TEXT DRIVEN LIPS ANIMATION SYNCHRONIZING WITH TEXT-TO-SPEECH FOR AUTOMATIC READING SYSTEM WITH EMOTION

Futoshi Sugimoto
Dept. of Information and Computer Sciences, Toyo University, 2100 Kujirai Kawagoe 3508585, Japan

Keywords: Lips animation, Text-to-speech, Phoneme context, Audio, Visual signal.

Abstract: We developed a lips animation system that is simple and suits the characteristics of Japanese. It adopts phoneme context that is usually used in speech synthesis. It only needs a small database that contains tracing data of movement of lips for each phoneme context. Through an experiment using subjects in which the subjects read a word from lips animation, we got result as correct as from real lips.

1 INTRODUCTION

Nowadays, the speech synthesis system is used in various areas of our daily life and many services are provided. However, since speech is the means of transmitting information only acoustically, the information could not be obtained when the speech is interfered or difficult to hear. If we can develop an interface which transmit information with not just speech but other media, the interface will be effective as the supplemental mean to transmit information of speech. There was a report showing that in an environment, where there exist multiple sounds, by showing a video display of a speaker increases the understanding of the speech content (Rudmann, Mccarley, and Kramer 2003). Thus, we are able to say that information can be transmitted more effectively and more precisely when both audio and visual signals are used synchronically.

We have been developing an automatic reading system that reads novels with emotion, and then we need lips animation for the reading system. There are many studies about lips animation (Ezzat and Poggio, 2000) (Kalberer, 2001) (Kim and Ko, 2001) (Bondy and Georganas, 2001). However, we developed a lips animation system that is simple and suits the characteristics of Japanese. It adopts phoneme context that is usually used in speech synthesis, and only needs a small database that contains tracing data of movement of lips for each phoneme context.

Through an experiment using subjects in which the subjects read a word from lips animation, we got result as correct as from real lips.

2 CONSTITUTE OF JAPANESE PHONEME

The smallest unit of an uttered sound is a consonant phoneme or a vowel phoneme. In Japanese, a syllable is formed by a vowel only or a combination of two phonemes of consonant and vowel. In addition, a word can be formed by a series of syllables. (Storm, 2000) Based on this fact, a word is a series of phonemes. Extraction from a part of the series is the phoneme context.

2.1 Phoneme Context

It is still in controversial whether or not when a word is spoken, each phoneme is affected by the phoneme in front and after it. In this study, we presumed that a phoneme is not influenced by the phoneme after it but only by the phoneme preceding it. A phoneme in Japanese is ended in a vowel. As a result, the preceding syllable is ended in a vowel. When we compared the utterance of a consonant and a vowel, we found that the vowel takes longer time to pronounce. The shape of lips when uttering a vowel gives the most influence in the change of lips shape when a syllable is pronounced. Also, the existence of consonants is for the purpose of adding

Sugimoto F. (2009).
TEXT DRIVEN LIPS ANIMATION SYNCHRONIZING WITH TEXT-TO-SPEECH FOR AUTOMATIC READING SYSTEM WITH EMOTION.
In *Proceedings of the 11th International Conference on Enterprise Information Systems - Human-Computer Interaction*, pages 194-197
DOI: 10.5220/0002156201940197

characteristic in lines of consonants. In terms of time, producing the consonant sound takes extremely short time. It ends in such an instant that it does not affect the change of lips shape. From the above, we assumed that the change of lips shape occurs based on utterance from a vowel to a vowel. And due to the consonants in between vowels, it creates some influences during the changes.

In this study, we assumed a link of a vowel + a consonant + a vowel (V + C + V) as a unit. Additionally, we assumed this unit as a phoneme context. Because there are two vowels in the same phoneme context, we called the former vowel as "the first vowel" and the latter as "the second vowel." The first vowel does not represent the whole phoneme and shall express only last shape of the lips at completion of utterance of the vowel. When there is no first vowel at the beginning of an utterance, thus, we added one more combination of a consonant + a vowel (C + V) to phoneme context. Of course, there is utterance of a vowel + a vowel. We considered such utterance a special case where the consonant does not exist and included it in the phoneme context of V + C + V Therefore, we have two phoneme contexts. These two types are enough for this study. (see Figure 1)

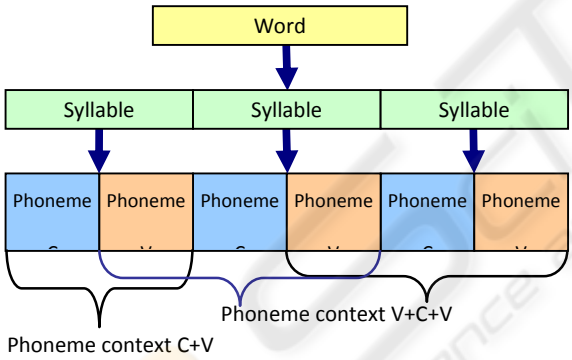

Figure 1: Phoneme context in Japanese.

2.2 Types of Phoneme Context

We presumed that there are 25 combinations of the first vowel and the second vowel in terms of macro taxonomy. There are several kinds of consonants sandwiched between the first vowel and the second vowel. In each macro taxonomy, depending on the consonant between vowels, the change of lips shape differs. In order to reproduce the change of lips shape when producing each phoneme context, we designed a trace data of the characteristic points of lips. There are 116 types of phoneme context in C + V; there are 565 types of phoneme context in V + C + V. As a consequence, almost all Japanese words can be spoken by combining these phoneme contexts together.

3 CHARACTERISTIC EXTRACTION OF LIPS SHAPE

In order to develop a lips animation, we must extract the change characteristic of lips shape. Therefore, we extracted the change characteristics from an image and a sound of change of lips shape at the time of a word spoken.

3.1 Recording both Image and Sound of Speaking Word

A high speed camera was used for recording. The shutter speed was set at 240 fps and the size was 256x256 pixels for a frame. In order to trace the characteristic points of the lips precisely after the recording, we printed eight characteristic points of the edge of lips of the subject as shown in Figure 2. We recorded 400 words which were spoken by the subject. The entire phoneme contexts were all included in these 400 words.

3.2 Image Extraction of Phoneme Context

For dividing the phonemes after recording, as shown in Figure 3, we analyzed the sounds which were recorded concurrently with the images. We used spectrogram as a clue and divided the image of the lips by each phoneme. The sound analysis software was used; we paid special attention to characteristic differences of formant phonemes and divided each phoneme in time axis. The time axis is applied as the time frame of lip images.

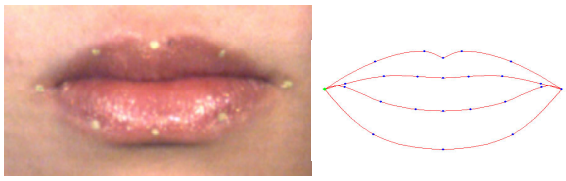

Figure 2: Lips of the subject and the basic shape of a two-dimensional computer graphic model of lips.

One phoneme context is from the endpoint of the first vowel to the endpoint of the second vowel. According to the sound analysis and the border of specified vowel and consonant as a clue, we were able to identify the part corresponding to the phoneme context mentioned in the section 2.2 from

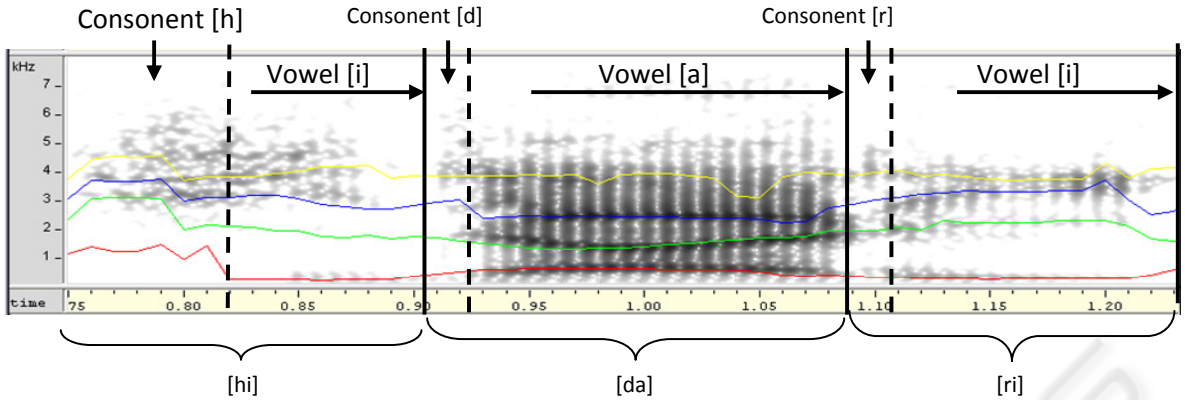

Figure 3: Sound analysis for dividing into the phonemes.

the images of speaking words. This enabled me to extract the image of the phoneme context.

3.3 Trace the Movement of a Characteristic Point

As mentioned previously, an endpoint of the first vowel to the endpoint of the second vowel is one phoneme context. The beginning part of latter phoneme context is the ending of the second vowel of the former phoneme context and the shape of lips at this instant becomes shape of lips of the first vowel of the latter phoneme context. From this point, a consonant is caught in between and reaching the second vowel and then a phoneme context is completed. From the recorded images, we traced movements of the characteristic points of lips corresponding to these series of flows. From the image of phoneme context extracted based on sound analysis, we used an image processing software that is able to get the data which was traced in two-dimensional movements of the printed characteristic points of lips of the subject along a time axis.

4 REALIZATION OF A LIPS ANIMATION

The basic shape of lips was drawn based on the lips image of the recorded subject. The basic shape of a two-dimensional computer graphic model of lips as shown in Figure 2 consists of four curves which are the outlines of the top and bottom part of upper lip and the top and bottom part of lower lip. These curves are drawn in Spline function. The numbers of the control point of the Spline function are as follows: 7 on the top part of upper lip, 9 on both the bottom part of upper lip and the top part of the lower lip, and 5 on the bottom part of the lower lip. Since the lips are symmetric, there are actually only 14 control points of the left half of the lips.

Because the beginning and the ending of a phoneme context becomes a perfect shape of lips when uttering a vowel, it is an extremely important factor for making lips animation. As shown in Figure 4, using the method mentioned above, based on the image of the subject we made shapes of lips on five vowels. When speaking a word, the lips surely become these shapes at the beginning and ending of a phoneme context. The movement of lips during the beginning and ending was realized, based on the trace data of the control points mentioned earlier. The bottom part of the upper lip and the top part of the lower lip, that have no trace data, were designed to interlock with the movements of top part of the upper lip and the bottom part of the lower lip respectively. In addition, regarding the teeth, only the bottom teeth were designed to link with the movement of the lower lip.

Table 1: Experimental result of reading word.

Japanese	English meaning	Lips animation	Real lips
Suzume	Sparrow	0.52	1.00
Yunomi	Teacup	0.56	0.88
Tesou	One's palm	0.65	0.65
Antena	Antenna	0.47	0.76
Izakaya	Bar	0.88	0.88
Setomono	Pottery	0.76	0.88
Everesuto	Everest	0.82	0.94
Kagurazaka	Kagurazaka(name)	0.56	0.56
Kasiopeya	Cassiopeia	0.65	0.76

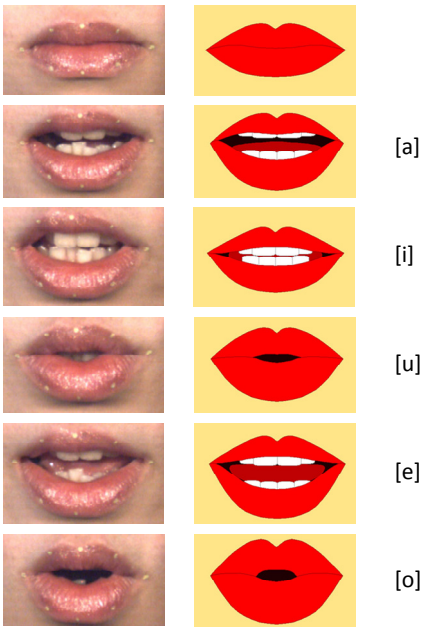

Figure 4: Shapes of lips corresponding to vowels.

5 EVALUATION AND CONCLUSIONS

We conducted an experiment to check how subjects can read a word from our lips animation. The subjects were one word among five ones after looking at an animation. The five words consist of one correct word and four dummy words that have the same number of syllables. The same experiment was done about real talking lips. Table 1 shows the average correction rate for 10 subjects. We can read a word from the lips animation a little bit less correctly than from real lips.

6 FUTURE WORKS

We developed a lips animation system that is simple and suits the characteristics of Japanese. We aim to apply this lips animation to our automatic reading system that reads novels with emotion, so the reality of the animation have to be improved on it's color and make it three-dimensional.

REFERENCES

Rudmann, D. S., Mccarley, J. S., Kramer, A. F., 2003. Bimodal displays improve speech comprehension in environments with multiple speakers, *Human Factors*, vol.45, no.2, pp.329-336.

Ezzat, T., Poggio, T., 2000. Visual Speech Synthesis by Morphing Visemes, *International Journal of Computer Vision*, Vol.38, No.1, pp.45-57.

Kalberer, G. A., 2001. Lip animation based on observed 3D speech dynamics, *Proceedings of SPIE, The International Society for Optical Engineering*, vol.4309, pp.16-25.

Kim, I. J., Ko, H. S., 200. 3D Lip-Synch Generation with Data-Faithful Machine Learning, *Computer Graphics Forum*, Vol.26, No.3, pp.295-301.

Bondy, M. D., Georganas, N. D., 2001. Model-based face and lip animation for interactive virtual reality applications, *Proceedings of the ninth ACM international conference on Multimedia*, pp.559-563.

Storm, H., 2000. *Ultimate Japanese*, Living Language Press.