<한국어 립씽크를 위한 3D 디자인 시스템 연구>

신 동 선[1] , 정 진 오[2]
성균관대학교 인지과학협동과정[1] 성균관대학교 영상학과[2]
ttpass@freechal.com[1] , jochung@skku.edu[2]

<A Study on 3D Graphics Design System for the Korean Lip-Sync Synthesis>

Dong Sun Shin[1], Jinoh Chung[2]
Dept. of Cognitive Science, SungKyunKwan University,
Dept. of Film, TV & Multimedia, SungKyunKwan University[2]

요 약

3 차원 그래픽스에 적용하는 한국어 립씽크 합성 체계를 연구하여, 말소리에 대응하는 자연스러운 립씽크를 자동적으로 생성하도록 하는 디자인 시스템을 연구 개발하였다. 페이셜애니메이션은 크게 나누어 감정 표현, 즉 표정의 애니메이션과 대화 시 입술 모양의 변화를 중심으로 하는 대화 애니메이션 부분으로 구분할 수 있다. 표정 애니메이션의 경우 약간의 문화적 차이를 제외한다면 거의 세계 공통의 보편적인 요소들로 이루어지는 반면 대화 애니메이션의 경우는 언어에 따른 차이를 고려해야 한다. 이와 같은 문제로 인해 영어권 및 일본어 권에서 제안되는 음성에 따른 립싱크 합성방법을 한국어에 그대로 적용하면 청각 정보와 시각 정보의 부조화로 인해 지각의 왜곡을 일으킬 수 있다. 본 연구에서는 이와 같은 문제점을 해결하기 위해 표기된 텍스트를 한국어 발음열로 변환, HMM 알고리듬을 이용한 입력 음성의 시분할, 한국어 음소에 따른 얼굴특징점의 3 차원 움직임을 정의하는 과정을 거쳐 텍스트와 음성를 통해 3 차원 대화 애니메이션을 생성하는 한국어 립싱크합성 시스템을 개발 실제 캐릭터 디자인과정에 적용하도록 하였다.
또한 본 연구는 즉시 적용이 가능한 3 차원 캐릭터 애니메이션뿐만 아니라 아바타를 활용한 동적 인터페이스의 요소기술로서 사용될 수 있는 선행연구이기도 하다. 즉 3 차원 그래픽스 기술을 활용하는 영상디자인 분야와 HCI 에 적용할 수 있는 양면적 특성을 지니고 있다. 휴먼 커뮤니케이션은 언어적 대화 커뮤니케이션과 시각적 표정 커뮤니케이션으로 이루어진다. 즉 페이셜애니메이션의 적용은 보다 인간적인 휴먼 커뮤니케이션의 양상을 지니고 있다. 결국 인간적인 상호작용성이 강조되고, 보다 편한 인간적 대화 방식의 휴먼 인터페이스로 그 미래적 양상이 변화할 것으로 예측되는 아바타를 활용한 인터페이스 디자인과 가상현실 분야에 보다 폭넓게 활용될 수 있다.

Keyword : Lip-sync, Facial Animation, Avatar, HCI

1. 서 론

1-1 연구목적

얼굴은 상호간의 대화 커뮤니케이션에 있어 매우 중요한 의미를 갖는 신체 부위이다. 인간 사이의 대화는 음성뿐만 아니라 얼굴 표정을 통하여 많은 정보를 주고 받는다는 것이 심리학 분야에서는 널리 알려져 있다. 이와 같이 표정은 사람의 내적 상태를 동시에 반영하는데, 이것은 정서 상태뿐만 아니라 정보의 인지과정, 사회적 상호작용 그리고 생리적 신호들을 포함한다[1].

정보를 전달하는 얼굴 표정에서 입 모양은 특히 중요한 역할을 한다. 전화로 하는 대화보다 상대방의 얼굴을 마주보면서 하는 대화로 훨씬 의사소통이 원활하게 이루어지는 것으로도 그 중요성을 쉽게 알 수 있다. 이와 같이 말소리가 명료하지 않을 때, 화자의 얼굴을 바라보는 것은 음성의

지각을 개선시킬 수 있는데(Sumby & Pollack, 1954), 이는 청각정보만으로는 알기 어려운 조음장소정보를 시각정보를 통해서 쉽게 파악할 수 있기 때문이다(Binie, Montgomery, & Jackson, 1977).

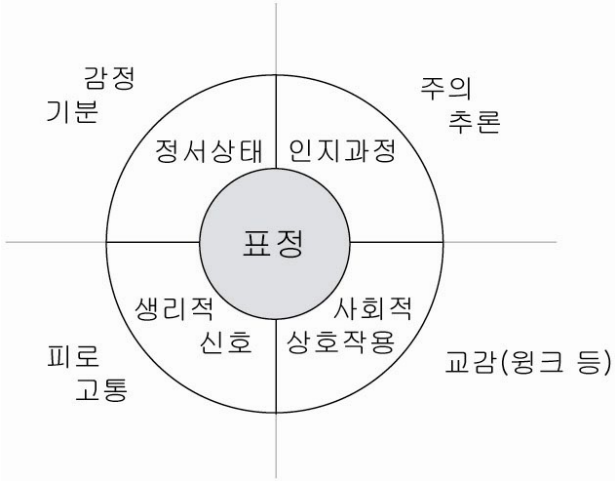


그림 1) 얼굴 표정이 반영하는 것들

이와 같은 지각의 문제로 인해 인간 간의 대화하는 느낌으로 아바타 혹은 매개 캐릭터와 대화할 수 있는 자연스러운 음성과 얼굴표정의 합성이 실시간 가능하다면 애니메이션 분야뿐만 아니라 HCI 에 커다란 진전을 보일 수 있을 것이다. 이러한 이유로 얼굴 표정 및 입 모양 합성에 관한 연구는 최근에 와서 얼굴 영상의 지적 부호화를 중심으로 그 연구가 활발해지고 있으며 해외에서는 이 같은 연구결과를 토대로 토킹헤드, 버추얼프랜드, 페이스이메일, 토크쇼, 버추얼 아나운서 등 많은 상업적 제품들이 개발되고 있다.

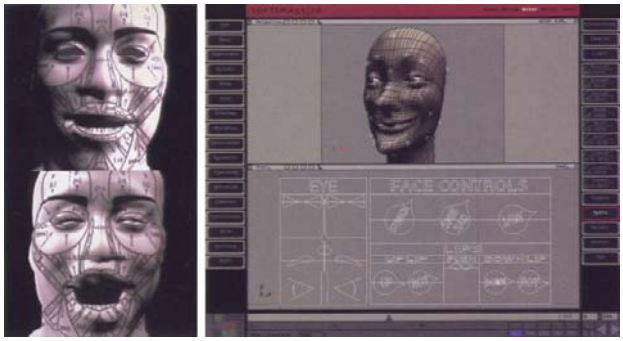


그림 2) 근육의 움직임과 표정변화 시뮬레이션

©Acclaim Entertainment

하지만 자연스러운 립씽크(Lip Sync)구현에는 아직 많은 어려움이 따르는데, 이는 음성과 자연스럽게 어울리는 입 모양을 표현하기 위해서는 많은 요소들이 적절히 조절되어야 하기 때문이다. 이러한 이유로 페이셜애니메이션을 위한 컴퓨터그래픽스 동기화 기술은 3 차원적인 세세한 변형 등 까다롭고 미세한 부분까지 신경 쓰지 않으면 안 되는 등 기술적인 난이도가 높아서, 현재까지는 완성된 연구 결과가 드물지만 다양한 시도가 이루어지고 있는 단계이다. 일반적으로 3 차원 애니메이션 클립을 만들 때 각각의 프레임을 수작업을 통하여 제어하거나 재사용이 불가능한 페이셜트래킹 방식의 모션캡쳐를 이용해왔다. 따라서 제작기간이 상당히 길고 고비용 저효율을 낳아 고가의 영화나 광고 동영상에 주로 활용되어왔다. 이러한 어려움을 감소시키기 위해 자동으로 음성에 영상을 동기화하는 페이셜애니메이션을 위한 연구들이 행해지고 있으며 본 연구에서는 3 차원 애니메이션에서 쉽게 재사용이 가능한 한국어 립씽크 합성 방법에 관한 연구를 하였다.

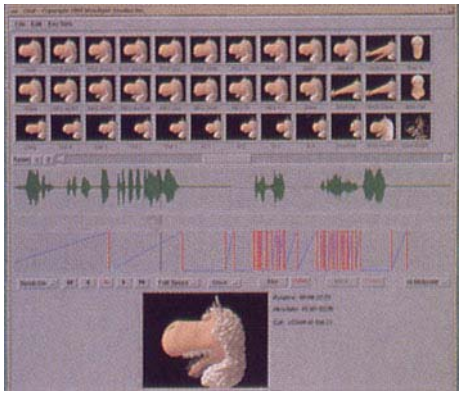


그림 3) 사운드웨이브를 이용한 립씽크 컨트롤

©CBS, Windlight Studio

1-2 연구방법 및 제약점

본 연구에서는 이와 같은 문제점들을 해결한 한국어 립싱크 합성을 위해 표기된 텍스트를 한국어 발음열로 변환하는 발음열 생성 모듈과 HMM 알고리즘을 이용한 음소 시분할 모듈, 그리고 음소에 따른 얼굴특징점의 움직임과 음성를 동기화하는 음성-영상 동기화 모듈을 통해 한국어의 특성을 고려한 한국어 립싱크합성시스템을 제안하였다. 현재 3 차원 입모양에 관한 표준은 아직 규정된 것이 없으므로 이와 같은 시스템은 몇 가지 제

약점을 가진다.

첫째, 표준발음에 대한 3 차원 입모양의 표준 규약이 없으므로 본 시스템을 통해 생성된 입모양은 표준이 아닌 개인에 특화된 입모양으로 합성이 될 수 밖에 없다.

둘째, 3 차원 애니메이션의 구현에 관한 법칙들([2][3])에 대한 고려가 이루어지지 않았다. 이 법칙들은 전통적인 2 차원 카툰 애니메이션을 위한 '디즈니 애니메이션의 12 원칙'을 3 차원 애니메이션에 적용한 것이다.

셋째, 본 연구는 표정을 제거한 상태에서의 입모양만을 고려했으므로 감정 및 상황에 대한 변이는 적용되지 않았다. 감정 및 상황에 대한 얼굴의 변화는 얼굴 근육의 움직임을 기호화하여 이들의 조합으로 얼굴 표정을 나타내도록 한 FACS(Facial Action Coding System)[4]의 적용을 통해 이루어질 수 있을 것이다. 이 같은 제약점들은 추후 연구를 통해서 보완할 예정이다.

2. 한국어 립싱크합성시스템

2-1 시스템 개요

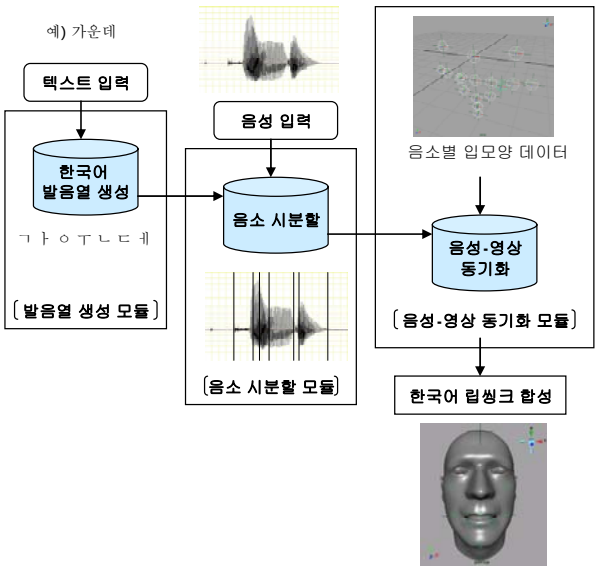


그림 4) 한국어 립싱크 합성시스템의 각 모듈

본 시스템은 그림 4)와 같이 세 가지 모듈로 구성되어 있다. 첫째는 음성에 해당하는 텍스트를 한국어 발음열로 변환하는 모듈, 둘째는 HMM 알고리즘을 이용해서 입력된 음성을 음소단위로 시분할하는 모듈, 셋째는 MPEG-4 에서 정의한 얼굴 움직임의 특징점에 따른 요소를 입력된 음소에 맞게 구동시켜주는 모듈이다. 이 세가지 모듈을 통해 한국어 립싱크가 이루어지며 각 모듈의 이론적 배경은 다음과 같다.

가. 한국어 자동발음열 생성

한국어는 표음문자임에도 불구하고 실제 표기와 발성되는 발음이 틀린 음운현상이 일어난다. 이러한 음운현상과 불규칙활용 등으로 인해 한국어 문서의 형태소 분석에 사용되는 형태소 사전을 음소의 시분할에 그대로 사용할 수는 없으며 음운변동을 표시할 수 있는 사전구조가 필요하다[5]. 음성인식이나 합성에서는 음소의 시분할을 위해 문자 형태의 말뭉치가 아닌 실제 발음 형태의 말뭉치가 사용된다. 이러한 발음열은 수작업으로 작성하기에는 많은 시간과 노력이 필요하며 일관성의 유지도 힘들기 때문에 정확한 발음열을 생성하기 위해서는 해당 언어가 가지는 특징뿐만 아니라 그 언어의 음운변화 규칙과 적용 양상에 대한 정확한 분석이 필요하다[6]. 본 논문에서는 이와 같은 발음열 생성의 문제를 해결하기 위해 문교부 표준어 규정의 표준 발음법에서 유도된 필수 및 수의적 음소 변동 규칙과 변이음 규칙의 단계적 적용 모델을 사용했다[6].

나. 음소 시분할

입력된 음성을 음소로 시분할하기 위한 방법으로 가변어휘(vocabulary-independent) 음소분할 방법을 사용했다. 여러 명의 화자로부터 다양한 음운현상이 충분히 반영된 대용량 음성 데이터베이스를 사용하여 음소모델을 훈련하고, 인식대상 어휘 모델들을 이러한 음소모델들의 연결로 구성하면 처음 한 번의 훈련으로 어떤 어휘라도 시분할이 가능하게 된다[7][8]. 시분할단계에서 음성이 입력되면 음성의 특징을 MFCC 특징파라미터(MFCC Feature Parameters)로 추출한 후 HMM 알고리듬을 사용한 인식네트워크를 통하여 시분할 과정을 수행하였다. 음성의 특징을 추출하고 음소

모델을 만들기 위한 훈련 데이터로 445 균일음소 분포단어(PBW; Phonetically Balanced Word)데이터 베이스를 사용했다.

다. 입모양의 분석

한글 표준어의 자음은 중자음을 포함한 21 개이고, 모음은 중모음을 포함한 19 개로 구성된다. 한글의 자소(초,중,종성)들의 조합으로 생성되는 문자는 총 14,364 자에 이르며, 현재 실용되는 한글 한자 표준코드(KSC5601)에서의 한글은 2350 자이다[9]. 그러므로 2350 자를 모든 음절에 대해서 한글의 표준 발음법에 준하여 입모양을 분류함으로써 모든 음절에서 표현될 수 있는 입모양을 만들 수 있다. 입모양의 분류에 관한 기존의 연구에서는 대부분 입 모양의 변화를 단지 정면에서 본 사진으로만 분류하였기 때문에 정면 외의 각도에서의 변화는 고려하지 않았다[10]. 하지만 본 연구에서는 입모양의 변화를 입체적으로 측정할 것이므로 한국어의 모든 모음과 자음을 분류하였다.

2-2 립씽크합성을 위한 음소모델링

한국어 립씽크 합성을 구현하기 위해서는 먼저 각 음소와 그에 따른 입모양이 정의되어야 한다. 본 연구에서 구현한 세부내용은 다음과 같다.

가. 특징점의 위치값 획득

본 연구에서는 한국어의 음소에 대응하는 입모양의 움직임을 측정하기 위해 위해 MPEG(Moving Picture Experts Group)의 MPEG-4 에서 정의한 FDP(Facial Definition Parameter)기반의 특징점을 사용하였다. FDP 는 눈, 눈썹, 코, 입, 턱, 뺨, 혀, 치아, 귀, 머리 그리고 얼굴의 회전 등을 나타내는 총 9 개의 그룹으로 나뉘어 정의되는데[11] 본 연구에서는 이중 조음동작에 관련된 턱, 입술, 뺨의 범주에 해당하는 11 개의 FAP 를 입 모양 합성을 위한 특징점으로 선정하였다. 이 때 코끝에 해당하는 9.3 특징점 1 개를 기준정렬을 위해 사용하였으며 10 개의 특징점은 각각의 범주에서 가장 큰 움직임을 보이며 입 모양합성시스템을 완성했을 때 사용자들이 얼굴모델에 직관적으로 적용할 수 있는 특징점으로 선정하였다(그림 3).

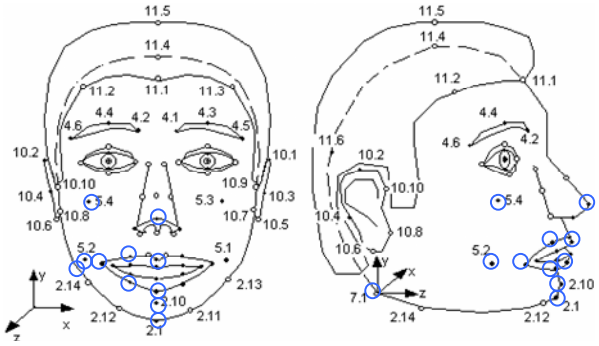


그림 5) 선정된 11 개의 FDP/FAP 특징점

선정한 특징점을 기반으로 실제 발성하는 입모양의 모델을 생성한 프로세스는 아래와 같다.

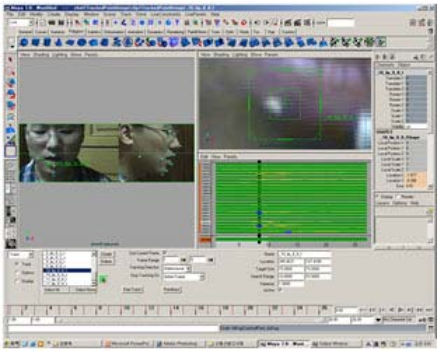
ㄱ) 촬영

피실험자의 얼굴에 트랙포인트 부착후 445 균일음소분포단어 발성하는 모습을 DV Cam 2 대로 정면 측면 동시촬영

ㄴ) 정렬 및 동기화

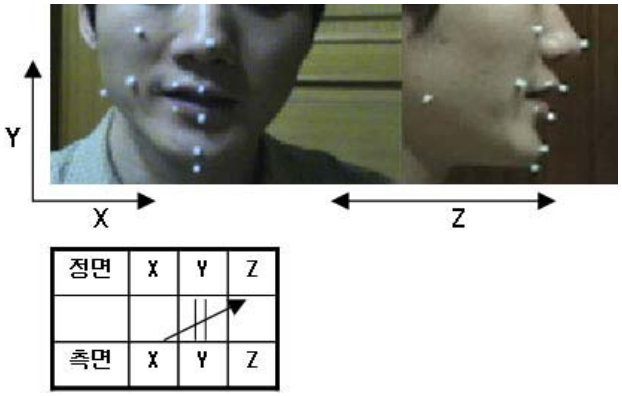
두 동영상을 Y 축을 기준으로 정렬 및 동기화

ㄷ) 트래킹(Tracking)



445 개 파일로 각각 랜더링 후 11 개의 특징점을 트래킹

ㄹ) 좌표획득

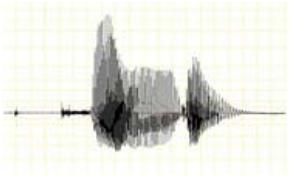


정면 영상과 측면 영상을 통해 완전한 X, Y, Z 축의 좌표획득

나. 음소의 위치값 정의

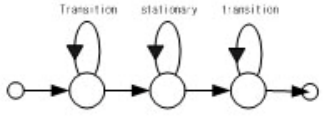
445 개 단어의 발성에서 얻은 3 차원 데이터와 동영상에서 추출한 음성파일을 가지고 MFCC 와 HMM 알고리듬을 통해 각 음소에 대응하는 3 차원 위치 데이터를 얻었으며 그 프로세스는 아래와 같다. 이 때 같은 음소라도 연음 관계에 따라 입 모양이 달라질 수 있으므로 음소에 따른 입 모양을 정의하기 위해 K-평균(K-Means)군집분석을 사용해 수 개에서 수 천 개의 음소 중에서 가장 중심이 되는 음소를 그 음소를 대표하는 입 모양으로 정의하였다. 이렇게 정의된 39 개의 입 모양에 대해 계층적 분류분석을 하고 이를 토대로 특징점의 움직임이 유사한 입 모양을 군집화(clustering)하였다. 군집의 개수를 7 로 했을 때 가장 고른 분포를 보였으며 각 군집에서 가장 거리가 가까운 음소를 그 군집의 대표값으로 정의하였다.

ㄱ) **좌표획득** (445 개의 동영상에서 음성을 추출)

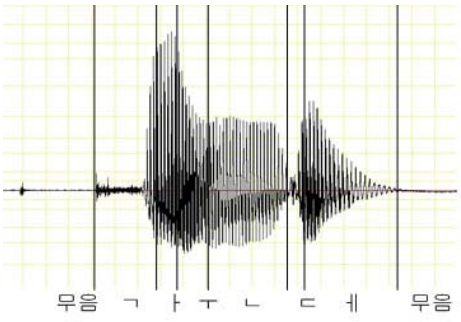


ㄴ) **음성에서 MFCC 추출**

ㄷ) **음소 시분할** (발음열 생성기로 생성된 발음열과 HMM 알고리듬을 통해 음소시분할)



ㄹ) **시분할** (음성을 음소구간에 맞춰 시분할)



다. 립씽크 합성프로그램의 구현

입력된 단어를 발성할 때의 발음열로 바꾸어 주는 발음열 생성 프로그램, 음성파일에서 MFCC 를 추출해는 프로그램, 발음열과 음성파일을 시분할하는 프로그램, 각 음소의 입 모양을 정의한 데이터를 가지고 한국어 립씽크 합성프로그램을 제작하였다. 각 단위프로그램들은 Microsoft 의 Visual C++ 언어로 제작되었으며 MAYA 에서 구동시키기 위해 MEL 로 인터페이스를 만들었다. 완성된 프로그램은 그림 6)과 같으며 프로그램의 기능은 다음과 같다.

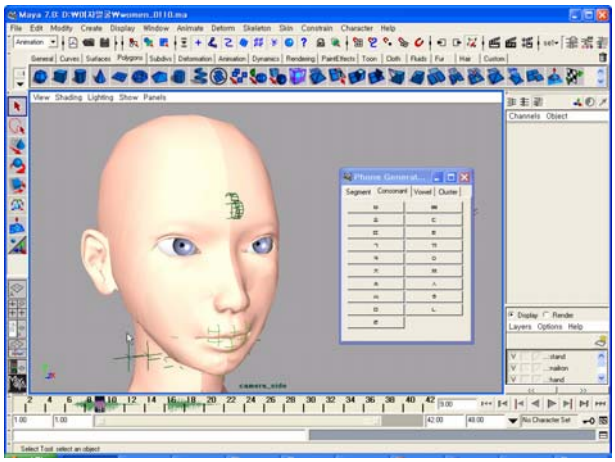
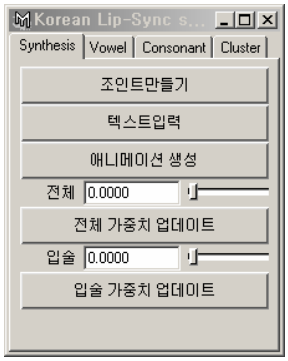


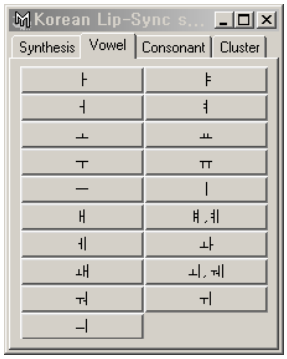
그림 6) 한국어 립씽크 합성프로그램의 실행모습

(1) Synthesis 메뉴의 '조인트 만들기'를 누르면 특징점 11 개에 해당되는 조인트가 생성되며 이렇게 생성된 조인트를 립싱크를 합성하고자 하는 모델에 바인딩 할 수 있다. '텍스트입력' 버튼을 클릭해 음성파일에 해당하는 단어를 입력하면 그 단어의 발음열을 생성한다. 그 후 생성하고자 하는 입 모양의 음성파일을 저장한 후 '애니메이션 생성' 버튼을 클릭하면 음성파일의 MFCC 추출, 입력한

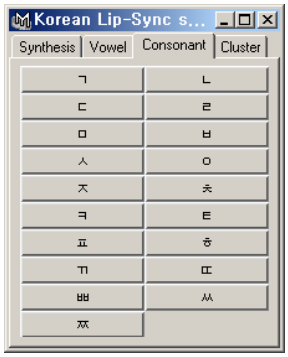
단어에 해당하는 발음열 생성, 음성파일의 음소 시분할단계를 거쳐 립싱크 애니메이션이 생성된다. ‘전체’ 항목의 슬라이드바를 통해 수치를 입력한 후 ‘전체 가중치 업데이트’ 버튼을 누르면 조인트들의 전체적인 크기를 조절할 수 있고 ‘입술’ 항목의 슬라이드바를 통해 수치를 입력한 후 ‘입술 가중치 업데이트’ 버튼을 누르면 입술의 움직임에 대한 가중치를 줄 수 있다.



(2) Vowel 메뉴에서는 각 모음이 표기된 버튼을 누름으로써 모음에 대해 정의한 입 모양을 시뮬레이션 할 수 있다.



(3) Consonant 메뉴에서는 각 자음이 표기된 버튼을 누름으로써 자음에 대해 정의한 입 모양을 시뮬레이션 할 수 있다.



(4) Cluster 메뉴에서는 모든 음소를 K-평균(K-Means) 군집분석을 사용해 7 개의 군집으로 분류하고 각 군집에서 가장 중심에서 가까운 음소를 처음에 표기하였으며 중심에서 떨어진 거리에 따라 군집의 나머지 음소들을 표기하였다. 이를 통해 각 군집을 대표하는, 움직임이 큰 음소를 시뮬레이션 해봄으로써 애니메이션 작업을 할 때 3D 모델과 조인트와의 세부적인 조절을 보다 빠르게 할 수 있다.

4. 실험 및 고찰

완성된 프로그램을 검증하기 위해 음소의 입 모양 정의에 사용했던 445 균일음소분포단어와 한국전자통신연구원(ETRI)에서 구축한 POW (Phonetically Optimized Word) 3848 단어 DB 의 목록에 있는 단어를 발성하는 모습을 실제로 녹화한 동영상과 여기에서 음성파일을 추출한 후 한국어 립싱크합성프로그램을 사용해서 만든 애니메이션과 비교하였다. 음소모델링의 구축에 사용했던 모션 트래킹을 사용해서 비교하였으며 각 프레임에서의 입 모양(그림 7)과 모든 구간에서 각 특징점의 궤적의 변화(그림 8)를 비교하였다.

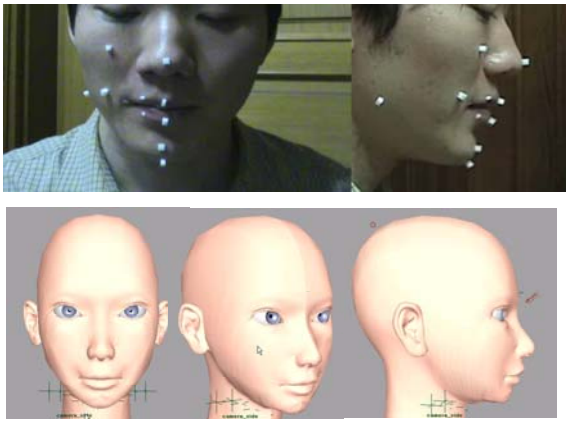


그림 7) ‘ㅃ’ 의 실제발성영상과 생성된 애니메이션

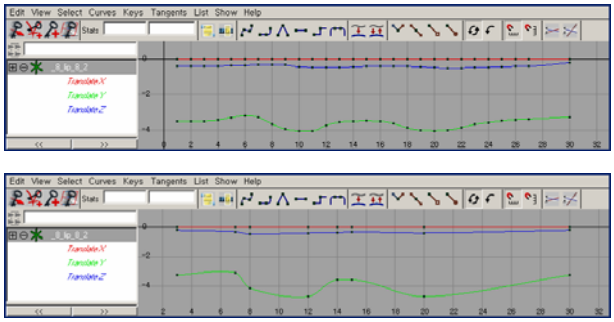


그림 8) 전 구간에서의 실제 발성(위)과
생성된 애니메이션(아래)의 궤적 (FDP 8.2)

한국어 립싱크합성프로그램으로 생성한 애니메이션은 본(Bone)구조와 형태의 차이로 음소모델링에서 사용했던, 실제 발성하는 모습의 입술 벌어지는 강도 및 턱의 움직임과 완전히 일치하지 않지만 음소에 대한 특징적인 입 모양과 음소 사이의 특징적인 입 모양은 일치하였다. 그리고 궤적의 패턴도 전 구간에서 일치하는 경향을 보였다. 본 프로그램으로 생성한 대표적인 음소는 그림 9)와 같다.

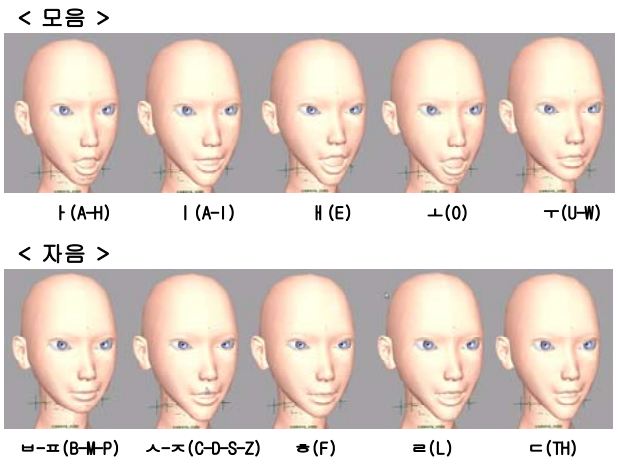


그림 9) 캐릭터애니메이션에 활용되는
대표적 음소에 최적화시킨 이미지

5. 결 론

본 연구의 결과로 모음 몇 개를 통해 입 모양을 수동으로 생성하던 기존의 방식의 한계를 넘어 보다 자연스러운 입 모양을 시간의 변화에 따라 정확하게 생성해낼 수 있었다. 하지만 생성된 입 모양애니메이션은 몇 가지 문제를 가지고 있다. 먼저 사람마다 골격과 근육이 다르고 이에 따라 실제 발성하는 입 모양도 틀리다. 본 연구에서는 실제로 발성하는 모습의 모션을 캡쳐 하였기 때문에 실제 사람의 골격과 애니메이션을 만들고자 하는 골격과의 차이가 심한 경우 움직임이 완전하게 정합되지 않으며 이는 수동으로 움직임의 가중치를 수정해주어야 한다. 또 사람에 따라 어떤 특정한 발성이 끝났을 때 입을 닫는 경우가 있고 벌린 상태를 그대로 유지하는 경우, 그 중간의 경우가 존재하며 이 역시 사람마다 다른 경향을 보였다. 따라서 특정인의 발화 습관을 그대로 모델링 하기 위해서는 이와 같은 발화 습관의 모델링도 이루어져야 할 것이다.

본 연구에서는 무표정한 상태에서 발성했을 때의 입 모양의 움직임만을 모델링 하였지만 여러 가지 상황에 따라 달라지는 입 모양의 변이까지 모델링 하여야 더욱 다양한 립씽크애니메이션을 생성할 수 있을 것이다. 더 나아가 무표정한 상태에서의 입 모양뿐 아니라 감정에 따라 달라지는 입 모양의 변화의 강약까지 모델링에 적용한다면 보다 상호작용성이 강조됨으로써 아바타를 활용한 HCI 와 가상현실 분야에 보다 폭넓게 활용될 수 있을 것이다.

참 고 문 헌

[1] B. Fasel and J. Luettin, “Automatic facial expression analysis: a survey”, Pattern Recognition 36, pp. 259-275, 2003.

[2] John Lasseter, "Principles of Traditional Animation Applied to 3D Computer Animation", Computer Graphics, pp. 35-44, 21:4, July 1987 (SIGGRAPH 87).

[3] George Maestri, "Digital Character Animation", New Riders Press, 1996.

[4] P. Ekman and W. V. Friesen, “Emotion in the Human Face System”, Cambridge University Press, San Francisco, CA, 2nd edition, 1982.

[5] 이건상, 양성일, 권영헌, “음성인식”, 한양대학교 출판부, 2001.

[6] 전재훈, “형태음운학적 분석에 기반한 한국어 발음열 자동 생성”, 서강대학교 전자계산학과 공학석사 학위논문, 1997.

[7] K. F. Lee, “Contex-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition”, IEEE Trans. Acoust., Speech, Signal Processing, vol.38, no.4, pp.599-609, Apr.1990.
[8] L. R. Rabiner, B. H. Juang, “Fundamentals of Speech Recognition”, Prentice-Hall Inc., 1993.
[9] 문교부 고시, 문화부 공고. “국어 어문 규정집”, 대한교과서주식회사, 1991.
[10] 이용동, 최창석, 최갑석, “휴먼인터페이스를 위한 음절분류에 따른 얼굴영상의 합성”, 대한전자공학회 학술대회 논문집(신호처리합동), 제 5 권 1 호, 433-437, 1992.
[11] Igor S. Pandzic, Robert Forchheimer, “MPEG-4:Facial Animation”, Wiley, 2002.