

DYNAMIC SPEECH DISCRIMINATION USING AN ARTICULATORY MODEL

Shun Ishizaki

Pattern Information Division
Electrotechnical Laboratory
2-6-1 Nagata-cho, Chiyoda-ku
Tokyo, JAPAN 100

The author proposes a dynamic processing method corresponding to the human one. It uses an articulatory model representing the tongue and the lip movements to extract speech features from speech waveforms. He uses a linear enhancement of the articulatory movements in order to estimate their target articulatory positions. When the target positions are used for vowel discrimination, the correct rate is improved fairly well. For example, ten males' symmetrical vowel $V_1V_2V_1$ are analyzed and V_2 discrimination correct rate is improved from 85% to 100%. Next, hearing tests show that such dynamic process corresponds well to the human auditory system.

1. INTRODUCTION

In a study on coarticulation using formant frequencies, there are various researches[1-5], including observation of vowel's formant movements, their model and coarticulation normalization using the first and the second formants. On the other hand, coarticulation is considered to be a phenomenon of various changes in articulatory conditions of the phoneme by the preceding and the succeeding phoneme influence. It seems that the changes in articulation will depend directly on the articulatory organ's movements, such as the tongue and the lips, and that a talker allows "movement reduction of articulatory organs" so that he might not spoil the spoken content clearness by hearing his speech through his auditory organs. Such hearing system shares a feedback loop with the speaking system.

In an articulatory domain, there are various studies based on X-ray photograph observations [6-8]. But very few studies apply them to the actual speech waveform analysis data. The author starts from approximately estimated vocal tract area functions from speech waveforms, and estimate the articulatory parameters expressing the tongue and the lip movements by using an articulatory model. Then, he proposes a dynamic processing method of coarticulation and discusses its correspondence with hearing tests of the sound segmented from natural voices.

2. ARTICULATORY PARAMETER EXTRACTION[10]

The speech wave analysis is made according to the procedure shown in Fig. 1. Vocal tract area functions are approximately estimated from the speech waves using an adaptive inverse filtering method of the second and the third order critical damping system[9]. Next, the vocal tract length is obtained by the "an information criterion" based on the maximum likelihood estimation by use of an autoregressive model[11]. Sizes of speaker's own articulatory organs are quantitatively estimated by using the estimated

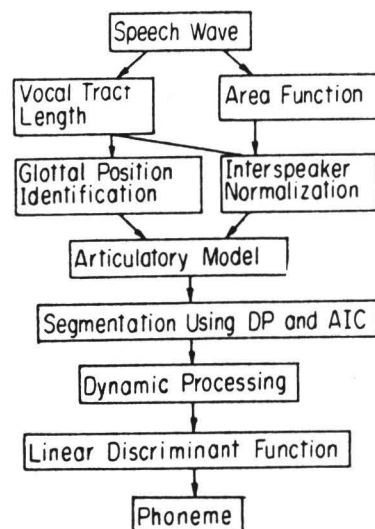


Fig. 1 Speech analysis system flow chart

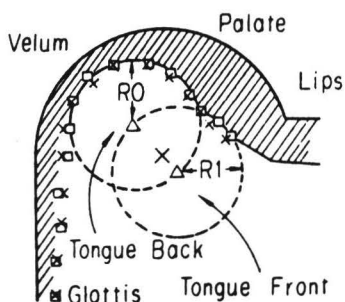


Fig. 2 Two-cylinder articulatory model
 x; area function estimated from speech wave
 o; area function identified with the model

vocal tract length, and individual differences are approximately normalized by uniforming the sizes to a certain constant value[10].

A vocal tract midsagittal cross sectional view is modeled as shown in Fig. 2. The "constriction position" between the upper jaw and a part of the tongue, is uniquely determined by this model, and is defined as the length measured from the glottis. By adding the "constriction area" and the "lip area" to the "constriction position", these three articulatory parameters are defined. The constriction position and the constriction area correspond well to the values measured from X-ray photographs, while the lip area corresponds well to the value measured from simultaneously recorded video tapes. Linear discriminant functions are calculated from these three articulatory parameters of isolated vowels, and discriminate them at 100% correct rate.

3. DYNAMIC PROCESS OF COARTICULATION

Coarticulation appears in almost all the phonemes of a continuous speech, and the phonetic characteristics are changed. These changed characteristics cause large error in speech recognition. When these phonemes are discriminated by only extracting static characteristics, discrimination algorithm cannot but become extensive. As mentioned later, it is known that hearing confusion phenomena are presented and caused by coarticulation. Namely, the same data can become a different phoneme by influence of the preceding and the succeeding phonemes.

Let a vowel sequence be $\dots, V_1, V_2, V_3, \dots$ and suppose that the influence of coarticulation is expressed by a linear combination of each articulatory parameters and is equal in the forward and backward directions of time. Let the values of articulatory parameters in quasi-stationary parts of triple continuous vowels V_1, V_2, V_3 be P_1, P_2 and P_3 , respectively, and the

target value $\hat{P}_2$ of P_2 based on the dynamic process is approximated with the following linear formulas.

$$\hat{P}_2 = P_2 + P_{21} + P_{22},$$

$$P_{21} = \alpha(P_2 - P_1),$$

$$P_{22} = -\alpha'(P_3 - P_2).$$

The α (α') is a function of both time constant of the lip or the tongue dynamic characteristics and a voicing time. For the articulatory organ which is easy to move because of a small time-constant, α is set to a small value, while for the articulatory organ not so easy to move, α is set to a large value. At the same time, α is a function of a voicing time. If there is a sufficient time to voice, the articulatory organs sometimes reach near the target positions and it is, therefore, natural that α should be made small. As there is a strong correlation between the movement distance and the voicing time of each articulatory organ, α is made small when the movement distance is beyond a threshold. Examples of dynamic process are shown in Fig. 3. It shows step function approximation obtained after the articulatory parameter segmentation, and Fig. 3(b) shows the target values after the dynamic process.

4. EXPERIMENTAL RESULT

Analyse are made on symmetrical triple-continuous vowel voiced at relatively fast

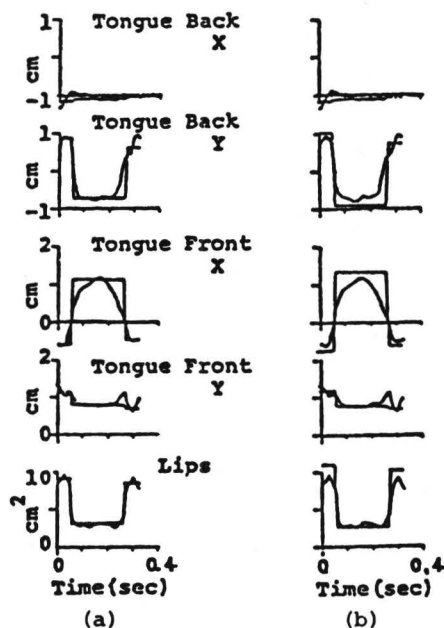


Fig. 3 /aia/'s articulatory parameters and
 (a) step functions obtained after segmentation
 (b) enhanced step functions with dynamic process

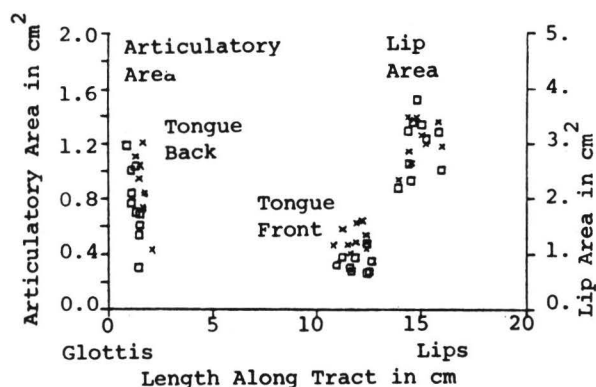


Fig. 4 Articulatory parameters of /i/ in /aia/ x; static model, o; dynamic model

Table 1 V_2 confusion matrices in $V_1 V_2 V_1$ voiced with relatively fast speeds. (a) matrix before operating the dynamic process, (b) matrix after it.

	a	i	u	e	o
a	35			4	1
i		32		8	
u	1		30		9
e	1			39	
o	1		5		34
correct rate = 85%					

	a	i	u	e	o
a	40				
i		40			
u			40		
e				40	
o					40
correct rate = 100%					

speeds by ten male adults. Fig. 4 shows their articulatory parameters of /aia/. The mark "x" denotes the value before dynamic process. The mark "o" denotes the value after the dynamic process. In Fig. 4, three vowels, /i/, are erroneously discriminated as /e/, with large articulatory areas of the tongue front. However, each articulatory area can be made sufficiently small with the dynamic process mentioned in Section 3. V_2 discrimination results in $V_1 V_2 V_1$ before and after the dynamic process are shown in Table 1.

5. CORRESPONDENCE WITH HEARING TEST

The fourth column from the left in Table 2 shows the hearing test where V_2 is segmented from $V_1 V_2 V_1$ and presented five times to ten subjects. From the hearing tests, there are several almost complete confusion phenomena, and as compared with the fifth column from the left in the table, the human dynamic function at the perceptual process are found. When the hearing tests are compared with the discrimination

results for segmented speech, most of the confusion tendencies agree with each other. In table 2, each /i/ after the dynamic process of /aia/, /uiu/ and /oio/ has common features corresponding well to [i] [13]. They are separated using a new category [i] and indicated as /i/ in Table 2.

ACKNOWLEDGEMENT

The author would like to thank Dr. Nakajima and Mr. Fuchi for their helpfull advice.

REFERENCES

- [1] Lindblom, B, JASA 35, 11, p.1773 (1963)
- [2] Ohman, S.E.G, JASA 39, pp.151-168 (1966)
- [3] Sato, Y. and Fujisaki, H, JASJ 34, 3 (1978)
- [4] Kasuya, H, Tohoku Univ. Doctoral Thesis (1973)
- [5] Kuwabara, H. and Sakai, H, JASJ 29, 2 (1973)
- [6] Ohman, S.E.G, JASA 41, 2, p.310 (1967)
- [7] Lindblom, B. and Sundberg, J, JASA 50, 4 (1971)
- [8] Kiritani, T, et al, JASJ 34, 3, p.132 (1978)
- [9] Nakajima, T, et al, Bul. Electrotech. Lab, 37, 4, p.462 (1973)
- [10] Ishizaki, S, Trans. IECE Japan, J61-A, 10, p.950 (1978)
- [11] Ishizaki, S, Trans. IECE Japan, E61, 5 (1978)
- [12] Ishizaki, S, 4-th IJCPR, p.1050 (1978)
- [13] Fant G, "Acoustic theory of speech production", Mouton (1960)

Table 2 Hearing tests and discrimination results of $V_1 V_2 V_1$ voiced with fairly fast speeds by a male.

Sample $V_1 V_2 V_1$	V_2 Discrimination		Hearing Test	
	Static Model	Dynamic Model	Stimulus V_2 (%)	Stimulus $V_1 V_2 V_1$ (%)
i a i	e	a	e: 98	a: 100
u a u	o	a	a: 44 o: 34	a: 100
e a e	e	a	e: 100	a: 100
o a o	o	a	a: 76 o: 24	a: 100
a i a	e	i	e: 58 i: 24	i: 100
u i u	u	i	u: 100	i: 94
e i e	i	i	i: 100	i: 100
o i o	e	i	i: 46 e: 30	i: 94
a u a	u	u	u: 92	u: 100
i u i	u	u	u: 100	u: 100
e u e	u	u	u: 100	u: 100
o u o	u	u	u: 54 o: 46	u: 98
a e a	e	e	e: 92	e: 80 i: 20
i e i	e	e	e: 68 i: 32	e: 100
u e u	u	e	e: 74 u: 14	e: 100
o e o	u	e	e: 46 u: 22	e: 86 i: 10
a o a	o	o	o: 78 a: 22	o: 88
i o i	u	o	u: 92 o: 8	o: 100
u o u	o	o	u: 84 o: 16	o: 100
e o e	u	o	o: 64 u: 34	o: 78 u: 18
Correct Rate	45%	100%	54.1%	95.9%