SPEECH TRAINING TOOLS BASED ON VOWEL SWITCH/VOLUME CONTROL AND ITS VISUALIZATION

Yuichi UEDA and Tadashi SAKATA

Graduate School of Science and Technology, Kumamoto University, Kumamoto, Japan
E-mail: { ueda, tadashi }@cs.kumamoto-u.ac.jp

ABSTRACT

We have developed a real-time software tool to extract a speech feature vector whose time sequences consist of three groups of vector components; the phonetic/acoustic features such as formant frequencies, the phonemic features as outputs on neural networks, and some distances of Japanese phonemes. In those features, since the phoneme distances for Japanese five vowels are applicable to express vowel articulation, we have designed a switch, a volume control and a color representation which are operated by pronouncing vowel sounds. As examples of those vowel interface, we have developed some speech training tools to display a image character or a rolling color ball and to control a cursor's movement for aurally- or vocally-handicapped children. In this paper, we introduce the functions and the principle of those systems.

Keywords: real-time processing, speech visualization, speech features, vowel switch

1. INTRODUCTION

We have developed various speech application systems so far: a speech visualizing system[1], a formant decomposing type of hearing aid[2], a speech coding method for the cochlear implant[3], a word speech recognizer using the compound speech parameters[4] and so on. Real time operations of those systems are indispensable for practical use, in particular. Those systems have the common speech processing where the formant frequencies[5], the fundamental frequency, the speech spectra and so on are estimated from real speech signal. In order to realize such practical systems, we need to realize a real-time software engine to extract those speech parameters. Therefore, we have integrated some fundamental processing techniques including our original ones and designed a software engine which was possible to build in speech application tools on a general-purpose PC.

In this paper, we describe a framework of the developed real-time engine and then propose new software tools for learning vowel articulation using the software engine. Since those tools aim at promoting the pronouncing abilities of the hearing impaired children or the speech disordered children, we have designed two kinds of software tool which had a switching mode with an ON/OFF state and a volume control mode with continuous state. Those states correspond to characteristics of vowel articulation, respectively. Section 2 gives the developed software engine's structure and functions; in section 3, we introduce the proposed software tools' functions.

2. REAL-TIME SOFTWARE ENGINE FOR SPEECH ANALYSIS

To estimate speech feature vectors in real-time, we have designed a new real-time software engine where many speech processing techniques for our application systems were integrated. The system's structure and functions are as follows.

2.1 Structure of Functions of the System

Figure.1 shows a functional diagram of the developed software engine. Speech signal is sampled at 12kHz of sampling frequency and stored into buffer memory in duration of 20ms. The buffered signal is processed every frame period of 10ms and three types of parameter groups are estimated shown in Fig.1.

2.1.1 Speech Parameters and Signal Processing

As shown in Fig.1, the parameter estimations consists of three hierarchical stages where the signal processing is as follows:

1) Phonetic/Acoustic features

The following parameters are estimated as phonetic/ acoustic features of frame signal.

a. Effective value (RMS)
b. Loudness Level (LOUD)
c. Zero crossing rate (ZCR)
d. LPC-Mel-cepstrum coefficients(MLCC); $C_0 \sim c_{12}$
e. Formant frequencies; $F_1 \sim F_4$
f. Fundamental frequency; F_0 (voice pitch)

Especially, the formant frequencies are estimated by the IFC-method (Inverse Filter Control method[5]) which was proposed and developed by A.Watanabe. The IFC based formant tracker can estimate from F_1 to F_5(F_6) for male (female) voice at high precision. To implement in real-time, the number of formant's order are fixed to fourth in the proposed engine. Moreover, the fundamental frequency, F_0, is estimated by picking up peak position of the autocorrelation function of the inverse filtered speech signal, that is, a differential glottal source signal.

2) Phonemic features

Phonemic features similar to the distinctive features of phoneme in phonetics are obtained as output of two neural networks (NNs); an integrated NN regarding source, articulation manner, and articulation place and a fricative NN for /s/, /z/, and /h/. Therefore, values of phonetic features have a real number from 0.0 to 1.0.

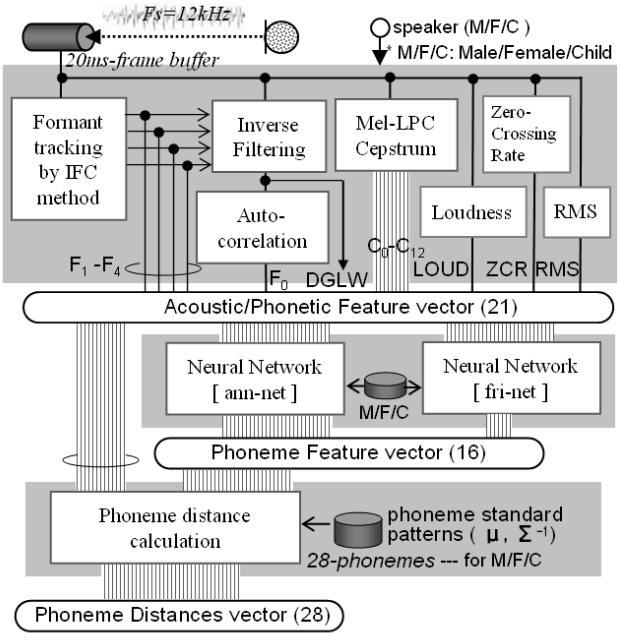


Figure.1 Functional diagram of a real-time software engine to estimate speech feature vectors

3) Phoneme distances[4]

Phoneme distances are the Bayes distances which are calculated for each Japanese phonemes of 28 standard ones in five parameter dimensions. Those distances were used to recognize the Japanese word speech based on the DP matching algorithm.

The number of parameters from 1) to 3) becomes to be 65 elements and we call a set of those parameters the speech feature vector.

2.1.2 Real-time Operation

Previously, we have designed the IFC based formant tracker on a DSP board and realized a real-time operation[6]. However, to utilize the formant tracking function in wider practical application, it needed to realize the real-time processing on a general purpose of PC. All parts of the developed software engine are written in C-language and those functions are capable to be embedded into various speech applications. The whole processing time of functional blocks in Fig.1 is about 4.9ms on a general purpose PC which operates single CPU chip of 2.6GHz, and equals to about 50% of a frame period, 10ms. Since the designed system is based on frame by frame processing, frame parameters are extracted at every 10ms of a frame period. As results, we confirmed that our software engine could operate in real-time.

2.2 Speech Feature Vectors

The elements of speech feature vector are listed in Table.1 where those elements are divided into three types of categories. We have selected some parameters(elements) in those elements and built those into auditory supplements such as a hearing aid and a cochlear implant processing, or into speech visualizations by composing speech features including formant frequencies, pitch, phoneme features and so on.

Table.1 Components of the speech feature vector defined in 65 dimension space.

1	F_0	Fundamental Freq.
2	F_1	First Formant Freq.
3	F_2	Second Formant Freq.
4	F_3	Third Formant Freq.
5	F_4	Fourth Formant Freq.
6	RMS	Effective value
7	LOUD	Loudness
8	ZCR	Zero crossing rate
9	C_0	Mel-LPC-Cepstrum coefficients
10	C_1	
11	C_2	
12	C_3	
13	C_4	
14	C_5	
15	C_6	
16	C_7	
17	C_8	
18	C_9	
19	C_{10}	
20	C_{11}	
21	C_{12}	

22	VOW	Vowel
23	NAS	Nasality
24	BUZ	Buzz
25	BUR	Plosive(Burst)
26	FRI	Fricative
27	FLP	Flap
28	VOI	Voiced sound
29	UNV	Unvoiced sound
30	SIL	Silence
31	LAB	Labial
32	ALV	Alvolar
33	VEL	Velum
34	N	Uvula nasal
35	/s/	Unvoiced fricative
36	/h/	Glottal fricative
37	/z/	Voiced fricative

38	a / A	Vowel/Long vowel
39	o / O	
40	u / U	
41	i / I	
42	e / E	
43	j	[j]
44	y	[Voiced]j :
45	Y	[Unvoiced]j
46	w	wa, wo
47	m	[m]
48	n	[n]
49	N	[N]
50	b	[b]
51	d	[d]
52	g	[g]
53	r	[r]
54	z	[z]
55	h	[h]
56	f	[Φ]
57	s	[s]
58	c	[t͡ʃ]
59	p	[p]
60	t	[t]
61	k	[k]
62	T	[ts]
63	G	[ŋ]
64	Q	silence befor plosive
65	q	silence

2.3 Examples of Speech Feature Vectors

Figure.2 shows temporal patterns of speech feature vectors which were estimated in a sentence uttered by a male, /tikarakurabeo simasita/ in Japanese. By observing the formant trajectories, we can confirm stability and reliability of formant tracking by the IFC method. Moreover, characteristics of each speech segments appear on both the pitch trajectory and the variations of ZCR or RMS. Using the upper group of parameters in Fig.2 (F1~F4, ZCR, RMS and so on), the phoneme features in Fig.2(a) are estimated by two NNs. The vowel distances and consonant distances are plotted in Fig.2(b) and Fig.2(c), respectively. By comparing the vowel distances with the phoneme sequence, we can see that the vowel phoneme with the minimum distance corresponds to one in phoneme sequences. Moreover, the phoneme features such as NAS(nasality) , BUR(burst), and FRI(fricative) correspond to consonant feature respectively.

In this study, we propose the speech training tools regarding vowel pronouncing by use of some features of the speech feature vector, selectively .

3. SPEECH TARINING TOOLS USING THE REAL-TIME ENGINE

In this section, we propose two kinds of speech training tools regarding vowel pronunciation. These tools consist of a voice switch and a voice control respectively, to be operated by features of vowel articulation.

3.1 Training tool based on a vowel switch

This tool was designed by using features regarding vowel in the speech feature vectors and had five kinds of switch modes corresponding recognition results of Japanese vowels. In designing this tool, we assumed that the severe hearing impaired children or the speech disordered children would make use of this tool to learn articulation manner of short vowels or sustained vowels. Therefore, the gaming characteristics were incorporated in feedback actions of recognition results.

3.1.1 Switching algorithm

A flow graph of the vowel switching algorithm is illustrated in Figure.3, where RMS(effective value), VOW(vowel feature), and vowel phoneme distances (*dv_n*; *n=1,2,3,4,5*) are used. We used the following parameters to change the stages of switch.

1) Pronouncing power (*Pow_th*: weak/normal/strong)
2) Pronouncing duration (*Dur_th*:0.1,0.5, 1.0, 2.0, 5.0 sec)
3) Duration of ON-state (*Dsp_th*)

where *Pow_th* and *Dur_th* are pre-set according to user's ability in pronouncing. *Dsp_th* gives a display time after turning ON of switch. The condition of change in switch's state (OFF→ON) is a case that the input speech segments are determined as the same vowel during a duration of Dur_th.

We proposed the following two display modes as responses after turning ON. First mode is that the vowel recognition result appears as one of the prepared five characters. We call this tool as "*Vowel Pong!*". Second one is that a cursor on a screen walks according to phonemes of vowel sounds and called as "*Vowel Walking*". Each mode is selected from a menu of the software tool. Figure.4 shows examples of display windows of two modes, (a) "*Vowel Pong*" and (b)"*Vowel Walking*". The main functions of each mode are outlined below.

3.1.2 Vowel Pong

Fig.4(a) shows an example of system operations, where a character corresponding to vowel/o/ has appeared after pronouncing a sustained vowel /o/ over the thresholds of both a pronouncing duration (>*Dur_th*) and a pronouncing power(>*Pow_th*). The character disappears after displaying during constant time (=*Dsp_th*) and then becomes to be stand-by mode. Purpose of speech training using this tool is to pronounce optional vowel at constant power over the preset duration.

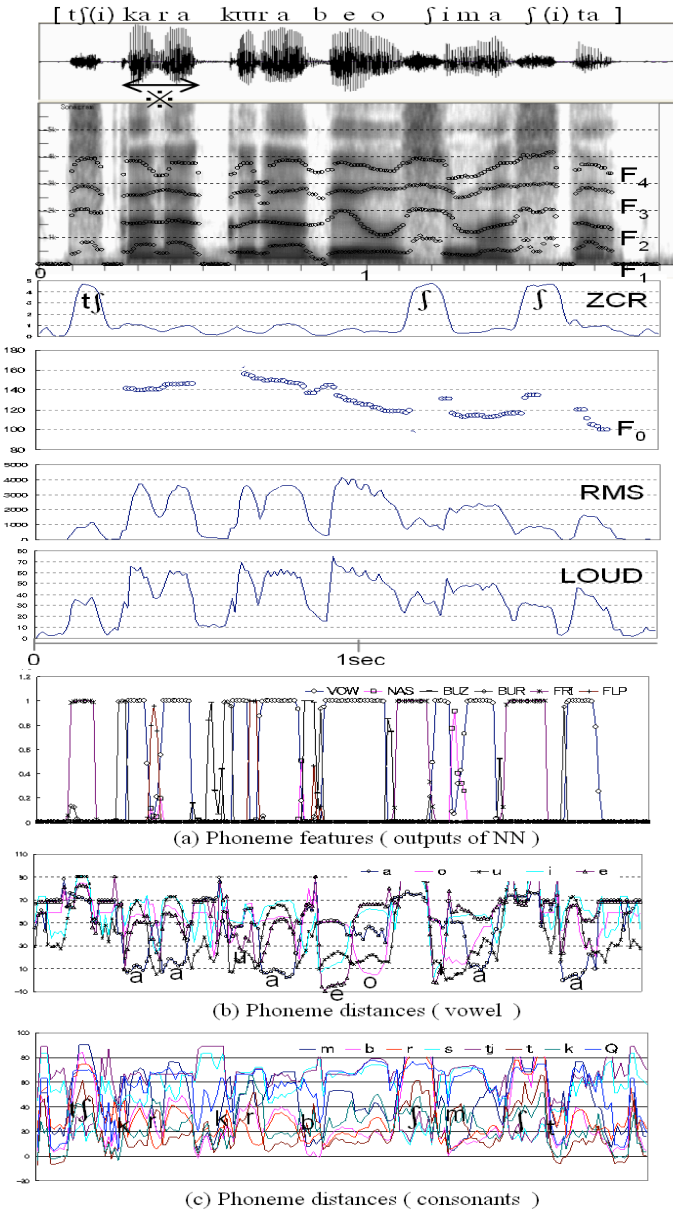


Figure.2 Examples of the extracted speech feature vectors of a sentence /tikarakurabewo simasita/ uttered by a

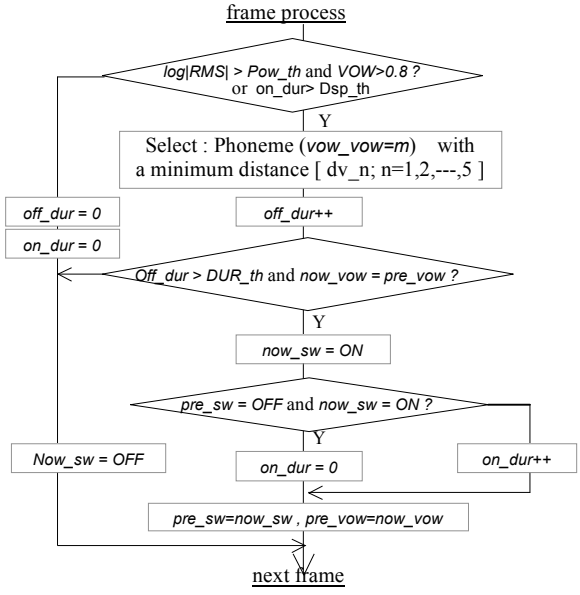
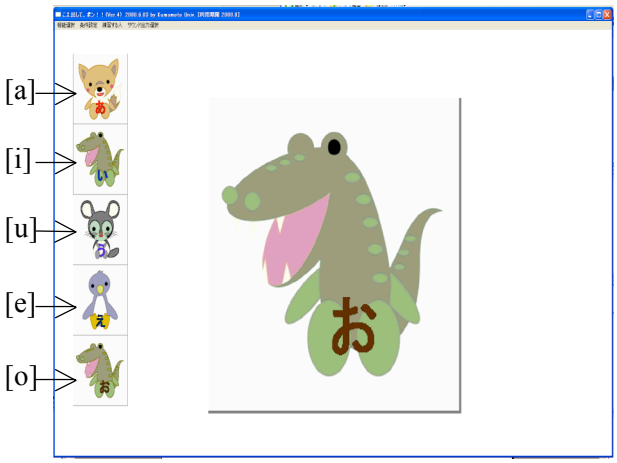
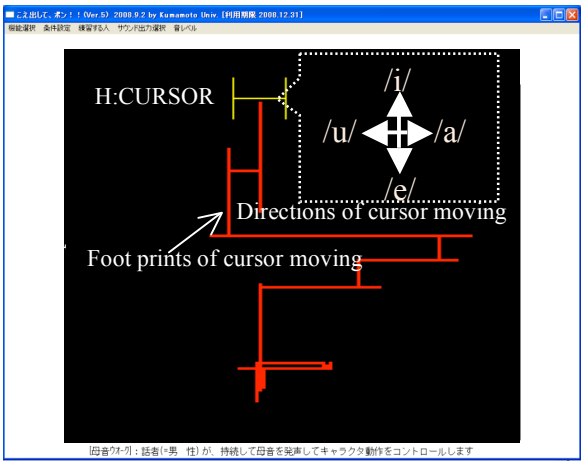


Figure.3 Flow graph to illustrate the vowel switching mode



(a) "*Vowel Pong*" in a case of recognition result of vowel /o/



(b) "*Vowel Walking*" in a case of pronouncing some sustained vowels

Figure.4 Two modes of the proposed speech training tool regarding vowel pronouncing based on the vowel switching.

3.1.3 Vowel Walking

In this mode, we assume that user is demanded to pronounce a vowel sound sustainably, with a power over a threshold (*Pow_th*) . Fig.4(b) indicates an example of real-time display in pronouncing vowel /i/. The cursor marked as a large 'H' has four directions corresponding to four vowel, /a/, /i/, /u/ and /e/, respectively. Therefore, in pronouncing the same vowel, the cursor continues to move("*walks*") in one direction. On the other hand, the cursor stops in a case of recognizing as non-vowel. We call this mode "*Vowel Walking*". As an application of this mode, we can design a gaming play; that is to say a user moves the cursor in the direction of goal along the preset path on screen, by repeating to pronounce some sustained vowels.

3.2 Training tool based on a vowel control

The second tool uses correspondence of the auditory vowel phoneme and the color vision and expresses in vowel sounds to color patterns.

3.2.1 Color Conversion

It is known that the auditory perception of vowel's phoneme varies continuously according to articulation movement. From this point of view, we have proposed new visual representation of vowel to reflect articulation movements in vowel pronouncing. In this method, the vowel sounds are converted into the colored patterns whose primary color signals (Red, Green and Blue) are calculated by using three lower formant frequencies (F_1, F_2 and F_3) known as vowel features. Figure.5 illustrates the color converting process from the formant space to color one. In Fig.5, three equations of R, G and B, are defined so that the vowel segments with the same formant ratios (F_1/F_2, F_2/F_3 and F_3/F_1) can be displayed as the same color. By such a processing, it is expected to normalize vowel phonemes aurally. On the other hand, since slight changes in articulation are reflected on difference in color directly, user can observe one's own pronouncing manner visually.

3.2.2 Auditory Visual Distribution of Japanese Vowel

Figure.6 shows F_1-F_2 diagram of the vowel formant distribution. The corresponding color diagram is painted on Fig.6, which converts the components of formant frequencies into the corresponding color image by using the converting equations; Formant-to-RGB in Fig.5.

3.2.3 Rolling Color Ball

To realize the color representation of vowel as a pronouncing tool, we have considered the following functions.

1) To relate auditory perception of vowel's phoneme to an angular position of the color circle in Fig.6(b).
2) To express vowel's phoneme as the corrsponding colored ball.
3) To relate a radius of the rolling ball to speech power, log(RMS).

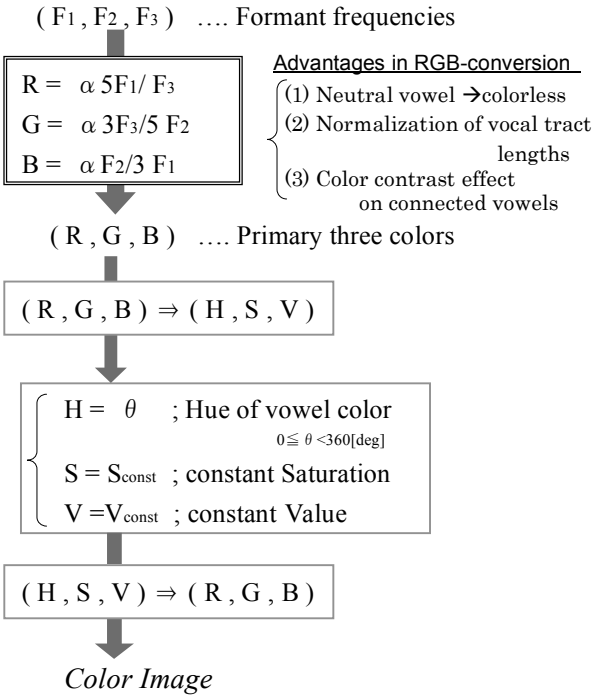
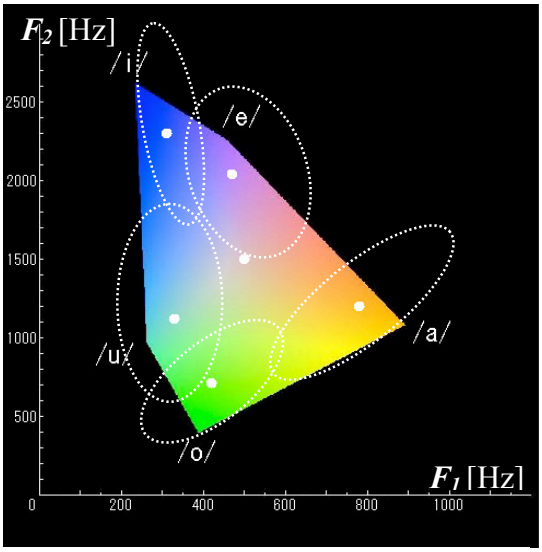
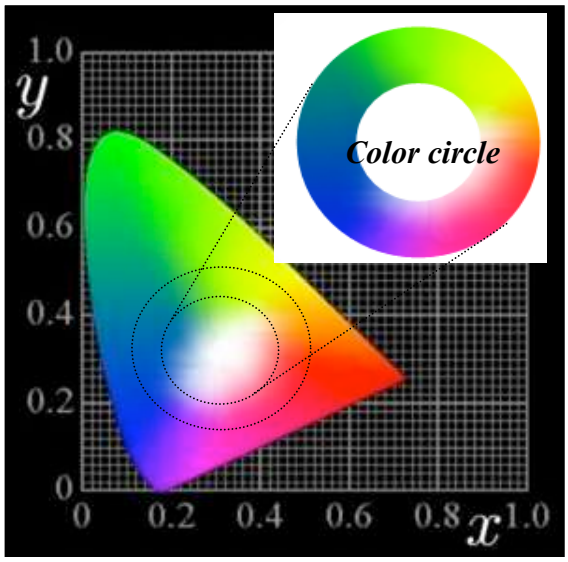


Figure.5 Flow of color conversions and color processing Formant-RGB-HSV



(a) Distribution of typical Japanese five vowels on F_1-F_2 plane and the color mapped vowel area



(b) CIE x-y chromaticity diagram

Figure.6 Illustration of association between the vowel formant distribution and the chromaticity diagram as results of color conversion by equations (Formants-to-RGB) in Fig.5.

Figure.7 shows an example of operating states in pronouncing a sustained vowel, /i/. If user's articulation varies with time, the rolling ball alters its own color and then rotates up to the hue position corresponding to its phoneme. We call this tool "Rolling Color Ball". The purpose of this tool is to learn the association between vowel's phoneme and color image and then learn consciously the differences in articulation manner.

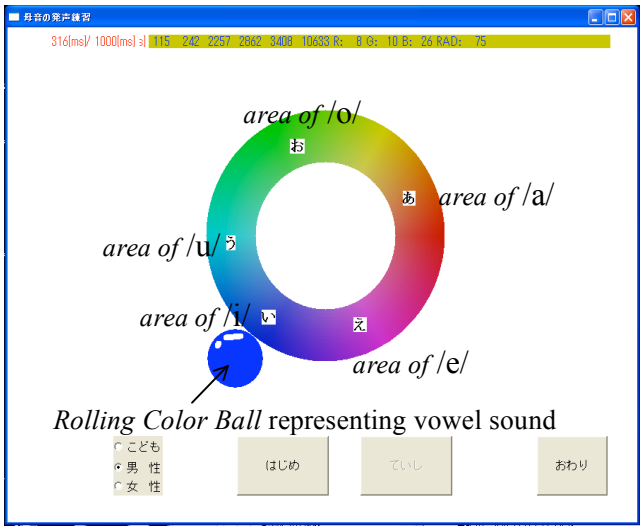


Figure.7 "*Rolling Color Ball*" which is a tool for learning vowel articulation.

4. CONCLUSION

In this paper, we described a real-time software engine to estimate speech feature vector which could be built into various speech application system. Those functions are realized by various real-time processing elements including our original ones. And, as applications of such feature vectors, we proposed a vowel switch based on ON-OFF decision and a volume control based on color representation. Since those tools aim at utilizing by the disordered children, those visual representations are designed by very simple manner. Additionally, by adding gaming functions and so on, those tools seem to improve the efficiency.

Acknowledgement

This work is supported in part by the foundation for Fusion of Science and Technology, Japan. And we thank Dr. Akira Watanabe for advice and offers in designing the real-time engine to estimate the speech feature vectors.

REFERENCES

[1] A.Watanabe, S.Tomishige and M.Nakatake, "Speech Visualization by Integrating Features for the Hearing Impaired, " IEEE Transactions on Speech and Audio Processing, Vol.8, No.4, pp.454-466,2000

[2] Y.Ueda, S.Hario, T.Sakata," Formant based speech enhancement for listening speech sound in noisy place," Proc. of ICSV-2008, SS0298, pp.1-8, 2008

[3] M.Sato,T.Sakata,A.Watanabe, Y.Ueda," Formant Peak Stimulating Method based on Phantom Sensation for Cochlear Implant System," Proc. of WESPAC IX, hu-2-5-370, 2006.

[4] N.Ikeda, T.Sakata, Y.Ueda, A.Watanabe," Word Recognition Method using Multiple Templates for Different Utterance Speeds," Proc. of ICA-2004, Tu.P2.16, 2004

[5] A.Watanabe, "Formant Estimation Method Using Inverse Filter," IEEE Trans. on Speech and Audio Proc., vol.9, pp.314-326, 2001

[6] Y.Ueda, T.Hamakawa, T.Sakata, A.Watanabe, "A real-time formant tracker based on the inverse filter control method," Acoustical Science and Technology, Vol.28 , No.4 , pp.271-274, 2007