DOI: 10.3745/KIPSTB.2008.15-B.3.179

가상대학에서 교수자와 학습자간 상호작용을 위한 지식기반형 문자-얼굴동영상 변환 시스템

김 형 근[†] · 박 철 하[††]

요 약

본 논문에서는 가상대학에서 교수자와 학습자간 상호작용을 위한 지식기반형 문자-얼굴동영상 변환(TTFSI : Text to Facial Sequence Image) 시스템에 관해 연구하였다. TTFSI 시스템의 구현을 위해, 한글의 문법적 특징을 기반으로 가상강의에 사용된 자막정보에 립싱크된 얼굴 동영상 합성하기 위하여 자막정보를 음소코드로 변환하는 방법, 음소코드별 입모양의 변형규칙 작성법, 입모양 변형규칙에 의한 얼굴 동영상 합성법을 제안한다. 제안된 방법에서는 한글의 구조분석을 통해 기본 자모의 발음을 나타내는 10개의 대표 입모양과 조음결합에서 나타나는 78개의 혼합 입모양으로 모든 음절의 입모양을 표현하였다. 특히 PC환경에서의 실시간 영상을 합성하기 위해서 매 프레임마다 입모양을 합성하지 않고, DB에서 88개의 해당 입모양을 불러오는 방법을 사용하였다. 제안된 방법의 유용성을 확인하기 위하여 텍스트 정보에 따른 다양한 얼굴 동영상을 합성하였으며, PC환경에서 구현 가능한 TTFSI 시스템을 구축하였다.

키워드 : 문자-얼굴동영상 변환, 음소코드, 입모양의 변형규칙, 얼굴 동영상 합성

Knowledge based Text to Facial Sequence Image System for Interaction of Lecturer and Learner in Cyber Universities

Hyoung-Geun Kim[†] · Chul-Ha Park[††]

ABSTRACT

In this paper, knowledge based text to facial sequence image system for interaction of lecturer and learner in cyber universities is studied. The system is defined by the synthesis of facial sequence image which is synchronized the lip according to the text information based on grammatical characteristic of hangul. For the implementation of the system, the transformation method that the text information is transformed into the phoneme code, the deformation rules of mouse shape which can be changed according to the code of phonemes, and the synthesis method of facial sequence image by using deformation rules of mouse shape are proposed. In the proposed method, all syllables of hangul are represented 10 principal mouse shape and 78 compound mouse shape according to the pronunciation characteristics of the basic consonants and vowels, and the characteristics of the articulation rules, respectively. To synthesize the real time facial sequence image able to realize the PC, the 88 mouth shape stored data base are used without the synthesis of mouse shape in each frame. To verify the validity of the proposed method the various synthesis of facial sequence image transformed from the text information is accomplished, and the system that can be applied the PC is implemented using the proposed method.

Key Words : Text to Facial Sequence Image, Code of Phoneme, Deformation Rule of Mouth Shape, Synthesis of Facial Sequence Image

1. 서 론

최근, 통신 및 네트워크 기술과 멀티미디어 기술의 발전에 힘입어 인터넷은 정치, 경제, 교육, 문화 등 여러 가지 새로운 응용분야를 창출하고 있다. 이중에서 교육 분야에서의 인터넷 응용은 매우 획기적인 발전을 하고 있다. 특히, 웹을 기반으로 한 가상강의는 시간과 공간적인 제약을 받지 않고, 교육 수요자의 필요에 의한 선별적 학습이 가능하다는 장점이 부각되면서 교육의 새로운 형태로 자리 잡아 가고 있다[1-2].

가상강의는 많은 장점에도 불구하고 여러 가지 문제점을 갖고 있다. 즉, 수강자와 교수자 사이의 상호 작용이 원활하지 못하므로 학습 효율이 떨어지고, 면대면(Face to Face)

※ 이 논문은 2006년도 전기 한국방송통신대학교 학술연구비 지원을 받아 작성된 것임
† 정 회 원 : 한국방송통신대학교 컴퓨터과학과 교수
†† 정 회 원 : 대불대학교 경찰학부 부교수
논문접수 : 2007년 12월 27일
수 정 일 : 1차 2008년 1월 22일, 2차 2008년 2월 11일
심사완료 : 2008년 2월 13일

환경에 익숙한 수강자에게 이질적인 학습 환경을 제공하여 학습 집중도를 떨어뜨리는 요인으로 작용하고 있다[3-4]. 이러한 문제점을 해결하기 위해 가상강의에서 환경에 기존의 면대면 환경의 장점만을 부가하기 위한 노력이 이루어지고 있다. 그중 하나의 방법으로 가상강의에서 학습 환경의 이질감을 해소하고 학습자에게 친숙한 인터페이스 환경을 제공하기 위한 휴먼인터페이스를 이용하는 방법이다.

음성과 얼굴표정을 이용한 휴먼 인터페이스의 구상은 NYIT(New York Institute of Technology)의 F. I. Parte, MIT의 A. Lippman, Badler 등에 의해 산발적으로 제안되었다[5]. 그러나 얼굴의 합성법이 단순한 그래픽기술에 의존하였기 때문에, 현실감이 떨어진 만화와 같은 얼굴영상이 얻어져, 현실감이 요구되는 휴먼 인터페이스에 대한 체계적인 연구로는 연결되지 못했다[6-7]. 그 이후, 일본 동경대학의 原島 博을 중심으로 한 연구그룹에 의해 2차원 얼굴사진을 이용하여 3차원 얼굴영상을 분석하고 합성하는 방법이 제안되어 이 분야의 연구가 활발해 졌다. 이 방법은 얼굴사진을 기본으로 하여 현실감이 뛰어난 얼굴영상을 합성할 수 있기 때문에, 다방면에 응용이 가능할 것으로 보여 크게 주목을 받고 있다.

相澤과 原島는 2차원 얼굴사진과 얼굴의 3차원 형상모델을 이용하여 사진속의 얼굴의 3차원 모델을 생성하는 방법을 제안하였다[8]. 또한, 얼굴의 3차원 모델을 이용하여 두부의 3차원 움직임을 얼굴영상으로 합성하는 방법과 얼굴표정의 합성법을 개발하였다[9].

한편, 林島와 原島는 얼굴합성법을 이용한 휴먼 인터페이스의 실현을 목적으로 일본어 및 영어발음에 대한 입모양합성법을 제안하였다[10]. 이 연구에서는 자소단위로 입모양 변형규칙을 정하여, 발음에 일치된 입모양을 동영상으로 합성하고, 그래픽엔진, 트랜스퓨터 등을 이용하여 실시간 얼굴영상 합성도 시도하였다.

그러나 이상의 연구에서는 얼굴근육의 움직임을 고려하지 않았기 때문에 합성된 얼굴표정이 부자연스러울 것으로 예상된다. 또한, 발음에 동기 된 입모양의 합성에서 일본어나 영어에 대해서 자소단위로 입모양의 변형규칙을 정하였다.

최근 국내에서도 얼굴에서 입모양의 분석 및 합성법에 대해 많은 연구가 이루어지고 있다[11]. 이러한 연구는 청각장애자를 위한 구화교육의 일환으로 얼굴 입모양을 합성하는 연구와 애니메이션 영화의 현실감을 증대하기 위해 음성에 동기된 입모양을 합성하는 방법 등이 소개되고 있다. 전자의 경우는 청각 장애자를 위한 구화교육을 위한 목적으로 개발되었기 때문에 정확한 입모양 생성을 위해 한글의 발음 및 발음기관의 구조적 특징들을 고려하여 입모양을 만들었고, 후자의 경우는 정확함보다는 입모양의 현실감과 자연스러움에 무게를 두어 사람의 얼굴의 입주위에 전자태그를 부착하고 말을 하면 3차원 위치스캐너가 대화를 하는 동안 입모양의 위치변화를 실시간으로 받아들여 위치 정보를 애니메이션 객체의 움직임 정보로 사용하는 방법이다.

본 논문에서는 가상대학에서 교수자와 학습자간의 상호작용을 위한 지식기반형 TTFSI 시스템을 설계한다. 가상강의에서 학습자의 학습효율을 최대화하고 학습 환경에 대한 이질감을 최소화하기 위해서는 오프라인 강의에서 나타나는 교수자와 학습자 사이의 상호작용과 면대면 환경을 구축하여 교수자의 강의영상을 직접 볼 수 있게 하는 것이 매우 중요하다.

저자 등은 교수자와 학습자의 상호작용을 증진시키기 위해서 가상대학에서 학습감독권에 대한 연구를 진행하였고, 본 논문은 후속 연구로서 면대면 환경을 구축하고 수강자에게 친밀감을 주기 위해서 교수자의 강의영상을 직접 볼 수 있게 하기 위한 연구이다.

그러나 일반적인 인터넷환경에서 교수자의 강의모습과 전자칠판의 내용을 둘 다 보여주기 위해서는 많은 양의 동영상 데이터를 네트워크를 통해 전송해야 하기 때문에 수강자의 네트워크 속도가 느리거나 컴퓨터의 성능이 높지 않은 경우에는 원만한 강의 수강이 이루어지지 않게 된다. 현재 이루어지는 거의 대부분의 가상강의는 전자 칠판의 내용과 교수자의 모습을 교대로 보내주거나, 전자칠판이나 교수자의 모습 중에 어느 하나만을 선택적으로 보여주고 있다.

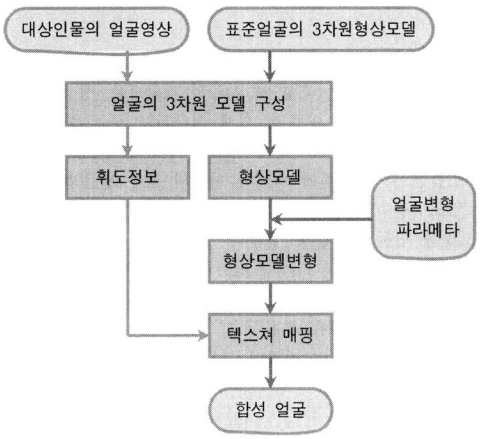
그리고 현재의 가상강의는 음성정보와 영상정보를 기반으로 작성되었기 때문에 정상인의 경우에는 강의를 수강하는데 아무런 문제가 없지만 청각 기능에 장애가 있는 장애자의 경우에는 효과적인 강의 수강이 어렵게 된다. 이러한 장애자를 위한 대책으로 가상강의에 강의 내용을 텍스트 기반의 자막정보와 함께 전송한다면 청각장애자의 경우 음성정보와 동기된 자막정보만으로도 충분히 강의를 수강할 수 있게 된다[12-13].

그러므로 앞에서 제기된 가상강의의 학습 환경의 이질감과 가상강의 전송데이터를 최소화 하고 장애자의 학습권을 보장하기 위해 전자칠판의 강의 내용과 자막정보만을 네트워크를 통해 전송 받고, 교수자의 강의모습은 문자-얼굴동영상 변환(TTFSI: Text to Facial Sequence Image)을 통해서 보여주며, 강의 음성은 문자-음성 변환(TTS:Text to Speech)을 통해서 들려주는 가상강의 수강시스템을 설계한다.

이하 가상강의에서 교수자의 강의모습을 텍스트로 이루어진 자막정보를 기반으로 얼굴근육의 움직임을 이용하여 얼굴영상을 합성하는 지식기반형 TTFSI 시스템의 설계과정을 간략하게 설명한다. 제1장에서는 가상강의의 문제점과 얼굴영상합성법에 대한 연구 동향을 살펴보고 제2장에서는 본 논문에서 사용한 얼굴의 근육운동을 이용한 얼굴합성법을 설명한다. 이 방법은 얼굴근육의 움직임을 고려하고 있기 때문에 입모양 변형규칙을 정하기가 용이하고, 입모양 이외에 표정합성과도 친화성이 좋을 것이라 생각된다. 제3장에서는 한글로 작성된 자막정보에서 입모양 합성을 위한 대표음소를 발음할 때 나타나는 입모양의 특징을 분류하고 제4장에서는 한글 대표음소의 입모양과 혼합 입모양으로 만들어지는 얼굴동영상 합성법을 설명한다. 마지막으로 제5장에 본 연구에서 얻어진 결론을 기술한다.

2. 얼굴근육운동을 이용한 얼굴합성

얼굴근육운동을 이용한 얼굴 합성은 얼굴의 3차원 형상모델을 변형하고 그것에 휘도정보를 부여함으로서 이루어진다. 여기서 어떠한 파라메타에 의해 얼굴형상을 변화시키는가가 중요한 문제가 된다. 얼굴표정은 얼굴근육의 움직임에 의해 변화되는 것이므로, 얼굴 근육의 움직임을 고려하여 얼굴영상을 합성하는 것이 바람직하다. 다시 말하면, 자연스러운 합성영상을 얻기 위해서 얼굴근육의 움직임에 따라 형상모델을 변형하고 변형된 형상정보에 휘도정보를 매핑하여 얼굴영상을 합성한다. 이와 같은 얼굴의 합성과정을 (그림 1)에 나타낸다. 이하, 얼굴 3차원 모델 생성과 근육의 움직임을 고려한 얼굴변형 파라메타 그리고 입에 대한 AU (Action Unit)의 변형규칙에 대해서 기술한다.



(그림 1) 얼굴영상 합성과정

2.1 얼굴의 3차원 모델

얼굴의 일반적인 3차원형상을 표현하기 위해서, 얼굴을 점과 선으로 근사시킨 얼굴의 3차원 형상모델을 준비한다. 이것을 3차원얼굴의 표준모델로 한다. 이것은 약 600개의 꼭지점과 약 700개의 삼각형 패치(patch)로 이루어져 있다. 그리고 여러 가지 두발이 근사 가능한 형상모델을 별도로 준비한다. 이 모델을 대상인물의 정면상에 정합하여 개인얼굴의 3차원형상모델을 얻는다. 이와 같이 개인에 대한 3차원 형상모델을 얻게 되면, 이 모델의 삼각형패치에 대해서 얼굴의 표면을 정의하여, 그 표면의 삼각형마다에 2차원 원영상의 휘도정보를 투영함으로서 얼굴의 3차원 모델이 구성된다.

얼굴의 3차원 모델은 2차원 얼굴영상에서 3차원 구조로 복원하는 작업이므로, 이 모델이 일단 구성되면 표정변화를 용이하게 실현할 수 있을 뿐 아니라, 음성에 동기된 입모양의 합성도 가능하다. 얼굴의 3차원 모델의 구성과정을 (그림 2)에 나타내었다.

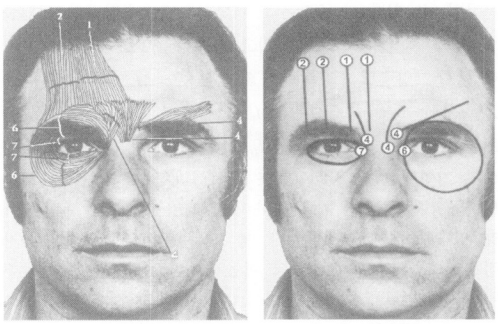
(a) 원영상 (b) 얼굴형상의 정합 (c) 개인얼굴의 3차원모델

(그림 2) 얼굴의 3차원 모델의 구성

2.2 얼굴변형 파라메타

심리학자인 Ekman과 Friesen은 얼굴근육의 위치 및 움직임의 방향을 해부학적으로 고려하고, 그 움직임을 44개의 기본동작으로 분해하였다. 나아가서, 그 기본동작을 이용해서 표정을 기술하였다. 이 시스템을 FACS(Facial Action Coding System)라 하고, 그 기본동작을 AU라고 부른다[14].

얼굴의 여러 가지 표정변화는 얼굴근육의 움직임에 의해 나타나기 때문에 AU의 조합으로 표현 가능하다[15]. 그리고 입부분에 관련하는 AU를 형상모델의 변형파라메타로 사용



(a) Muscular Anatomy (b) Muscular Action

(그림 3) 얼굴 상부의 근육과 AU[14]

〈표 1〉 AU목록

no.	AU 기능	no.	AU 기능
1	눈썹 내측을 올린다	20	입술양단을 옆으로 끈다
2	눈썹 외측을 올린다	23	입술을 강하게 다문다
4	눈썹을 내린다	24	입술을 상하로 누른다
5	윗눈꺼풀을 올린다	25	턱을 내리지않고 아랫입술을 내린다
6	뺨을 올린다		
7	눈꺼풀을 긴장시킨다	26	턱을 내리면서 아랫입술을 내린다
8	입술을 서로 접근시킨다		
9	코에 주름을 잡는다	27	입을 크게 벌린다
10	윗입술을 올린다	28	입술을 빨아들인다
11	콧볼 옆 주름을 깊게 한다	29	아래턱을 내민다
12	입술양단을 끌어 올린다	30	턱을 좌우로 이동시킨다
13	입술양단을 예리하게 끌어 올린다	32	입술을 깨문다
		35	볼을 빨아들인다
14	보조개를 만든다	41	윗 눈꺼풀을 내린다
15	입술양단을 내린다	42	눈을 가늘게 뜬다
16	아랫입술을 내린다	43	눈을 감는다
17	턱을 올린다	44	눈을 작게 뜬다
18	입술을 좁힌다	45	눈을 껌뻑인다
		46	윙크
			안구의 회전

함으로서 음성에 동기한 입모양의 움직임을 자연스럽게 표현할 수 있을 것이다. 이러한 관점에서, 본 논문에서는 얼굴근육의 움직임을 고려하기 위해서 AU를 형상모델의 변형파라메타로 사용하였다. 이 AU의 목록을 <표 1>에 나타낸다.

그리고 얼굴근육의 위치와 움직이는 방향을 (그림 3)에 나타낸다. 그림에서 번호는 AU의 번호이고 위치는 근육이 움직이는 방향을 나타내고 있다. 이 AU를 사용하여 얼굴표정을 합성하기 위해서 각 AU에 대한 변형규칙을 정했다. 이 변형규칙은 얼굴표정 합성뿐만 아니라 음성에 동기된 입모양도 쉽게 합성할 수 있다.

2.3 얼굴하부 AU의 변형규칙

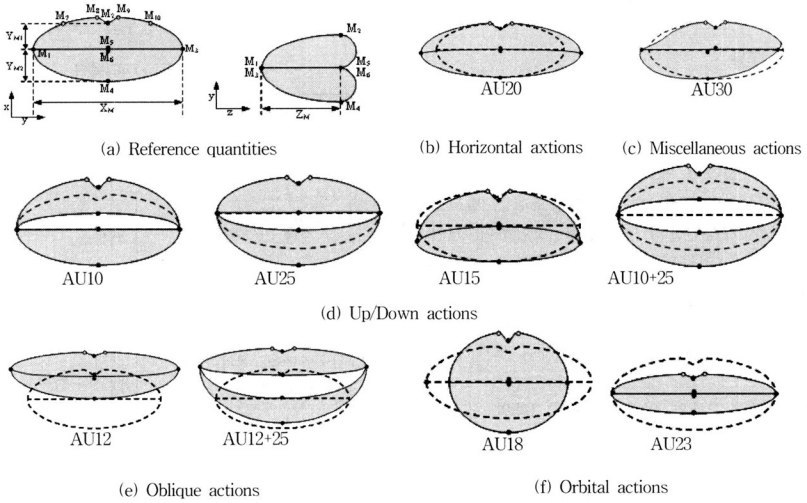
얼굴은 상부와 하부로 나눌 수 있다. 상부얼굴은 눈썹과 눈에로 관계하는 AU가 있고 하부의 얼굴에는 입술 움직임과 턱의 회전을 포함한 많은 AU가 있다. 입부분의 변형은 특징점을 이동하고, 관련하는 근방 꼭지점의 위치는 직선 또는 2차 곡선으로 근사한다. (그림 4)는 입술의 움직임에 관계하는 AU들에 대한 변형규칙의 예를 보인다. 입술은 상/하, 수직, 수평, 경사적인 움직임, 궤도적인 움직임, 5가지로 분류할 수 있다. 기준량은 수평방향 X_M, 수직방향 윗입술 Y_M과 아래 입술 Y_{M2}, 그리고 깊이방향으로 Z_M이 사용된다. 윗 입술의 좌측 좌표는 $M_1-M_8-M_9$, 우측은 $M_3-M_9-M_8$, 중심선은 $M_1-M_5-M_3$, $M_1-M_6-M_3$을 포물선 근사한다. 점 M_7-M_{10}의 움직임은 이동점 M_1-M_3의 움직임에 따라서 이동하도록 하였다. 이들의 AU에 대한 변형규칙은 눈썹과 눈에 대한 것들과 유사하다. 턱은 귀 아래 점을 중심으로 회전한다. 뺨 부분에서 형상모델의 꼭지점은 눈, 입술 그리고 턱의 이동점과 조화시키기 위해서 3차원 어핀변환을 이용한다.

3. 한글과 입모양의 분류

일반적으로 한글에 있어서 모든 음절에 대한 입모양을 정할 필요가 없다고 생각한다. 그것은 음절에는 음이 달라도 같은 형태의 입모양을 나타내는 패턴이 많이 존재하기 때문이다. 또한, 음절 발성 시 입모양은 복수의 입모양의 연결이다. 그래서 인간이 음절을 발성할 때 입모양이 이루어지기까지 각 프레임으로 연결되어 있는데 입모양 패턴이 이루어지기 위한 중간과정은 굳이 입모양 패턴으로 분류할 필요가 없다. 따라서 모든 음절에 대해서 입모양을 정할 것이 아니라 대표적인 음소에 대해서만 입모양을 정하여, 이 입모양의 결합으로 모든 음절의 입모양을 나타내는 것을 생각한다. 이렇게 함으로써, 비교적 적은 수의 입모양의 변형규칙을 정하여도 얼굴동영상의 합성이 가능하다. 본 장에서는 한글의 구조와 모든 음절을 표현할 수 있는 대표음소의 입모양을 분류하고자 한다.

3.1 음절에 따른 입모양 분류[16-17]

한글은 자소(초성, 중성, 종성)들의 조합으로, 생성되는 문자는 총 14,364자에 이르며, 현재 실용되는 한글 한자 표준코드(KSC5601)에서의 한글은 2350자이다. 그러므로 2350자를 포함한 모든 음절에 대해서 한글의 표준 발음법에 준하여 입모양을 분류함으로서 모든 음절에서 표현될 수 있는 입모양을 만들 수 있다. 한글의 구조형태에서 받침이 있는 음절과 없는 음절로 나눌 수 있는데 여기에서는 이 두 가지로 나누어서 각각 입모양을 분류하였다. 먼저, 받침이 없는 음절에 따른 입모양을 분류하고, 이 분류된 입모양에 받침이 결합하는 경우를 고려하면 모든 음절을 분류를 하는 것이 된다.



(a) Reference quantities (b) Horizontal axtions (c) Miscellaneous actions

(d) Up/Down actions

(e) Oblique actions (f) Orbital actions

(그림 4) 입에 대한 AU의 변형규칙

3.1.1 초성 자음에 의한 분류

음절에 따른 입모양패턴은 대부분 모음에 의존하지만 순음「ㅁ, ㅂ, ㅃ, ㅍ」과 조합되는 경우와 설단음「ㄷ, ㅅ, ㅌ, ㄸ, ㅆ, ㄴ, ㄹ」과 조합된 경우에는 발음초기의 입모양이 다르다. 예를 들면 자음이 모음「ㅏ」와 결합할 경우에 「바」 등과 같이 순음과 결합할 때는 입술이 닫쳐진 모양에서 시작되고, 「다」 등과 같이 설단음과 결합할 때는 입술이 약간 열리고 혀끝으로 이 사이를 막은 형태로 시작한다. 그리고 나머지 경구개음, 연구개음, 성문음은 입모양 형성에 영향을 미치지 않는다. 따라서 음절에서 자음의 영향을 받아 표현되는 입모양은 순음과 설단음에 의한 두개의 입모양이 있다.

3.1.2 종성 자음에 의한 분류

한글의 표준발음법에서 받침소리로는「ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅇ」의 7개 의 대표자음으로만 발음 한다. 즉, 한글의 구성형태에서 어느 자음이 받침으로 올지라도 발음은 7가지로만 발음한다. 따라서 받침이 있는 경우의 입 모양을 분류하기 위한 음절은 7개의 자음이 받침으로 오는 경우만 고려하면 된다. 이 경우 또한 모음에 지배되어 입모양은 기본형 입모양으로 나타나고, 최종 생성되는 입모양은 이미 분류한 순음「ㅁ, ㅂ」과 설단음「ㄴ, ㄷ, ㄹ」의 입모양이 나타난다. 그리고 연구개음「ㄱ, ㅇ」이 종성 자음에 사용될 때는 입모양의 기본 형태는 직전 모음에 의해 결정되지만 입을 닫고 끝내려는 종성 자음의 특성 때문에 기본형 모음의 형태에서 50%정도 입을 닫은 형태를 보여준다.

3.1.3 모음에 의한 분류

자음은 발음초기에는 영향을 미치지만 입모양패턴을 지배하는 것은 모음이므로 본 절에서는 모음에 대한 분류를 하고자 한다. 여기에서 모음은 중모음과 단모음으로 구분할 수 있지만 입모양패턴의 분류는 입모양의 가로와 세로의 벌어진 형태에 의한 분류이므로 이것은 동영상 합성 시 고려하기로 한다. 먼저, 분류의 편의를 위해 다음의 말을 정의한다.

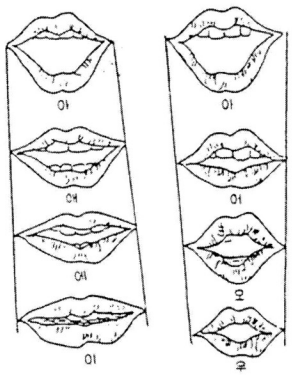
- 기본형모음 : 발음에 있어서 기본이 되는 입모양패턴을 형성하는 모음
- 조합형모음 : 두개의 기본형모음의 결합으로 이루어진 모음

이 정의에 따라 모음을 분류한 것을 <표 2>에 나타낸다. 단, 여기에서 괄호안의 모음「ㅑ, ㅕ, ㅛ」은 길게 발음해 보면 그 입모양 자체가 각각「ㅏ, ㅓ, ㅗ」의 입모양패턴으로 나타나지만 초기에 나타나는 입모양이「ㅣ」의 입모양에 영향을 받는다. 그러므로 이들 모음들은 초기 입모양에 모음「ㅣ」의 형태를 반영하여 나타난다.

「ㅠ」는「ㅜ」와 같은 입모양을 나타내므로 따로 분류하지 않았다. 그리고 각 모음에 대한 입모양패턴을 살펴보면, 기본형 모음은 (그림 5)와 같이 입술의 가로와 세로로 벌어진 정도에 따라 각각 표현된다. 이에 반해 조합형모음의 입모양패턴은 기본형 모음의 입모양패턴이 조합한 것으로 나타난다. 예를 들면,「ㅘ」의 입모양은「ㅗ」와「ㅏ」의 입모양의 결합이다.

〈표 2〉 입모양에 따른 모음분류

모음	기본형모음	ㅏ(ㅑ), ㅓ(ㅕ), ㅗ(ㅛ), ㅜ(ㅠ), ㅡ, ㅣ, ㅔ, ㅐ
	조합형모음	ㅘ, ㅙ, ㅚ, ㅞ, ㅝ, ㅟ, ㅢ



(그림 5) 모음의 입모양 비교

3.2 자음의 입모양 특징

자음은 모음의 음소들과는 달리 각각의 음소들마다 독특한 입모양을 갖지 않고 아주 짧은 동안 발음된다. 그러나 자음의 음소들은 조음위치에 따라 입모양의 분류가 가능하며, 이러한 입모양은 초성자음의 시작 입모양이나 종성자음의 끝 입모양을 결정하는 중요한 요소이다.

조음위치에 따른 분류된 순음, 설단음, 경구개음, 연구개음, 성문음 중에서 순음은 연속된 단어 중에서 초성에 사용할 때는 반드시 입을 다문 상태에서 출발하고, 종성에 사용할 때는 입을 닫고 끝내는 특징을 갖고 있다.

설단음은 입을 약간 벌리고 치아가 약간 노출되며, 이 치아 사이를 혀가 막고 있는 형태를 보이는데 순음과 마찬가지로 초성에서는 시작부분에서, 종성에서는 끝나는 입모양을 결정한다.

구개음, 연구개음, 성문음에 포함된 자음의 음소들은 발음할 때 입모양 변화에 유의할만한 변화를 보이지 않기 때문에 입모양 특징을 규정하지 않는다.

그러므로 자음의 입모양 특징은 순음과 설단음의 2가지 입모양 특징만 나타나게 된다.

3.3 모음의 입모양 특징

자음은 발음의 초기(초성+중성)나 끝(초성+중성+종성)에는 영향을 미치지만 발음되는 시간이 아주 짧을 뿐만 아니라 각각의 음소별로 구분된 입모양을 나타내지는 않는다. 그러므로 입모양은 결정짓는 주요 요소는 모음이라 할 수 있다.

모음은 조음위치의 변화여부에 따라 단모음과 중모음으로 구분할 수 있고 중모음은 단모음의 조합으로 나타낼 수 있

기 때문에 따로 입모양 특징을 정의하지 않고 단모음에 대해서만 입모양 특징을 정의한다. 그러나 단모음 중에 4개 모음 「ㅑ, ㅕ, ㅛ, ㅠ」은 길게 발음해 보면 그 입모양 자체가 각각 「ㅏ, ㅓ, ㅗ, ㅜ」의 입모양패턴으로 나타나므로 굳이 분류하지 않았다.

다음의 <표 3>은 20인의 학생들이 모음 음소를 발음할 때 나타나는 턱, 입술, 이, 혀의 모습을 사진 촬영하여 입모양에 영향을 주는 특징을 기본형 모음 각각에 대해서 분류한 것으로 숫자가 작을수록 해당 움직임이 크다는 것을 의미하고, x는 해당 움직임이 거의 없다는 것이다. 이상과 같이 모음에 대한 입모양의 특징을 살펴본 결과, 기본형 모음에서 나타나는 8개의 입모양패턴으로 모든 모음을 표현할 수 있음을 알 수 있다.

〈표 3〉 모음을 발음할 때 나타나는 특징

특징 \ 모음	ㅏ	ㅓ	ㅗ	ㅜ	ㅐ	ㅔ	ㅣ	ㅡ
턱의 움직임(AU17, AU25, AU26)	1	5	2	x	3	4	x	x
입의 크기(AU18, AU20)	4	5	6	7	3	2	1	1
윗 이의 노출정도	6	5	x	x	3	4	1	2
아랫 이의 노출정도(AU16)	1	2	x	x	5	6	3	4
윗입술을 올림(AU10)	2	1	x	x	x	x	x	2
혀의 노출정도	2	2	x	x	1	1	x	x

4. 텍스트에 립싱크된 얼굴동영상 합성

4.1 텍스트를 음소코드로 변환

가상대학에서 사용하는 강의에 대한 자막정보는 텍스트정보와 음성정보와의 동기를 위해서 시간정보를 중심으로 HTML 태그를 변형하여 사용하였다.

이러한 자막정보의 텍스트에 립싱크 된 얼굴동영상을 합성하기 위한 음소코드를 생성하기 위해 다음과 같은 방법을 사용한다.

① 텍스트에서 문장기호(쉼표, 마침표)를 시간정보로 변환하고, 각각의 단어들을 음절로 분할한다.
② 완성형으로 작성된 음절을 자소분할 프로그램에 의해 음소단위로 분할한다.
③ 자음의 발음법칙을 적용하고 종성 자음을 7개의 대표자음으로 변환한다.
④ 입모양에 영향을 주는 자음을 2개의 그룹(순음:A, 설단음:B)으로 코드화하고, 입모양 생성에 영향을 주지 않는 자음은 모두 삭제한다.
⑤ 중모음을 발음할 때 나타나는 입모양은 중모음을 구성하는 단모음 입모양이 연속해서 나타나는 경우가 거의 대부분이기 때문에 중모음을 단모음으로 분할한다.
⑥ 단모음중 입모양의 구분이 거의 어려운 모음을 8개의 대표모음으로 변환한다.

이러한 과정을 통해 텍스트에 립싱크 된 얼굴동영상 합성을 위한 음소코드로 만들어진다. 이것을 (그림 6)에 나타낸다.

(그림 6) 음소코드의 생성과정

4.2 음소코드별 입모양의 변형규칙

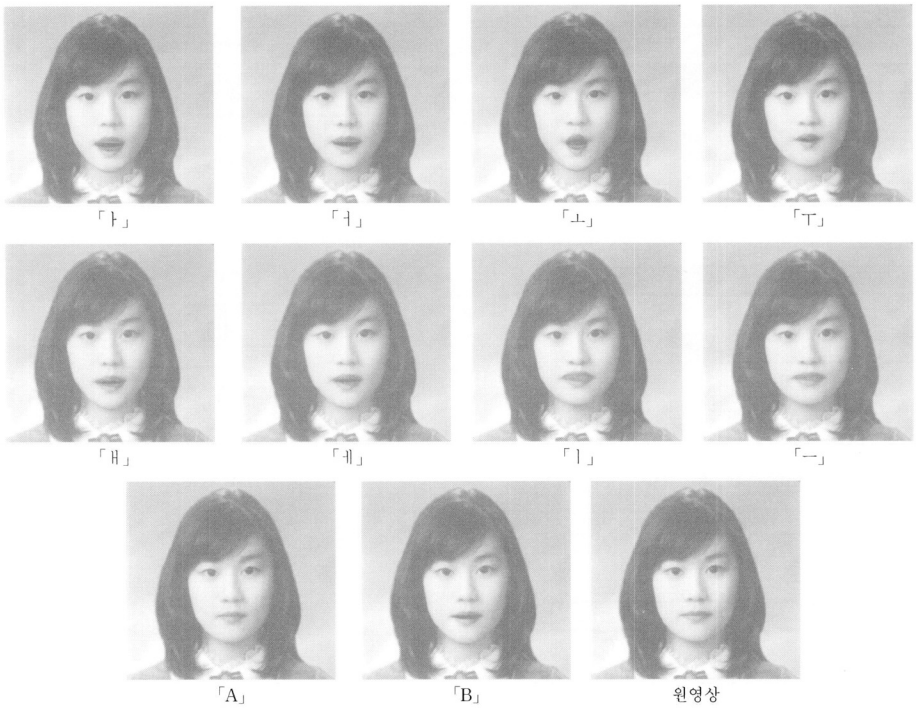
음소에 따른 입모양 합성시스템을 실현하기 위해 음소에 따른 입모양을 분류하였다. 그 결과 모든 음소의 입모양은 8개의 모음 입모양과 2개의 자음 입모양으로 표현할 수 있었다. 이처럼 분류된 10개의 입모양에 대해 (그림 5)의 입모양의 표준패턴과 같은 모양이 되도록 AU를 선택하고 AU의 강도를 부여하여 음소코드별 입모양의 변형규칙을 결정하였다. 이 결과를 <표 4>에 나타낸다.

이와 같이 얼굴근육의 움직임을 고려하여 입모양 변형규칙을 정하면 얼굴영상의 크기 및 개인차가 있어도 같은 규

〈표 4〉 음소코드별 입모양의 변형규칙

입모양	AU no.	AU의 강도
「ㅏ」	AU15	0.3
	AU26	0.5
	AU27	0.1
「ㅓ」	AU10	0.1
	AU15	0.2
	AU26	0.3
	AU29	0.2
「ㅗ」	AU10	0.2
	AU15	0.3
	AU18	0.6
	AU24	0.5
	AU26	0.6
「ㅜ」	AU10	0.1
	AU15	0.1
	AU18	0.6
	AU26	0.1

입모양	AU no.	AU의 강도
「ㅐ」	AU15	0.2
	AU26	0.4
「ㅔ」	AU10	0.1
	AU15	0.1
	AU26	0.3
「ㅣ」	AU10	0.2
	AU20	0.3
	AU25	0.2
「ㅡ」	AU10	0.3
	AU20	0.2
	AU25	0.3
A (순음)	AU18	0.3
B (설단음)	AU26	0.2



(그림 7) 음소코드별 입모양 합성영상

칙을 적용할 수 있다. <표 4>의 변형규칙에 따라 얻어진 음소코드별 입모양 합성영상을 (그림 7)에 나타낸다.

4.3 입모양 변형규칙에 의한 얼굴 동영상 합성

자연스러운 얼굴동영상 합성을 위해서는 글자를 발음하는 시간이 사람들이 일상적으로 발음하는 시간과 일치해야 한다. 사람들이 일상적인 발음시간을 측정하기 위해서 KBS 뉴스화면을 분석한 결과 가장 정확한 발음을 한다고 생각되는 뉴스진행자의 발음 시간은 하나의 글자를 발음하는데 평균 0.15초의 시간을 소비했다. 그러므로 합성된 립싱크 얼굴영상에서도 하나의 글자를 발음하는데 0.15초의 시간이 주어져야 하고 이것은 동영상을 초당 30프레임으로 제작했을 때 5프레임에 해당한다.

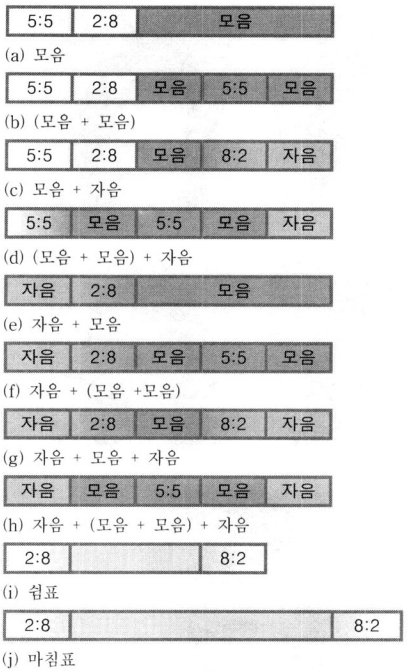
또한 자음과 모음으로 구성된 단어의 음성파형을 분석해 보면 자음영역과 모음영역의 시간 비율이 3:7 정도로 모음영역의 발음 시간이 상대적으로 크다는 것을 알았다. 이러한 사실을 근거로 하나의 글자를 5프레임으로 구성하고 한글의 4가지 구성형태에 대해 립싱크 동영상 합성규칙을 만들었다. 이것을 (그림 8)에 나타낸다.

(그림 8)의 (a), (c), (e), (g) 모음영역에 단모음이 사용된 경우로 (a)는 모음만으로 이루어진 경우로 모든 프레임이 모음의 영향을 받으며 앞의 2프레임은 이전 글자의 마지막 키 프레임과 모음의 키 프레임을 5:5와 2:8의 비율로 반영하여 합성한 프레임이다. (c)의 경우는 모음 + 자음의 경우로 종성에 입술소리나, 혀끝소리의 자음이 오는 경우이다. 이때는 앞의 3프레임은 (a)의 경우와 같지만 4번 프레임은 모음과 자음의 반영비율을 8:2로 혼합한 프레임이고 5번 프레임은 자음의 프레임이다. (e)의 경우는 자음 + 모음의 구조로 1번 프레임에 자음의 키 프레임을 2번에 혼합프레임, 3~5번 프레임은 모음의 키 프레임이 들어간다. (g)의 경우는 자음 + 모음+ 자음의 경우로 3개의 키 프레임과 2개의 혼합프레임으로 구성된다.

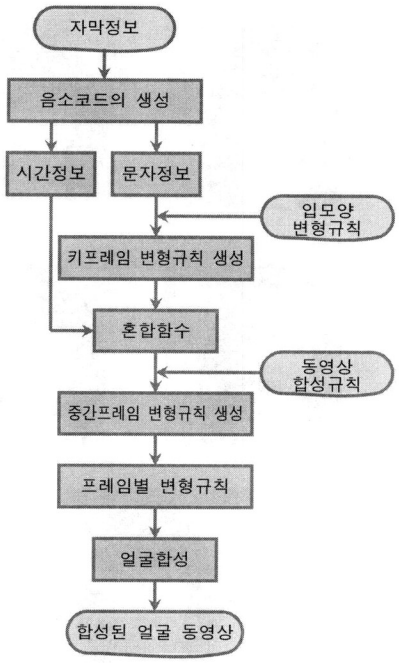
(그림 8)의 (b), (d), (f), (h)는 모음영역에 복모음이 사용되는 경우로 본 논문에서는 복모음의 입모양을 따로 정의하지 않고 복모음을 구성하는 2개의 단모음을 연속해서 발음하는 경우이다.

(그림 8)의 (i), (j)는 쉼표와, 마침표가 나타났을 때 동영상 합성규칙으로 시작과 끝 프레임에 이번 음소와 다음 음소의 입모양을 20% 반영하였고 나머지 구간의 프레임에는 원 영상을 내보낸다. 그리고 자막정보에서 주어진 시간 지연태그는 하나의 프레임을 0.03초로 계산하여 주어진 시간만큼 원 영상을 내보내게 된다.

가정용 PC환경에서 립싱크 얼굴 동영상을 초당 30프레임으로 합성하는 데는 많은 계산량으로 인한 문제가 나타날 수 있다. 더욱이 가상대학 수강 시 립싱크 얼굴동영상 제작뿐만 아니라 자막정보의 처리 및 음성정보처리와 강의를 구



(그림 8) 한글의 구성형태별 립싱크 동영상 합성규칙



(그림 9) 동영상 합성과정

성하는 화면 정보까지 처리해야만 한다.

그러므로 PC의 성능이 아주 높다 하더라도 이러한 많은 일을 동시에 처리하기에는 무리가 따른다. 따라서 본 논문에서는 PC환경에서 구현 가능한 실시간 립싱크 영상의 합성을 위하여 얼굴영상에서 얼굴 입모양과는 상관이 없는 배경부분과 얼굴 상부를 분리하고, 얼굴 하부에 대해서만 립싱크 영상을 작성하는 방법을 사용 하였으며, 88개의 립싱크된 얼굴 하부 영상을 프로그램 환경설정 시에 미리 작성하여 DB화하여 보관한 후에 각각의 프레임에서는 이전프레임의 배경과 얼굴 상부 영상에 DB에서 가져온 얼굴 입모양 영상을 합하여 화면에 내보내기 때문에 PC의 CPU의 부하를 거의 주지 않은 상태에서 초당 30프레임의 동영상 합성이 가능하다.

실시간 제작을 위하여 DB에 보관된 얼굴 하부의 입모양 영상은 대표음소의 입모양 10개(대표자음 2개, 대표모음 8개) 와 자음과 모음의 혼합 입모양 32개(자음 2:8 모음 ⇒ 16개, 자음 5:5 모음 ⇒ 16개), 모음과 모음의 혼합 입모양 28개(모음 5:5 모음 ⇒ 20개), 원영상과 모음의 혼합입모양 16개(원영상 2:8 모음 ⇒ 8개, 원영상 5:5 모음 ⇒ 8개), 원영상과 자음의 혼합입모양 2개(자음 2:8 원영상 ⇒ 2개) 총 88개이다. (그림 9)에 동영상 합성과정을 나타내었다.

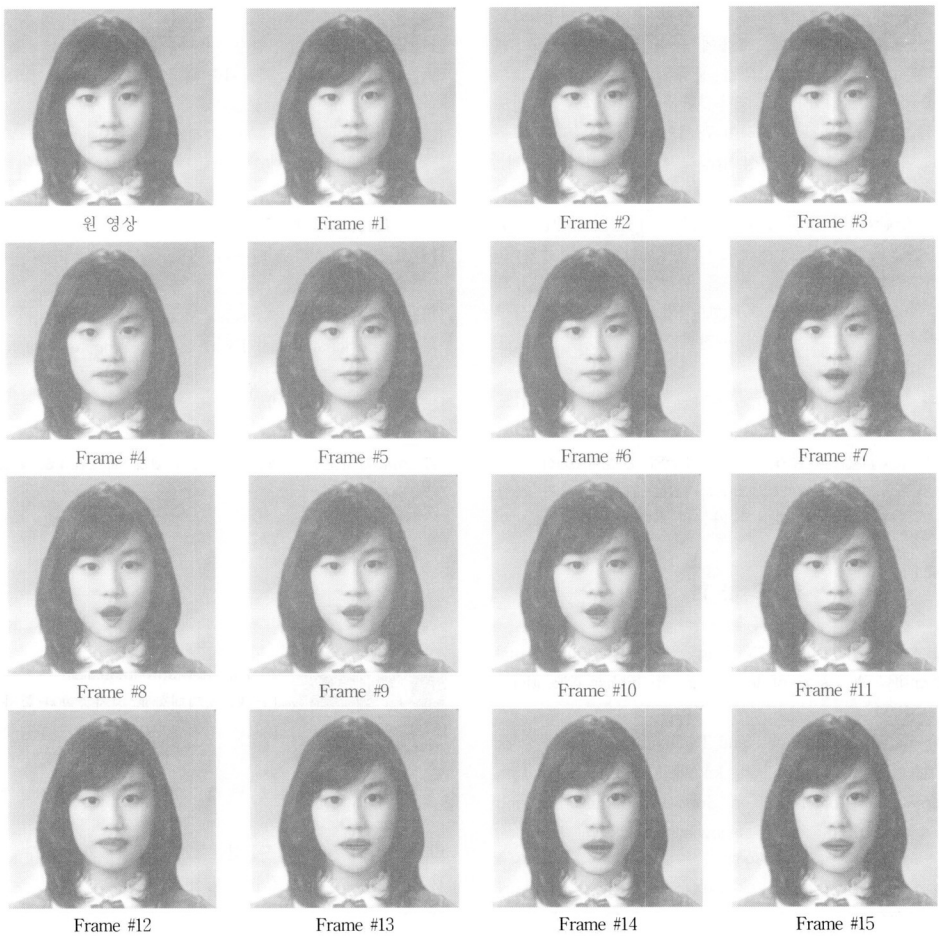
립싱크된 얼굴 동영상 합성의 실효성을 입증하기 위해서 "입모양"이라는 단어를 발음하는 동영상을 제작하였다. 합성된 동영상은 3개의 음절로 구성되었기 때문에 각 음절 당 5프레임씩, 총 15프레임으로 합성된다. <표 5>에 음소코드별 입모양 변형규칙과 구성형태별 동영상 합성규칙에 의해 만들어진 AU의 조합과 강도를 프레임별로 나타내었다.

(그림 10)에는 <표 5>에서 만들어진 프레임별 입모양의 AU조합과 강도에 의해 합성된 얼굴 동영상을 나타내고 있다. 여기에서 치아의 모양은 윗니의 경우는 위치를 고정하고, 아랫니는 하악골의 움직임에 따라 같이 움직이게 하였으며 입술의 움직임에 따라 노출정도가 결정된다.

얼굴 합성의 고속화를 위해 얼굴 하부만을 합성하여 배경과 결합하면 합성된 경계 부분에 눈에 띄는 불연속부분이

〈표 5〉 단어 「입모양」의 동영상 생성을 위한 AU조합과 강도

AU Num. / Frame Num.	AU10	AU15	AU18	AU20	AU24	AU25	AU26	AU27	합성 규칙
#0	0.0			0.0		0.0			시작
#1	0.1			0.15		0.1			5:5
#2	0.16			0.24		0.16			2:8
#3	0.2			0.3		0.2			ㅣ
#4	0.16		0.06	0.24		0.16			8:2
#5			0.3						ㅁ
#6			0.3						ㅁ
#7	0.16	0.24	0.54		0.4		0.48		2:8
#8	0.2	0.3	0.6		0.5		0.6		ㅗ
#9	0.2	0.3	0.6		0.5		0.6		ㅗ
#10	0.2	0.3	0.6		0.5		0.6		ㅗ
#11	0.2	0.15	0.3	0.15	0.25	0.1	0.3		5:5
#12	0.2			0.3		0.2			ㅣ
#13	0.1	0.15		0.15		0.1	0.25	0.05	5:5
#14		0.3					0.5	0.1	ㅏ
#15		0.24					0.4	0.08	8:2



(그림 10) 단어 「입모양」의 동영상 합성의 예

발생하여 자연스러운 동영상 합성을 저해한다. 그러므로 합성된 얼굴하부 영상의 외곽부분에 완만한 경사를 갖는 흐림처리를 하였고, 입을 벌림으로서 나타나는 영역의 경계에 대해서도 같은 방법을 사용하였다.

이러한 방법은 고품질의 자연스러운 립싱크 영상의 제작을 가능하게 하였지만 얼굴에 표정변화나, 안구의 움직임, 머리 움직임, 몸통의 움직임 등이 배제되었기 때문에 부자연스러움을 유발한다. 이는 현재의 PC환경에서 립싱크 영상의 합성을 구현하기 위한 고육지책이지만, PC환경이 좀 더 발전한다면 입모양만이 아니라, 표정, 두부나 몸통의 움직임 등도 얼굴영상 합성에 반영하여야 할 것이다.

본 논문에서 제안한 방법은 얼굴을 3차원으로 취급하는 기술이기 때문에 얼굴영상을 DB에서 불러오지 않고 매 프레임을 직접 합성하는 방법을 쓴다면 기술적인 수정 없이 앞에 언급한 문제점을 모두 해결할 수 있다.

5. 결 론

가상대학에서 교수자와 학습자간 상호작용을 위한 지식기반형 TTFSI 시스템을 구현하기 위해, 한글 발음법칙을 분석하여 입모양에 영향을 미치는 음소들을 추출하고 이들 음소의 발음에 따라 변형된 얼굴영상의 합성과 음소별 얼굴영상과 한글 구성의 형태를 이용하여 립싱크 얼굴 동영상을 합성하는 방법을 연구하였다.

한글의 문법적 특징을 기반으로 가상강의에 사용된 자막정보를 얼굴 입모양 합성을 위한 음소코드로 변환하는 방법을 제안하였다.

한글의 모든 음절을 대상으로 입모양의 가로와 세로로 벌어진 형상을 토대로 음소를 발음할 때 나타나는 입모양 특징을 정의하고 음소코드별 입모양 변형규칙을 작성하여 음소코드별 입모양 얼굴영상을 합성하는 방법을 제안하였다.

음소코드별 입모양은 음절에서 자음이 입모양에 미치는

영향과 모음이 미치는 영향을 고려하여, 자음에 의한 입모양으로 순음과 설단음 2개, 모음에 의한 입모양으로는 8개로서, 모두 10개의 대표 입모양을 분류하였다.

음소코드의 입모양 사이에 조음결합에 의해 나타나는 입모양을 한글 구성형태별로 분류하여 78개의 혼합 입모양 패턴을 작성하고 이를 바탕으로 동영상 합성규칙을 정의하고 하였다.

PC환경에서 초당 30프레임의 실시간 영상을 합성하기 위해서 매 프레임마다 입모양 패턴을 합성하지 않고, 대표 입모양 10개와 혼합 입모양 78개를 DB에 보관하고 각각의 프레임에서는 DB에서 해당 입모양을 불러오는 방법을 사용하였다. 얼굴근육의 움직임을 고려하여 입모양 패턴을 합성한 결과, 얻어진 입모양은 실제의 사진과 같이 현실감 있게 표현되는 것을 확인하였다.

이상의 결과로부터, 음절이 주어지면 그 음절에 따른 입모양과 동시에 표정이 변화된 현실감이 넘치는 TTFSI를 합성할 수 있고, TTS 시스템과 동기를 맞추어 통합하면 저 비트 전송이 가능한 텍스트 정보만으로도 음성과 동영상이 지원되는 휴먼 인터페이스의 실현이 가능하리라고 생각된다.

참 고 문 헌

[1] W.Baker, A.Gloster., "Moving towards the virtual university; A vision of technology in higher education", 17(2), 1994.

[2] A.Yoshitaka, M.Hirakawa, and T.Ichkawa, "A frame work for query processing utilizing knowledge," Proceeding 15th International Conference on Software Engineering and Knowledge Engineering, Knowledge System Institute, Skokie, Illinois, pp.1-10, 1993.

[3] 조은순, "최상의 학습 성과를 위한 e-러닝의 활용", 한국능률협회, 2002.

[4] 송상호, "동기적으로 적응적인 인터넷 기반수업 서례방안의 고찰", 교육공학연구, 16(2), pp.37-57, 2000.

[5] F.I.Parke, "A Parameterized Model for Facial Animation", IEEE Computer Graphics&Applications, 2, 9, pp.61-68, Nov., 1982.

[6] 김남수, "휴먼 인터페이스 기술", Telecommunications Review 제5권, 1호 pp.228-239, 2000.

[7] 정동춘, 이상용, "전문가 시스템을 위한 휴먼인터페이스: 가상현실", 공주대학교 생산기술연구소 논문집, 제6권, pp.54-61, 1998.

[8] 相澤, 原島 博 외 2인, "知的畵像符號化のための顔3次元モデルの構成について", 日本畵像符號化シンポジウム(PCSJ), pp.57-58, 1986.

[9] 林島, 原島 博 외 2인, "知的インタフェースのための顔の表情合成法の一檢村", 日本 電子情報通信學會論文誌, J73-D-II, 3, pp.351-359, 1990.

[10] 林島, 原島 博, "畵像と音聲の知的インタラクティブ符號化の構想", 日本畵像符號 化シンポジウム(PCSJ), 5, 6, 1987.

[11] 조성업, 오범수, 임철수, "청각장애아동의 구화교육을 위한 애니메이션 콘텐츠 제작기술 개발", 한국멀티미디어학회지, 제9권, 제3호, pp.36-45, 2005.

[12] 이정훈, 이은주, 민홍기, "청각장애인을 위한 인터넷 자막 방송의 구현", HCI2000 학술대회, pp.1-6. 2000.

[13] 한국정보통신기술협회, "한국 텔레비전 자막방송 표준 TTA.KO-07.0010", 1997.

[14] P. Ekman, W. V. Friesen, "Unmasking the Face", Prentice-Hall, 1975.

[15] K.Waters, "A Muscle Model for animating Three-Dimensional Facial Expression", Computer Graph., 15, 3, pp.17-24, 1987.

[16] 송철의, "자음의 발음", 새 국어 생활, 제13권 1호, 국립국어연구원, 1993.

[17] 이승재, "모음의 발음", 새 국어 생활, 제13권 1호, 국립국어연구원, 1993.

김 형 근

e-mail : hgrikim@knou.ac.kr

1982년 명지대학교 전자공학과(학사)

1984년 명지대학교 대학원 전자공학과 (공학석사)

1991년 명지대학교 대학원 전자공학과 (공학박사)

1993년~현 재 한국방송통신대학교 컴퓨터과학과 교수

관심분야 : 패턴인식, 컴퓨터 비젼 등

박 철 하

e-mail : pch@mail.daebul.ac.kr

1988년 명지대학교 전자공학과(학사)

1990년 명지대학교 대학원 전자공학과 (공학석사)

1994년 명지대학교 대학원 전자공학과 (공학박사)

1996년~2000년 대불대학교 전자계산소 소장

1994년~현 재 대불대학교 경찰학부 부교수

관심분야 : 영상신호처리, 음성신호처리, 멀티미디어 등