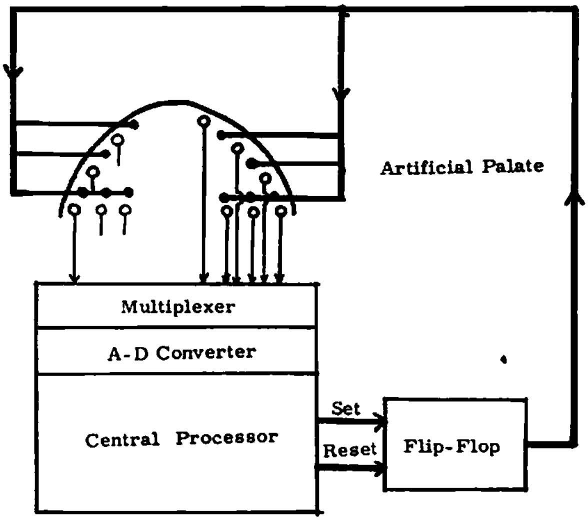
PHONEME IDENTIFICATION WITH DYNAMIC PALATOGRAPHY

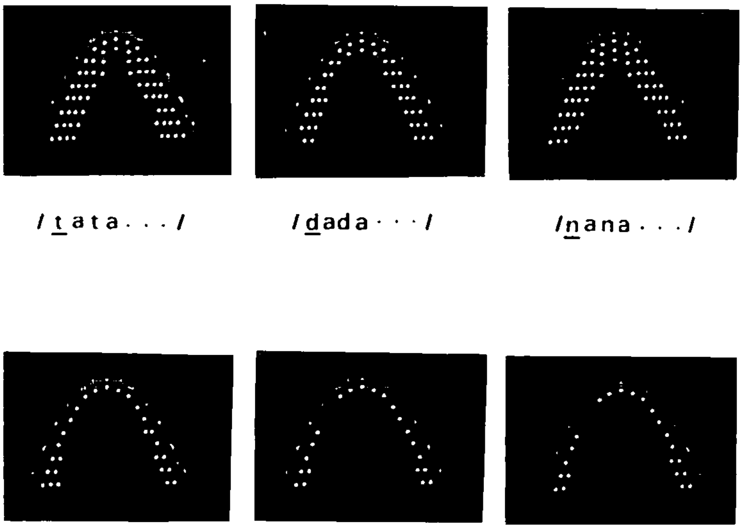
Itaru Fujii

Phoneme identification experiments based on acoustic signals have shown the overwhelming complexity and difficulty of the problem in general. Our interest is to see how different the phoneme identification problem might be if we use articulatory information. A thin artificial palate[1)] with four banks of electrodes implanted along the tooth curvature was used. The banks of electrodes labeled groups 1, 2, 3 and 4, numbered from the outer rim in, have 19, 17, 15 and 13 pairs of electrodes, respectively. A pair of electrodes consists of an input electrode and an output electrode. The input electrodes are connected to a bus-line and the 64 output electrodes to a 64-channel multiplexer. The signals from the different output channels are stored in a computer (see Figure 1).



Block Diagram Figure 1.

Pulses obtained from a PDP-9 computer, trigger a flip-flop, producing a rectangular waveform with 10,000 pulses per second. The signals appear at those output electrodes of the palate where the tongue shorts the pairs of electrodes. The palatal signals thus obtained are quantized into 6 bit levels and sampled every 10 msec. In this way palatographic data covering a 2 sec. interval can be stored in the 8 k word core memory of the PDP-9. The stored signals are then dichotomized in reference to threshold levles, which are set separately for individual electrodes before processing. The contact patterns between the tongue and the hard palate can be displayed on an oscilloscope display as a slow motion picture for preliminary inspections. Figure 2 compares patterns for selected frames of the utterances /tatata. . . /, /dadada . . . / and /nanana. . . /.



/tata. . . / /dada. . . / /nana. . . /

/tata. . . / /dada. . . / /nana. . . /

Figure 2.

Phonemes are identified by time series of contact patterns. For each phoneme /P/ a maximum pattern MAX[P], a minimum pattern MIN[P], and the minimum time duration in terms of the frame number DUR[P] are determined from the visual inspection of palatospectrograms.[2)]

Nonsense words of the form /VCV/ without accent nuclei were spoken by a native male speaker of Japanese (Tokyo dialect). Each of the consonants [t, d, n, s, z, ʃ, ʒ, ts, dz, tʃ, dʒ] was embedded in all the possible vowel contexts using the five vowels [i, e, a, o, u]. Using utterances for all of the /VCV/ samples as reference patterns, the maximum and minimum patterns and the minimum time durations were determined for the C's and V's as follows.

The sustained portion of a given phoneme /P/ is first defined as its "inherent segment" by visual inspection of each palatospectrogram which contains the phoneme /P/ (see Figure 3, the middle segment of the palatospectrogram designated as /t/). In the case of the dental stop consonants, a complete contact of all group 1 electrodes defines the closure period of /t/ as its inherent segment. Since it is not easy to give criteria for defining the inherent segments of such phonemes as the vowels [i, e, a, o, u] and the fricative consonants [s, z, ʃ , ʒ] from their contact patterns alone, the sound spectrograms, which also are displayed in the palatospectrograms, were used to assist in their definitions.

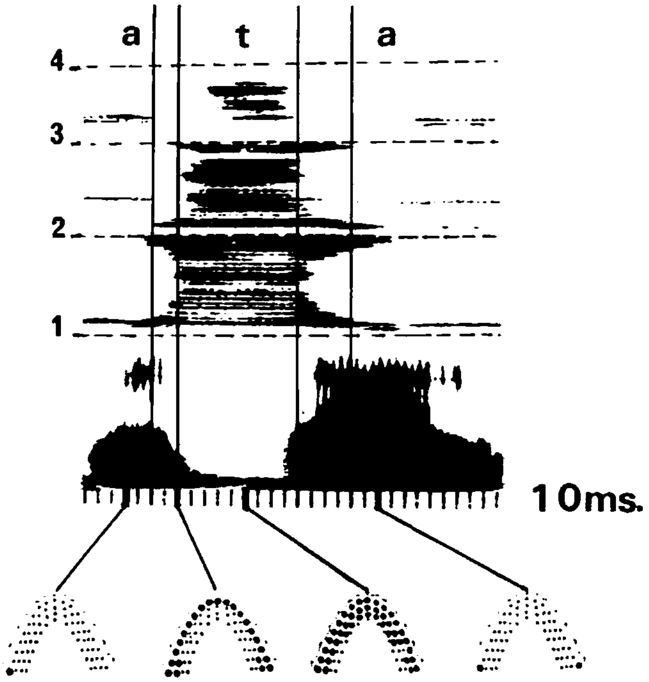


Figure 3.

After the inherent segments had been determined for a /VCV/ sample, let us say sample i, for a given phoneme, $\max_i(P)$ and $\min_i(P)$ were obtained for the pertinent phoneme from a set of pattern frames for the sample utterance. Thus among the many pattern frames contained within the inherent segment of /t/ in the case of Figure 3, the maximal pattern $\max_i(t)$, was obtained as the palatal pattern that contained the highest number of contact points (thick dots in Figure 3) and the minimal pattern $\min_i(t)$ as the pattern with the fewest contact points. In a similar way max(t)'s and min(t)'s of other samples containing the same phoneme /t/ in different vowel contexts were obtained and stored in the computer memory. The maximum pattern MAX(t) and the minimum pattern MIN(t) of the phoneme /t/ were then defined by use of the max(t)'s and min(t)'s:

$$\mathrm{MAX}(t) = \max_1(t) \vee \max_2(t) \vee \ldots \vee \max_n(t) \qquad (1)$$

$$\mathrm{MIN}(t) = \min_1(t) \wedge \min_2(t) \wedge \ldots \wedge \min_n(t) \qquad (2)$$

where n represents the total number of samples containing the phoneme /t/, and $\wedge$, v designate logical AND and OR, respectively. Then the time durations of the inherent segments for the phoneme /t/ were measured for all relevant samples and the minimum among them was selected as DUR(t). Similar procedures were taken for the other consonants and vowels.

Based on the MAX(P), MIN(P) and DUR(P) for all P's, time series of the unknown input pattern frames were automatically examined as follows. In pattern i, frame X_i of an input series is a "candidate frame" for a phoneme /P/ when the following condition is satisfied.

$$\mathrm{MAX}(P) \geq X_i \geqq \mathrm{MIN}(P), \qquad (3)$$

where the relation $A \geqq B$ signifies that the pattern B of the contact points is included in (or equal to) the pattern A. When succeeding input pattern frames of a series $\chi = X_1, ---X_n$ are identified as candidate frames for the phoneme /P/ continuously for n frames, and if this number of frames $n(\chi)$ is larger than or equal to DUR(P), then this series of patterns χ is identified as the phoneme /P/.

In Table 1, the utterance list for the identification experiment is shown. These samples were uttered by the same speaker employed in determining MAX(P), MIN(P) and DUR(P) in a different session. The vowels [i, e, a, o, u] and the consonants [t, s, ts, ʃ, tʃ] were identified in this experiment. Other phonemes, in particular [p, b, k, g], were omitted because in

[t, d, n]		/ita/	/uta/	/eta/	/ota/
[s, z]		/isa/	/usa/	/esa/	/osa/
	/asu/	/isu/	/usu/	/esu/	
[ts, dz]	/aca/	/ica/	/uca/	/eca/	/oca/
	/acu/	/icu/	/ucu/	/ecu/	
[tʃ, dʒ]	/acja/		/ucja/	/ecja/	/ocja/
	/aci/		/uci/	/eci/	
[ʃ, ʒ]	/asi/	/isi/	/usi/	/esi/	
	/asja/	/isja/	/usja/	/esja/	/osja/

Table 1.

Utterance List

the case of [p, b] there is no characteristic palatal contact, and the arrangement of the electrodes for the palates employed in this study was not suitable for identifying [k, g]. In Table 1, such a pair as [s, z] will have to be distinguished from each other based on other kinds of information, for instance, simple characteristics of the acoustic waveform, since there was observed no characteristic difference between the contact patterns of [s] and [z]. Although a consistent difference was observed in the patterns of [t] and [d], if we compared them in the same context, they were treated as constituting a single class since contexts (i. e., the preceding and following vowels) were not considered in this experiment. A treatment of /r/ would require an approach based on characteristic temporary changes of the pattern, which is beyond the scope of this report.

Frame No.	**Contact Pattern**
1. . . . 5	0
6. . . . 22	i
23	sj
24. . . 36	t
37	s
38. . . 42	u/e
43	0
44. . . 66	a

Table 2.

Result / i t a /

Table 2 shows a sample of a computer output for the identification of a phoneme. The series of numbers in the left column shows the numbering of the input pattern for the utterance [ita], and the candidate frame for the input pattern is given to the right, where zero indicates rejection, that is, the input pattern was judged not to be a candidate for any phoneme. At the bottom of Table 2, the phoneme identification made in this manner is shown. In this table, a series of input patterns appearing less than DUR(P) for any phoneme is neglected in the final judgment. The phonemes [ts] and [tʃ] were treated as the sequences of the phonemes [t] and [s], and [t] and [ʃ] respectively.

OUTPUT

INPUT

	a,o	i	u	e	t	ts	s	tʃ	ʃ	total	%
a,o	35									35	100
i		14								14	100
u			10				8			18	56
e				9						9	100
t					4					4	100
ts						9				9	100
s							9			9	100
tʃ								7		7	100
ʃ									9	9	100

Confusion Matrix

Table 3.

The results of the phoneme identification tests conducted in this manner are shown in Table 3 in a form of confusion matrix. All phonemes except [u] were recognized correctly, but all[u]'s in the position [-su] and [-tsu] were identified as [s], because [u] of this position showed the same patterns as[s]. It is recognized by phoneticians that the articulation of this vowel in these contexts is quite similar to that of the preceding consonant.

From this pilot study of phoneme identification with an extremely simple and straightforward algorithm, it may be clear that phoneme identification on the articulatory level is much simpler than that on the acoustic level for some particular distinctions that tend to make the acoustic approach formidably complex. We also hope that this approach will be helpful in constructing a successful modeling of the speech dynamics.

References

1) S Shibata, "A Study of Dynamic Palatography," Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo), No. 2, 28-36 (1968).

2) O. Fujimura and S. Shibata, "A Study of Dynamic Palatography," Reports of the 6th International Congress on Acoustics, B-1-7 (1968)