RELATIONSHIP BETWEEN LIP SHAPES AND ACOUSTICAL CHARACTERISTICS DURING SPEECH

Keisuke MORI and Yorinobu SONODA***

*Information Processing Educational and Research Institute, Kyushu Kyoritu University,
Jiyugaoka 1-8 , Yahatanishi 807-8585, Japan
**Faculty of Engineering, Kumamoto University Kurokami 2-39-1, Kumamoto 860, Japan.

ABSTRACT

A quantitative knowledge of the articulatory characteristics is necessary for understanding the dynamics of speech production. Accordingly, it is expected that observations of the shape of mouth will provide useful data for the study on articulatory behaviors in speech production.

This paper describes characteristic changes in the shape of the mouth on the basis of processed image data taken by high-speed video recorder, and studies recognition tests jointly using articulatory behavior of lips and sound pattern during speech.

The speech material used in this paper were nonsense words of form /eCVCe/ (V: a, i, u, e, o, C: p, b, m). Subject were four adult males, all of them were native speaker of Japanese.

In recognition of the consonant, consonant was more closely related with shape pattern than formant pattern. These results show effect of consonant (/p/, /b/, and /m/) on the middle vowel of utterance.

1. INTRODUCTION

Speech is the most fundamental means of human communication. Since acoustic speech signals are produced as a result of skillful coordinated movements of various articulatory organs, it is highly important to investigate their articulatory behaviors on a physiological basis as well as speech sounds on an acoustic one. A quantitative knowledge of the articulatory characteristics is necessary for understanding the dynamics of speech production. Until now, we have been studying articulatory dynamics of tongue and jaw movements on the basis of articulatory measurements[1][2]. Among various movements of speech organs, changes in the shapes of mouth can be observed from the external configurations which play an important role regarding acoustic properties of the speech sound. Accordingly, it is expected that these visible observations will provide useful data for the study on articulatory behaviors in speech production [3][4]. Lip shapes, one of the usual phonetic subject, have been one of stronger interest in the knowledge-based synthesis of person's facial expression. This paper describes characteristic changes in the shape of the mouth in the production of Japanese on the basis of processed image data taken by high-speed video recorder, and studies recognition tests jointly using articulatory behavior of lips and sound pattern during speech.

2. MEASURING SYSTEM AND SPEECH MATERIAL

Lip movements were observed by using a high-speed video recorder and then recorded image data were processed by a small computer system equipped with an image processor unit. In order to observe rapid changes of lip shapes, it is necessary to monitor articulatory movements of lip at a fast frame rate. Image data were sampled at 200 frames per second. The mouth of speaker was illuminated by a stroboscopic light. To improve the accuracy of analysis and eliminate the time consumption for image processing, color marks were painted on the lip of the subjects. The recorded image data were quantized and digitized into 256 x 256 pixels with six bits (64 gray level), and stored into the magneto-optical disk. At the same time, speech sound was recorded on digital recorder. The recorded speech sound were quantized and digitized into 48 KHz sample with 16 bit.

The speech material used in this paper were nonsense words of form /eCVCe/ (V: a, i, u, e, o, C: p, b, m, 15 words). Subject were four adult males, HF, KM, KO. and YN, all of them were native speaker of Japanese. Words were randomly arranged and repeated three times for each word at slow (normal) speaking rate (45 samples at each Subjects).

3. DETECTION AND DESCRIPTION OF LIP SHAPES AND SPEECH SOUND

Measurements of various dimensions relevant to the mouth opening were made to provide some quantitative description of the articulatory process. After the contours of lips were detected through effective and appropriate ways to process image data, they were approximated by polynomial function curves and transformed to the parametric forms. From preliminary experiments, the following fourth-polynomial function was adapted [5][6];

$$f(x) = a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0 \qquad (1)$$

Each of these coefficients was estimated by minimizing squared error of differences between detected curves of lip contours and calculated ones from above equation. Figure.1. shows examples of approximate lip contours with raw images. The number of each frame indicate a sequence of lip movement. Then two geometrical parameters, area of lip opening and width of lip edges, were obtained as a time pattern from approximated contours with a fourth-degree polynomial. Figure.2. shows examples of time-varying patterns of various geometric dimensions relevant to the lip articulation on the subject HF during the speech of [epape].

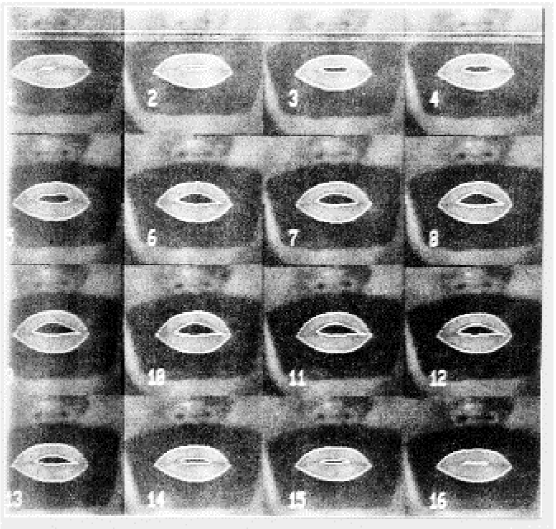
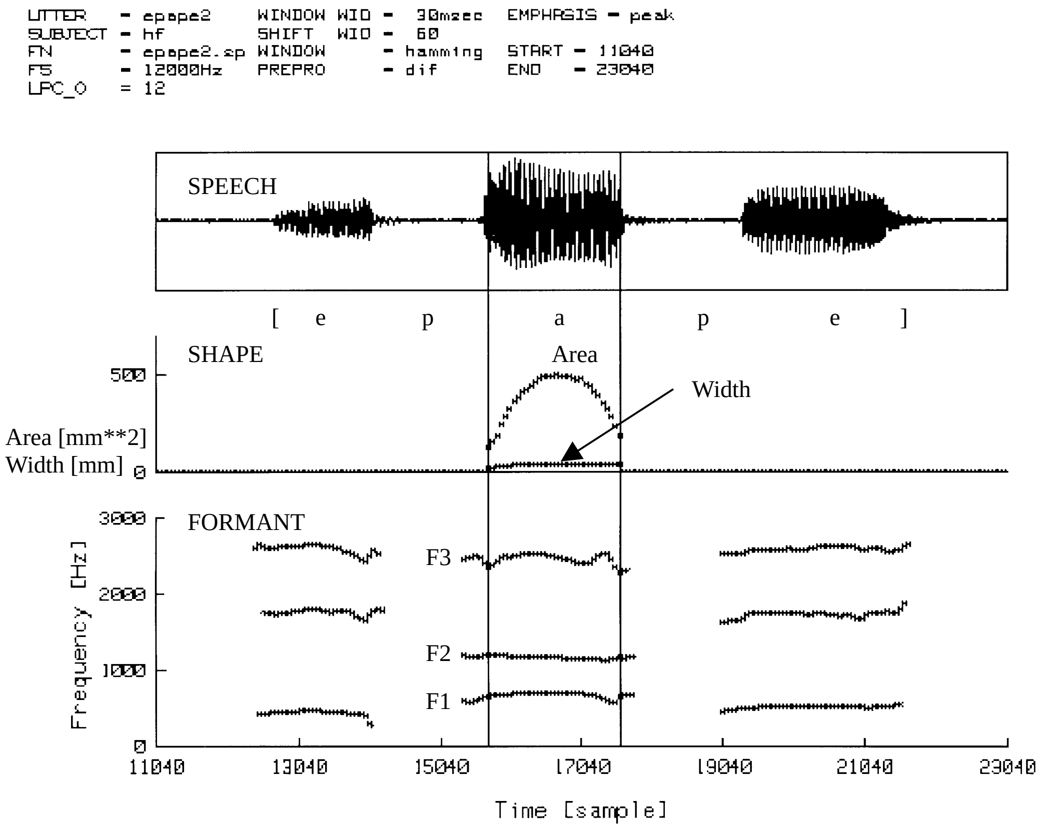


Figure.1. Examples of approximate lip contours with raw images. The number of each frame indicate a sequence of lip movement.



```
UTTER    - epape2     WINDOW WID -  30msec   EMPHASIS - peak
SUBJECT  - hf         SHIFT  WID -  60
FN       - epape2.sp  WINDOW     - hamming   START - 11040
FS       - 12000Hz    PREPRO     - dif       END   - 23040
LPC_O    = 12
```

Figure.2. Examples of time-varying patterns of various geometric dimensions relevant to the lip articulation on the subject HF during the speech of [epape].

Upper: Acoustic signals

Center: Time patterns of lip articulation

Lower: Acoustical patterns of three formants patterns (F1, F2 and F3)

Center part of the figure shows movement patterns of lip articulation approximated by fourth-polynomial function (Area: area of mouth opening and Width: horizontal extent of lip separation). Upper part of the figure shows acoustic signals. It has been found that immediately following the plosion of [p], the rate of increase in the mouth opening was relatively high, and the amount of its opening area (Area) primarily depended on its vertical separation as compared to horizontal extent (Width) Recognition tests were examined by using of acoustical parameters of three formant patterns (F1,F2 and F3) as well as the articulatory parameters. Formant patterns were detected by using linear predictive coding (LPC) method at the time when the lip image was sampled.

4. EXPERIMENTS

An architecture of artificial neural networks is motivated by the computational style found in biological nervous system. Neural network used for recognition test of identifying

(1) morae (CV),

(2) vowel (V),

(3) consonant (C)

from the time patterns of articulatory movements and acoustical sounds in -CVC-. Figure.3. shows experiments patterns of recognition test. Training words were selected two sets of speech materials that were selected from among repeated three sets of each speech materials. And recognition words were selected one sets of speech materials not including training sets.

For each recognition tests, training data were selected .

(a) shape pattern only,

(b) formant pattern only and

(c) both shape pattern and formant pattern

A network was trained enough with each training sets. After the network was trained, it then was tested on the recognition set of the same speaker.

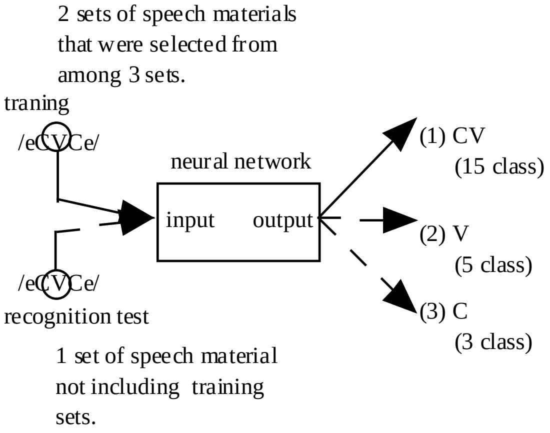


Figure.3. Experiments patterns of recognition test. Training words were selected two sets of speech materials that were selected from among repeated three sets of each speech materials. And recognition words were selected one sets of speech materials not including training sets.

5. RESULTS

At recognition test of the morae, recognition scores using three types of parameters (a), (b) and (c) above mentioned are almost equal and their values are about 65 %.

In the case of the vowel, by appending the articulatory parameter to the acoustic features (formant frequency), the speech recognition score increased form about 80 % using acoustic features alone to about 90 %.

In recognition tests of the morae(CV) and the vowel(V), the recognition performance increases in the case of jointly using both shape pattern and formant pattern.

At recognition test of the consonant, recognition value using shape pattern only is about 70 % to 90%. It using acoustical pattern and both two pattern above mentioned are almost equal and their values are 50 %. These result shows that consonant ([p], [b] and [m]) is more closely related with geometrical parameter than acoustical parameter on the middle vowel of utterance.

Table.1. shows confusion matrices of recognition test for consonants. The table shows the degree of confusion with other classes. Each number in the table represents the frequency of the recognized output class when using shape pattern or format pattern only. The y axis represents the true class while the x axis show the classifier output.

When using shape pattern only. In the case of recognition test of [p] at Subject KO and YN, all recognition tests were correct. On the other hand, at Subject HF and KM, test words were apt to be recognized class [b].

When using formant pattern only. At Subject KM, KO and YN, output classes were scattered in disorder. On the other hand, at Subject HF, test words were inclined to be recognized class [p].

6. CONCLUSION

On the basis of image data taken by the high-speed video recorder, we have studied the relationship between lip shapes and acoustical characteristics during speech.

In order to provide the parametric representation on the lip shapes, contours of the lips were estimated by fourth-polynomial equation. In recognition of the vowel. vowel was closely related with both shape pattern and formant pattern as well. In recognition of the consonant. Consonant was more closely related with shape pattern than formant pattern. These results show effect of consonant (/p/, /b/, and /m/) on the middle vowel of utterance.

7. REFERENCES

1. Y.Sonoda, "Articulatory characteristics of tongue and jaw point movements in connecting sound of Japanese", Trans. IEICE Jpn., Vol.62, pp.555-562, Sept.(1979).

2. Y.Sonoda, "Effect of speaking rate on articulatory dynamics and motor event", J.Phonetics, Vol.15, pp.145-156(1987)

3. O.Fujimura, "Bilabial stop and nasal consonants: a motion picture study and its acoustical implications", J.Speech.Hear.Res., Vol.4, pp.233-247, Sept.(1961)

4. Y.Fukuda and S.Hiki, "Characteristics of the mouth shape in the production of Japanese-stroboscopic observation", J.Acoust.Soc.Jpn.,Vol.(E)3, pp.75-91, Feb.(1982)

5. Y.Sonoda et al., "Articulatory characteristics of lip shape during the production of Japanese", ICSPL 90,11.6.1, (1990)

6. K.Mori and Y.Sonoda, "Relationship between lip shapes and acoustical characteristics during speech", Acoustcal Society of America and Acoustical Society of Japan, Third Joint meeting , 2pSC22, pp.879-882, Dec.(1996).

Table.1. Confusion matrices of recognition test for consonants. The table shows the degree of confusion with other classes. Each number in the table represent the frequency of the recognized output class. The y axis represents the true class while the x axis show the classifier output.

(a) Subject HF

[C] Subject HF		Output class		
shape formant		b	m	p
Input class	b	8 4	4 0	2 11
	m	3 0	11 2	1 13
	p	4 0	0 1	11 14

(b) Subject KM

[C] Subject KM		Output class		
shape formant		b	m	p
Input class	b	10 7	0 1	5 7
	m	1 5	12 3	2 7
	p	4 5	0 1	11 9

(c) Subject KO

[C] Subject KO		Output class		
shape formant		b	m	p
Input class	b	11 8	2 2	2 5
	m	2 2	11 8	0 5
	p	0 4	0 3	15 8

(d) Subject YN

[C] Subject YN		Output class		
shape formant		b	m	p
Input class	b	14 6	1 5	0 4
	m	0 5	13 7	2 3
	p	0 5	0 2	15 8