STUDIA PHONOLOGICA I (1961)

The Phonetic Typewriter: Its Fundamentals and Mechanism

Toshiyuki SAKAI

What is the Phonetic Typewriter?

The phonetic typewriter is the device which is to convert the human voice into typed letters. However, at present, it has not yet been put to practical use.

Man has made effort to realize such kinds of devices, because in human society voice is very important as a medium of communication, for its speed and causing relatively slight fatigue. At present it is easy to record voice, but to convert voice into letters, the intermediation of man is required. The relation may be shown in Fig. 1. When man intervenes, the speech sound wave caught by him is analyzed at the auditory organs, and, after going through some more repeaters, is recognized at the brain, which, as a result, gives orders to the hands to strike the keys of a typewriter or write letters.

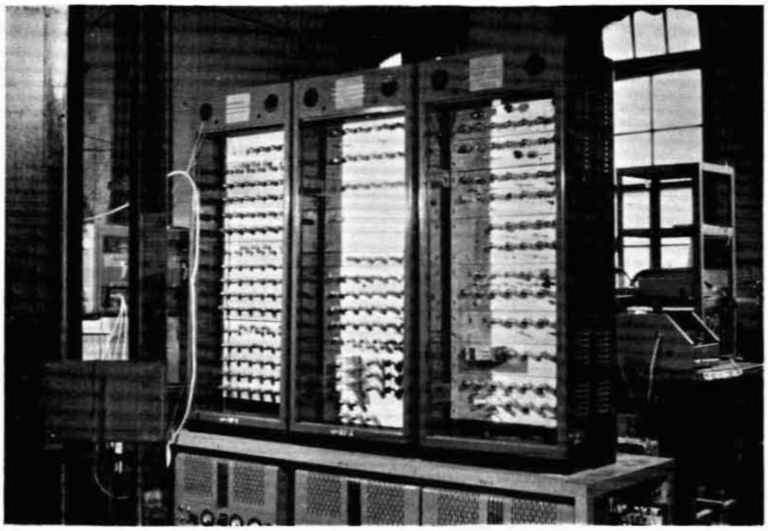
Though amplitudes, timbers, and durations of the speech wave coming into the ear are various, these speech sounds are changed into a definite letter set when the hands strike the keys of a typewriter. Thus in converting the voice into corresponding letters, it is fundamentally necessary to change a continuous quantity in time region into a discrete quantity; on the other hand, it must be considered to convert the speech sound into a code system like telegraph tape.

Fundamental Requirements for the Phonetic Typewriter

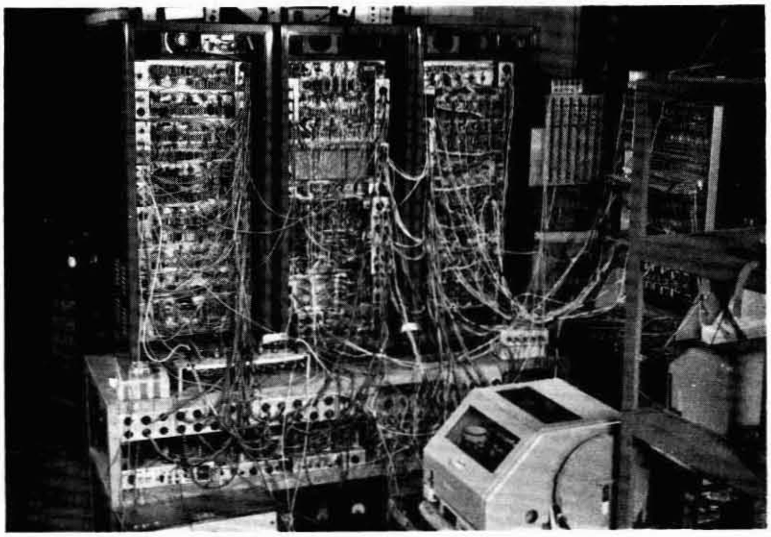
The output letter system: An ideal phonetic typewriter will strike out sentences successively corresponding to the conversational speech sound by whatever man or in whatever manner it may be uttered. As compared with Japanese writing, European writings have no such different kinds of letters as the Japanese alphabet (KANA letters) and the Chinese character, while the phonetic symbols and the spellings are considerably different. Because the Japanese language has the fundamental phonemic structure such as "a consonant plus a vowel", the correspondence between the speech sound and the letters is simple. Though, as the output letter system, there is no intrinsic difference between the Japanese alphabet and the Roman alphabet, the letters of the Japanese alphabet can be two or three times more rapidly struck out by a mechanical typewriter.

Extraction of controlling signals: If the speech sound is limited to the frequency range lower than W c/s, the waveform is correctly represented by the time series of the amplitudes sampled every 1/2W second. Assuming that those amplitudes

Toshiyuki SAKAI 坂井利之: Professor of Electrical Engineering Department, University of Kyoto.



Front View of the First Phonetic Typewriter



Back View of the First Phonetic Typewriter

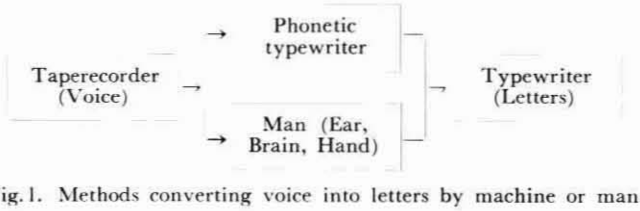


Fig. 1. Methods converting voice into letters by machine or man.

are digitalized to 2^H stages (for convenience of discussion the amplitude distribution is assumed to be uniform), every sample has H bits; considering that the number of samples produced is 2W per second, the entropy speed of the speech sound (information source) is $H'=2W\cdot H$ bits/second; if $W=8{,}000$ c/s and $H=4$, $H'=64{,}000$ bits/second is given. It is nearly impossible for the present electronic computer to discriminate such a great amount of information in real time, before the phenomenon comes to an end.

It may be considered that when we hear a sound, our ears do not take notice of such details of the sound but catch it more vaguely, more roughly; for, to say nothing of the difference between the voices of old men and children, men and women, even the waveforms of the same speech sound uttered by the same person at the same time are not absolutely the same from the standpoint of acoustic analysis and yet we can identify the phoneme.

Therefore, instead of the above mentioned mechanical, detailed sampling, it is considered suitable to extract necessary cues for discrimination from the speech wave to be analyzed. However, it is a difficult problem to decide which part of an unknown input speech sound is to be sampled and analyzed.

Based on the principle of discriminating phonemes which are the elements of every conversational speech sound, the phonetic typewriter of Kyoto University is provided with generality. However, for simplicity of the control system, monosyllables have been chosen as the object at the first stage. Incidentally it may be remarked that this kind of device announced by the present time (the end of 1960) is still in the stage of treating some specific words.

Determination of the sampling interval: For the analysis, it is necessary to determine which part of the speech wave is to be sampled; unstability of the sample has a bad influence on the analyzed results. The Japanese speech sound is composed of monosyllables having the phonetic structure of a consonant followed by a vowel. The duration of the consonant part is about 10 ms～200 ms. Vowels, having duration of more than 50 ms, are predominant over consonants; so that it has come to be considered desirable to treat consonants separated from the vowel part and independent of the predominance of the vowel.

Analysis: The analysis is to investigate the quality of the sampled speech sound; the frequency analysis and the zero-crossing wave analysis are very important analyses from the standpoint of real time processing.

The Frequency Analysis and the Zero-Crossing Wave Analysis

It has been known for a long time that after analyzing the human speech sound by some frequency band-pass filters, the corresponding sound can be reproduced from the analyzed signal. So that the attempts have been made to discriminate the speech sound by the method of frequency analysis. Each filtered

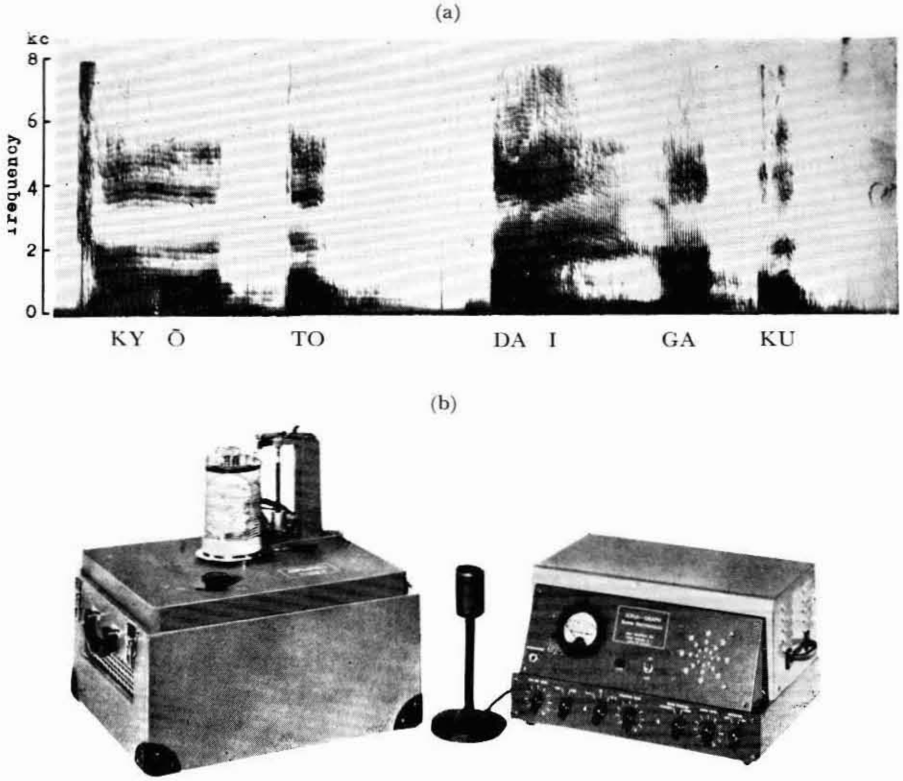


Fig. 2. Sonagram of a Japanese word (a) and Sonagraph (b)

component wave can be recorded by the electro-magnetic oscillograph; moreover the Sonagraph can make visible the three-dimensional analysis of the time, frequency and intensity, as in Fig. 2. In this figure, the time (within 2.4 sec.) is represented by abscissa, the frequency (within 8 kc) by ordinate, and the intensity is proportional to the blackness; thus the visible pattern peculiar to each sound is obtained.

However, from the standpoint of analysis, band width of analyzing filters (300 c/s or 75 c/s) must be narrow; on the other hand if it is too narrow it becomes unsuitable for the analysis of the consonant part changing rapidly, according to the uncertainty principle of frequency and time (the product of band width and time-resolving power is constant). Uttered rapidly, consonants including plosives of short duration have higher frequency components; moreover, making one hundred monosyllables by their combinations with the five vowels, the analysis of consonants is extremely important.

Thus we began the zero-crossing wave analysis for the following reasons: the amount of information given by the speech sound being very large, some simplified method must be adopted for its mechanical recognition; both sides of the speech sound, articulation and naturalness, are not always necessary.

The Principle of the Zero-Crossing Wave Analysis

In order to describe the waveform of the speech sound, the fundamental frequency (pitch), spectrum of higher harmonics (timber), amplitude, duration, and the form of the envelope(the growth and decay times) are required. It can be considered that, the minimum demand in communication is satisfied when articulation, at least, is preserved and the content of speech is understood even if naturalness is lost which identifies the individuality of the voice. Especially, in the phonetic typewriter converting the speech sound into typed letters, naturalness is not necessary.

The zero-crossing wave, discovered by J. C. R. Licklider, has 90% of articulation for a monosyllable. As shown in Fig. 3, the zero-crossing wave is the rectangular wave having, as the switching point, the zero point of sound pressure in the original waveform, and the constant amplitude. Let us describe, qualitatively, with the five Japanese vowels, why such a simplified waveform has high articulation. In uttering the Japanese vowels, the oral cavity, much modified by the tongue, makes a large resonant cavity. This characterizes the vowels "u", "o", "a", by the low resonance frequency called the first formant. Further there are the two vowels, "i" and "e" with the characteristic resonance frequencies, the first formant and the second formant, when the oral cavity is divided by the tongue into two larger and smaller resonant cavities.

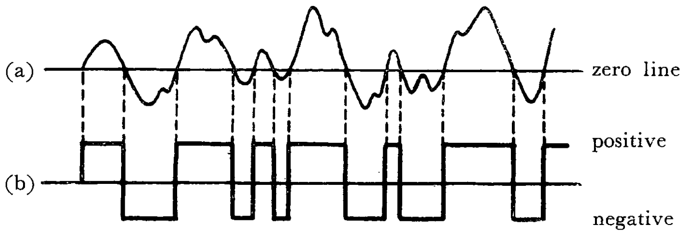


Fig. 3. Examples of (a) original speech wave
(b) zero-crossing wave

Now, as shown in Fig. 4, waveform of a vowel is a train of damped oscillation repeated at every pitch frequency of the vocal cords, which includes some numbers of formant oscillations (higher harmonics) in the pitch period. A vowel predominated by the first formant is displayed in the figure (a); a vowel by the first and second formants in the figure (b), where the waveforms of the first formant and that of the second formant overlap and usually the oscillatory amplitude of the latter is smaller. Therefore, it is easily understood from the figure that in the zero-crossing wave, the smaller oscillation becomes remarkable in the neighbourhood where the level of the first formant oscillation becomes smaller. Thus, as a result, both rectangular waves corresponding to the first and the second

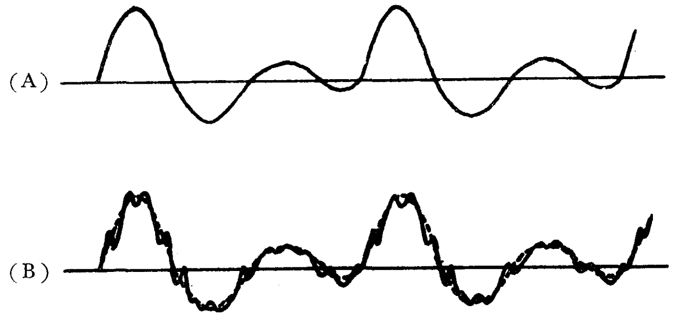


Fig. 4. Simplified waveforms of Japanese vowels : (a) "a", "o", "u", (b) "i", "e"

formants are included in the zero-crossing waveform.

Next, the zero-crossing wave is the rectangular wave whose switching point corresponds to the time point of zero level of the original wave. Experimentally the crossing level may be a certain level, and this is called the constant-level-crossing wave.

In rectangular wave the object to be measured and analyzed is its width and amplitude. In the zero-crossing wave, the amplitude, being constant, is insignificant as an information; so that the measurement of the rectangular pulse width and the time of its occurence is all the information included in the zero-crossing wave.

The measurement of the time interval is carried out by counting standard clock pulses. For example, in Fig. 5, starts the standard clock pulse oscillator at the same time with the front edge of the rectangular wave. By counting the number of clock pulses produced by the time corresponding to the end of the rectangular wave, the time interval is measured in the digitalized form.

The channel plan for the analysis is made according to the number of clock pulses (for example, as shown in Table 1) and is determined considering the following points: how many channels are necessary for the analysis of the speech sound, and

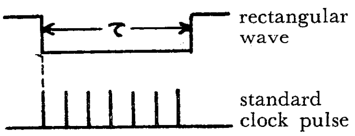


Fig. 5. Measurement of a rectangular width

Table 1. Channel classification of analyzer

Channel No.	No. of pulse count	Frequency (cps)	Channel No.	No. of pulse count	Frequency (cps)
1	2	11000-5500	9	10, 11	1210-1000
2	3	5500-3700	10	12, 13	1000- 830
3	4	3700-2800	11	14, 15	830- 740
4	5	2800-2200	12	16-19	740- 590
5	6	2200-1800	13	20-23	590- 485
6	7	1800-1560	14	24-31	485- 360
7	8	1560-1360	15	32-47	360- 240
8	9	1360-1210	16	48-63	240- 160

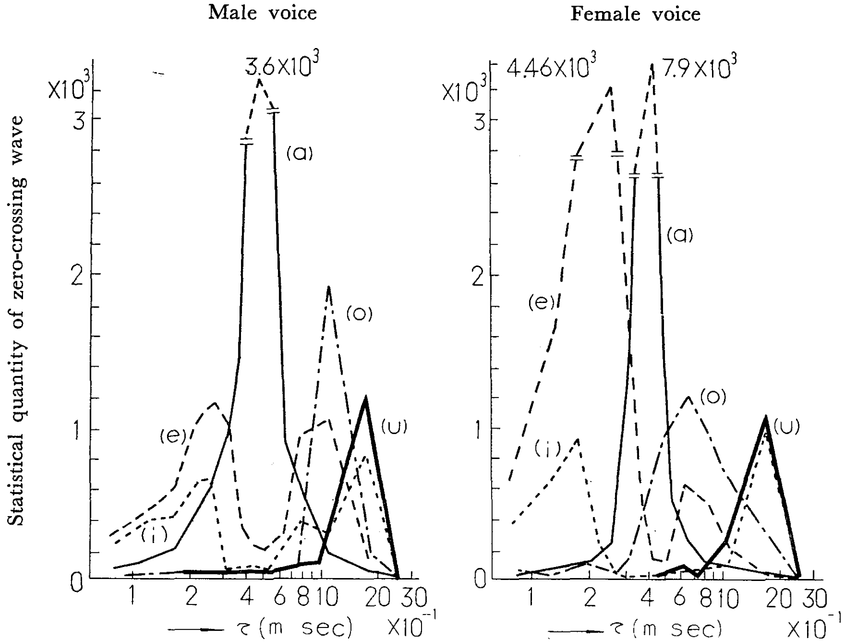


Fig. 6. Results of Japanese vowels by the zero-crossing analysis

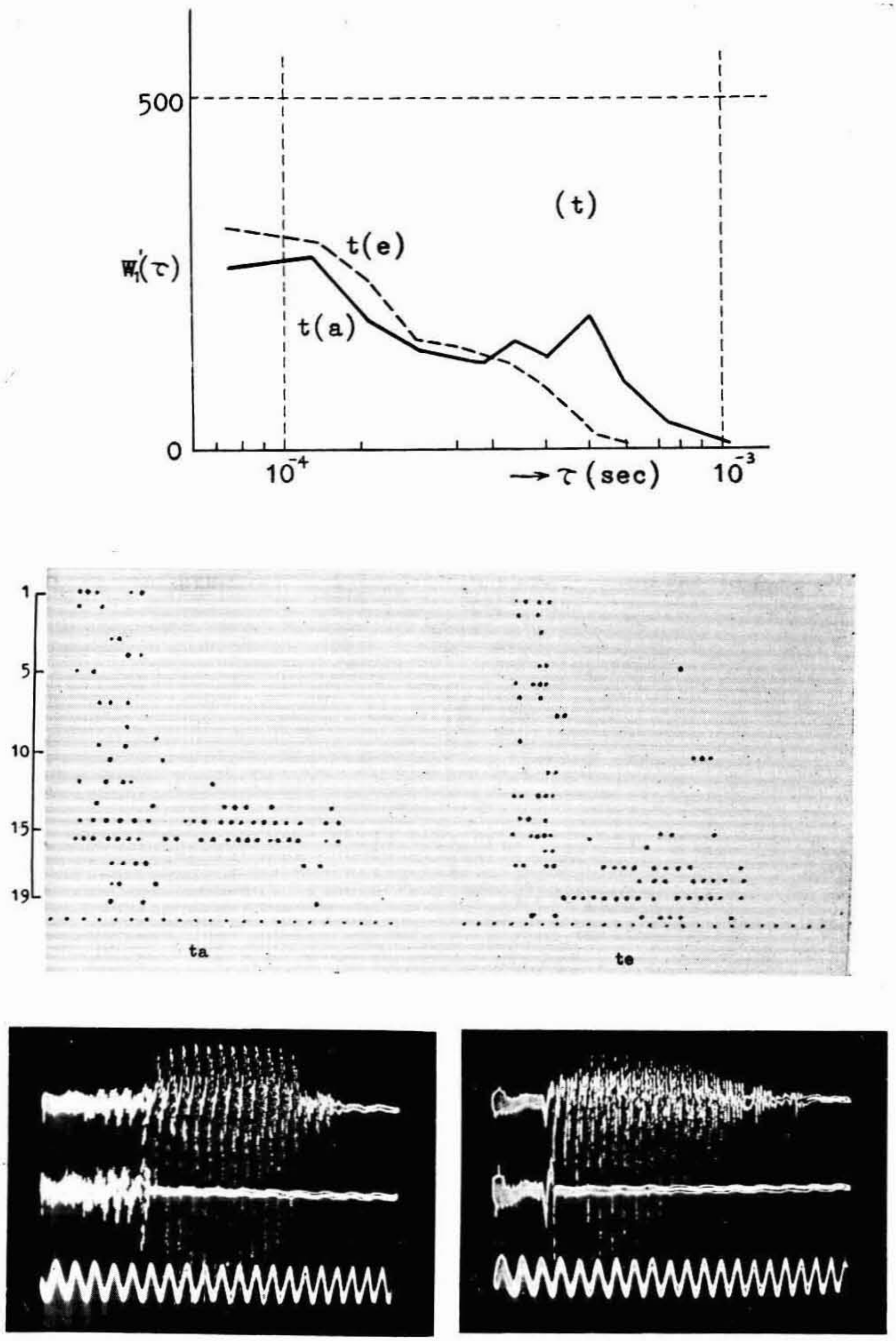
what frequency range is assigned to each channel.

Fig. 6 shows the distribution of the Japanese vowels, each uttered by man and woman, obtained by the method of zero-crossing wave analysis. Further, in Fig. 7 are shown the zero-crossing wave analysis of the consonat part, the time series of a zero-crossing wave, and the original waveform of consonant part, when consonant and vowel are separated by a mechanical device.

Merits and Demerits of the Zero-Crossing Wave Analysis

The zero-crossing wave analysis is convenient because the result is obtained without delay from the end of each rectangular wave and because the classification assignment and the number of channels are easily modified. Though the analysis is not limited by the uncertainty principle, it is not an easy work to grasp the whole pattern of the analysis, when each point indicating the existence of classified pulses is made visible as by a Sonagraph.

What is still more essential to be noticed is that as the waveform has been extremely simplified at first, it is lacking in some informations as speech sound, but there is no defect in the processing of zero-crossing wave. In the zero-crossing wave, the small signals are amplified to the comparative level with the large signals and the S/N ratio comes near to unity; so that the results of its analysis are influenced by noises. Therefore the integration of the count by the zero-crossing analysis and the preliminary filtering of the input sound by frequency filters before the analysis are the practical treatment in consideration of the regularities of signals and the randomness of noises.



t(a)

t(e)

Fig. 7. Various kinds of representation of speech sound

The Structure of the Research Model of Phonetic Typewriter (Sonotype)

The research model of the phonetic typewriter was constructed in order to investigate if the realization of the phonetic typewriter might be possible, and if possible, how it should be designed. Therefore it has been so constructed that it may be provided with functions as a typewriter and may work at the conversational speech sound in future as well as at Japanese monosyllables and that it may be utilized for languages other than Japanese. Further, the speech sound being continuous and complex analogue quantity, it would not be too much to say that the method and reference value to be adopted in digitalizing the speech sound have been scarcely established. In order that the experiments of the speech sound digitalization may be easily carried out, our phonetic typewriter is provided with the logical circuits network on the plug board system and is so designed as to be convenient for determining the threshold value and confirming the code pattern.

The phonetic typewriter is, roughly speaking, composed of three functional parts, the phoneme classifier, the sampling-control circuits, and the analyzing circuits. One hundred Japanese monosyllables, which are the information source for this phonetic typewriter, are classified according to the classification of phonemes as in Table 2. Assuming that the speech sound waveform, which is the representation in time region of what has been originated from the articulation of the vocal organs, includes traces of the manner of articulation; the phoneme classifier recognizes, by the combinational circuits of several digitalized outputs from the distinctive feature extractors, whether the input speech sound is one of vowels, unvoiced sounds, voiced sounds, plosives, nazals.

Table 2. Classification of Japanese phonemes

- Japanese monosyllable
  - voiced
    - vowel ; |a|, |i|, |u|, |e|, |o|
    - consonant
      - fricative ; |z|, |ʒ|
      - plosive ; |b|, |d|, |g|,|r|
      - nasal ; |m|, |n|
      - semi-vowel ; |w|, |j|
  - unvoiced
    - fricative ; |s|, |ʃ|, |h|
    - affricate ; |ts|, |tʃ|
    - plosive ; |p|, |t|, |k|

The reasons for providing the phoneme classifier are as follows. First, it was reasoned from a large number of experimental data that it was almost impossible to put Sonagraph patterns or analyzed patterns of the zero-crossing wave in one-to-one correspondence with one hundred monosyllables. Secondly, in order that the device may have sufficient merginal region against the difference of the speakers or utterances, the phonetic typewriter must be so designed that precise, detailed analysis may not be necessary; thus it is considered that the rough classification of sounds according to the manner of articulation is a useful method.

The complexity of the analyzed patterns is due to the fact that the sound utterances are various and not similar, so that it is unavoidable for us to find difficulty in recognizing all the differences by simple analysis of one kind.

Next, the main purpose of the sampling-control circuits is to determine what part of the input speech should be sent to the analyzing circuits. In this case the control needs to be changed according to voiced sounds, unvoiced sounds, or vowels.

The reason is this: in voiced sounds, caused by the oscillation of the vocal cords, the oscillation called the buzz comes out first, and as it has no information for the articulation it is insignificant to sample this section. When an unknown sound comes in, though whether it is voiced or unvoiced is not known to the device, its treatment must be changed accordingly. So that the sampling-control circuits and the phoneme classifier are correlated and reliability of the analysis influenced by the suitability of the sampling-control circuit.

In the conversational speech sound the control becomes further complicated; signals must be given, showing the beginning and the end of a syllable or a vowel section in the continuous speech sound which has no distinction of the beginning and the end. For reference, an oscillogram and a Sonagram of a conversational speech sound are shown in Fig. 8. Moreover, the Japanese accent, which is not

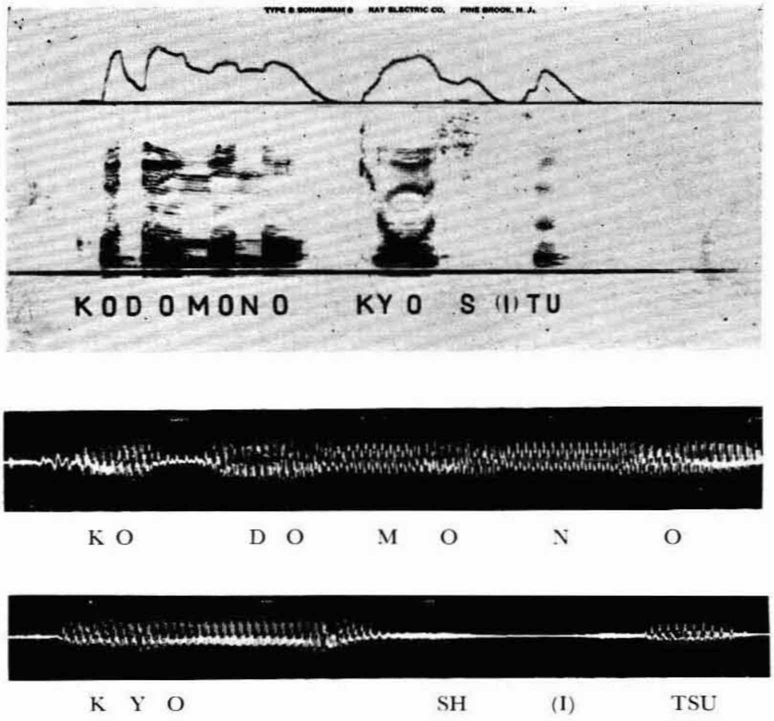


Fig. 8. An example of conversational speech (KO DO MO NO KYO SHI TSU)

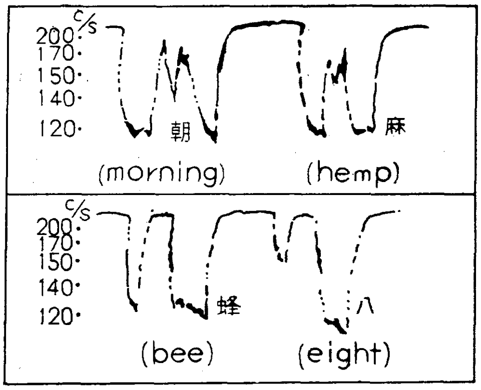


Fig. 9. Musical accent of Japanese

the stress accent as in European languages, is necessary to distinguish homonyms written in KANA letters. It will be understood from Fig. 9, that the pitch frequency changes according to the accent.

Next, let us describe the analysis. Though our phonetic typewriter incapable of detailed analysis of the zero-crossing wave, the analysis required in discriminating each phoneme need not be so detailed. The more detailed the analysis is, the more influence is exerted on the analyzed results by individuality and difference of the utterance. The analyzed results necessary to discriminate a sound belonging to the same phoneme classification are identified with the statistics of the utterance of the one hundred Japanese monosyllables by a large number of people. For example, as unvoiced plosive sounds, there are three "k", "t", "p" series; when "a" is given as the following vowel, only the analyzed data enough to distinguish "ka", "ta", "pa" are required. The discrimination to this extent is extremely simple, and the optimum classification of detailed channels and setting up of counts are determined by the statistical data of many utterances. Those characteristic channels are obtained by collecting some channels in the original detailed zero-crossing analysis, by means of "OR" logical circuits, and reorganizing them into new, optimum analyzing channels for the phonemes. The benefit of the zero-crossing wave analysis is that no overlap nor loss of counting is caused in it by the channel reorganization. In these reorganized channels the threshold value of the counts are set up; the analyzed number more than that is digitalized as "1" and the number less than that as "0"; the threshold is determined, of course, by the statistics of the analyzed results. The reason we judged the consonant part including the influence of the following vowel is this: it is evident from analyses that the consonant part undergoes a remarkable influence of the vowel part. In a plosive, as it must be uttered rapidly, the place of articulation is so set beforehand that the following vowel may be easily uttered. That is, from the standpoint of phonetic analysis, "ka", "ki"…"ko" expressed in Roman letters have no common and fixed acoustic quality of "k"; therefore they are called phonemes on the KANA letter system. Whereas "sa", "si"…"so" have the common and fixed acoustic quality of "s", and they are called phonemes on the Roman letter system. Thus the threshold value of digitalization based upon the statistics is not available in common with different phonemes; so that the threshold value is so designed that it may be most suitable for every group of phonemes and every monosyllable.

For the reorganization of those channels the logical wiring circuits on the plug board system have been provided so that the channels may be employed in treating, not only the utterance of standard Japanese speech sound, but also dialects and foreign languages. The logical circuits mean the circuits of logical variables, taking either "1" or "0", and have such circuits as "AND", "OR", and "NOT". The electronic computer is composed of their combinations and they are utilized everywhere in our phonetic typewriter.

Fig. 10 shows the block diagram and the photograph of the structure of our phonetic typewriter. When the monosyllables uttered through a microphone or recorded on a recording tape are sent into the phonetic typewriter, the instructions are given about the manner of articulation and the sampling section of the input speech sound. According to the instructions the analysis is carried out by the cascaded circuits of filter and zero-crossing analyzer respectively for vowels, unvoiced sounds, voiced sounds, and nasals. In the channel corresponding to the width of each rectangular wave composing the zero-crossing wave series, one output pulse comes out as soon as the rectangular wave ends. In the level selector, the number of classified pulses is counted by the integration counter in every reorganized channel, by sending a constant electric current to the condenser for a (certain) time; when the number is more than the threshold value the digitalized pulse "1" comes out. The pulse showing the existence of "1" or "0" is sent into the holding circuits and is retained till the judgement of the relevant monosyllable is finished and the instruction of reset is given for the next sound.

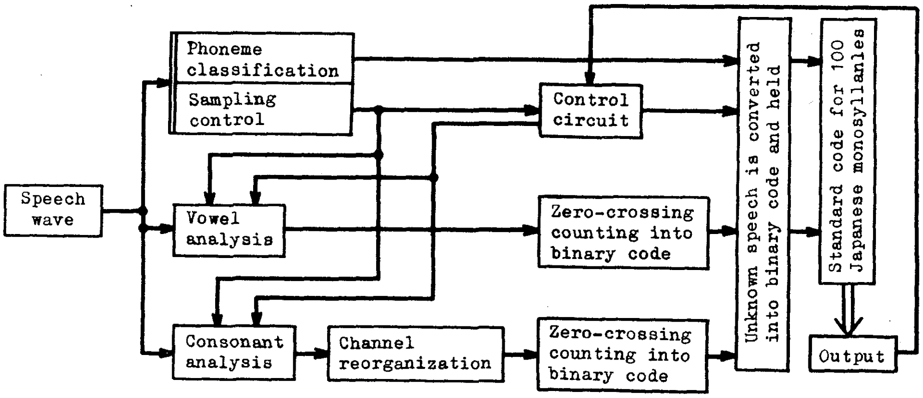

Phoneme classification
Sampling control
Control circuit
Speech wave
Vowel analysis
Zero-crossing counting into binary code
Consonant analysis
Channel reorganization
Zero-crossing counting into binary code
Unknown speech is converted into binary code and held
Standard code for 100 Japanese monosyllanles
Output

Fig. 10. Block diagram of phonetic typewriter

The discrimination of the vowels exerts a great influence on the discrimination score; the vowels are discriminated on the plane of the first formant and the second formant; and the result of discrimination, together with the phoneme classification of speech sounds, is also displayed by lamps. Various kinds of the extractors (questions), analyzing circuits, etc., provided parallely beforehand, work at the same time with the startpoint of a monosyllable, sending their results into the main judging circuits. In other words, an unknown input speech sound is so conditioned that it may say "YES" or "NO" to the questions stored in the "device". The questions concern the classification of vowels, voiced sounds, unvoiced sounds, plosives, nasals, contracted sounds, etc., the existence of high frequency components, the length of the duration, which are necessary for the classification of speech sounds, and also the existence of the output in the analyzing channels for discrimination of each monosyllable. In the device seventy items of questions are allowed and the answers for these questions are identified with the standard answers for each monosyllable (the standardized pattern of a monosyllable based upon the statistics) which are built in as a pattern (code) by inserting diodes horizontally into the main judging circuits. Therefore, in the vertical direction in the main judging circuits, there is capacity for one hundred monosyllables and some special codes.

When an unknown speech sound is put in, "YES" or "NO" to the questions is automatically sent by the above method to the relevant position in the main judging circuits; when the answer accords with a stored standard pattern it is judged that the monosyllable equivalent to the standard pattern has been sent in as the input.

There are one hundred and more possibilities of judgement, that is more than 2^6 and less than 2^8; so that they are transformed into signals on eight units code of paper tape and printed in KANA letters. The special codes include the prolonged vowel signal, and the space signals; when the analyzed pattern does not accord with the prepared signals, "?" can also be printed.

We have described the phonetic typewriter (Sonotype) research in Kyoto University; however, the device has not been accomplished and there will be modifications and reconstructions in future. The practical phonetic typewriter will be realized when our research emerges from the present method of pattern recognition in which the absolute value is used on frequency, amplitude, duration and the total number of zero-crossing, and when the pattern comes to be recognized certainly and relatively. It may be noticed that this method is the same with the fundamental transforming circuits employed in compressing the frequency band to one or two hundreds c/s in the telephone communication system.