the adde advantage of great flexibility in carrying out perceptual experiments on the spot. This facility is particularly suitable in experiments involving variation of the time parameter which as such plays an important part in the perception of speech.

Actually, we have found that the proper dimensioning of the time parameter makes it possible to neglect quite a number of the details of formant information, which is usually claimed to paramount importance by other investigators.

Dr.A.Cohen,

J.'t Hart,