International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:3, No:2, 2009

A Talking Head System for Korean Text

Sang-Wan Kim, Hoon Lee, Kyung-Ho Choi, and Soon-Young Park

Abstract**—**A talking head system (THS) is presented to animate the face of a speaking 3D avatar in such a way that it realistically pronounces the given Korean text. The proposed system consists of SAPI compliant text-to-speech (TTS) engine and MPEG-4 compliant face animation generator. The input to the THS is a unicode text that is to be spoken with synchronized lip shape. The TTS engine generates a phoneme sequence with their duration and audio data. The TTS applies the coarticulation rules to the phoneme sequence and sends a mouth animation sequence to the face modeler. The proposed THS can make more natural lip sync and facial expression by using the face animation generator than those using the conventional visemes only. The experimental results show that our system has great potential for the implementation of talking head for Korean text.

Keywords**—**Talking head, Lip sync, TTS, MPEG4.

I. Introduction

THE comprehension of speech depends not only on the auditory information, but also on the visual information such as lip movements or facial expression. The audial and visual multi-modal synthesis involves many applications: increase of understanding rate in noisy environments, enhancement of accessibility to human computer interaction (HCI), decrease of the bit-rate in coding schemes such as MPEG-4 [1], and automatic dubbing and realistic avatar animation [2]. There has been a large amount of research on incorporating bimodality of a speech into HCI interfaces [3]. The 2D or 3D modeling of realistic talking heads is one of the research topics in this area [4]. The talking head is defined as animated faces with lip shape mapping to specific synthesized speech. Virtual human's lip shape is helpful to viewers by providing visual information of speech sounds. We usually use TTS system to generate a sequence of phonemes from an input text. A phoneme is the basic unit of the acoustic speech. A visual representation of the phoneme is called viseme. Many phoneme sounds are visually ambiguous while being pronounced [5]. Therefore, one viseme can be used to represent the corresponding several phonemes. The conventional lip sync method is first to decompose the synthesized speech from a text sequence or the real speech into the phonemes. Then mapping between phonemes in the speech signal and visemes in the lookup table is carried out to construct the character's lip shape. MPEG-4 is an object-based multimedia compression standard that allows for encoding of differential audiovisual objects in the scene independently. MPEG-4 specifies 84 feature points on the neutral face, to provide spatial reference for defining FAP (facial animation parameter). The 68 FAPs represent a complete set of basic facial actions including head motion, tongue, eye and mouth control [1]. Among the FAPs, there are two high-level parameters which are visemes and expressions. MPEG-4 defines only 14 static visemes in the standard set. The shape of the mouth of a speaking human is not only influenced by the current phoneme, but also the neighboring phonemes.

In this paper, the THS is proposed by consisting of SAPI compliant TTS engine and MPEG-4 compliant face animation generator to animate the face of a speaking 3D avatar effectively. Especially, the proposed THS can make more natural lip sync and facial expression by using the coarticulation rules to the phoneme sequence extracted from Korean text.

Authors are with School of Information Engineering, Mokpo National University, Republic of Korea (e-mail: swkim, hoonlee, khchoi, sypark@mokpo.ac.kr).

II. The Proposed Talking Head System

There are 24 phonemes in the Korean alphabet: 14 consonants and 10 vowels. The phonemes are combined together into syllable blocks. The shapes of the consonants g/k, n, s, m and ng are graphical representations of the speech organs used to pronounce them. Other consonants were created by adding extra lines to the basic shapes.

Fig. 1 shows the block diagram of the proposed THS. The input to the THS is a unicode text that is to be spoken with synchronized lip shape. The TTS engine generates a phoneme sequence with their duration and audio data. The THS keeps the audio data in allocated memories and applies the coarticulation rules the phoneme sequence. Then a mouth animation sequence among FAPs is generated and sent to the Face modeler. The THS plays the audio data synchronizing the speech and the visemes. For the natural lip sync, the coarticulation extractor considers the dominance of the present and neighboring phonemes and the FAP converter generates the mouth animation sequence by mapping the Korean phonemes to the visemes of MPEG-4 standard set.

Some visemes for only English phonemes such as /F/, /V/, /Th/, /Sh/ are excluded in the MPEG-4 FAP set since the Korean alphabet does not pronounce those phonemes. The visemes for Korean double vowels such as ㅑ[ja], ㅠ[ju], ㅙ[wa] etc. are constructed by adding two visemes for each vowel consecutively. For example, the double vowel ㅑ[ja] are mapped to the viseme for ㅣ[i] followed by the viseme for ㅏ[a]. The rules for mapping the Korean phonemes to the visemes in the MPEG-4 are follows.

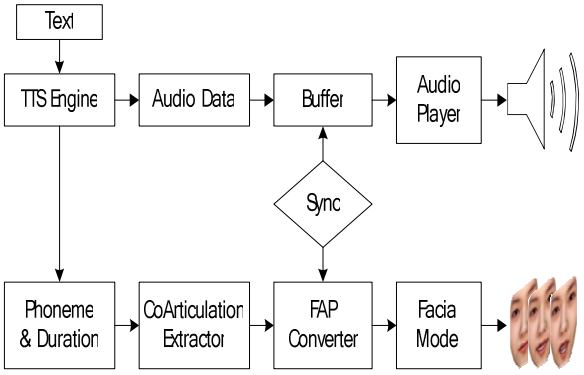


Fig. 1 Block diagram of the propose THS

A. Rule 1

Korean consonants ㄱ[g],ㅋ[k],ㅇ[-ŋ],ㅎ[h], and ㄹ[l]" located at the beginning or the end of a syllable do not need the visemes. Only the vowels in the middle of the syllable are mapped to their visemes.

B. Rule 2

Korean consonants ㄴ[n], ㄷ[d], ㅌ[t], ㅅ[s],ㅆ[s^*], ㅈ[ʧ], ㅉ[ʧ*] and ㅊ[ʧʰ] connected to the front or end of vowels ㅗ[o], ㅛ[jo], ㅜ[u], ㅠ[ju], ㅡ[ɨ], ㅣ[i], ㅢ[ɨj], and ㅟ[wi]" do not need their visemes. Only the vowels in the middle of the syllable are mapped to their visemes.

C. Rule 3

Korean consonants ㄴ[n], ㄷ[d], ㅌ[t], ㅅ[s],ㅆ[s^*], ㅈ[ʧ], ㅉ[ʧ*] and ㅊ[ʧʰ] which are following (at the end of a present syllable or beginning of a next syllable) vowles ㅑ[a], ㅒ[ja], ㅓ[ʌ], ㅕ[jʌ], ㅐ[æ], ㅖ[jæ], ㅘ[wa], ㅙ[wæ], ㅚ[we], ㅝ[wʌ], ㅞ[we] are mapped to their visemes.

D. Rule 4

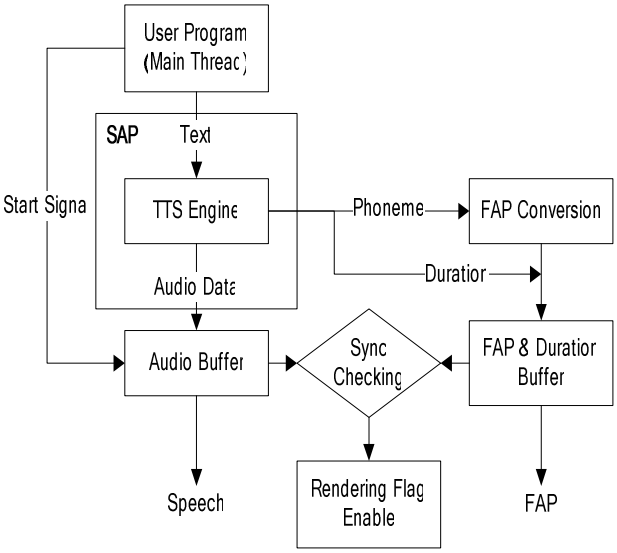
Double vowels need transition from one viseme to the next which are defined by blending two visemes with a weighting factor.

. ㅒ[ja]→ ㅣ[i] + ㅏ[a]
. ㅕ[jʌ]→ ㅣ[i] + ㅓ[ʌ]
. ㅖ[jæ] → ㅣ[i] + ㅐ[æ]
. ㅖ[je]→ ㅣ[i] + ㅔ[e]
. ㅘ[wa] → ㅗ[o] + ㅏ[a]
. ㅙ[wæ] → ㅗ[o] + ㅐ[æ]
. ㅚ[we] → ㅗ[o] + ㅔ[e]
. ㅝ[wʌ] → ㅜ[u] + ㅓ[ʌ]
. ㅞ[we] → ㅜ[u] + ㅔ[e]
. ㅟ[wi] → ㅜ[u] + ㅣ[i]

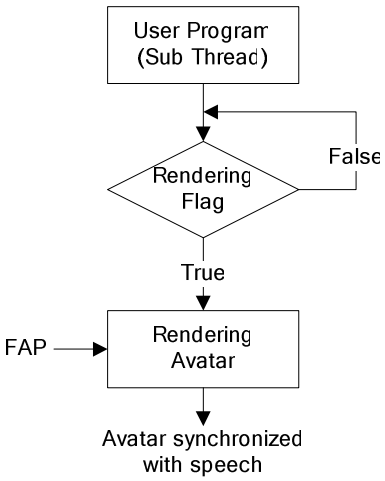
III. Experiments

The THS system has been developed with Visual C++ on MS windows. For Korean TTS, we used voice Text of Voiceware and MS Windows Speech API 5.1 system. The animation was not the topic of the interest in this work as it has already been implemented in the toolkit. In this paper, we used a MPEG-4 compliant face animation toolkit called FaceGen of Singular Inversions.

Fig. 2 shows the structure of the user program consisting of main and sub threads. The main thread sends a text to the SAPI controlled TTS engine. The audio data is allocated to the audio buffer and phoneme sequence is mapped to the visemes in the FAP by considering the coarticulation. Then Visemes and their durations are kept in the corresponding buffer. The user program for the main thread sends start signal for the audio buffer to play the audio data. The sync checking generates rendering flag by synchronizing the elapsed time of sounds and duration of visemes. The sub thread manages the rendering process by checking the rendering flag. If the rendering flag is true, then the THS generates the 3D avatar synchronized with speech.



(a) Main thread



(b) Sub thread

Fig. 2 The structure of the user program consisting of main and sub threads

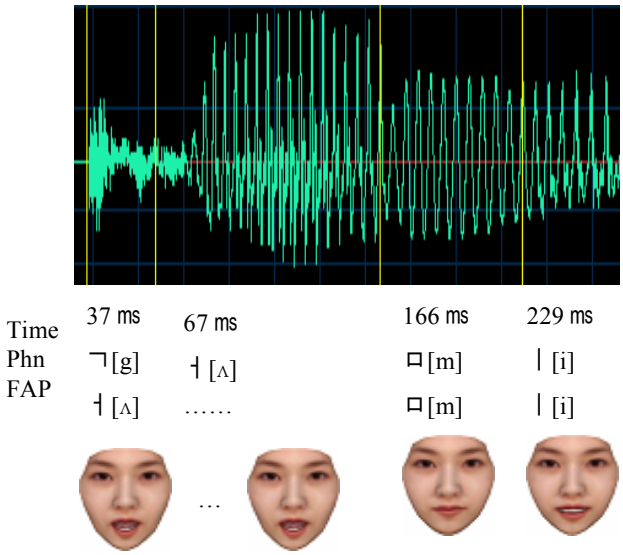


Fig. 3 Result of lip sync for '거미'/[gʌ-mi]/

Fig. 3 shows the result of lip sync for Korean word' 거미'/[gʌmi]/ using Rule 1. The viseme for ㅓ[ʌ] is appeared in the location of ㄱ[g] and is held to the end of the syllable without the viseme for ㄱ[g] while the viseme for ㅁ[m] is appeared at the beginning of the syllable.

Fig. 4 shows the result of lip sync for Korean word'아는' anun[a-n-eu-n] using Rule 2 and 3. The consonant ㄴ [n] following the vowel ㅏ[a] is mapped to its viseme (Rule 3) while ㄴ [n] connected at the end of vowel ㅡ[eu] does not need the viseme (Rule 2).

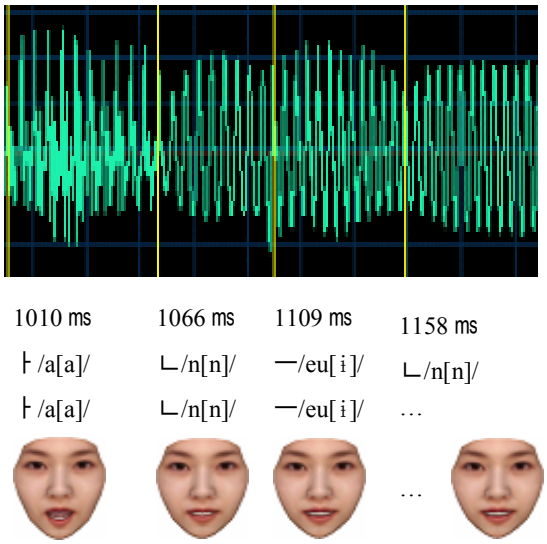


Fig. 4 Results of lip sync for '아는'/anun[a-n-eu-n]/

Fig. 5 shows the result of lip sync for Korean '왕' /[oa-ŋ]/ Double vowels ㅘ[wa] takes transition from one viseme to the next by blending two visemes with a weighting factor.

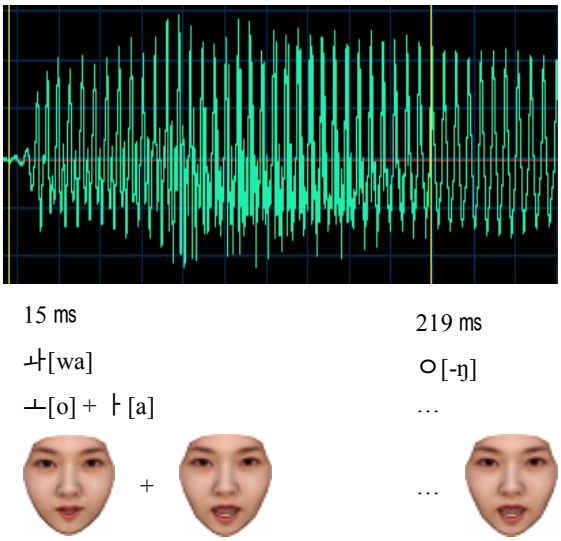


Fig. 5 Results of lip sync for '왕'/[oa--ŋ]/

The MPEG-4 compliant face animation generator employed in this experiment can also animate the six primary facial expressions. Fig. 6 shows the anger and surprise expressions when the avatar animate speech of a Korean text '언'/[u-n]/

(a) Anger expression

(b) Surprise Expression

Fig. 6 Facial expressions during speech of a Korean text '언'/[u-n]/

IV. Conclusion

This paper presented the THS to synchronize the lip movements of a speaking avatar with the Korean speech synthesized from Korean text. The proposed system consisted of SAPI compliant TTS engine and MPEG-4 compliant face animation generator. The input to the THS was a unicode text that was to be spoken with synchronized lip shape. For the natural talking head, the coarticulation extractor considered the dominance of the present and neighboring phonemes and the FAP converter generated the mouth animation sequence by mapping the Korean phonemes to the visemes of MPEG-4 standard set. The experimental results showed that our system had great potential for the implementation of talking head for Korean text.

Acknowledgment

This research was financially supported by the MEST and the KOTEF through the Human Resource Training Project for Regional Innovation.

References

[1] I. S. Pandzic and R. Forchheimer, Edited, *MPEG-4 Facial animation*, Wiley, England, 2002.
[2] C. Pelachaud, E. Magno-Caldognetto, "Modelling an Italian Head", Audio-visual speech processing, Scheelsminde, Denmark 2001
[3] E. Cosatto, J. Ostermann, H.P. Granf, "Lifelike talking faces for interactive services", Proc. IEEE91(9), 1406-1428, 2003.
[4] S. Morishima and S. Nakamura "Multimodal translation system using texture-mapped lip-sync images for video mail and automatic dubbing applications", EURASIP Journal on Applied Signal processing, pp. 1637-1647, 2004.
[5] G. Zoric and I.S. Pandzic, "Real-time language independent lip sysnchronization method using a genetic algorithm", Signal processing 86, pp. 3644-3656, 2006.