We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900
Open access books available

185,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



20

Vowel Judgment for Facial Expression Recognition of a Speaker

Yasunari Yoshitomi, Taro Asada and Masayoshi Tabuse
Kyoto Prefectural University
Japan

1. Introduction

To better integrate robots into society, a robot should be able to interact in a friendly manner with humans. The aim of our research is to contribute to the development of a robot that can perceive human feelings and mental states. A robot that could do so could, for example, better take care of an elderly person, support a handicapped person in his or her live, encourage a person who looks sad, or advise an individual to stop working and take a rest when he or she looks tired.

Our study concerns the first stage of the development of a robot that has the ability to detect visually human feelings or mental states. Although a mechanism for recognizing facial expressions has received considerable attention in the field of computer vision research (Harashima et al., 1989; Kobayashi & Hara, 1994; Mase, 1990, 1991; Matsuno et al., 1994; Yuille et al., 1989), currently it still falls far short of human capability—especially from the viewpoint of robustness under widely varying lighting conditions. One of the reasons for this is that the nuances of shade, reflection, and localized darkness—as the result of the inevitable changes in gray levels—influence the accuracy of the discernment of facial expressions.

To develop a robust method of facial expression recognition applicable under widely varied lighting conditions, we do not use a visible ray (VR) image, instead we use an image produced by infrared rays (IR), which show temperature distributions of the face (Fujimura et al., 2011; Ikezoe et al., 2004; Koda et al., 2009; Nakano et al., 2009; Sugimoto et al., 2000; Yoshitomi et al., 1996, 1997a, 1997b, 2000, 2011a, 2011b; Yoshitomi, 2010). Although a human cannot detect IR, a robot can process the information contained in the thermal images created by IR. Therefore, as a new mode of robot vision, thermal image processing is a practical method that is viable under natural conditions.

The timing for recognizing facial expressions also is important for a robot because processing can be time consuming. We adopted an utterance as the key to expressing human feelings or mental states because humans tend to say something to express their feelings (Fujimura et al., 2011; Ikezoe et al., 2004; Koda et al., 2009; Nakano et al., 2009; Yoshitomi et al., 2000; Yoshitomi, 2010). In conversation, we utter many phonemes. We have selected vowel utterances for use as timings to recognize facial expressions because the number of vowels is very limited, and the waveforms of vowels tend to have a bigger amplitude and a longer utterance period than consonants. Accordingly, the timing range of each vowel can be relatively easily decided by a speech recognition system.

In this paper, we briefly look at a proposed method (Koda et al., 2009) for recognizing the facial expressions of a speaker. For this facial expression recognition, we select three image timings: (i) just before speaking, and speaking (ii) the first vowel and (iii) the last vowel in an utterance. To apply the proposed method (Koda et al., 2009), three subjects spoke 25 Japanese given names that provide all combinations of first and last vowels. These utterances were used to prepare the training data and then the test data.

2. Speech recognition system

We use a speech recognition system called Julius (Kawahara et al., 2010b) to save as a wav file the timing positions of the start of speech, and the first and last vowels (Koda et al., 2009; Yoshitomi, 2010; Fujimura et al., 2011; Yoshitomi et al., 2011a).
Julius has been widely used by researchers and engineers, especially in Japan. Julius can achieve typically real-time dictation of a 20,000-60,000 word vocabulary with an accuracy of about 90% on a PC (Kawahara et al., 2010a). In the references (Kawahara et al., 2010a, 2010b; Lee & Kawahara, 2009), Julius is explained in detail. Based on these references, we briefly explain the characteristics of Julius.
Julius has been developed as a research software for large-vocabulary continuous speech recognition (LVCSR) since 1997, and is distributed under an open license together with source codes. Julius is an open-source software for Japanese LVCSR. Word N-gram, context-dependent Hidden Markov Model (HMM), tree lexicon, N-gram factoring, crossword context dependency handling, enveloped beam search, Gaussian pruning, and Gaussian selection are used as the main techniques in Julius. According to the references (Kawahara et al., 2010a, 2010b; Lee & Kawahara, 2009), the main characteristics of Julius are:

- Real-time, high-speed, recognition based on a two-pass strategy.
- Live audio input recognition via microphone/input socket.
- Less than 32 M Bytes required for work area in memory.
- Supports language model (LM) of N-gram, grammar, and isolated words.
- Any LM in standard ARPA format and acoustic models in HTK ascii hmmdefs format can be used.
- Set various search parameters. Alternate decoding algorithm can be chosen.
- Triphone HMM/tied-mixture HMM/phonetic tied-mixture HMM with any number of states, mixtures and models are supported in HTK.
- Most mel-frequency cepstral coefficients and their variants are supported in HTK.

Figure 1 shows examples of outputs by Julius. The figure shows the timing position at the start of speech, and each trimming range of the first and last vowels for the utterance of "Shinnya" pronounced by Subject A while expressing the emotions "angry," "happy," "neutral," "sad," and "surprised."

	Angry	Happy	Neutral	Sad	Surprised
Silent	0-35 frame	0-12	0-64	0-43	0-48
Consonant	36-56	13-38	65-75	44-66	49-67
Vowel(/i/)	57-73	39-53	76-83	67-74	68-76
Consonant	74-91	54-69	84-110	75-99	77-97
Vowel(/a/)	92-103	70-84	111-129	100-117	98-117
Silent	104-191	85-191	130-191	118-191	118-191

Fig. 1. Examples of outputs by Julius

3. Method for recognizing facial expressions

Figure 2 is a flowchart of the proposed method. We have two modules in our system. The first is for speech recognition and dynamic image analysis, and the second is for learning and recognition. In the module for learning and recognition, we embedded the module for for front-view face judgment, which is not described in this paper because it is not directly related to speech recognition. The procedure used, except for the pre-processing module for front-view face judgment (Fujimura et al., 2011), is explained below.

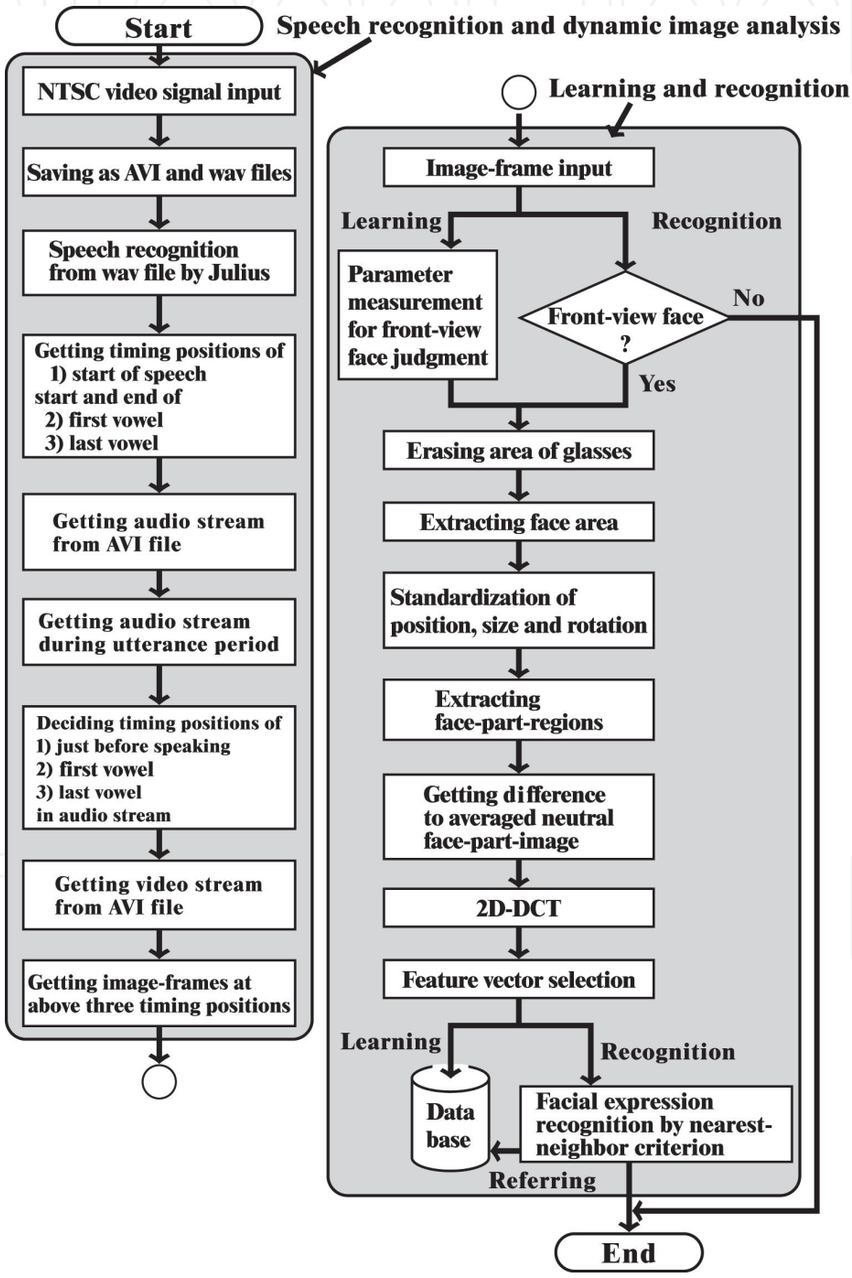


Fig. 2. Flowchart of proposed method (Fujimura et al., 2011)

3.1 Speech recognition and dynamic image analysis

Figure 3 shows the waveform of the Japanese given name "Taro;" the timing position of the start of speech, and the timing ranges of the first vowel (/a/) and the last vowel (/o/) were decided by Julius. By using these three timing positions obtained from a wav file, three thermal image frames are extracted from an AVI file. For the timing position just before speaking, we use 84 ms, as determined in a previously reported study (Nakano et al., 2009). As the timing position of the first vowel, we use the position where the absolute value of the amplitude of the waveform is the maximum while speaking the vowel. For the timing position of the last vowel, we apply the same procedure used for the first vowel.

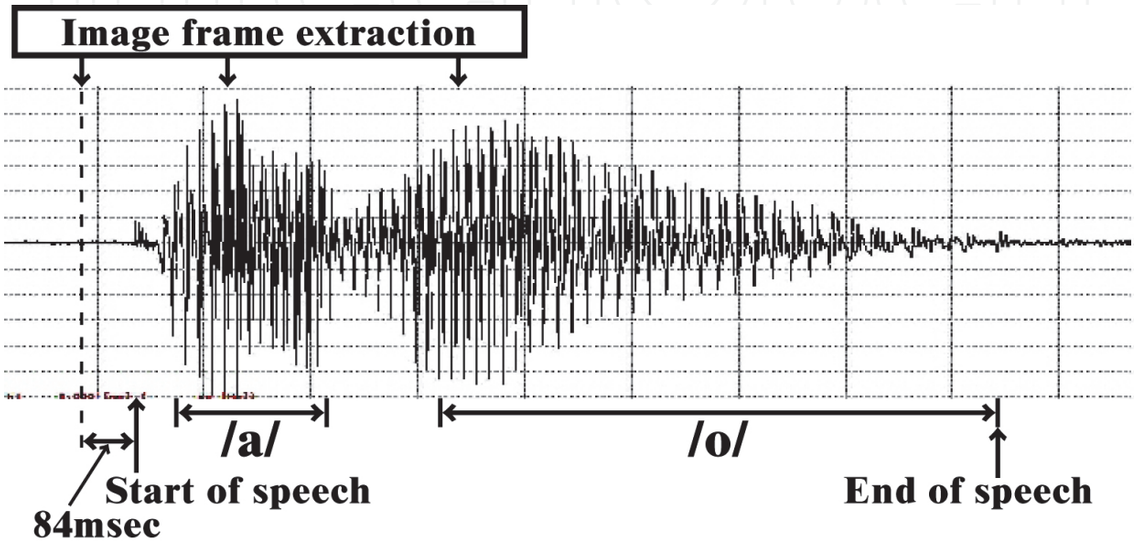


Fig. 3. Speech waveform of "Taro" and timing positions for image frame extraction (Koda et al., 2009)

3.2 Learning and recognition

For the static thermal images obtained from the extracted image frames, the process of erasing the area of the glasses, extracting the facial area, and standardizing the position, size, and rotation of the face are performed according to the method described in our previously reported study (Nakano et al., 2009). Figure 4 shows the blocks for extracting the facial areas in a thermal image of 720 × 480 pixels. In the next step, we generate difference images between the averaged neutral face image and the target face image in the extracted facial areas to perform a two-dimensional discrete cosine transform (2D-DCT). The feature vector is generated from the 2D-DCT coefficients according to a heuristic rule (Ikezoe et al., 2004; Nakano et al., 2009).

The Julius speech recognition system used in our study sometimes makes a mistake in recognizing the first and/or last vowel(s). For example, /a/ for the last vowel is sometimes misrecognized as /o/. We correct this misrecognition for the training data, however, corrections cannot be made for the test data. For example, in the experiment described later, when Julius correctly judges the first vowel of the utterance of "Ayaka," but misjudges the last vowel as /o/, the training data in speaking "Taro" are used for recognition instead of those for speaking "Ayaka." The facial expression is recognized by the nearest-neighbor criterion in the feature vector space by using the training data just before speaking and while uttering the phonemes of the first and last vowels.

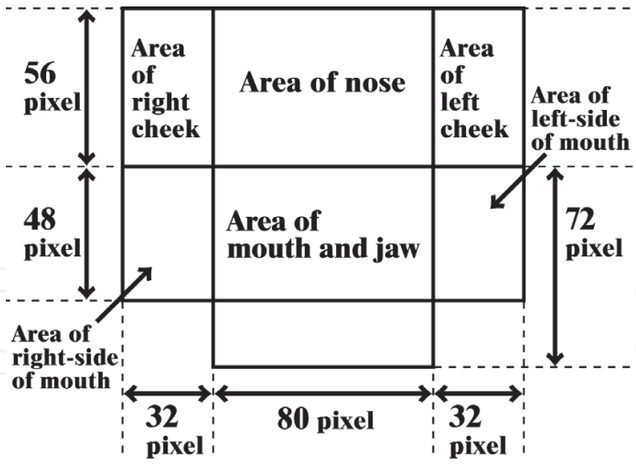


Fig. 4. Blocks for extracting facial areas from a thermal image

4. Experiments

4.1 Conditions

The thermal image produced by a thermal video system (Nippon Avionics TVS-700) and the sound captured by an electret condenser microphone (Sony ECM-23F5), amplified by a mixer (Audio-Technica AT-PMX5P), were transformed into a digital signal by an analogue/digital converter (Thomson Canopus ADVC-300) and input into a computer (DELL Optiplex GX620, CPU: Pentium IV 3.4 GHz, main memory: 2.0 GB, OS: Windows XP (Microsoft)) with an IEEE1394 interface board (I·O Data Device 1394-PCI3/DV6). We used Visual C++ 6.0 (Microsoft) as the programming language. To generate a thermal image, we set the conditions so the thermal image had 256 gray levels for the detected temperature range. This range was decided independently for each subject in order to best extract the facial area. We saved the visual and audio information in the computer as a Type 2 DV-AVI file, in which a frame had a spatial resolution of 720 × 480 pixels and an 8-bit gray level, and the sound was saved in a PCM format as stereo, 48 kHz, 16-bit file. The version 4.0 of Julius was used in the current study.

All subjects exhibited, in alphabetical order, each of the intentional facial expressions of "angry," "happy," "neutral," "sad," and "surprised," while speaking the semantically neutral utterance of each of the Japanese given names listed in Table 1. There were three subjects. Subject A was a male without glasses. Subject B was a male with glasses. Subject C was a female without glasses. Figures 5, 6, and 7 show images of Subjects A, B, and C, respectively.

		First vowel				
		a	i	u	e	o
Last vowel	a	ayaka	shinnya	tsubasa	keita	tomoya
	i	kazuki	hikari	yuki	megumi	koji
	u	takeru	shigeru	fuyu	megu	noboru
	e	kaede	misae	yusuke	keisuke	kozue
	o	taro	hiroko	yuto	keiko	tomoko

Table 1. Japanese given names used in the experiment (Yoshitomi et al., 2011a)

	Just before speaking	In speaking firstvowel (/i/)	In speaking last vowel (/a/)
Angry			
Happy			
Neutral			
Sad			
Surprised			

Fig. 5. Thermal images of Subject A expressing all facial expressions when speaking "Shinnya"

	Just before speaking	In speaking first vowel (/i/)	In speaking last vowel (/a/)
Angry			
Happy			
Neutral			
Sad			
Surprised			

Fig. 6. Thermal images of Subject B expressing all facial expressions when speaking "Shinnya"

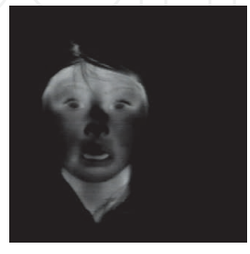
	Just before speaking	In speaking first owel (/i/)	In speaking last vowel (/a/)
Angry			
Happy			
Neutral			
Sad			
Surprised			

Fig. 7. Thermal images of Subject C expressing all facial expressions when speaking "Shinnya"

In the experiment, all subjects kept intentionally the front-view faces in the AVI files saved as both the training and test data. Accordingly, the pre-processing module for front-view face judgment (Fujimura et al., 2011) was not used in the experiment. We assembled 20 samples as training data and 10 samples as test data. From one sample, we obtained three images at the timing positions of just before speaking, and just speaking the phonemes of the first and last vowels. We obtained training data for all combinations of vowel type of the first and last vowels.

4.2 Results and discussion

The mean values for the recognition accuracy of Subject A in speaking 25 names with five emotions were 94.1% for the first vowel and 87.0% for the last vowel. Those of Subject B were 87.4% for the first vowel and 80.4% for the last vowel. Those of Subject C were 84.7% for the first vowel and 70.8% for the last vowel. For all subjects, Julius recognized the first vowel more accurately than the last vowel. Tables 2, 3 and 4 show the recognition accuracy for both the first and last vowels of Subjects A, B, and C, respectively. In Tables 2, 3 and 4, the recognition accuracy means the ratio in percentage of cases in which both the first and last vowels are correctly recognized. The mean values for the recognition accuracy for both the first and last vowels of Subject A in speaking 25 names with each emotion were 82.0% for “angry," 87.2% for “happy," 82.0% for “neutral," 83.6% for “sad," and 76.4% for “surprised." Those of Subject B were 54.2% for “angry," 74.5% for “happy," 98.0% for “neutral," 69.2% for “sad," and 62.0% for “surprised." Those of Subject C were 54.4% for “angry," 47.4% for “happy," 74.0% for “neutral," 63.6% for “sad," and 55.6% for “surprised." For both Subjects B and C, the mean value of the recognition accuracy for both the first and last vowels in speaking 25 names with the emotion of “neutral” was higher than that with other emotions, while the difference of emotion did not have much of an influence on the mean value of the recognition accuracy for both the first and last vowels in speaking 25 names in the case of Subject A, who could clearly pronounce the names selected in the study with all of the emotions.

Table 5 shows the mean values for the recognition accuracy for both the first and last vowels in speaking each name while expressing all five emotions. The highest value, 98.7%, as the mean value of all subjects was obtained for the name “Shinnya” where the first and last vowels are /i/ and /a/, respectively, while the lowest, 42.0%, was obtained for “Yuki” where the first and last vowels are /u/ and /i/, respectively. Moreover, the mean values of the recognition accuracy for both the first and last vowels with five emotions depended remarkably on the name to be pronounced, especially with Subject C. Figure 8 shows the waveforms for “Shinnya” pronounced by Subjects A, B, and C while expressing each emotion. All of the first and last vowels whose waveforms are shown in Fig. 8 were correctly recognized by Julius. Figure 9 shows the waveforms of “Koji” pronounced by Subjects A, B, and C for each emotion. In Fig. 9, all of the first and last vowels pronounced by Subject A were correctly recognized by Julius. As shown in Fig.9, some of the first and last vowels pronounced by Subject B and C were misrecognized by Julius. Julius tends to correctly recognize an utterance when it is clearly pronounced.

Table 6 shows facial expression recognition accuracy as mean values over all combinations of first and last vowels. The mean recognition accuracy of the facial expressions of all subjects was 79.8%. The mean recognition accuracy of the facial expressions was 85.5% for Subject A, 74.1% for Subject B, and 79.7% for Subject C. As stated in Section 3.2, Julius sometimes makes a mistake in recognizing the first and/or last vowel(s). For example, /a/ for the last vowel is sometimes misrecognized as /o/.

First-last vowel	Angry	Happy	Neutral	Sad	Surprised	Mean
a-a	50	100	100	80	0	66.0
a-i	100	100	100	90	80	94.0
a-u	90	0	100	100	70	72.0
a-e	80	100	100	90	100	94.0
a-o	80	60	100	90	80	82.0
i-a	100	100	100	100	100	100.0
i-i	100	80	100	100	60	88.0
i-u	100	80	100	100	90	94.0
i-e	30	100	70	40	0	48.0
i-o	100	100	30	20	100	70.0
u-a	80	70	40	70	60	64.0
u-i	100	90	20	10	100	64.0
u-u	90	100	70	70	100	86.0
u-e	100	100	100	100	100	100.0
u-o	100	100	70	100	100	94.0
e-a	50	100	80	100	10	68.0
e-i	100	100	100	100	100	100.0
e-u	30	40	100	100	90	72.0
e-e	50	80	100	90	90	82.0
e-o	100	100	60	100	90	90.0
o-a	80	80	100	100	60	84.0
o-i	100	100	100	100	100	100.0
o-u	100	100	100	100	100	100.0
o-e	100	100	100	100	70	94.0
o-o	40	100	10	40	60	50.0
Mean	82.0	87.2	82.0	83.6	76.4	82.2

Table 2. Accuracy (%) of speech recognition for Subject A (Yoshitomi et al., 2011b)

First-last vowel	Angry	Happy	Neutral	Sad	Surprised	Mean
a-a	30	100	100	100	60	78.0
a-i	40	60	100	70	33	60.6
a-u	50	90	100	56	80	75.2
a-e	67	90	90	70	60	75.4
a-o	30	90	100	60	20	60.0
i-a	100	100	100	100	100	100.0
i-i	40	90	100	90	20	68.0
i-u	20	40	100	78	0	47.6
i-e	70	75	100	70	67	76.4
i-o	0	90	90	90	60	66.0
u-a	50	100	100	100	80	86.0
u-i	70	10	90	10	70	50.0
u-u	100	60	100	90	100	90.0
u-e	50	78	100	67	100	79.0
u-o	10	40	100	20	90	52.0
e-a	80	90	90	90	80	86.0
e-i	100	90	100	70	100	92.0
e-u	0	40	100	50	30	44.0
e-e	89	80	90	90	90	87.8
e-o	50	60	100	90	90	78.0
o-a	80	100	100	80	0	72.0
o-i	60	30	100	10	60	52.0
o-u	40	90	100	10	50	58.0
o-e	30	90	100	80	100	80.0
o-o	100	80	100	90	10	76.0
Mean	54.2	74.5	98.0	69.2	62.0	71.6

Table 3. Accuracy (%) of speech recognition for Subject B

First-last vowel	Angry	Happy	Neutral	Sad	Surprised	Mean
a-a	100	78	100	50	80	81.6
a-i	80	100	100	100	40	84.0
a-u	0	0	40	40	0	16.0
a-e	60	60	90	30	100	68.0
a-o	89	40	100	80	100	81.8
i-a	100	100	100	80	100	96.0
i-i	0	56	100	10	30	39.2
i-u	0	0	0	50	0	10.0
i-e	90	90	100	80	50	82.0
i-o	70	20	100	80	80	70.0
u-a	70	100	100	90	90	90.0
u-i	10	0	50	0	0	12.0
u-u	0	0	0	10	0	2.0
u-e	70	80	100	80	50	76.0
u-o	40	0	60	80	10	38.0
e-a	90	60	100	60	100	82.0
e-i	100	100	100	100	100	100.0
e-u	0	0	10	70	10	18.0
e-e	90	100	100	100	90	96.0
e-o	40	30	100	100	100	74.0
o-a	90	70	100	90	50	80.0
o-i	0	0	0	0	0	0.0
o-u	0	0	10	10	30	10.0
o-e	100	20	90	100	90	80.0
o-o	70	80	100	100	90	88.0
Mean	54.4	47.4	74.0	63.6	55.6	59.0

Table 4. Accuracy (%) of speech recognition for Subject C

First-last vowel	Sbject A	Subject B	Subject C	Mean
a-a	66.0	78.0	81.6	75.2
a-i	94.0	60.6	84.0	79.5
a-u	72.0	75.2	16.0	54.4
a-e	94.0	75.4	68.0	79.1
a-o	82.0	60.0	81.8	74.6
i-a	100.0	100.0	96.0	98.7
i-i	88.0	68.0	39.2	65.1
i-u	94.0	47.6	10.0	50.5
i-e	48.0	76.4	82.0	68.8
i-o	70.0	66.0	70.0	68.7
u-a	64.0	86.0	90.0	80.0
u-i	64.0	50.0	12.0	42.0
u-u	86.0	90.0	2.0	59.3
u-e	100.0	79.0	76.0	85.0
u-o	94.0	52.0	38.0	61.3
e-a	68.0	86.0	82.0	78.7
e-i	100.0	92.0	100.0	97.3
e-u	72.0	44.0	18.0	44.7
e-e	82.0	87.8	96.0	88.6
e-o	90.0	78.0	74.0	80.7
o-a	84.0	72.0	80.0	78.7
o-i	100.0	52.0	0.0	50.7
o-u	100.0	58.0	10.0	56.0
o-e	94.0	80.0	80.0	84.7
o-o	50.0	76.0	88.0	71.3
Mean	82.2	71.6	59.0	70.9

Table 5. Accuracy (%) of speech recognition for Subjects A, B, and C

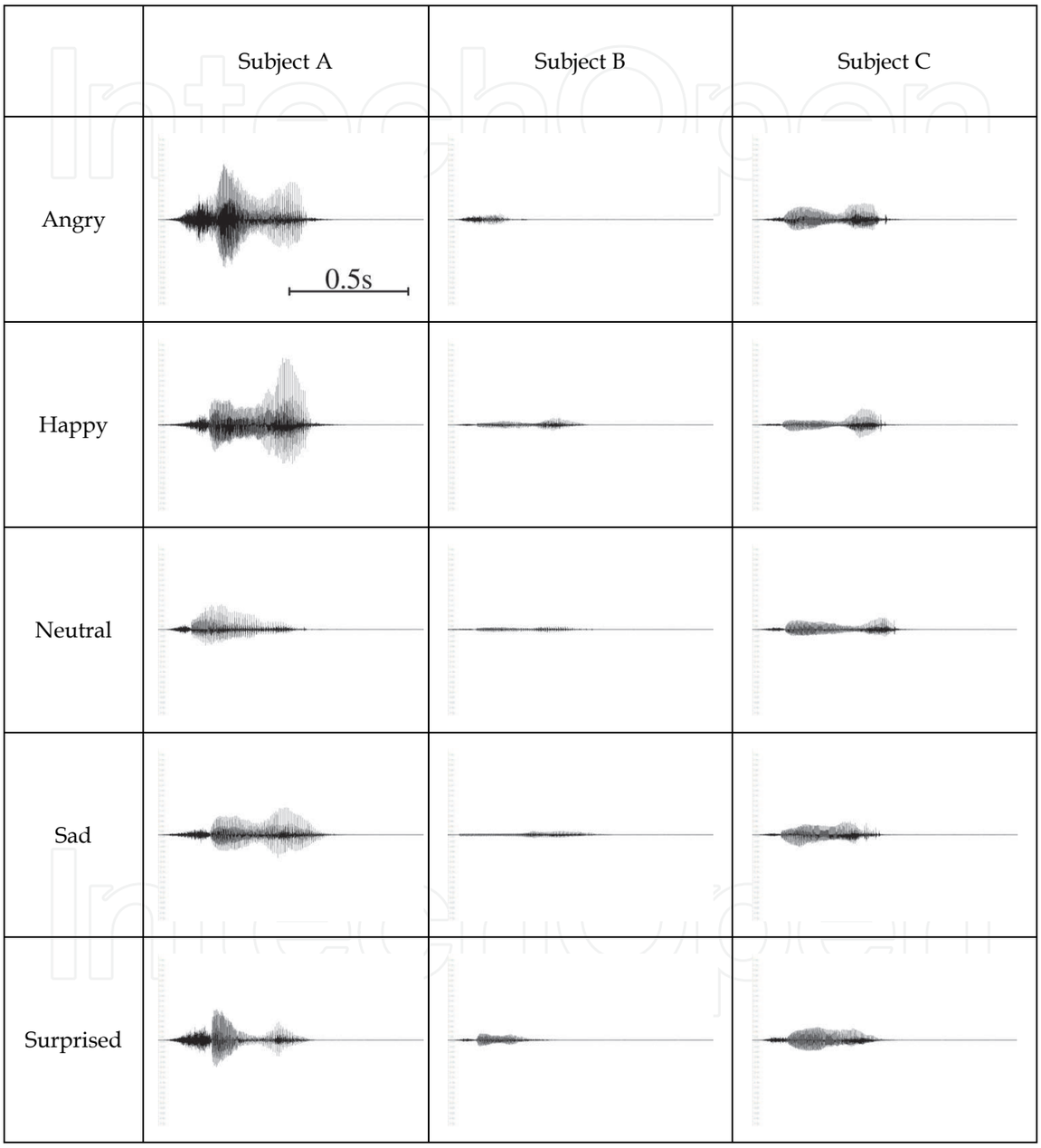


Fig. 8. Waveforms for each subject when speaking “Shinnya”, whose vowels are /i/ for the first and /a/ for the last, while expressing each emotion

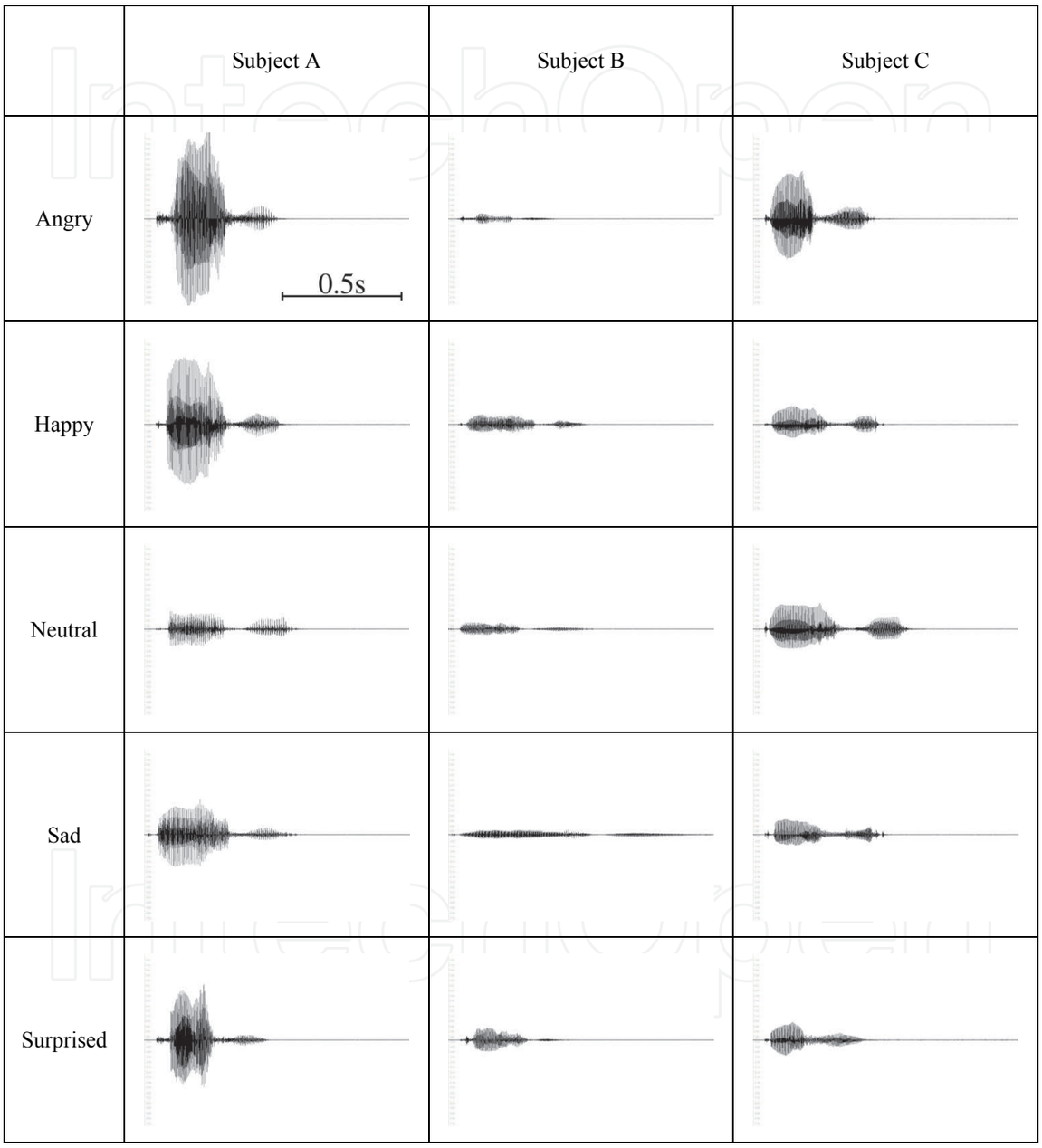


Fig. 9. Waveforms for each subject when speaking “Koji”, whose vowels are /o/ for the first and /i/ for the last, while expressing each emotion

In order to estimate the effect of improving the recognition accuracy of vowel(s), we manually corrected the misrecognition of vowel(s) when Julius made a mistake in recognizing the first and/or last vowel(s). In this case, for example, when Julius correctly judged the first vowel at the utterance of "Ayaka" but misjudged the last vowel as /o/, the training data in speaking "Ayaka" were used for facial expression recognition after manually correcting /o/ into /a/on the speech recognition of the last vowel. Table 7 shows the accuracy in recognition of facial expressions as mean values for all combinations of first and last vowels after correcting the misrecognition of vowel(s). Each value in Table 7 means one obtained in the case of perfect recognition of both the first and last vowels. In such a case, the mean recognition accuracy of the facial expressions of all subjects was 87.0%, and the mean recognition accuracy of the facial expressions was 89.7% for Subject A, 82.5% for Subject B, and 88.7% for Subject C. Accordingly, improving the recognition of the first and last vowels would improve the mean value of facial expression recognition by up to 7.2%.
The mean values of the recognition accuracy of all subjects in speaking 25 names while expressing all five emotions were 88.7% for the first vowel and 79.4% for the last vowel. The recognition accuracy of vowels pronounced while expressing various emotions might be high enough to decide the timing of facial expression recognition using the speech recognition system. Accordingly, as a continuation of our work, we will use the proposed method for recognizing facial expressions in daily conversation.

Subject A		Input facial expression				
		Angry	Happy	Neutral	Sad	Surprised
Output	Angry	90.4	0.4	0.0	6.0	2.4
	Happy	3.6	94.8	9.6	4.8	7.2
	Neutral	0.0	0.0	87.2	0.0	0.0
	Sad	0.4	2.8	0.4	78.8	14.0
	Surprised	5.6	2.0	2.8	10.4	76.4
Subject B		Input facial expression				
		Angry	Happy	Neutral	Sad	Surprised
Output	Angry	64.1	5.2	5.2	6.7	9.2
	Happy	10.8	77.2	0.8	7.3	13.4
	Neutral	0.0	4.0	83.6	0.0	0.0
	Sad	7.2	7.6	2.0	78.0	9.7
	Surprised	17.9	6.0	8.4	8.0	67.7
Subject C		Input facial expression				
		Angry	Happy	Neutral	Sad	Surprised
Output	Angry	78.8	4.4	7.2	5.2	0.8
	Happy	2.4	77.1	3.6	2.4	0.0
	Neutral	10.8	2.0	71.9	3.2	2.0
	Sad	2.8	11.3	12.9	78.8	5.2
	Surprised	5.2	5.2	4.4	10.4	92.0

Table 6. Accuracy (%) of facial expression recognition (Yoshitomi et al., 2011b)

Subject A		Input facial expression				
		Angry	Happy	Neutral	Sad	Surprised
Output	Angry	91.6	0.4	0.0	4.8	1.2
	Happy	3.2	94.8	2.4	4.8	3.2
	Neutral	0.0	0.0	97.6	0.0	0.0
	Sad	0.4	2.8	0.0	79.2	10.4
	Surprised	4.8	2.0	0.0	11.2	85.2
Subject B		Input facial expression				
		Angry	Happy	Neutral	Sad	Surprised
Output	Angry	79.2	2.8	4.4	3.7	3.6
	Happy	5.6	85.2	2.8	3.6	11.9
	Neutral	0.0	4.0	83.6	0.0	0.0
	Sad	2.4	4.0	1.2	89.1	9.2
	Surprised	12.8	4.0	8.0	3.6	75.3
Subject C		Input facial expression				
		Angry	Happy	Neutral	Sad	Surprised
Output	Angry	88.4	2.4	4.8	3.6	0.4
	Happy	0.4	92.8	2.8	1.6	0.0
	Neutral	4.8	0.8	79.1	2.0	0.4
	Sad	1.6	1.6	10.5	86.0	2.0
	Surprised	4.8	2.4	2.8	6.8	97.2

Table 7. Accuracy (%) of facial expression recognition for perfect speech recognition of both first and last vowels

5. Conclusion

We have developed a method for recognizing the facial expressions of a speaker by using thermal image processing and a speech recognition system. To implement the proposed method, three subjects spoke 25 Japanese given names that provided all combinations of first and last vowels. These subjects were used to prepare training data and then test data for all combinations of the first and last vowels. The mean values of the recognition accuracy of all subjects in speaking 25 names while expressing five emotions were 88.7% for the first vowel and 79.4% for the last vowel. Using the proposed method, the facial expressions of three subjects were discernable with an accuracy of 79.8% when the subject exhibited one of the intentional facial expressions of “angry,” “happy,” “neutral,” “sad,” and “surprised.” Improving the recognition of the first and last vowels could improve the mean value of facial expression recognition by up to 7.2% The recognition accuracy of vowels pronounced

with various emotions might be high enough to decide the timing of facial expression recognition using the speech recognition system. We expect the proposed method to be applicable for recognizing facial expressions in daily conversation.

6. Acknowledgment

We would like to thank Mr. K. Shimada of Nova System Co. Ltd for his valuable cooperation while he was a student at Kyoto Prefectural University (Yoshitomi et al., 2011a, 2011b). We would like to thank all the subjects who cooperated with us in the experiments. This work was supported by KAKENHI (22300077).

7. References

Fujimura, T.; Yoshitomi, Y.; Asada, T. & Tabuse, M. (2011). Facial Expression Recognition of a Speaker Using Front-view Face Judgment, Vowel Judgment, and Thermal Image Processing, *Proceedings of 16th International Symposium on Artificial Life and Robotics*, pp. 219-224, ISBN 978-4-9902880-5-1, Beppu, Oita, Japan, January 27-29, 2011

Harashima, H.; Choi, C. S. & Takebe, T. (1989). 3-D Model-based Synthesis of Facial Expressions and Shape Deformation, *Human Interface,* Vol.4, pp. 157-166 (in Japanese)

Ikezoe, F.; Ko, R.; Tanijiri, T. & Yoshitomi, Y. (2004). Facial Expression Recognition for Speaker Using Thermal Image Processing, *Transaction of Human Interface Society,* Vol.6, No.1, (February 2004), pp. 19-27 (in Japanese)

Kawahara,T.; Lee, A.; & Kiyohiro Shikano, K. (2010). Julius: Open-source software toolkit for large vocabulary continuous speech recognition. In S.Itahashi and C-Y.Tseng, editors, *Computer Processing of Asian Spoken Languages*, Consideration Books, pp.305-308. 2010

Kawahara, T. et al. (December 2010). Open-Source Large Vocabulary CSR Engine Julius Julius rev.4.1.5.1 http://julius.sourceforge.jp/

Kobayashi, H. & Hara, F. (1994). Analysis of Neural Network Recognition Characteristics of 6 Basic Facial Expressions, *Proceedings of 3rd IEEE International Workshop on Robot and Human Communication*, pp. 222-227, ISBN 0-7803-2002-6, Nagoya, Japan, July 18-20, 1994

Koda, Y.; Yoshitomi, Y.; Nakano, M. & Tabuse, M. (2009), Facial Expression Recognition for a Speaker of a Phoneme of Vowel Using Thermal Image Processing and a Speech Recognition System, *Proceedings of 18th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 955-960, ISBN 978-4244-5081-7, ISSN 1944-9445, Toyama, Japan, September 29 - October 1, 2009

Lee, A. & Tatsuya Kawahara, T. (2009), Recent Development of Open-Source Speech Recognition Engine Julius, *Proceedings of Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 131-137, Sapporo, Japan, October 4-7, 2009

Mase, K. (1990). An Application of Optical Flow – Extraction of Facial Expression, *Proceedings of IAPR Workshop on Machine Vision and Application*, pp. 195-198, Kokubunji, Tokyo, Japan, November 28-30, 1990

Mase, K. (1991). Recognition of Facial Expression from Optical Flow. *Transaction of IEICE*, Vol.E74, No.10, (October 1991), pp. 3474-3483

Matsuno, K.; Lee, C.; & Tsuji, S. (1994). Recognition of Facial Expressions Using Potential Net and KL Expansion, *Transaction of IEICE,* Vol.J77-D-II, No.8 , pp. 1591-1600 (in Japanese)

Nakano, M.; Ikezoe, F.; Tabuse, M. & Yoshitomi, Y. (2009). A Study on the Efficient Facial Expression Using Thermal Face Image in Speaking and the Influence of Individual Variations on Its Performance, *Journal of IEEJ,* Vol.38, No.2, (March 2009), pp. 156-163, ISSN 0285-9831 (in Japanese)

Sugimoto, Y.; Yoshitomi, Y.; & Tomita, S. (2000). A Method for Detecting Transitions of Emotional States Using a Thermal Face Image Based on a Synthesis of Facial expressions, *Journal of Robotics and Autonomous Systems,* Vol.31, No.3, (May 2000), pp. 147-160, ISSN 0921-8890

Yoshitomi, Y.; Kimura, S.; Hira, E. & Tomita, S. (1996). Facial Expression Recognition Using Infrared Rays Image Processing, *Proceedings of the Annual Convention IPS Japan*, Vol.2, pp. 339-340, Osaka, Japan, September 4-6, 1996

Yoshitomi, Y.; Kimura, S.; Hira, E. & Tomita, S. (1997). Facial Expression Recognition Using Thermal Image Processing, *IPSJ SIG Notes*, Vol.CVIM103-3, pp. 17-24, Kyoto, Japan, January 23-24, 1997

Yoshitomi, Y.; Miyawaki, N.; Tomita, S. & Kimura, S. (1997). Facial Expression Recognition Using Thermal Image Processing and Neural Network, *Proceedings of 6th IEEE International Workshop on Robot and Human Communication*, pp. 380-385, ISBN 0-7803-4076-0 (Softbound Edition), 0-7803-4077-9 (Microfiche Edition), Sendai, Japan, September 29 - October 1, 1997

Yoshitomi, Y.; Kim, S.-Ill; Kawano, T., & Kitazoe, T. (2000). Effect of Sensor Fusion for Recognition of Emotional States Using Voice, Face Image and Thermal Image of Face, *Proceedings of 6th IEEE International Workshop on Robot and Human Interactive Communication*, pp. 178-183, ISBN 0-7803-6273-X, Osaka, Japan, September 27-29, 2000

Yoshitomi, Y. (2010). Facial Expression Recognition for Speaker Using Thermal Image Processing and Speech Recognition System, *Proceedings of 10th WSEAS International Conference on Applied Computer Science*, pp. 182-186, ISBN 978-960-474-231-8, ISSN 1792-4863, Appi Kogen, Iwate, Japan, October 4-6, 2010

Yoshitomi, Y.; Asada, T. ; Shimada, K. & Tabuse, M. (2011). Facial Expression Recognition of a Speaker Using Vowel Judgment, and Thermal Image Processing, *Proceedings of 16th International Symposium on Artificial Life and Robotics*, pp. 225-230, ISBN 978-4-9902880-5-1, Beppu, Oita, Japan, January 27-29, 2011

Yoshitomi, Y.; Asada, T. ; Shimada, K. & Tabuse, M. (2011). Facial Expression Recognition of a Speaker Using Vowel Judgment, and Thermal Image Processing, *Proceedings of Journal of Artificial Life and Robotics*, Vol. 16, to appear

Yuille, A. L.; Cohen, D. S.; & Hallinan, P. W. (1989). Feature Extraction from Faces Using Deformable Templates, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 104-109, ISBN 0-8186-1952-X, San Diego, California , USA, June 4-8, 1989

IntechOpen

IntechOpen



Speech Technologies

Edited by Prof. Ivo Ipsic

ISBN 978-953-307-996-7
Hard cover, 432 pages
Publisher InTech
Published online 23, June, 2011
Published in print edition June, 2011

This book addresses different aspects of the research field and a wide range of topics in speech signal processing, speech recognition and language processing. The chapters are divided in three different sections: Speech Signal Modeling, Speech Recognition and Applications. The chapters in the first section cover some essential topics in speech signal processing used for building speech recognition as well as for speech synthesis systems: speech feature enhancement, speech feature vector dimensionality reduction, segmentation of speech frames into phonetic segments. The chapters of the second part cover speech recognition methods and techniques used to read speech from various speech databases and broadcast news recognition for English and non-English languages. The third section of the book presents various speech technology applications used for body conducted speech recognition, hearing impairment, multimodal interfaces and facial expression recognition.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Yasunari Yoshitomi, Taro Asada and Masayoshi Tabuse (2011). Vowel Judgment for Facial Expression Recognition of a Speaker, Speech Technologies, Prof. Ivo Ipsic (Ed.), ISBN: 978-953-307-996-7, InTech, Available from: http://www.intechopen.com/books/speech-technologies/vowel-judgment-for-facial-expression-recognition-of-a-speaker

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2011 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License, which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.

IntechOpen

IntechOpen