THE ACOUSTICS OF CONSONANTS

WIKTOR JASSEM

An oscillographic or a spectrographic display of the speech wave shows the following basic types of events; (1) (near-)periodic, (2) aperiodic, (3) aperiodic superimposed on periodic, and (4) pulse-like. Besides longer stretches of zero energy (corresponding to pauses) there are also gaps of limited duration rarely exceeding 150 msec. In a good spectrogram of a syllable, word or any stretch of connected speech there are points along the time axis at which either of the following phenomena – or both – are discernable: (a) a rapid change from one kind of disturbance to another, e.g. from a periodic to an aperiodic; (b) a maximum of the rate of change of the spectrum envelope. In a curve displaying changes of overall level, such as is produced by a high-speed level recorder or the "amplitude display" unit (an accessory of the Sona-Graph), points along the time axis may be determined at which the rate of change of the overall level (with an integration eliminating the periodicity of the fundamental) is maximum. These points mostly coincide with those named before. Fig. 1 shows a spectrogram and the amplitude display of the Polish word /momenˈtalni/ *momentalny*. All the three types of changes are easily seen there. Those particular points at which one or more of the three specified changes occur are *boundaries* between *acoustic segments*. The acoustic segments which have thus been demarcated can be correlated with phonetic units based on the consideration of auditory sensations, articulations and structural linguistic principles. This does not mean that an acoustic segment always corresponds to one phoneme. The segmental acoustic correlate of a phoneme may consist of two or more segments, but the number is always integral and probably never greater than eight. One segment may occasionally correspond to two or even three phonemes, if a single phonetic unit (on the level of perception and/or articulation) has for structural reasons to be interpreted as representing two or three consecutive phonemes (such as the Southern British English vocalic triphthongs). The segmentation of the speech wave, then, reduces a continuous time function to a discrete function. Thus, every particular sound of a syllable or a longer stretch of connected speech can be correlated with one or more segments on a spectrogram, although some cue or cues for its auditory recognition may be contained in a segment which is primarily correlated with an adjoining sound. For instance formant

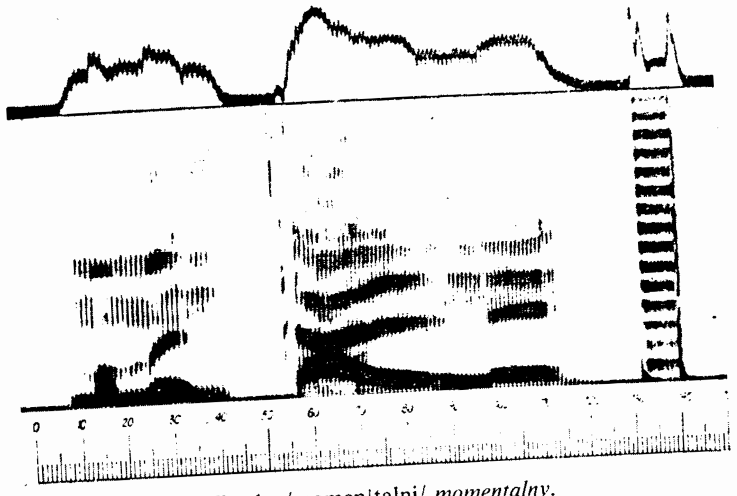

Fig. 1. /momenˈtalni/ *momentalny*.

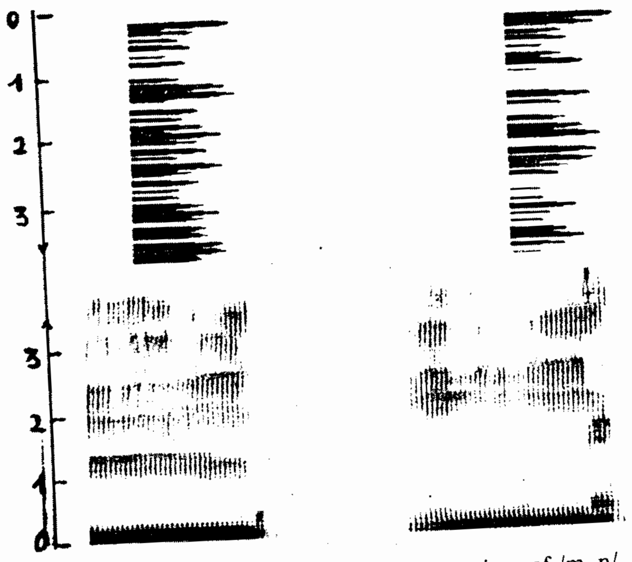

Fig. 2A. Spectrograms and sections of /m, n/.

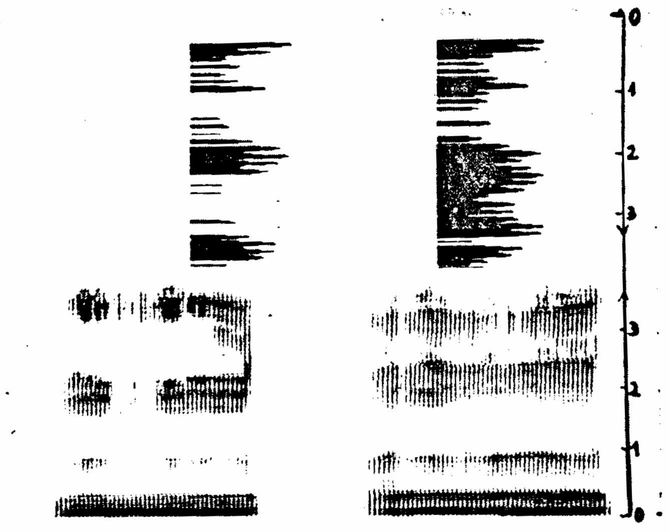

Fig. 2B. Spectrograms and sections of /ɲ, ŋ/.

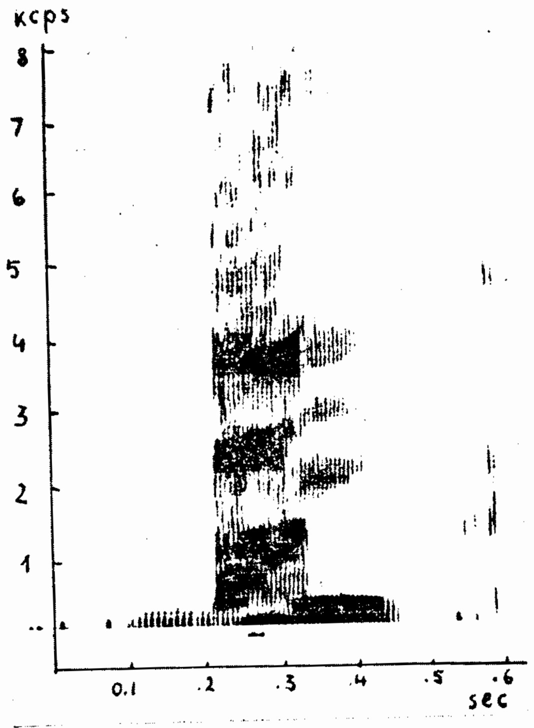

Fig. 4. /baŋk/ *bank*.

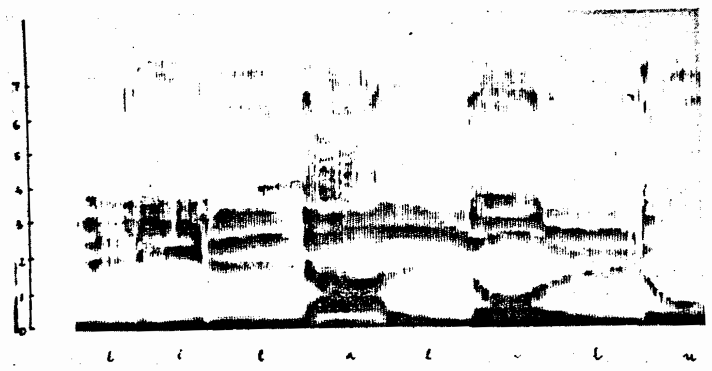

Fig. 3. [lilalolu].

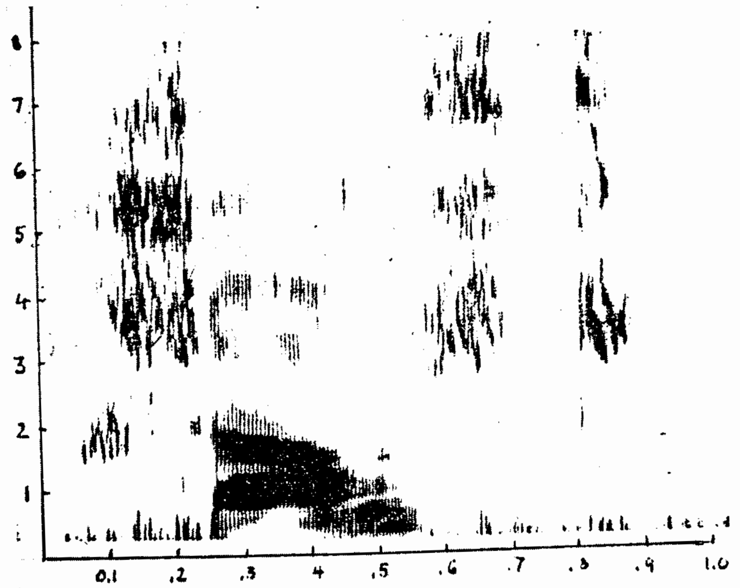

Fig. 5. /xʃonʃtʃ/ *chrząszcz* (fem.).

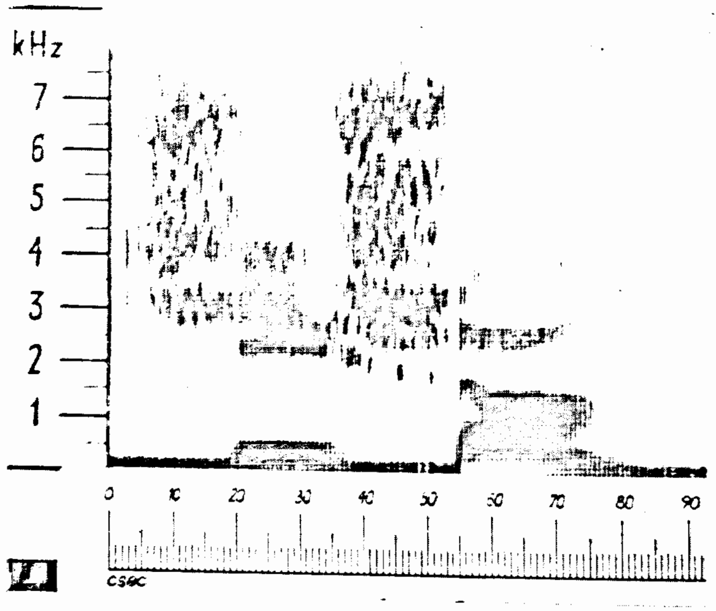

Fig. 6. /ˈt͡ɕiʃa/ *cisza.*

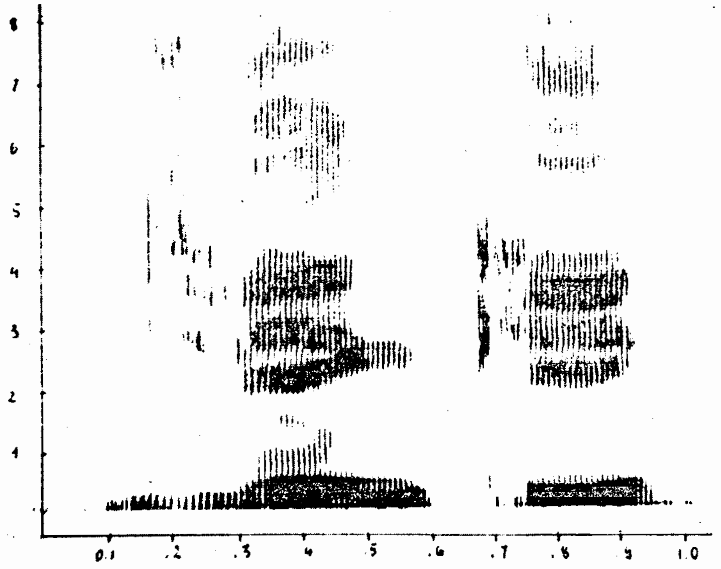

Fig. 8 /ˈd͡ʑvjeŋci/ *dźwięki.*

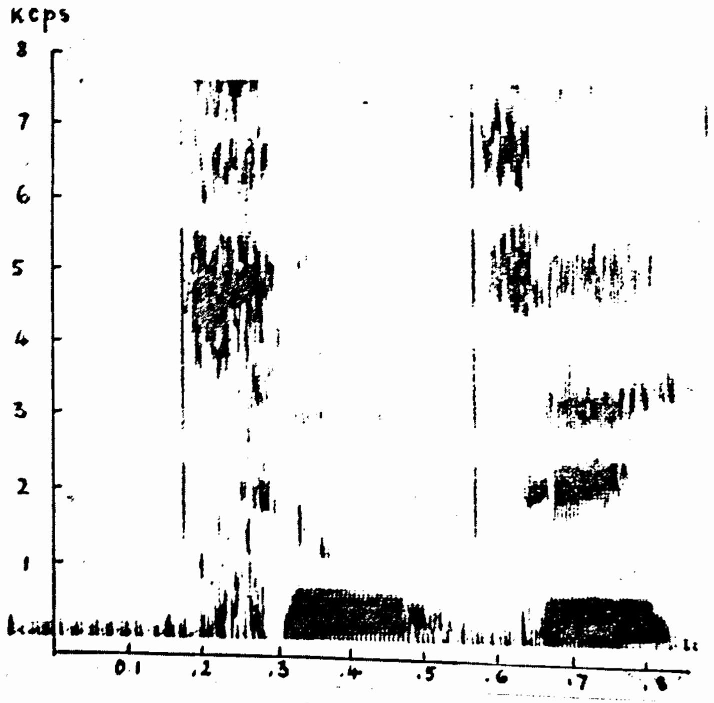

Fig. 7. /ˈt͡sud͡zi/ *cudzy* (*fem.*).

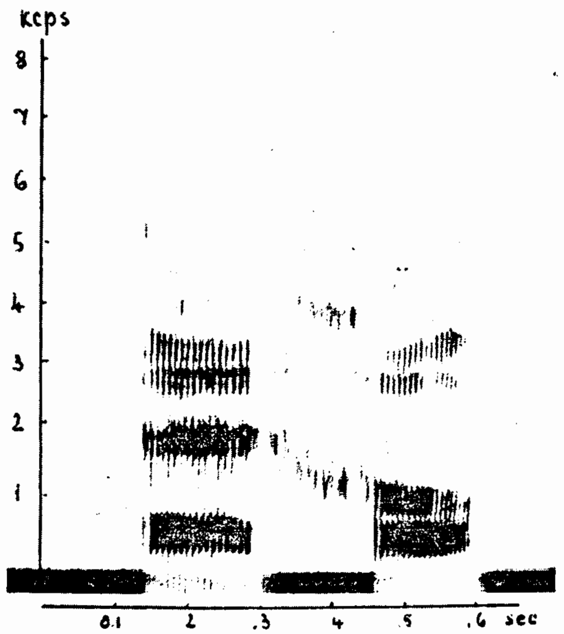

Fig. 9. /ˈexo/ *echo.*

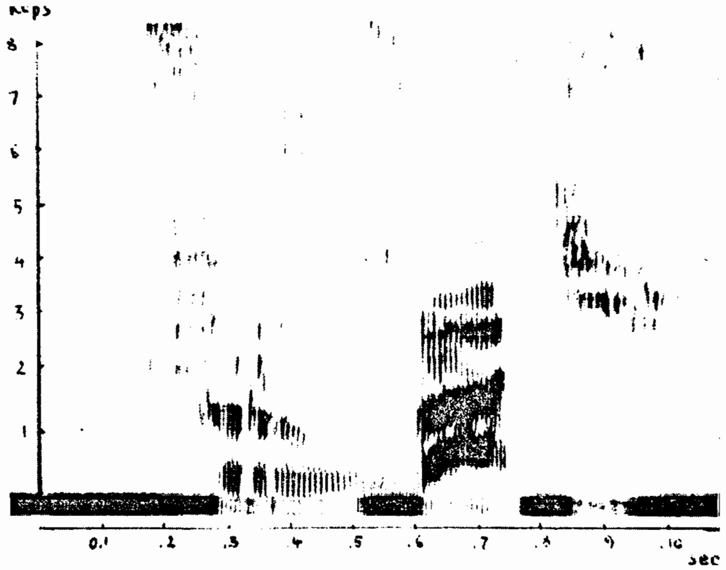

Fig. 10. /ˈfruvatc/ *fruwać*.

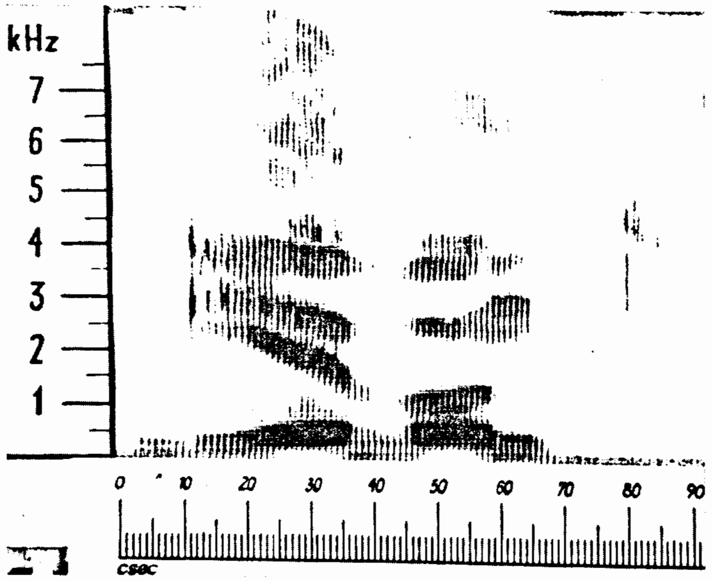

Fig. 11. /ˈɟevont/ *Giewont*.

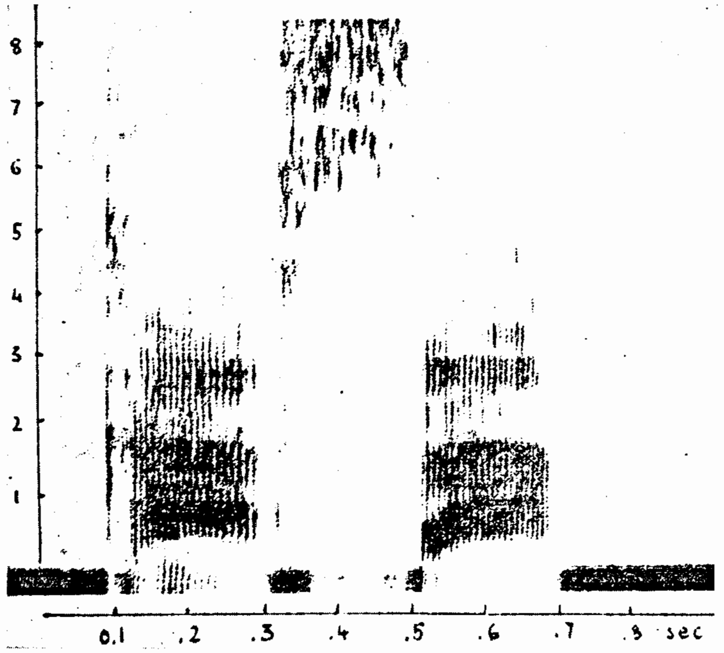

Fig. 12. /ˈkasa/ *kasa*.

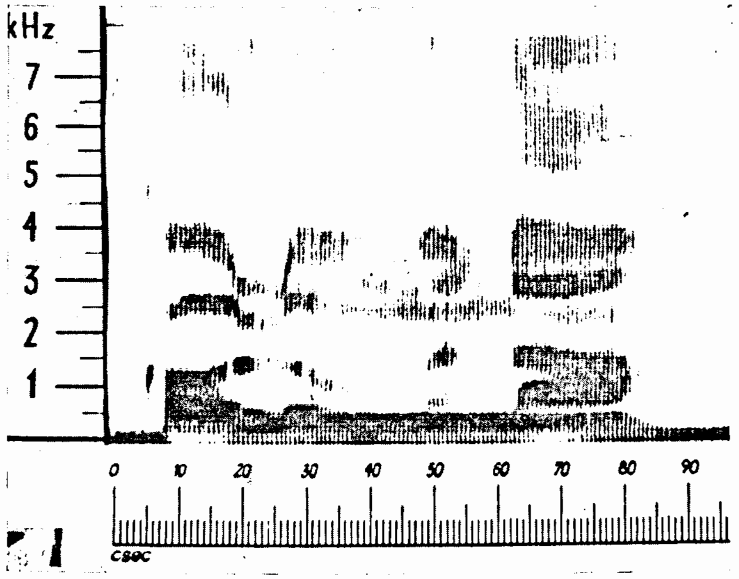

Fig. 13. /koˈlumna/ *kolumna*.

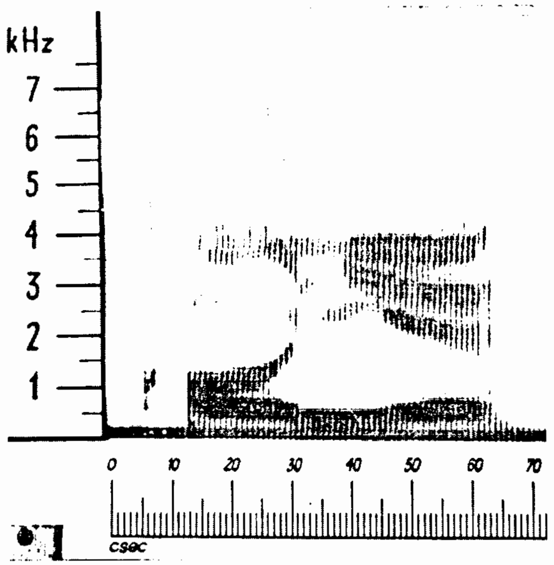

Fig. 14. /ˈkoɲe/ *konie.*

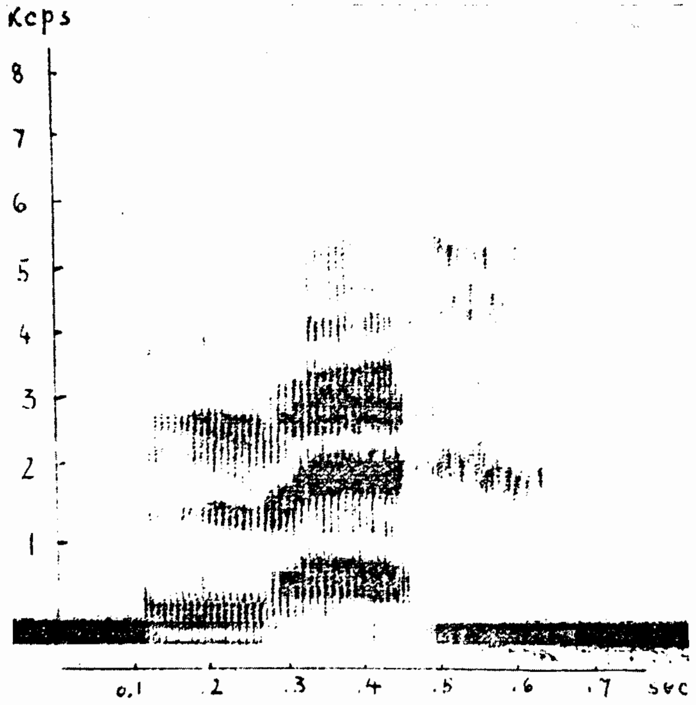

Fig. 15. /mex/ *mech.*

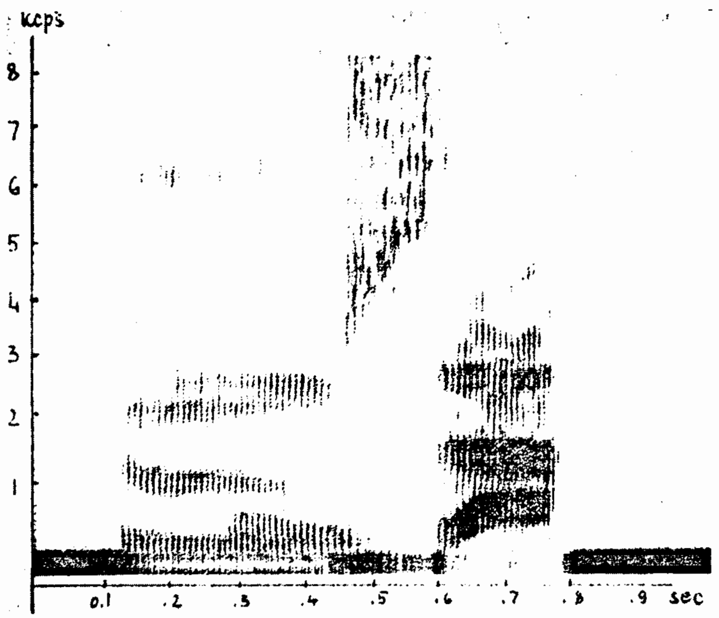

Fig. 16 /ˈmuza/ *muza.*

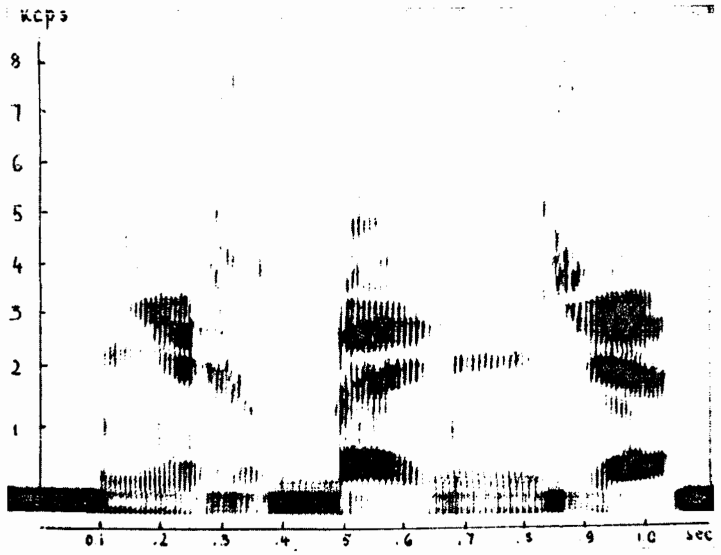

Fig. 17. /ɲexˈbeŋdze/ *niech będzie.*

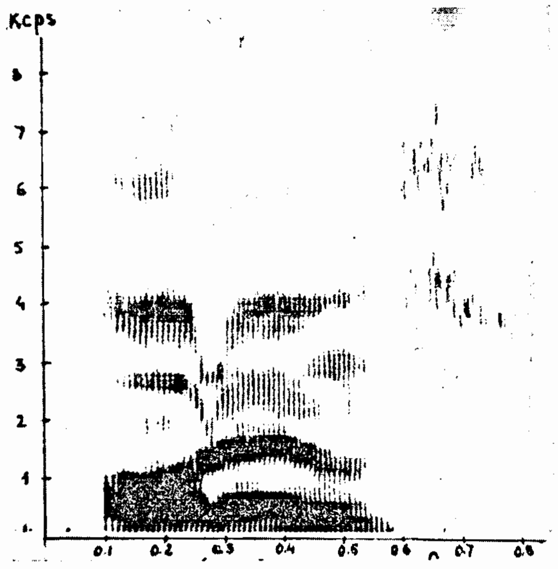

Fig. 18. /ˈoreŋʃ/ *oręz.*

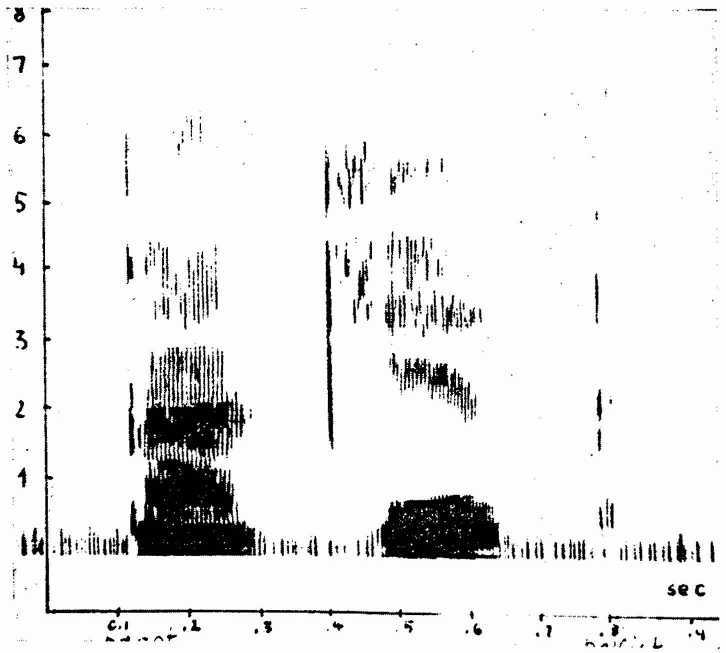

Fig. 19. /ˈpacet/ *pakiet* (fem.)

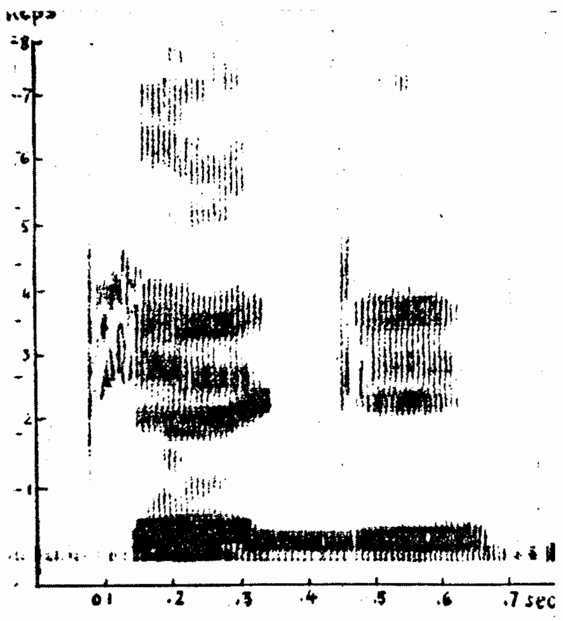

Fig. 20. /ˈpjeɟi/ *picgi.*

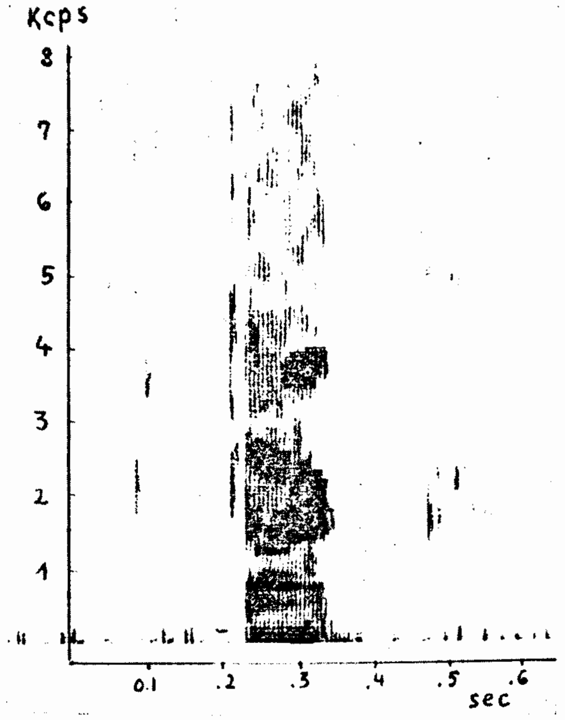

Fig. 21. /ptak/ *ptak.*

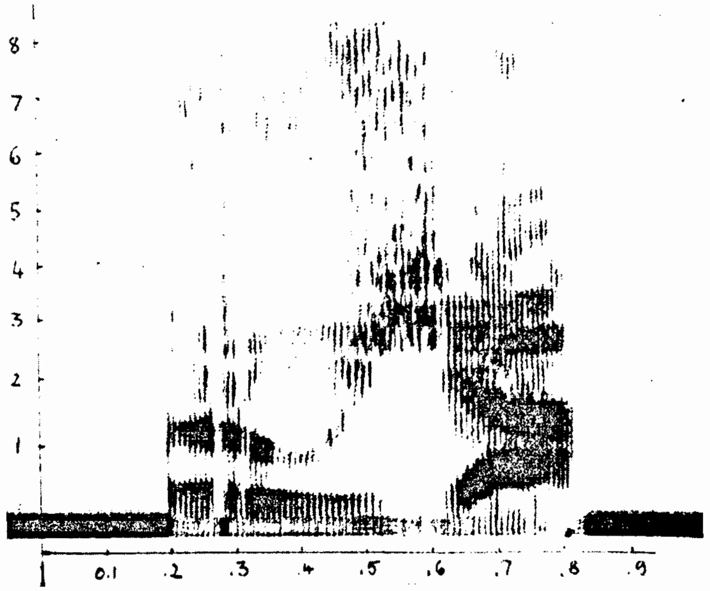

Fig. 22. /ˈruza/ *Rózia.*

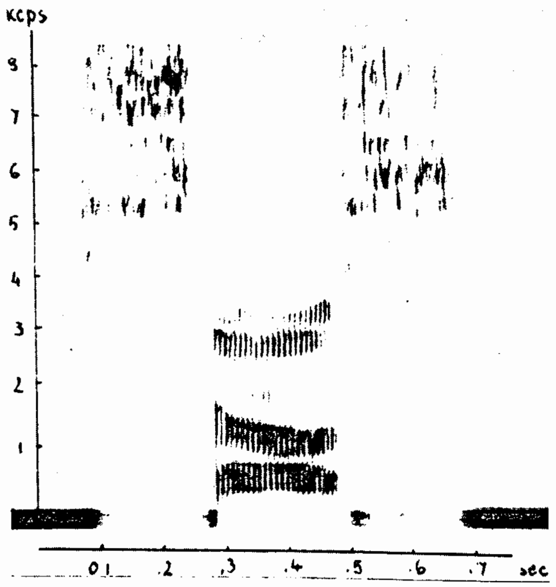

Fig. 23. /sos/ *sos.*

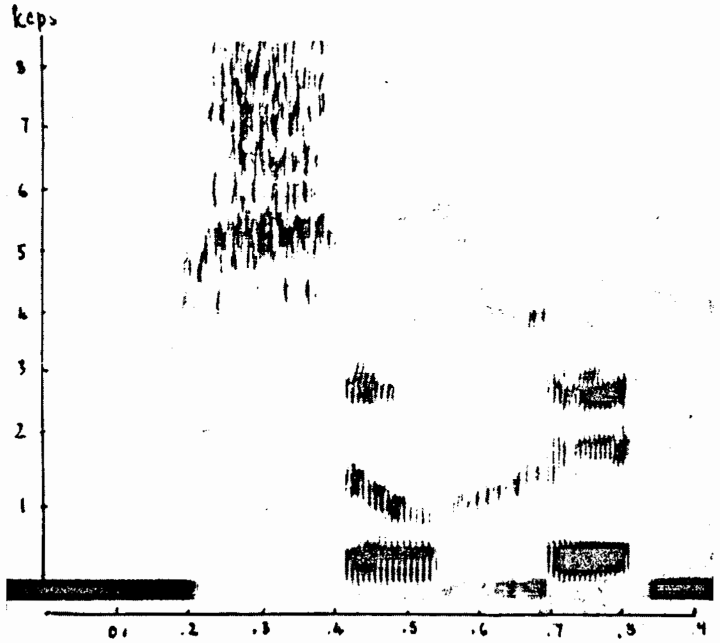

Fig. 24. /ˈsuxi/ *suchy.*

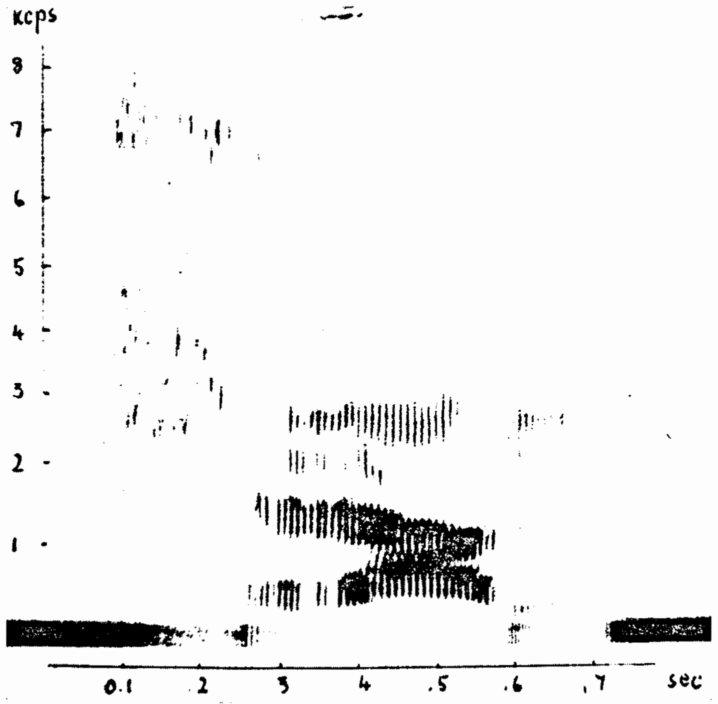

Fig. 25. ʃron/ *szron.*

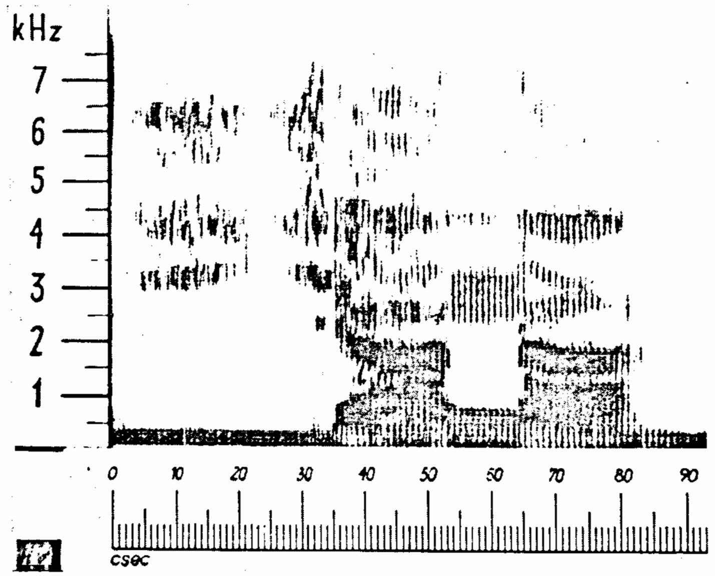

Fig. 26. /ˈçt͡çana/ *ściana*.

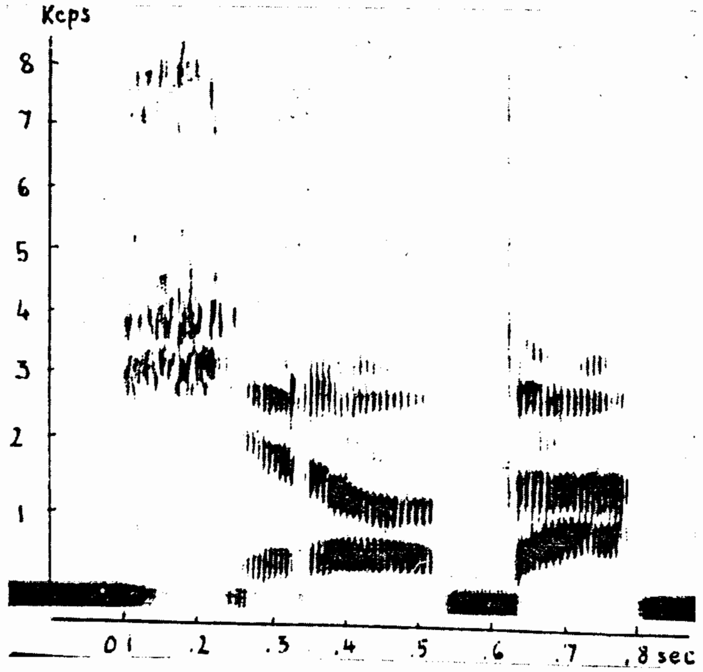

Fig. 27. /ˈçroda/ *środa*.

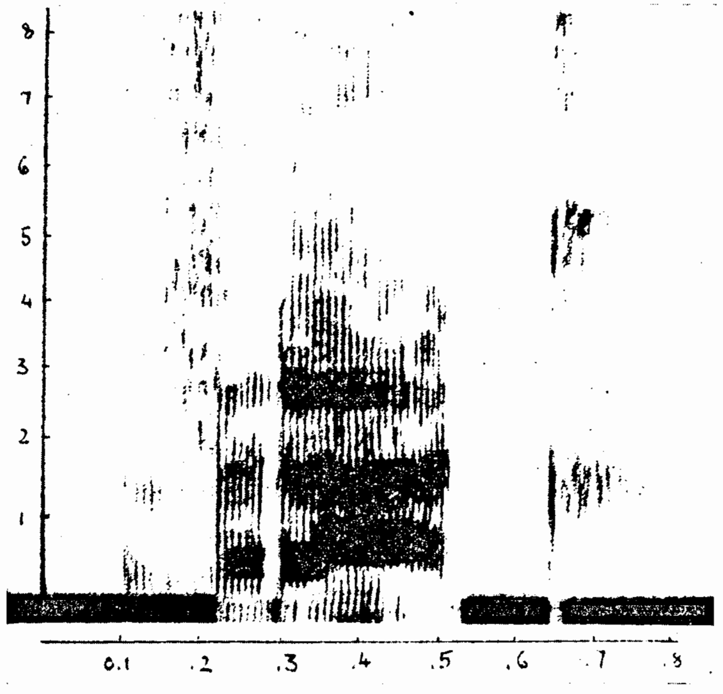

Fig. 28. /vrak/ *wrak*.

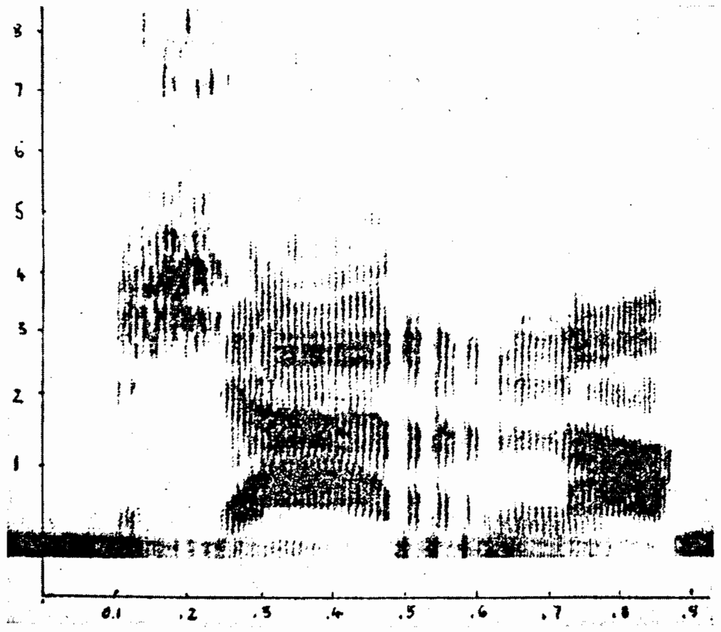

Fig. 29. /ˈz̦arno/ *ziarno*.

transitions in vowel segments are known to be important cues for the recognition of neighbouring plosives and nasals.[1]

The division between the two classes of elements referred to as vowels and consonants may be based on a purely phonetic description, or else it may be structural, i.e. phonemic. It is often necessary to state clearly whether the terms "vowel" and "consonant" are being used with their phonetic or their phonemic implications. The convenient terms "vocoid" and "contoid" introduced by K. L. Pike are based on the analysis of articulations without reference to the phonemic status of the speech sounds. Naturally, then, if we deal with speech sounds disregarding their function in any particular language, we can only mean vocoids and contoids, but it may be permissible to use the traditional and widely current terms "vowels" and "consonants" instead, provided we make it perfectly clear that those terms are being used without structural implications. We shall here be primarily dealing with consonants as phonetic elements, not as phonemic units.

Although the correlations between specific auditory features and specific acoustic features are not always simple, certain large classes of speech sounds are found to be represented by segments having quite clearly definable acoustic features in common. We will therefore proceed in our description classifying our material according to traditional distinctions based on articulatory features.

The number of the different consonants occurring in various languages is very large and many of them have not yet been investigated acoustically. We shall here limit ourselves to the consonants occurring in Standard Polish. Whilst this is a severe limitation in a general perspective, we hope that at least the most important problems will come under discussion because this language has a well-developed system of consonants. The observations are based on several hundred short utterances (mostly single words) produced by eleven speakers (male and female). The recordings were all made in Poznań, Poland, and the spectrograms were made in the Speech Transmission Laboratory of the Royal Institute of Technology in Stockholm, at the Phonetics Department of the University of Edinburgh, Scotland, and at the Acoustic Phonetics Laboratory of the Polish Academy of Sciences. I wish to thank Dr. Gunnar Fant and Mr. David Abercrombie for the opportunity of working in their respective Institutions.

The nasal consonants are represented by periodic segments. They have a distinct formant structure. Below 4 kcs up to seven formants can be distinguished. All nasals have one formant near the first harmonic, two around 2 kcs and two (or sometimes one) just above 3 kcs. Although some systematic differences in the exact position of those formants may be observed (e.g. the two above 3 kcs are closer together in the velar than in the others), the four Polish nasals: [m], [n], [ɲ] and [ŋ]

[1] On segmentation see: L. Lisker, "Linguistic Segments and Synthetic Speech", *Language*, Vol. 38, No. 3 (1957), pp. 370–374. More recently: G. Fant, *Acoustic Theory of Speech Production*, ('s-Gravenhage, 1960), pp. 207–208, and G. E. Peterson and I. Lehiste, *Duration of the syllable nuclei in English*, Univ. of Michigan Speech Laboratory Report No. 4, Ann Arbour, 1–49, esp. 4–16.

mainly differ in the shape of that part of the spectrum which lies between approx. 0.3 kc and 1.8 kcs. The labial has a strong formant whose position varies somewhat according to the phonetic context, but is contained between 0.7 and 1.2 kcs. It also has a faint trace of a formant at approx. 0.4 kc. The dental has two rather weak formants, one at 0.7 and the other at 1.5 kcs. The palatal and velar nasals both have one formant only in this region which lies at about 0.8 kc. They seem to differ chiefly in that the two formants just above 2 kcs tend to be closer together in the palatal than in the velar, whilst the two above 3 ksc are closer together in the velar than in the palatal. Spectrograms and 'sections' of isolated nasals are shown in Fig. 2. Not all of the named formants are always visible on spectrograms of complete words (but most can be seen in Figs. 15 and 16 ([m]), 25 and 29 ([n]), 14 ([ɲ]) and 4 ([ŋ]). Our data on the formants of the labial, dental and palatal nasals are in good agreement with those given by Fant for the three Russian nasals (*Acoustic Theory of Speech Production*, pp. 139–161). A comparison of Figs. 15 and 16 shows how the exact position of the formant around 1 kc in [m] depends on the neighbouring vowel. The figures also illustrate another important feature common to all nasal consonants, viz. the relative stability of the formants with time. The relations in the F_2 and F_3 transitions of neighbouring vowels are complex and require further study. Only a general and very brief statement can be made here: Next to [m] F_3 is usually negative. The F_3 transition is negative in front vowels and absent in back vowels. Next to the dental, front vowels have no transitions and back vowels have a positive F_2 transition. Both F_2 and F_3 transitions are positive next to [ɲ] except in [i] which has none. F_3 is negative and F_2 positive next to [ŋ] in front vowels. Both are negative or absent in back vowels (cf. Figs. 4 and 8).

The laterals. Acoustically, the laterals resemble the vowels more than any other consonants. If pronounced in isolation or if considered without regard to a neighbouring vowel, a lateral has not yet received a clear acoustic definition which would unambiguously distinguish it from a vowel sound. Probably the positions of the formants, especially the higher ones, relative to each other show some peculiarities. Fant, for instance, points to the clustering of the higher formants (*loc. cit.*, p. 167). Another possibility is a strong antiresonance below 2 kcs. Fig. 3 which represents the sequence [lilalolu] shows that the formant frequencies of a lateral vary largely according to the context. In agreement with the findings of O'Connor, Gerstman, Liberman, Delattre and Cooper,[2] the lowest formant of [l] has been found somewhat higher than that of the nasals, and lies around 0.4 kc. Our data for the higher formants differ slightly from those given by L. Lisker[3] chiefly in that F_2 and F_3 are closer together in our materials than in Lisker's. The greatest separation between F_2 and F_3 in [l] occurs next to [a] and is of the order of 1 kc. Next to [i] and [u] the F_2 and F_3 frequencies differ by about 0.5 kc. F_2 is contained, according to the neighbouring vowel, between 1.3 and 1.8 kcs. F_4 is approx. 0.5 kc higher than F_3 and lies near 2.5 kcs next to [u] and at 3.0 or somewhat higher next to front vowels. Next to back vowels, then, F_3 and F_4 of an [l] are relatively low and the transitions of the vowel formants are negative and very rapid (they may be even under 50 msec in duration). Cf. Fig. 13.

The flapped and the rolled [r]. Both the flapped and the rolled tip-of-tongue voiced [r] are used in Polish (without phonemic distinction, by the way). Intervocalically a flapped [r] is represented by a single segment, lasting about 20–25 msec, showing considerable reduction of overall level, with some formantlike energy concentration near the fundamental and at 1.5, 2.9 and between 3.5 and 4.0 kcs. Otherwise a flapped [r] consists of two segments, the one just described and another one which has all the features of a neutral, i.e. shwa-like vowel except that it is usually rather short (from 20 to 50 msec.). Sometimes, however, this segment exceeds 50 msec. (100 is the highest value in our materials) and it then reached the status of a phonetic shwa. A rolled [r] is an alternating succession of the two kinds of segments. Up to eight segments have been found in naturally pronounced words. An interesting feature of the flapped [r] is that there is a rapid and very marked minus transition of both F_3 and F_4 of the neighbouring vowels. Those transitions are clearly visible in Fig. 18. (Good examples of flapped and rolled [r]s also occur in Figs. 10, 22, 27 and 29).

The plosives. In most cases a plosive sound is represented by a sequence of acoustic segments. A common feature of all Polish plosives in all positions is a pulse-like segment. On spectrograms, this appears in the form of a single vertical stroke. An ideal pulse would give a stroke whose breadth would only depend on the time constants of the analysing and registering circuits, and which would stretch along the entire frequency range under analysis. Since in the actual speech wave the segments here considered are only approximations to pulses, the duration is of the order of 15 msec, the stroke is rarely seen along the whole frequency range, and it is usually broken, thus showing some energy concentrations. The pulse-like segment may be seen on spectrograms of both the voiceless and the voiced plosives. If it is absent from a spectrogram of a plosive, this only means that the sound is too weak relative to others in the word or syllable. With sufficient gain on reproduction (often with overload on vowels) the pulse-like segment can always be detected. It is seen in Figs. 19, 20, 21 and 27. The spectrogramm of a voiceless plosive in a non-initial position shows a gap lasting between 120 and 180 msec. In voiced plosives the pulse is preceded by a near-periodic segment with only a few low harmonics. The pulse is followed, almost always in voiceless plosives, but rarely in voiced plosives except [ɟ], by an aperiodic segment. In the bilabial, dental and velar plosives this segment shows energy concentrations in rather narrow frequency bands. In the dental and velar plosives there is one 'burst' between 1 kc and 2.5 kcs and another between 4 and 5 kcs. The latter is stronger in the dentals than in the velars. The labials also

[2] J. D. O'Conner, L. J. Gerstman, A. M. Liberman, P. C. Delattre, F. S. Cooper, "Acoustic Cues for the Perception of initial /w j r l/ in English", *Word*, Vol. 13, No. 1 (1957), pp. 24–43, esp. p. 32.

[3] Lisker, "Minimal Cues for Separating /w, r, l, y/ in Intervocalic Position", *Language*, Vol. 34, pp. 256–267.

have a burst in the region between 1 kc and 2.5 kcs. but the one above 4 kcs is very weak or not detectable. The bilabials have another burst between 3 and 4 kcs and sometimes a weak one below 1 kc. The exact position of the burst between 1 kc and 2.5 kcs depends on the context, and in the labials there are often two peaks rather than one in this region. In the palatals the main energy concentration has a broader frequency band between 2.5 and 4 kcs and is longer in duration, often exceeding 100 msec in [c] as compared with about 50 msec in [p, t, k]. The Russian plosives are phonetically equivalent to the Polish plosives and our data on the frequency of the bursts are in good agreement with those published by Fant (*loc. cit.*). E. Fischer-Jørgensen ("Acoustic Analysis of Stop Consonants", *Miscellanea Phonetica*, II, pp. 42–59, 1959), and later M. Halle. G. W. Hughes and J. P. Radley ("Acoustic Properties of Stop Consonants", *JASA*, Vol. 29, No. 1, 1957, pp. 107–116) showed that the problem of vowel transitions preceding or following a stop consonant is much more complex than would have appeared from earlier publications. The most recent the most extensive study of vowel transitions as revealed by spectrograms is that by I. Lehiste and G. E. Peterson, *Transitions, Glides and Diphthongs*, (Report No. 4 of the Speech Research Laboratory, University of Michigan, Ann Arbour, pp. 50–88, 1960). Our investigations of the transitions have not been as detailed, but the results seem to support Lehiste and Peterson's findings. We shall here have to limit ourselves to the general statement that the transitions depend both on the type of plosive and the type of vowel as well as on the position of the consonant relative to the vowel.

The fricatives. Every fricative sound is represented by a segment which is either aperiodic (voiceless consonant) or periodic-aperiodic (a superposition–voiced consonant). There are four pairs of fricative phonemes in Polish, each with a voiced and a voiceless member and another phoneme which has no distinctive feature of voicing, its allophones being either voiceless (in most positions) or voiced (in specific positions). The four pairs are: labiodental, postdental, palatoalveolar and alveolo-palatal. The voiceless fricatives are monosegmental, and so are the voiced fricatives except in utterance-initial. In this position the periodic-aperiodic segment is preceded by a periodic one with just a few low harmonics and some traces of energy at higher frequencies. The duration of this segment is of the order of 50 msec and it is not easily perceived as a separate auditory unit unless attention is drawn to it. The energy concentration in the low frequencies is greater in the first element of the bisegmental fricative than in the second. An example of a bisegmental voiced fricative is seen in Fig. 28. The differences among the four pairs are in the frequency range of the noise and the position of energy maxima in the spectrum. The noise spectra of these fricatives are as follows : The frequency range extends from approx. 4 or 5 kcs well beyond 8 and even 10 kcs (as was shown by additional analysis) in the postdentals and from approx. 2.5 to 8 kcs in the palato-alveolars. The lower limit lies about 2.8 in the alveolo-palatals, but the upper limit varies largely between 6 and 8 kcs. In the labiodentals the range is 1.5 to 10 kcs. In the last-named fricatives the energy is distributed fairly evenly, though slight maxima may be observed between 1 and 3 kcs and above 8 kcs. In the other fricatives the maxima are very distinct. In the dentals their frequencies vary largely, but there are usually two of them at least between 5 and 8 kcs. Another regular energy concentration in the [s–z] pair lies at 1.5 kcs. It is therefore a separate narrow band of noise. Its level is between 25 and 35 db below the peak at the higher frequencies so that it is rarely detectable in ordinary Sonagrams. Our data for the labio-dentals and postdentals are quite similar to those obtained for the corresponding English sounds by G. W. Hughes and M. Halle ("Spectral Properties of Fricative Consonants", *JASA*, Vol. 28, No. 2, 1956, pp. 303–310). An analogous, stronger band of noise is also found in the palato-alveolars, but its frequency largely depends on the F_2 of adjoining vowels. Fig. 6 represents the word /ˈtɕiʃa/, and this noise band is seen to link the second formant of [i] with the second formant of [a]. Further maxima in [ʃ] and [ʒ] are found just below 3 kcs, between 3 and 4 kcs, and there are one or two more at less fixed frequencies, between 4 and 7 kcs. The alveolo-palatals differ from the palato-alveolars in having a considerably weaker low-frequency noise, rarely present on Sonagrams. In lies about 0.5 kc higher than in [ʃ, ʒ]. The other maxima, too, are approx. 0.5 kc higher in the alveolo-palatals than in the palato-alveolars. Further examples of the fricatives just discussed appear in Figs. 10 and 28 (the labials), 12, 16, 23 and 24 (the dentals), 25 (the voiceless palato-alvolar) and 22, 27 and 29 (the alveolo-palatals). The voiceless velar fricative is represented by two narrow bands of noise. The higher-frequency noise has a narrow band with a center at about 4 kcs. The central frequency of the lower band depends on the neighbouring sounds. If these are vowels, this noise band has a frequency which is near that of F_2. Figs. 9 and 24 show this very clearly because here the [x] is in an intervocalic position, so that the centre of the noise band moves up or down making a link between the differing second formants of the vowels. There is usually also a third, very narrow band of noise with a centre at the frequency of F_1 of a neighbouring vowel. With no vowel preceeding or following the two lower bands of noise are at 0.4 and about 1.3 kcs. The rare voiced variant of /x/ i.e. [ɣ] is shown in Fig. 17. A detailed study of fricatives was recently published by P. Strevens ("Spectra of Fricative Noise in Human Speech", *Language and Speech*, Vol.3, pp. 1, 32–49, 1960) and our finding are in good general agreement with his data. Strevens also investigated differences in overall level and experiments with speech synthesis show that these differences are important in the process of auditory recognition.

The affricates. Polish has three pairs of affricates, corresponding to the postdental, palatal-alveolar and alveolo-palatal fricatives. They are essentially combinations of varieties of [t] or [d] with the appropriate fricative. Usually, the pulse-like segment typical for the plosives, is here immediately followed by an aperiodic segment whose spectral pattern is the same as that of the corresponding fricative. The duration of the fricative portion of an affricate is less than that of a fricative consonant in comparable circumstances. If an affricate is preceded by a homorganic fricative there may be, instead of a gap and a pulse, a gradual decrease of the noise energy followed by a gradual increase. This is shown in Fig. 26. Other examples of affricates are found in Fig. 5, 6, 7 and 17.

The figures which we have given are typical values for male voices. In female voices the values are 10–20 per cent higher.

The above description has of course been very fragmentary and sketchy, but it should be noted that our knowledge of the acoustics of speech is in any case very far from complete, one important reason for this being that the number of languages which have been analyzed by the methods of acoustic phonetics is still severely limited. It is quite essential for the further development of the various branches of General Phonetics that the number of institutions in various countries applying electro-acoustic analysis techniques should quickly increase.