論文98-35S-7-12

가상현실을 위한 합성얼굴 동영상과 합성음성의 동기구현

(Synchronization of Synthetic Facial Image Sequences and Synthetic Speech for Virtual Reality)

崔 昌 石 *, 李 基 榮 **

(Chang-Seok Choi and Ki-Young Lee)

요 약

본연구에서는 합성된 얼굴동영상과 합성된 음성의 동기를 실현하는 방법을 제안한다. 음성은 LP-PSOLA에 의해 반음절단위로 합성한다. 한글음성을 무제한으로 합성할 수 있도록 3,040개의 반음절을 준비하고 있다. 얼굴 동영상 합성을 위해, 한글 발음에 필요한 입모양을 자음과 모음으로 나누어, 총 11개의 기본형 패턴으로 정의하고 있다. 입력 한글 텍스트의 음절별로 초성, 중성, 종성에 기본형 입모양을 키프레임으로 할당하여, 이들 사이에 자연스럽게 변해가는 입모양을 중간프레임에서 보간하고 있다. 얼굴 동영상과 음성의 동기를 맞추기 위해, 합성음성의 음절별 지속시간을 초성과 중성사이, 중성과 종성사이로 구간을 나누어서, 동영상의 보간 프레임 수를 산출하고 있다. 음성 합성에 있어서는 3,040개의 반음절을 저장하기 위한 메모리가 필요하다. 그러나, 동영상 합성에 있어서는 1매의 무표정 얼굴을 이용하여 전 프레임을 합성하고 있기 때문에, 컴퓨터 메모리는 거의 필요로 하지 않는다. 이와 같은 방법으로 한글문장을 합성된 음성과 얼굴 동영상으로 낭독하는 시스템을 구축하여 동기를 구현하고 있다.

Abstract

This paper proposes a synchronization method of synthetic facial image sequences and synthetic speech. The LP-PSOLA synthesizes the speech for each demi-syllable. We provide the 3,040 demi-syllables for unlimited synthesis of the Korean speech. For synthesis of the Facial image sequences, the paper defines the total 11 fundermental patterns for the lip shapes of the Korean consonants and vowels. The fundermental lip shapes allow us to pronounce all Korean sentences. Image synthesis method assigns the fundermental lip shapes to the key frames according to the initial, the middle and the final sound of each syllable in Korean input text. The method interpolates the naturally changing lip shapes in inbetween frames. The number of the inbetween frames is estimated from the duration time of each syllable of the synthetic speech. The estimation accomplishes synchronization of the facial image sequences and speech. In speech synthesis, disk memory is required to store 3,040 demi-syllable. In synthesis of the facial image sequences, however, the disk memory is required to store only one image, because all frames are synthesized from the neutral face. Above method realizes synchronization of system which can read the Korean sentences with the synthetic speech and the synthetic facial image sequences.

* 正會員, 明知大學校 電子情報通信工學部
(Myongji University)
** 正會員, 關東大學校 電子通信工學部
(Kwandong University)

※ 본 연구는 과학재단 특정기초(96-0102-15-01-3)의 지원에 의해 이루어 졌음.
接受日字:1998年1月20日, 수정완료일:1998年6月17日

Ⅰ. 서 론

멀리 떨어져 있는 개인간의 대화 또는 여러사람의 회의에 있어서 임장감을 높이기 위해서 가상현실, 가상공간, 가상회의 등에 대한 관심이 고조되고 있다. 실제로는 멀리 떨어져 있는 사람이 가상적으로 같은 공간에서 대화하고 있는 느낌을 주는 가상현실을 위해서는 고품질의 얼굴 동영상과 음성을 합성하는 기술이 필수적이다.

합성음성과 합성얼굴표정을 이용하는 구상은 Parke나 Lippman 등에 의해 산발적으로 제안되었다.[1-2] 그러나, 얼굴의 합성법이 단순한 그래픽기술에 의존하였기 때문에, 현실감이 떨어진 만화와 같은 얼굴영상이 얻어져, 현실감이 요구되는 가상현실에 대한 체계적인 연구로는 연결되지 못했다. 그후, 일본 동경대학의 原島 博교수그룹에 의해 2차원 얼굴사진을 이용하여 3차원 얼굴영상을 합성하는 방법이 제안되어 이 분야의 연구가 활발해 졌다[3-7]. 이 방법은 얼굴사진을 기본으로 하여 현실감이 뛰어난 얼굴영상을 합성할 수 있기 때문에, 다방면에 응용이 가능할 것으로 보여 주목을 받아 왔다. 同 그룹의 森島는 이러한 연구의 일환으로 일본어에 대한 입모양합성법을 검토하였다.[5-6] 나아가서, 金子 등은 일본어문장에 대한 합성얼굴 동영상과 합성음성의 동기를 제어할 수 있는 시스템을 구현하였다.[8] 한편, 필자 등은 얼굴근육의 움직임을 고려하여[10] 근육의 움직임에 따라 변화해가는 다양한 얼굴표정의 합성법을 개발하였다[6-7]. 나아가서, 한국어의 자모음에 대한 입모양을 분석한 후, 자모음에 대한 11종류의 기본형 입모양 패턴을 정의하여, 한글문장을 얼굴 동영상으로 변환할 수 있는 기틀을 마련하였다.[11]

한편, 음성합성은 기계가 인간의 음성을 합성하는 기술로서 인간의 발성모델을 토대로 연구되고 있으며 성도의 전달함수와 성대의 진동특성을 모델링하여 구현되어 왔다[12]-[14]. 이러한 모델링에 의한 합성방식의 대표적인 것으로 LPC 계열의 파라메타 합성방식[15]-[17]이 주류를 이루어 오고 있다. 그러나, 음성파형을 직접 이용하는 PSOLA합성방식[18]-[20]으로부터 개발된 LP-PSOLA 합성방식은 LPC 잔차파형을 pitch-synchronous 하게 분석하므로 운율조절을 위한 처리도 용이하며, LPC 계열의 합성방식보다 명료성, 자연성이 향상된 합성음성을 얻을 수 있다.

본 논문에서는 한국어문장에 대한 얼굴 동영상의 합성, 무제한 음성합성, 이들의 동기구현에 대한 방법을 제안한다. 얼굴동영상의 입모양은 한국어문장의 각 음절을 초성, 중성, 종성으로 분해하여, 각각에 대해 기본형 입모양을 키프레임으로 대응시킨 후, 음절의 지속시간에 따라 키프레임을 보간하며 중간 프레임을 합성한다. 고품질의 무제한 음성합성를 실현하기 위하여 LP-PSOLA 합성방식을 이용하여 반음절 단위로 합성한다. 나아가서, 한국어 음절에 대한 지속시간을 조사하여, 얼굴동영상과 음성의 동기를 제어한다.

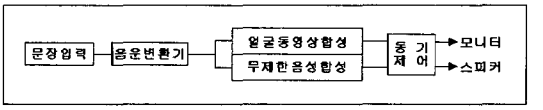


그림 1. 시스템의 구성도
Fig. 1. System Configuration.

Ⅱ. 시스템 개요

인간이 실제 말하는 것과 같이 현실감 있는 대화를 위해서 그림 1과 같은 시스템을 구상하고 있다. 이 시스템에서는 문장 입력으로부터 자연스런 얼굴 동영상과 음성을 합성해서, 모니터와 스피커를 통해서 디스플레이 한다. 먼저, 입력된 문장은 한글의 표준 발음법에 따라 소리나는 대로 음운을 변화해야한다. 음운 변화된 문장에 대하여, 각 음절별로 얼굴 동영상과 음성을 합성한다. 합성된 영상과 음성은 기기의 성능에 따라 동기를 제어해서 디스플레이 하면, 실제 인간이 말하는 것 같은 현실감을 얻을 수 있다. 이하에서 이러한 과정에 대해 구체적으로 기술하기로 한다.

Ⅲ. 음운변환기

한글은 자소(초, 중, 종성)들의 조합으로 생성되는 총 14,364자의 방대한 갯수로 이루어져 있으나, 실용화된 표준코드(KSC5601)에서는 2350자이다. 이 음절들에 대해서 한글의 표준발음법에 따라 얼굴의 입모양과 음성단위를 연관지어 분류해 준다면, 모든 음절에 대한 입모양과 음성을 동시에 표현할 수 있다. 그러나, 한글에서는 음절의 끝소리규칙, 자음동화, 구개음화, 자음축약, 연음규칙, 경음화 등으로 구분되는 음운변화가 음절을 중심으로 초성이나 종성에 위치하는 자음성분에서 발생한다. 이 때문에 입력된 문장에 대해 얼굴

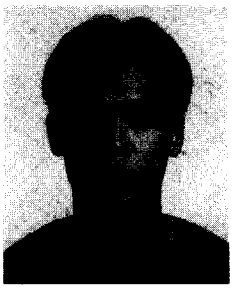
동영상과 음성이 올바르게 합성될 수 있도록 음운변화가 필요하다. 또한, 받침이 없는 음절에 대한 입모양과 음성을 분류하였다가, 분류된 입모양이나 음성단위에 받침이 결합하는 경우를 고려하면 모든 음절의 입모양과 음성을 합성해 낼 수 있다. 특히, 한글의 표준발음법에서는 받침소리로는 'ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅇ'의 7개 자음으로 대표되기 때문에, 한글의 어느 자음이 받침으로 올지라도 이상의 7개의 자음이 받침으로 소리나게 된다. 이상의 음운변화외에도 한국어 모음에서 한글로 표시된 철자와 달리 발음되는 경우로는 /져, 쪄, 쳐/를 /저, 쩌, 처/로 발음하는 경우와 자음을 초성으로 하는 음절의 /의/는 /이/로 발음되는 경우가 있으며, 이중모음에서 /웨/는 단모음/외/와 같이 발음되는 경우 등이 있다. 또한 한국어에서는 모음의 장단을 구별하여 발음한다. 그러나, 단어의 장단은 각 단어마다 독특하여 규칙적으로 처리할 수가 없기 때문에, 장음에 해당하는 한글은 예외사전에 넣어 처리한다. 이런 처리를 수행하기 위해서는 입력된 한글문장의 의미가 분석되어야 하므로 본 연구에서는 고려하지 않고 있다.

Ⅳ. 얼굴 동영상 합성방법

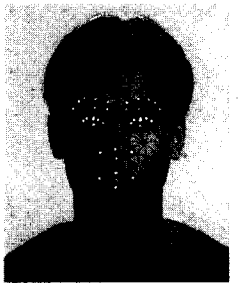
1. 얼굴의 3차원 모델

얼굴 동영상을 자연스럽게 합성하고, 컴퓨터 메모리의 절약을 위해서 얼굴의 3차원 형상모델을 이용하고 있다. 이 모델을 그림 2에 나타낸다. 이 모델은 얼굴의 일반적인 3차원 형상과 얼굴표정을 합성이 가능하도록 눈, 코, 입 등의 얼굴부위를 표현하고 있다. 나아가서, 이 모델을 개인의 무표정 얼굴에 정합하면, 자연스런 표정을 합성할 수 있는 기반이 된다. 정합방법은 문헌 [6]-[7], [9]에 소개하고 있다.

(a) 얼굴의 3차원 형상모델



(b) 무표정 얼굴



(c) 모델의 정합

그림 2. 얼굴의 3차원 모델과 무표정 얼굴에의 정합
Fig. 2. A 3-D Facial Model and Adjustment of the Neutral Face.

2. 한글 자모음에 대한 기본형 입모양

자연스런 얼굴 동영상을 합성하기 위해서는 얼굴표정을 자유자재로 변화시킬 수 있는 방법이 필요하다. 필자 등은 얼굴표정은 얼굴근육의 수축, 이완에 따라 변화한다는 것에 착안하여[10] 얼굴표정을 자유자재로 변화시킬 수 있는 방법을 개발했다.[6-7] 이 방법을이용하여 한글 발음에 필요한 기본형 입모양을 11종류의 패턴으로 분류했다. 한글은 「자음 + 모음」 또는 「자음 + 모음 + 자음」으로 구성되어 있기 때문에, 자음과 모음에 있어서 기본형 입모양을 정한 후, 음운변화된 문장에 기본형 입모양을 대응시키는 방식이다. 자음에 대한 기본형은 입술소리 (ㅁ, ㅂ, ㅃ, ㅍ)와 그 외 소리의 2종류로 분류하고 있고, 모음은ㅏ, ㅓ, ㅗ, ㅜ, ㅡ, ㅣ, ㅔ, ㅐ의 8종류로 분류하고 있으며, 묵음시 다물어진 입모양을 1개 추가하여, 총 11종류의 입모양으로 분류하고 있다. 이들을 표1에 정리한다.

표 1. 한글발음에 필요한 기본형 입모양의 분류
Table 1. Classification of the Fundermental Lip Shapes Required to Korean Pronunciation.

자음	입술소리(ㅁ, ㅂ, ㅍ)와 그외소리(ㄱ, ㄴ, ㄷ, ㄹ 등)
모음	ㅏ, ㅓ, ㅜ, ㅡ, ㅣ, ㅔ, ㅐ
묵음	띄어쓰기 또는 마침표

이 외의 모음의 입모양은 이들 기본형에 준하거나, 기본형의 조합으로 표현이 가능하기 때문에 따로 분류하지 않고 있다(상세한 것은 문헌 [11] 참조). 이들 기본형 입모양을 그림 3에 나타내고 있다.

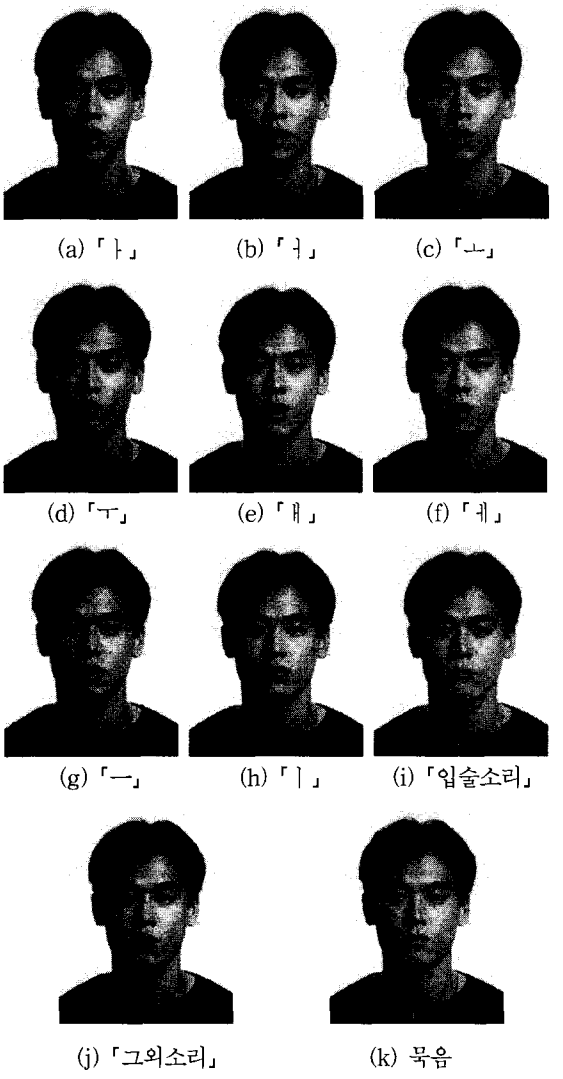


그림 3. 기본형 입모양 패턴
Fig. 3. The Fundermental Lip Shape Patterns.

그림 3의 기본형 입모양 패턴은 그림 2(b)의 무표정 얼굴 영상과 그림 2(c)의 정합된 형상모델을 이용하여 AU(Action Unit)의 종류와 강도에 따라 합성한 것이다. 합성방법은 문헌 [6]-[7],[9]에 소개되어 있다. 필자등의 그룹에서는 AU종류와 강도를 입력하면, 표정이 자동으로 애니메이션 될 수 있는 표정 시뮬레이터를 개발 해 놓고 있다. 이들의 표정변화를 AU로 정리하면 표2와 같다. 괄호안의 숫자는 AU의 강도를 나타낸다. AU는 독립적으로 움직일 수 있는 근육의 수축, 이완에 따라 변화하는 표정의 단위를 코드화한 것이다.[10] 문헌 [10]에 의하면, 얼굴에는 44개의 근육이 있고, 이들 근육의 움직임에 따라 표정이 변화하게 된다. 이들 근육의 독립적인 움직임을 AU로 정의하고 있다. 본 논문에서는 AU의 강도를 0에서 1까지로 정의하고 있으며, 0은 근육이 움직이지 않은 상태를 나타내고, 1은 근육이 최대한으로 움직인 상태를 나타낸다. 얼굴 표정을 모델화 하는데 있어서, 얼굴근육의 움직임이 표정변화의 기본이며, AU는 표정변화를 가장 자연스럽게 할 수 있는 단위이기 때문에 AU를 선택하고 있다.

표 2. 기본형 입모양의 AU조합과 강도
Table 2. The AU Combinations and Intensities of the Fundermental Lip Shapes.

입모양	AU의 조합과 강도
「ㅏ」	AU12(0.3), AU20(0.2), AU26(0.7)
「ㅓ」	AU26(0.3)
「ㅗ」	AU18(0.6), AU26(0.3)
「ㅜ」	AU18(1.0), AU25(0.2)
「ㅐ」	AU12(0.3), AU26(0.3)
「ㅔ」	AU10(0.1), AU20(0.3), AU25(0.2)
「ㅣ」	AU15(0.2), AU20(0.2), AU25(0.4)
「ㅡ」	AU20(0.4), AU25(0.1)
입술소리	AU23(0.5)
그외소리	AU26(0.3)
묵음	무표정

AU명	표정의 변화
AU12	입술양단을 비스듬이 올린다
AU10	윗입술을 올린다
AU15	입술양단을 내린다
AU18	입술을 오므린다
AU20	입술을 좌우로 늘인다
AU25	아래입술을 내린다
AU26	턱을 내리면서 입을 벌린다

3. 동영상 생성을 위한 기본형 입모양 사이의 중간 프레임 생성

입모양의 변화는 자음에서 시작하여 모음에서 최고조에 달하여, 자음에서 마무리 된다고 생각하여, 이들 AU의 강도를 선형적으로 변화시키고 있다. 즉, 한글 텍스트의 음절별로 자음과 모음의 입모양을 키프레임(Key Frame)으로 하여, 키프레임 사이의 중간프레임은 음절의 지속시간에 따라, AU파라미터상에서 선형보간한다. 그림 4는 "갑"음절에 대한 AU의 종류와 강도변화를 나타내고, 그림 5는 그림 4의 AU의 종류와 강도에 따라 변화된 입모양을 프레임 별로 합성한 영상을 나타낸다. 그림 4와 그림 5의 fr은 동영상의 프

레임번호를 나타내고 있다. 이와같이, 그림2의 무표정 얼굴과 정합된 3차원 형상모델을 이용하면, 동영상의 전프레임을 합성할 수 있기 때문에, 동영상을 저장할 필요가 없어서 컴퓨터 메모리는 거의 필요로 하지 않는다. fr#0, fr#4, fr#7이 각각 "갑"의 자모음 "ㄱ", "ㅏ","ㅂ"에 대한 키프레임이고, fr#2, fr#6은 각각 "ㄱ"과 "ㅏ"사이, "ㅏ"와 "ㅂ"사이의 보간된 중간 프레임이다. 이러한 방법을 통해 입모양이 자연스럽게 변화 해 가고 있는 것을 알 수 있다. 여기에서 초성->중성->종성의 변화에 필요한 지속시간에 대해서는 Ⅵ장에서 상술한다.

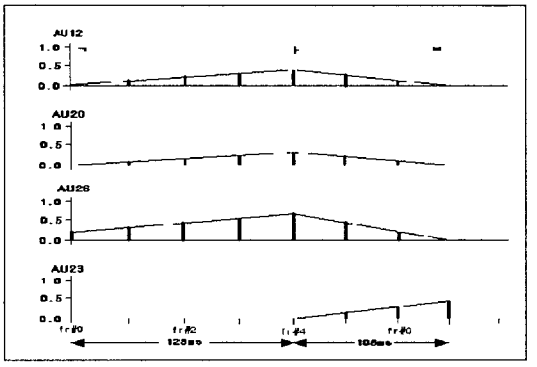
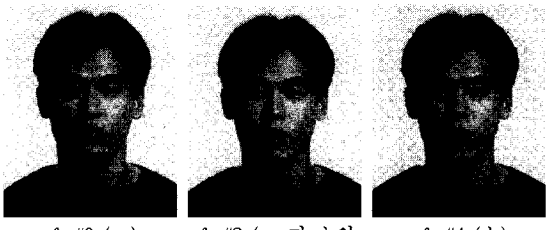


그림 4. "갑"의 음절에 대한 입모양의 AU 파라미터의 변화

Fig. 4. Changes of the AU Parameters of the Lip Shapes for the Korean Syllable "Kap".



fr #0 (ㄱ) fr #2 (ㄱ 과 ㅏ의 중간프레임) fr #4 (ㅏ)

fr #6 (ㅏ 와 ㅂ의 중간프레임) fr #7 (ㅂ)

그림 5. "갑"의 음절에 대한 입모양의 변화

Fig. 5. Changes of the Lip Shapes for the Korean Syllable "Kap".

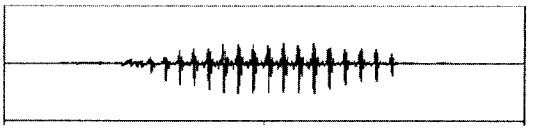


그림 6. 음절 '갑'의 반음절단위 LP-PSOLA 합성음성

Fig. 6. LP-PSOLA Synthetic Speech for Each Demi-Syllable of the Korean Syllable "Kap".

Ⅴ. LP-PSOLA 음성합성

본 연구에서는 무제한 음성합성방법으로 LP-PSOLA 합성방식을 사용한다. 이 방식은 parametric 합성방식으로서 TD-PSOLA 방식보다 다소 음질은 떨어지는 대신 합성단위의 접속(concatination)으로 인한 음성의 불연속성을 피하기 위한 주파수 영역의 처리와 시간영역의 처리가 가능할 뿐만아니라 잔차파형을 pitch-synchronous하게 분석하므로 운율조절을 위한 처리도 용이하며, LPC 계열의 합성방식보다 명료성, 자연성의 향상된 합성음성을 얻을 수 있는 방식이다. 따라서 이 방식에서는 원음성을 먼저 LPC 분석하고 여기서 얻은 잔차파형을 pitch-synchronous하게 분해하며 합성을 위해 분석해 두었던 LPC필터를 사용한다. 그림 7은 LP-PSOLA 방식을 이용하여 데이터베이스를 구축하는 분석과정 및 합성과정을 보이고 있다

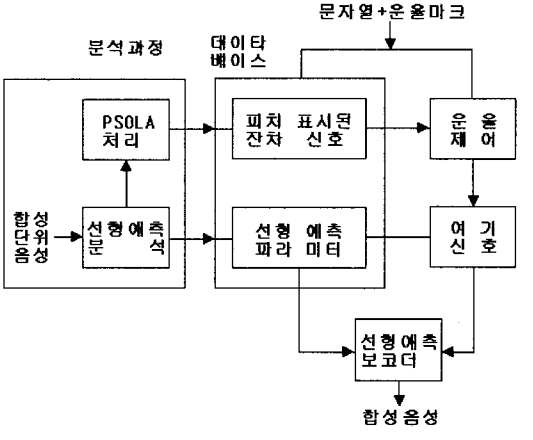


그림 7. LP-PSOLA 방식의 분석과정 및 합성과정

Fig. 7. Analysis and Synthesis Procedures of the LP-PSOLA Method.

Ⅵ. 동기구현을 위한 합성음성의 지속시간

합성될 얼굴동영상의 입모양은 초성, 중성, 종성으로

구성되는 음절을 중심으로 이루어지기 때문에 음절의 중성에 해당하는 모음의 입모양이 대부분을 차지하며 초성과 종성에 의해 다물어질 입모양이 결정된다. 따라서, 합성된 음성과 얼굴동영상의 입모양과 동기가 되도록 구현하기 위하여 각 음절의 지속시간을 측정하였다. 조합형의 한글 3,040자에 대한 합성단위를 반음절로한 경우의 합성음절의 지속시간을 측정한 결과 평균 283.7 ms 와 표준편차 23.5 ms 이었다. 대부분 지속시간이 긴 것은 비음을 종성에 포함하는 음절과 중성에 이중모음을 포함하는 음절이었다. 표3에는 3,040자의 모든 음절과 비음을 포함한 음절 및 이중모음을 포함한 음절에 대한 각각의 평균 및 표준편차를 나타내었다. 이들은 음성을 한음절씩 발음하여 얻은 결과로, 연속음을 발음할 경우는 다소 긴 편이다. 차후, 연속음절에 대한 지속시간을 얻을 예정이다.

표 3. 반음절의 지속시간 비교 [ms]
Table 3. Comparison of Duration Times for Demi-Syllables.

종 류	갯 수	평 균	표준편차
초성이 비음인 음절	480	265	51
초성이 비음이 아닌 음절	2560	287	48
종성이 비음인 음절	1520	321	29
종성이 비음이 아닌 음절	1520	246	36
초성과 종성이 비음인 음절	240	302	33
초성종성이 비음이 아닌 음절	2800	282	50
중성이 단모음인 음절	1368	263	51
중성이 이중모음인 음절	1672	299	41
이중모음과 종성비음인 음절	836	331	27
전음절	3040	284	24

표 3에서 합성음절이 비음을 초성으로 갖는 경우 지속시간이 평균 265 ms로 초성이 비음이 아닌 경우의 평균 287 ms 보다 짧게 나타났다. 그러나, 종성, 즉, 받침이 비음인 경우에는 지속시간이 평균 318 ms로서, 받침이 비음이 아닌 경우의 평균 269 ms 보다 길게 나타났으며, 중성의 모음이 이중모음인 경우에는 지속시간이 평균 299 ms 로서 단모음인 경우의 평균 269 ms 보다 길게 나타났다. 여기서 반음절단위 합성음절의 경우 지속시간이 제일 긴 것은 중성의 모음이 이중모음이고 종성이 비음인 음절로 평균 331 ms 인 것으로 나타나고 있다. 그림6은 음절 '갑'의 합성음성으로 전반부 반음절은 128.0 ms, 후반부 반음절은 108.8 ms 로서 지속시간은 총 236.8 ms 이다.

이러한 합성음절의 각 음절의 지속시간을 이용하여 합성된 얼굴동영상과의 동기를 구현한다. 즉, 합성음성의 각 음절의 지속시간을 전반부와 후반부 반음절로 나누어 초성부터 중성의 최고조까지를 전반부, 그 이후의 중성부터 종성까지를 후반부 반음절과 대응시킨다. 그림 4과 같이 얼굴동영상을 위한 AU의 변화는 합성음절의 초성, 중성, 종성의 순서에 따라, 이들 간의 지속시간을 33ms로 나누어서, 동영상의 중간 프레임을 생성하고 있다.

Ⅶ. 실 험

(1) **실험환경** : 펜티엄 프로 PC에 구축되어 있으나, 일반적인 펜티엄PC에서도 실험이 가능하다.

(2) **음운변환과 자모음분석** : 한글텍스트가 입력되면, 소리나는 대로 음운을 변환한 후, 음절별로 초성, 중성, 종성의 자모음을 분석한다.

(3) **음성합성** : 음운변환된 음절의 초성, 중성, 종성에 따라 반음절단위로 LP-PSOLA를 이용하여 음성을 음절별로 합성한다. 무제한 한글발음을 위해 3,040개의 반음절 DB(데이터베이스)를 준비하고 있다.

(4) **얼굴 동영상 생성** : 음운변환된 음절별로, 초성, 중성, 종성에 따라, 입모양 AU를 선택하여, 얼굴 동영상의 키프레임으로 사용한다. 이들 키프레임의 AU강도를 초성과 중성사이, 중성과 종성사이의 지속시간을 33ms로 나누어 중간 프레임 수를 산출한다. 산출된 프레임 수에 따 라, 프레임별 AU의 종류와 강도를 보간한다.(그림 4~그림5 참조) 키프레임 및 중간 프레임 은 AU의 강도에 따라 무표정 얼굴과 3차원 형상모델을 변형하여 합성하고 있다.

(5) **사용 메모리** : AU에 따라 동영상의 각 프레임을 무표정 얼굴을 변형하여 생성하고 있기 때 문에, 동영상의 각 프레임을 저장하지는 않는다. 다만, 무표정 얼굴의 1개의 영상, 1개의 정합 된 3차원 형상모델과 3,040개의 반음절 DB(데이터 베이스)를 저장할 메모리는 필요하다.

(6) **음성과 동영상의 동기 확인** : 얼굴 동영상의 실시간 합성을 위해서는, 펜티엄 PC의 CPU 속도는 느린 편이다. 그러나, 음성을 합성한 후, 얼굴 동영상을 약 100프레임 정도를 주메모리 에 합성 저장해 놓고, 동시에 디스플레이 하면, 음성과 얼굴 동영상의 동기를 확인할 수 있 다. 이와 같은 이유로, 1회에 입력

할 수 있는 문장의 단어의 종류는 제한 하지 않으나, 단어 의 수는 15~20단어 정도로 제한하고 있다. 단어 수를 무제한으로 할 수 있는 얼굴동영상의 실시간 합성에 대해서는 추후 보고할 예정이다.

Ⅷ. 결 론

본 논문에서는 한국어 문장에 대한 얼굴동영상의 합성, 무제한 음절에 대한 음성합성, 얼굴 동영상과 음성의 동기 구현에 대한 방법을 제안하였다. 무제한 음성합성 방법을 반음절 단위 LP-PSOLA를 이용하고 있고, 한글 발음을 위해 3,040개의 반음절 DB를 준비하고 있다. 얼굴 동영상 합성을 위해, 한글 자모음에 대한 11개의 기본형 패턴을 정의하고 있다. 한글 텍스트가 입력되면, 자모음에 따라 기본형 입모양을 키프레임으로 대응시킨 후, 키프레임 사이를 보간하여 동영상을 합성하고 있다. 동영상의 프레임 수는 음성 지속시간에 따라 산출되고 있다. 이 방법에 따라 한글문장을 합성된 음성과 얼굴동영상으로 낭독하는 시스템을 펜티엄 PC에서 구축하여 동기가 구현됨을 확인하였다. 이 방법을 이용하면, 한국어 문장에 대해, 무제한으로 얼굴 동영상 합성과 음성합성이 가능하기 때문에, 가상현실 휴먼인터페이스 등에서 개인간의 대화, 회의 등에 이용이 가능할 것으로 생각된다.

참 고 문 헌

[1] J. P. Lewis and F. Parke, "Automated Lip_Synch and Speech Synthesis for Charactor Animation", ChI+GI 1987 conf. Proc., pp. 143-147, 1987.

[2] A.Lippman, "Semantic Bandwidth Compression: Speech-maker", Picture Coding Symposium, pp. 29-30, 1981.

[3] 原島 博, "知的映像符號化と知的通信", 日本テレビジョソ學會誌, vol. 42, no. 6, pp. 519-525, 1988

[4] 森島, 岡田, 原島, "知的インタフェースのため表情合成法の一檢討", 日本電子情報通信學會論文誌, vol. J73- D-II, no. 3, pp. 351-359, 1990

[5] S.Morishima, K. Aizawa and H. Harashima, "An Intelligent Facial Image Coding Dirven by Speech and Phones", IEEE ICASSP, 39M8.7, pp. 1795-1798, 1989.

[6] 崔昌石, 原島 博, 武部 幹, "顔の3次元モデルに基づく表情の記述と合成", 日本 電子情報通信學會論文誌, vol. J73-A, no. 7, pp. 1270-1280, 1990

[7] C.S.Choi, K.Aizawa, H. Harashima, T. Takebe, "Analysis and Synthesis of Facial Image Sequences in Model-Based Coding", IEEE Trans. Circuit. Sys. Video Tech., vol. 4, no. 3, pp. 257-275, 1994.

[8] 金子 正秀, 小池, 淳, "テキスト情報に對應した口形形象變化する顔動畵像の合成",日本 電子情報通信學會論文誌D-II, vol. J75, no. 2, pp. 203-215, 1992.

[9] K.Aizawa, H.Harashima, T.Saito, " Image Communication", Elsevier, vol. 1, no. 2, 1989.

[10] P. Ekman and W. V. Friesen, "Facial Action Coding System", Consulting Psychologist Press, 1977.

[11] 이 용동, 최 창석, 최 갑석, "휴먼인터페이스를 위한 한글 음절의 입모양 합성", 통신학회논문지, vol. 19, no. 4, pp. 614-623, 1994.

[12] J.L.Flanagan, Speech analysis, Synthesis and Perception, 2nd. Ed., Springer Verlag, Berlin, 1972.

[13] S.Furui, M.M.Sondhi, Advances in Speech Signal Processing, Marcel Dekker, Inc., New York. Basel.Hong Kong, pp. 741-853, 1992.

[14] Eric Keller,Fundamentals of Speech Synthesis and Speech Recognition, John Woley & Sons, pp. 69-127, 1995.

[15] B.E.Caspers, B.S.Atal, "Changing pitch and duration in LPC synthesized speech using multipulse excitaion,"J.Acoust. Soc. Amer., suppl., vol. 73, no. 1, pp. S5, Spring, 1983.

[16] T.Takagi, T.Umeda, "Voice quality convertsion with correction of spectral distortion by pitch manipulation, and its subjective evaluation," the Transactions of the Institute of Electronics,Information and Communication Engineers A vol.

J73-A, no. 3, pp. 387-396, Mar. 1990.

[17] T.Takagi, "Voice quality conversion," Trans. Televison, vol. 47, no. 12, pp. 28-32, 1993.

[18] F. Chapentier, M. G. Stella, "Diphone SynthesisUsing Overlap-add Technique for Speech Waveforms Concatination," ICASSP 86, pp. 2015-2018, 1986.

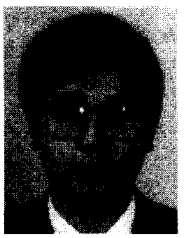
[19] E. Moulines, F. Charpentier, "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones," Speech Communication 9, pp. 453-467, 1990.

[20] Kiyoung Lee, Changseok Choi, "Synchronized Realization of Synthetic Speech and Synthetic Facial Image Sequences", in Proc. ICSP 97, pp. 187-190, 1997.

저 자 소 개

崔 昌 石(正會員)

1954년 7월 15일생. 1978년 2월 홍익대학교 전자공학과 졸업. 1988년 2월 일본 가나자와 대학원 전기정보공학과 석사과정 졸업(공학석사). 1991년 2월 일본 가나자와 대학원 전기정보 공학과 박사과정 졸업(공학박사). 1984년 1월 ~ 1992년 2월 산업기술 정보원 책임 연구원. 1993년 3월 ~ 현재 명지대학교 정보통신공학과 부교수



李 基 榮(正會員)

1961년 5월 7일생. 1984년 2월 명지대학교 전자공학과 졸업. 1986년 2월 명지대학교 전자공학과 석사과정 졸업(공학석사). 1992년 2월 명지대학교 전자공학과 박사과정 졸업(공학박사). 1993년 3월 ~ 현재 관동대학교 전자정보통신공학부 부교수