Journal of Digital Contents Society Vol. 9 No. 2 June. 2008(pp. 331-339)

한국어 음소를 이용한 자연스러운 3D 립싱크 애니메이션

정일홍*, 김은지**

요 약

본 논문에서는 3D 립싱크 애니메이션에 필요한 키 데이터를 생성하는 효율적이고 정확한 시스템 개발을 제안한다. 여기서 개발한 시스템은 한국어를 기반으로 발화된 음성 데이터와 텍스트 정보에서 한국어 음소를 추출하고 분할된 음소들을 사용하여 정확하고 자연스러운 입술 애니메이션 키 데이터를 계산한다. 이 애니메이션 키 데이터는 본 본문에서 개발한 3D 립싱크 애니메이션 시스템뿐만 아니라 상업적인 3D 얼굴 애니메이션 시스템에서도 사용된다.

전통적인 3D 립싱크 애니메이션 시스템은 음성 데이터를 영어 음소 기반으로 음소를 분할하고 분할된 음소를 사용하여 립싱크 애니메이션 키 데이터를 생성한다. 이러한 방법의 단점은 한국어 콘텐츠에 대해 부자연스러운 애니메이션을 생성하고 이에 따른 추가적인 수작업이 필요하다는 것이다. 본 논문에서는 음성 데이터와 텍스트 정보에서 한국어 음소를 추출하고 분할된 음소를 사용하여 자연스러운 립싱크 애니메이션을 생성하는 3D 립싱크 애니메이션 시스템을 제안한다.

Natural 3D Lip-Synch Animation Based on Korean Phonemic Data

Il-Hong Jung*, Eun-Ji Kim**

Abstract

This paper presents the development of certain highly efficient and accurate system for producing animation key data for 3D lip-synch animation. The system developed herein extracts korean phonemes from sound and text data automatically and then computes animation key data using the segmented phonemes. This animation key data is used for 3D lip-synch animation system developed herein as well as commercial 3D facial animation system.

The conventional 3D lip-synch animation system segments the sound data into the phonemes based on English phonemic system and produces the lip-synch animation key data using the segmented phoneme. A drawback to this method is that it produces the unnatural animation for Korean contents. Another problem is that this method needs the manual supplementary work. In this paper, we propose the 3D lip-synch animation system that can segment the sound and text data into the phonemes automatically based on Korean phonemic system and produce the natural lip-synch animation using the segmented phonemes.

Keywords : 립싱크, 애니메이션, 음소 데이터, 음성 인식

1. 서론

최근 3D 캐릭터 애니메이션 산업은 과거에 비해 비약적인 발전과 함께 산업계의 주목을 받고 있다. 3D 캐릭터 애니메이션 중에서도 얼굴 애니메이션은 자연스러운 표정과 감정의 표현을 위주로 발전하였는데 아직 3D 게임이나 애니메이션에 등장하는 캐릭터가 자연스러운 얼굴 표정을 연출하고 있다고는 보기 어렵다[1]. 애니메이션에 등장하는 캐릭터의 자연스러운 얼굴 표정 연출의 가장 큰 문제점은 음성 신호와 입술 움직임의 동기화이다. 국내에서 제작되어지는 애니메이션 캐릭터들은 한국어를 말하면서 영어 형식으로 분할된 음소에 맞추어 동기화가 이루어지기 때문에 자연스러운 립싱크가 이루어지지 않는 경우가 많았으며, 한국어에 맞추어 음소 분할을 하더라도 많은 시간과 노력을 투자하여야

※ 제일저자(First Author) : 정일홍
접수일자:2008년04월23일, 심사완료:2008년05월02일
* 대전대학교 컴퓨터공학과,
ijung@dju.kr
** 대전대학교 컴퓨터공학과

하기 때문에 실제 애니메이션을 만드는 데 한계가 있다.

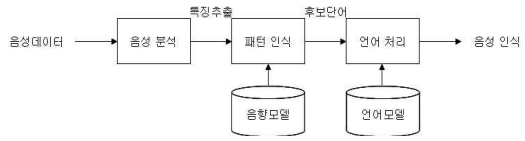
일반적인 음성인식 및 립싱크 애니메이션 시스템의 문제점을 3가지로 분석할 수 있다. 첫째, 기존 시스템의 음성 인식률에 대한 문제이다. 기존 시스템은 대체로 음성 인식을 음성 데이터에만 의존하였다. 현재의 기술 수준으로 화자 독립의 음소인식 성능이 60~70% 정도 밖에 되지 않음을 감안할 때, 실제 애니메이션 제작을 위한 립싱크 애니메이션 시스템에 사용하기에는 자동 분할 및 레이블링 후 수작업을 통해 많은 수정을 거쳐야 하므로 적합하지 않다. 따라서 기존 방법보다 인식률을 높이는 알고리즘이 필요하며, 음성 데이터 외에 유저가 일반적으로 사용가능한 음성 인식을 보조할 수 있는 수단이 따로 요구된다. 둘째, 한국어 기반의 음소분할에 대한 문제이다. 대부분의 애니메이션 제작에서는 한국어 애니메이션임에도 불구하고 영어기반으로 음소가 분할되어 립싱크가 자연스럽지 않거나 제작된 애니메이션을 수작업으로 재 작업해야 하는 경우가 많다. 따라서 한국어 전용 립싱크 애니메이션을 위해서는 음소를 분할할 때 한국어의 음운론과 발음법에 맞는 방법을 사용하여야 할 것이다. 셋째, 음절에 따른 입모양 패턴 분류에 대한 문제이다. 한국어의 경우 모든 음절에 따라 각기 다른 입모양을 갖는다. 그러나 음절의 음은 달라도 같은 형태의 입모양 패턴이 많이 존재한다. 따라서 모든 음절에 대해서 입모양을 정할 것이 아니라 대표적인 음절에 대해서만 입모양을 정하여 이 입모양의 결합으로 다른 음절의 입모양을 나타내는 것이 그동안 한국어 립싱크 애니메이션 연구에서 많이 사용되던 방법이다. 그러나 이러한 대표 입모양은 보통 40개 전후로 정해져 유저가 일일이 입모양을 만들어내기에는 사실상 많은 수이다. 그러므로 입모양 패턴의 모양을 유저가 제작 가능한 개수이내가 되도록 효과적으로 분류하는 알고리즘이 필요하다.

본 논문에서는 한국어 발화의 음성 데이터와 발화된 음성 데이터의 텍스트 정보를 이용하여 음소를 자동 분할하고 동기화 정보를 추출하여 상용화된 3D 모델링 툴에서 적용 가능한 한국어 체계에 맞는 정확하고 자연스러운 키 데이터를 추출하는 한국어 기반의 립싱크 애니메이션 시스템을 제안한다.

2. 관련 연구

2.1 음성 인식 시스템

음성 인식 시스템은 크게 음성 분석 단계와 패턴 인식 단계, 또 언어처리 단계로 나눌 수 있다[2],[3]. 음성 분석 단계에서는 사용자가 발성한 음성으로부터 인식에 필요한 특징 벡터를 추출하고, 패턴 인식 단계에서 음성 데이터베이스로부터 훈련한 기준 패턴과의 비교를 통해서 인식 결과를 얻게 된다. 보다 복잡한 구조의 음성을 인식할 때에는 언어모델을 이용한 언어 처리 과정을 통해 최종 인식 결과를 출력한다. (그림 1)는 일반적인 음성인식 시스템의 구성도이다[4].



(그림 1) 일반적인 음성 인식 시스템의 구성

2.1.1 음성 분석 단계

음성 분석 단계는 크게 끝점 검출 단계와 특징 추출 단계로 나눌 수 있다. 먼저 끝점 검출은 입력되는 신호로부터 음성 구간만을 검출하는 처리 과정이다. 일반적으로 사용되는 방법은 입력 신호의 매 구간에서 에너지 값을 구하여 통계에 의해 미리 결정된 임계값과의 비교를 통해서 음성구간과 묵음구간을 판별하는 방법이다[3]. 이 방법은 비교적 높은 신호잡음비(SNR)환경에서는 뛰어난 성능을 보이지만, 잡음이 많은 환경에서는 그 성능이 현저하게 떨어진다. 주변 잡음은 시간이 지남에 따라 그 특성이 변화하므로 이에 맞게 임계값을 적응적으로 조절해주는 과정이 반드시 요구된다.

음성의 특징 추출이란 음성으로부터 의미 정보를 나타내어주는 특징을 추출하는 것으로 일종의 음성 압축 부분이며 한편으로 인간의 발성기관을 모델링하는 부분이라고 생각할 수 있다. 음성 인식 시 특징 파라미터로 사용되는 것은 여러 가지 형태가 있으나 대부분 성도 특성을 나타낸다. 인간의 발음은 입 모양과 혀의 위치,

기타 유·무성 분류, 에너지 변화패턴 등에 의해 결정되며 이중 가장 중요한 요소인 입 모양과 혀의 위치에 의해 성도 특성이 결정된다. 발음 특성을 잘 나타내는 파라미터가 음성 인식용으로 좋은 파라미터이며 대표적인 것으로 선형 예측 분석에 의해 추출하는 LPC(Linear Predictive Coefficients)[4], 귀의 인지 특성을 고려한 MFCC(Mel Frequency Cepstral Coefficients)[4] 등이 있으며 이 외에 여러 가지 변형된 형태들의 파라미터들도 사용된다.

2.1.2 패턴 인식 단계

패턴 인식은 음성을 일종의 패턴으로 간주하여 등록되어 있는 패턴과 입력되는 패턴과의 유사도를 측정하여 인식하는 방법이다. 음성 인식에 사용되는 가장 일반적인 방법은 템플릿 기반의 패턴 매칭 방법을 이용하는 DTW(Dynamic Time Warping)[2],[3] 와 통계적 패턴 인식을 이용한 HMM(Hidden Markov Model)[2],[3]이 대표적으로 많이 이용된다. 신경 회로망을 이용한 방법도 음성 인식에 이용되고 있으나 많은 계산량 등의 문제로 최근에는 HMM과 결합한 형태를 많이 취하고 있다.

HMM은 수학적인 배경에서 개발된 알고리즘으로 음의 상태가 한 상태에서 다음 상태로 바뀌는 것을 천이 확률로 표현한다. HMM은 음성 신호의 시간적인 통계적 특성을 이용하여 훈련 데이터로부터 이들을 대표하는 모델을 구성한 후 실제 음성 신호와 유사도가 높은 확률 모델을 인식 결과로 채택하는 방법이다. 이 방법은 단독음이나 연결음, 연속음 인식에까지 구현이 용이하며 좋은 인식 성능을 나타내어 화자 독립, 대화체 음성 인식 등 여러 가지 응용 분야에 많이 이용되고 있다. 또한 대어휘에서 DTW보다는 계산량이 적은 장점을 갖고 있다. 그러나 학습 데이터가 부족할 경우, 모델간의 변별력이 부족하고 음성 신호간의 연관성을 무시하는 경향이 있다. 인공신경망은 입력과 출력사이의 비선형 함수관계를 학습하는 특성을 가지고 있으므로 HMM의 단점을 보강하는 수단으로 많이 연구되고 있다. 실제로 음성 인식 기술이 대중화된 계기는 HMM의 등장이라고 할 수 있다.

2.1.3 언어 처리 단계

연속 음성인식기의 예로 보면, 출현 가능한 단어 수를 10만 개라고 했을 경우 인식기는 매 단어가 입력될 때마다 10만 개의 후보 중에서 한 개를 골라내야 한다. 이는 많은 오류율을 발생시키며 연산양도 매우 크다. 사람의 경우는 어떤 단어가 인식되었을 경우 다음에 따라 나올 단어의 종류는 문법상, 의미상 제한 조건에 의해 그렇게 많지 않다. 그 이유는 사람의 경우 뇌에 저장되어 있는 지식 정보를 이용하기 때문이다. 음성인식 과정에도 이와 같은 지식 정보를 초보적인 수준이나마 적용시키고 있는데 이를 언어 모델(Language Model)이라 한다[4].

대용량 연속 음성인식기의 경우 언어 모델로써 단어 천이 확률을 적용시킨다. 즉 인식 대상 단어를 N개라 했을 경우 임의의 N개의 단어가 연속하여 발생할 확률을 미리 데이터베이스화하여 갖고 있다. 가장 간단한 경우, 즉 N이 2인 경우의 확률 모델을 바이그램이라 하는데 N×N개의 가지 수가 생기며, N이 3인 경우 트라이그램이라 하며 N×N×N개의 가지 수가 생긴다. 이러한 확률 모델을 이용하여 연속으로 발생할 후보 단어의 범위를 축소시킴으로써 인식률도 향상시키고 연산양도 감축시킨다.

2.2 기존의 립싱크 애니메이션의 기술 분석

2.2.1 기존의 영어 기반 시스템의 문제점 분석

대부분의 상용 프로그램들은 립싱크를 위하여 비슷한 기능들을 포함하고 있지만 음소 분할에 있어서는 한국어의 음소 분할과 전혀 다르다. 많은 립싱크 애니메이션 프로그램에서 참조하고 있는 Preston Blair Phoneme Series에 의하면 영어기반의 음소 분할은 <표 1>과 같다[5][6].

영어 기반 립싱크 애니메이션 시스템은 우선 영어와 한국의 발음 체계가 다르다는데 문제가 있다. 영어의 경우에는 모음 위주로 발음 현상이 나타나는 한국어와는 다른 아예 발음할 수 없거나 표현 방식이 어려운 자음 발음인 /F/, /Th/, /Sh/등을 감안하여 발음 구분을 해야 한다. 모음 역시 한국어에서 구분된 단모음을 모두 분류하지 않는다. 이로 인해 영어 기반의 립싱크 애니메이션 시스템을 이용하여 한국어를 발음하였을 때는 한국어에서는 구분할 필요가 없는 /F/나 /Th/의 입모양을 표현하게 된다.

<표 1> 영어기반 음소 분할의 예

음소	A, I	E	O	U	C, D, G, K, N, R, S, Th, Y, Z
입모양					
예	apple, dive	free, peach	off, goat	fund, treasure	roach, kill
음소	F, V (sometimes D or Th)	L (sometimes D or Th)	M, B, P	W, Q	Th
입모양					
예	forest, very	election, leg	bear, put	how, quick	the, they

또한 이중모음의 처리에서도 한국어와 영어는 사용하는 이중모음이 서로 달라 한국어의 경우에는 '과', '거', '귀', '게' 등이 대표적이지만 영어의 경우에는 'ai', 'ei', 'ow'등을 이중모음으로 분류한다. 결과적으로 영어 기반의 립싱크 애니메이션 시스템을 이용하여 한국어 립싱크 애니메이션을 만들었을 때는 한국인의 발음과 달리 음소 분할이 이루어지기 때문에 자연스러운 립싱크 애니메이션을 구성할 수 없다.

2.2.2 기존의 한국어기반 시스템의 문제점 분석

음성 인식 방법에는 다양한 방법이 연구되었는데 일반적인 음소 인식 방법으로는 HMM 알고리즘과 SVM 알고리즘이 가장 많이 사용된다. HMM 알고리즘은 SVM 알고리즘에 비해 인식 속도가 느리다는 단점을 가지고 있지만 상대적으로 인식 성능은 우수하다. 3D 애니메이션 제작 등에 사용할 때에는 보다 정확한 음소 인식 방법이 요구되므로 HMM 알고리즘을 이용하는 것이 유리하다[3][7].

국내에서 연구되고 있는 대부분 한국어 기반 립싱크 시스템의 음소 분할 방법은 모음의 발음을 기준으로 분류하는 것은 일치하지만 모음의 개수나 이중모음의 처리 등에서는 차이를 보인다. 한국어의 단모음은 학자에 따라 7개에서 10개로 구분하며 'ㅏ', 'ㅔ', 'ㅣ', 'ㅗ', 'ㅜ', 'ㅡ', 'ㅓ'의 7개의 단모음으로 분류하는 방법이 일반적이다. 그러나 립싱크 애니메이션을 구현할 때 7개의 모음만으로 음소 분할하여 구분한 경우, 이중모음이나 초성 및 종성에 오게 되는 자음에 의한 발음의 표현을 고려하지 못하며 음소를 재분할해야 하는 문제가 생길 수 있다. 또, 음성 데이터만을 입력받아 음소 분할을 한 경우 음성의 인식 성능에 따라 부정확한 음소 분할이 이루어질 수 있다. 따라서 3D 애니메이션 제작을 위해서는 음소 분할의 정확성을 높이고, 최대한 실제 입모양에 근접한 음소 병합 방법이 연구되어야 한다.

3. 한국어 음소분할 알고리즘을 이용한 3D 립싱크 애니메이션 시스템

3.1 한국어 기반의 자동 음소 분할 알고리즘

3.1.1 음절에 따른 입 모양 분류

한국어의 음절에 따른 입 모양은 우선 자음과 모음의 영향으로 달라질 수 있다. 그러나 립싱크를 하는 경우 실제 사람의 입모양을 모두 표현하기에는 어려움이 있기 때문에 모음을 기준으로 단순화할 필요가 있다.

음절에 따른 입 모양 패턴은 대부분 모음에 의존하지만, 입술소리와 조합되는 경우에는 발음 초기의 입 모양이 다르다. 또한 입술소리의 받침이 오는 경우 역시 발음의 마지막 순간의 입 모양이 닫힌 입 모양으로 변하게 된다. 따라서 음절에서 자음의 영향을 받아 표현되는 입 모양은 입술소리(ㅁ, ㅂ, ㅃ, ㅍ)와 그 외의 소리에 의한 2개의 입 모양으로 분류할 수 있다.

자음은 발음 초기와 발음을 마칠 때 영향을 미치지만 입 모양 패턴을 지배하는 것은 모음이므로 모음에 대한 입 모양의 분류는 매우 중요하다. 입 모양 패턴의 분류는 입 모양이 가로와 세로의 벌어진 형태에 의한 분류이므로 애니메이션의 입장에서는 단모음과 이중모음으로 <표 2>와 같이 분류할 필요가 있다.

<표 2> 입 모양에 따른 모음의 분류

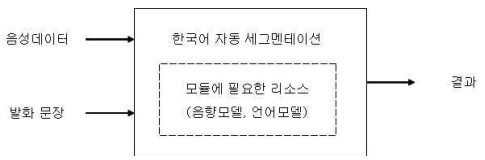
단모음	ㅏ(ㅑ), ㅔ(ㅖ), ㅗ(ㅛ), ㅜ(ㅠ), ㅡ, ㅣ, ㅐ, ㅒ
이중 모음	ㅘ, ㅙ, ㅚ, ㅞ, ㅝ, ㅟ, ㅢ

자음과 모음에 속하지는 않지만, 입 모양 분류의 관점에서 볼 때, 묵음에 대한 고려 또한 필요하다. 묵음은 입술소리와 조합되는 발음에서도 발음 초기 또는 발음의 마지막에서 사용되는 입 모양으로 발음하지 않은 상태의 무표정한 얼굴을 묵음으로 간주한다[13][14].

3.1.2 한국어 자동 세그먼테이션

음성 인식에서는 음소와 같은 작은 음성단위로 구분하는 것이 인식부에서의 연산처리의 효율 증가와 표준패턴을 불특정화자에 대비하여 용이하게 작성하고 수정할 수 있는 시스템을 구축하는데 중요하다. 이와 같이 연속된 음성신호를 동일한 음운특성을 갖는다고 생각되는 소구간으로 나누어 주는 것을 세그먼테이션이라고 한다.

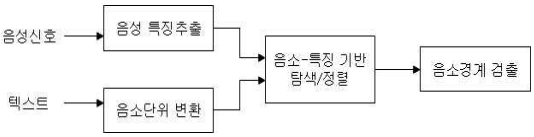
본 논문에서 제안한 한국어 자동 세그먼테이션(KAS)은 한국어로 발화된 음성데이터에 대한 음소단위 시간정보를 추출하는 기법이다. 이 모듈은 한국어로 발음된 음성 데이터와 이 음성데이터의 텍스트 정보를 입력으로 받아 음소 단위 발화 시간 및 음절 단위 발화 시간 정보를 출력하도록 구성하였다. (그림 2)는 한국어 자동 세그먼테이션 모듈 구성도를 나타낸다.



(그림 2) 한국어 자동 세그먼테이션 모듈

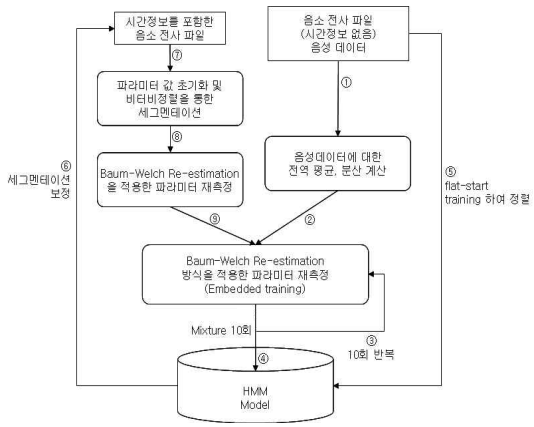
본 논문에서 사용한 알고리즘은 먼저 언어학적 정보를 사용하는 HMM 방식을 이용하여 음성을 분할한 뒤, 분할 결과를 보정하는 2단계 자동 음성 분할 기법을 제안한다. 첫 번째 단계에서는 모노폰 및 다양한 음운현상을 고려한 트라이폰을 분할단위로 설정한 후 HMM 모델을 생성하여 자동 음성 분할을 수행하고, 두 번째 단계에서는 유성음/무성음/묵음 특징을 추출 후 분할 결과를 보정하여 음성 분할 결과를 향상시킨다[7][8][9][10].

언어학적 정보를 사용하는 음성 분할 과정은 발화자가 음소의 빈도수를 고려한 문장을 발화하여 생성된 음성 데이터에서 각각의 음소들에 대한 통계적 모델을 생성한다. (그림 3)는 HMM 모델을 이용한 음성 분할 방식으로 음소의 경계를 검출하는 흐름도이다.



(그림 3) HMM을 이용한 자동 음소 경계 검출

본 논문에서는 HMM 모델 생성과정으로 2단계를 거쳐 처리하는 방식을 (그림 4)와 같이 제안한다. 첫 번째 단계에서는 Flat Start방식을 사용하며 1차 HMM 모델을 생성하고 Bootstrap 방식을 통한 두 번째 단계에서 각 음소에 대한 HMM 최종 모델을 생성한다. 최종 생성된 모델을 기반으로 비터비 디코딩 과정을 통해 최종 음성 세그먼테이션 정보를 얻게 된다.



(그림 4) 2단계를 통한 HMM 모델 생성 과정

자동으로 분할된 음소는 분할 결과를 보정하는 과정이 필요한데, 유성음/무성음/묵음의 구간을 분리하는 특징 파라미터를 이용하여 경계 위치를 보정한다[11][12].

3.1.3 립싱크 입 모양의 대체 알고리즘

가장 자연스러운 립싱크 표현을 위해서는 물론 17개 모음과 무표정의 묵음발음의 입모양 등 18개의 얼굴 모양을 모두 모델링 하는 것이 바람직하다. 그러나 모델링 되지 않은 입 모양이

있을 경우 모델링 된 입 모양만으로 최대한 자연스러운 립싱크가 이루어지도록 입 모양 대체 순서가 필요하다. 본 연구에서 개발한 프로그램에서는 유저가 설정하지 않은 입 모양의 경우 대체 ID를 정하여 이미 설정된 입 모양으로 대체하도록 하였다. <표 3>은 입 모양 대체 구현에 사용된 대체 순위 표이다.

<표 3> 입 모양 대체 순위

ID	기호	모음	대체 순위 ID																
1	axe	(ㅏ)	1	4	15	2	5	6	16	3	7	9	17	8	10	12	13	14	11
2	jac	(ㅑ)	2	1	5	16	4	15	3	6	8	10	7	9	17	13	14	12	11
3	wac	(ㅘ)	3	1	6	17	7	9	4	2	8	10	5	15	16	13	14	12	11
4	eoc	(ㅓ)	4	5	1	6	15	2	16	17	9	7	10	8	3	13	14	12	11
5	jeoc	(ㅕ)	5	4	2	6	3	1	15	16	8	10	7	9	17	13	14	12	11
6	woc	(ㅝ)	6	3	17	4	9	5	10	8	7	13	16	2	1	15	14	12	11
7	oxc	(ㅗ)	7	8	9	10	6	17	3	4	15	5	16	2	1	13	14	12	1
8	joc	(ㅛ)	8	7	10	9	17	6	3	5	16	15	4	13	2	11	14	12	1
9	uxc	(ㅜ)	9	10	7	6	8	3	17	15	4	13	5	16	11	14	12	2	1
10	juc	(ㅠ)	10	9	8	7	6	17	13	3	5	16	15	4	2	11	14	12	1
11	euc	(ㅡ)	11	14	13	12	9	17	10	15	16	6	7	8	5	4	3	1	2
12	ixc	(ㅣ)	12	14	13	11	15	16	9	8	10	7	17	6	5	4	3	2	1
13	wic	(ㅟ)	13	14	12	11	9	10	8	7	17	6	16	15	3	5	4	2	1
14	euic	(ㅢ)	14	13	12	11	16	15	17	10	9	6	8	7	3	5	4	2	1
15	aec	(ㅐ, ㅔ)	15	16	17	3	6	4	1	5	2	7	9	8	10	13	14	12	11
16	jec	(ㅖ)	16	17	15	3	6	4	2	5	8	10	7	9	1	13	14	12	11
17	wec	(ㅞ, ㅙ, ㅚ)	17	16	6	15	3	13	14	10	9	5	4	8	7	11	12	2	1

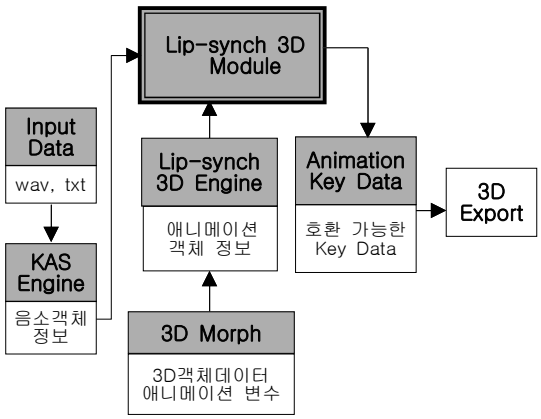
단, 3D 모델링의 발음은 'ㅏ, ㅔ, ㅣ, ㅗ, ㅜ, ㅡ, ㅓ'와 묵음인 default 발음의 최소 8개의 오브젝트를 가지는 것을 권장한다.

본 논문에서 제안한 오브젝트의 대체 순서는 단모음의 경우, 포먼트 주파수에 따른 입 모양의 유사도 정도를 이용하여 대체 순위를 작성하였으며, 이중모음의 경우에는 두 단모음의 조합으로 립싱크 애니메이션을 적용하게 된다.

3.2 한국어 기반 3D 립싱크 애니메이션 시스템

본 논문에서 개발된 립싱크 애니메이션 시스템은 wav 파일로 구성된 음성 데이터와 이에 해당하는 텍스트 데이터가 한국어 음소 분할 시스템인 KAS엔진의 입력으로 사용되어 분할된 음소 데이터를 추출하게 된다. 더불어 3D 모델링 툴인 Lightwave에서 제작된 얼굴 모델은 분할된 음소 데이터와 매핑되며 선형 보간법을 사용한 립싱크 모션 알고리즘에 의해 3D 애니메이션 키 데이터가 생성된다. 이 때 분할된 음소 데이터에 비하여 얼굴 모델이 모두 제작되지 않은 경우에는 입 모양 대체 알고리즘에 의하여 자동 대체되며 이중모음의 경우에는 제작된 단모음의 연결로 실행하게 된다. 또한, 이 시스템은 자동 분할된 음소의 동기화 데이터와 상용 3D 모델링 툴에서 만들어진 3D 데이터를 Import하여 애니메이션 키 데이터를 생성하고 Export기능도 추가되어 기존 3D 툴과의 데이터 호환성을 높였다.

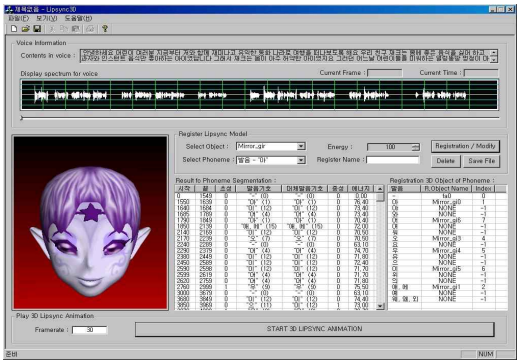
다음의 (그림 5)는 본 논문에서 제안된 립싱크 애니메이션 시스템의 전체적인 개요도로, 음소 인식 및 음소 분할을 담당하는 KAS(Korean Auto Segmentation) 엔진과 렌더링 및 애니메이션을 담당하는 Lip-synch3D 엔진으로 구성되어 있음을 볼 수 있다.



(그림 5) 3D 립싱크 애니메이션 시스템 개요도

(그림 6)은 최종적으로 구현된 한국어 음소 전용 3D 립싱크 애니메이션 시스템의 실행화면이다. 제안된 시스템은 입력된 음성의 텍스트, 오디오 파일에 해당하는 스펙트럼과 이에 따른 슬라이드 바, 그리고 현재 립싱크 되고 있는 프레임과 시간을 표시할 수 있고 음소 분할된 결과는 발화 시간과 초, 종성의 유무, 음성의 에너지 값을 나타낸다. 3D 파일을 Import하여 립싱크에 사용할 모델을 가져온 후 17개 모음과 묵음으로 구분된 음소와 매핑하게 된다. 18개의 음소 중 매핑되지 않은 음소는 앞에서 기술한 대체 ID 테이블에 의하여 가까운 입모양을 자동으

로 찾게 된다. 음소와 입모양 오브젝트의 매핑이 끝나면 시스템의 성능과 요구하는 애니메이션의 수준을 고려하여 프레임 수를 조절하여 립싱크 애니메이션을 실행한다.



(그림 6) 제안된 시스템의 실행 화면

4. 성능 분석 및 비교 평가

4.1 음성 인식 및 음소 분할 성능 평가

본 논문에서는 한국어의 음성 인식 및 자동 음소 분할을 위하여 HMM 알고리즘을 사용하였다. 시스템에 적용한 HMM을 이용한 자동 음소 분할 및 보정 알고리즘의 결과를 살펴보면, 모노폰으로 훈련된 HMM 모델을 사용한 음소분할 결과보다 단순화된 트라이폰 모델을 사용한 음소 분할 결과가 최대 8%까지 성능이 향상된 것을 볼 수 있고 보정 후에는 최대 100%까지 향상되었다. 특히 유성음과 무성음의 경계와 같이 경계 구분이 뚜렷한 음운환경의 경우는 다른 음운 환경의 경계보다 더 정밀한 결과를 도출할 수 있었다.

이 결과에 유성음/무성음/묵음 구간 분류 알고리즘을 이용하여 보정한 결과를 살펴보면 구간의 구분이 용이한 무성음과 유성음의 경계와 묵음과 음성의 경계(Silence+Speech)에 보정 결과는 크게 향상되었으나, 유성음과 유성음 및 무성음과 무성음의 경계에서는 보정 결과 향상 정도가 미흡하였다.

또한, 자음과 모음간의 분할 결과를 살펴보면, 모노폰으로 훈련된 HMM 모델을 사용한 음소분할 결과보다 단순화된 트라이폰 모델을 사용한 경우 4%~6%의 성능 향상을 가져 왔으며 보정 후에는 최대 90%까지 향상되었다. 그러나 모음+모음의 분할에서는 유성모음 또는 무성모음간의 분할 성능이 상대적으로 미흡하여 결과 보정 후 분할 결과가 떨어짐을 알 수 있었다.

다음의 <표 4>는 본 논문에서 제안한 자동 음소 분할 및 보정 알고리즘과 다른 알고리즘을 적용한 후 결과를 비교한 표이며 오차 범위는 10ms로 규정하였다.

<표 4> 음성 인식 및 음소 분할 알고리즘 비교

알고리즘 / 음운현상	모노폰	단순화된 트라이폰	제안한 알고리즘
Silence+Speech	92%	93%	100%
무성음+무성음	72%	80%	81%
무성음+유성음	76%	80%	91%
유성음+무성음	77%	83%	93%
유성음+유성음	72%	76%	81%
자음+자음	76%	80%	90%
자음+모음	73%	79%	86%
모음+자음	74%	79%	85%
모음+모음	74%	78%	80%

4.2 기존의 한국어 기반 립싱크 시스템과 비교

본 절에서는 다음의 <표 5>와 같이 기존에 연구된 한국어 기반 립싱크 애니메이션 시스템과 본 논문에서 제안한 시스템과의 비교를 통해 시스템의 성능을 평가한다. 기존 연구는 편의상 '기법 1'("실시간 Lip-Synch 시스템을 위한 음소 인식, 음소 합병 및 입술모양 보정 알고리즘", 강용성)[15], '기법 2'("음성인식을 이용한 실시간 립싱크 애니메이션 시스템 개발", 이형호)[16], '기법 3'("RBF를 이용한 자연스러운 립싱크 구현", 김명곤)[17], '기법 4'("3차원 모델의 움직임 보정을 통한 Speech Animation에 관한 연구", 고락신)[18]로 구분하여 비교한다.

음소 인식 및 분할 알고리즘은 '기법 1'의 경우에는 HMM에 비해 음소 분할 속도가 빠른 SVM을 사용하였고 '기법 2'의 경우 HMM 알고리즘을 사용하였다. '기법 3'과 '기법 4'의 경우에는 음성의 입력 없이 텍스트의 입력으로 음소의 분할을 실행하였다. 본 논문에서 제안된 시스템의 경우 HMM 알고리즘을 2회 실행하는 방법으로 음소 인식 및 분할의 정확성을 높였다.

입모양에 따른 음소 병합의 측면에서 보면 '기

법 1'과 '기법 2'의 경우는 7개의 단모음을 기준으로 분류하였으며 '기법 1'은 단모음을 포먼트 주파수를 이용하여 3개의 클래스로 재분류하였다. '기법 3'은 33개의 기본음절을 구성하여 음소를 병합하였으며 '기법 4'는 음소를 병합하지 않고 초, 중, 종성의 조합형 매핑 테이블을 이용하였다. 그에 비하여 제안된 시스템은 최대 17개의 단모음 및 이중모음을 기준으로 음소를 병합하였으며 기본적으로 'ㅏ', 'ㅖ', 'ㅣ', 'ㅗ', 'ㅜ', 'ㅡ', 'ㅕ'와 입술소리 및 묵음을 처리할 수 있는 입모양의 8개만을 선택하여도 애니메이션을 제작하는데 있어 기존 시스템에 비해 향상된 성능을 보였다.

<표 5> 한국어 립싱크 애니메이션 시스템 비교

구분	기법 1	기법 2	기법 3	기법 4	제안된 기법
입력	음성	음성	텍스트	텍스트	음성, 텍스트
음소인식 알고리즘	SVM	HMM	음성입력 없음	음성입력 없음	2단계의 HMM
입모양에 따른 음소병합	3개의 클래스로 나눔	ㅏ, ㅖ, ㅣ, ㅗ, ㅜ, ㅡ, ㅕ	33개의 비즘으로 분류	조합형 매핑 테이블 이용	최대 17개의 모음 및 묵음
보간방법	선형	선형	쌍선형	베지어 커브	선형
애니메이션	2D	2D	2D	3D	3D
입술소리 처리	불가능	불가능	가능	가능	가능
이중모음 처리	불가능	불가능	가능	가능	가능
비고	• 실시간 애니메이션 • 포먼트를 이용한 음소분할 보정	• 실시간 애니메이션	• 비즘별 아바타에 우세함수를 적용하여 애니메이션	• 마커를 이용하여 모션데이터 취득	• 포먼트를 이용한 입모양 대체 알고리즘 사용 • Export 가능

한국어 기반의 립싱크에서 고려해야할 사항 중 입술소리와 이중모음의 처리는 '기법 1'과 '기법 2'의 경우 입술소리와 이중모음 모두 처리할 수 없었으며 '기법 3'과 '기법 4'에서는 입술소리와 이중모음을 모두 고려하였다. 본 시스템은 초성과 종성의 존재 여부를 구분하여 입술소리를 처리하였으며 단모음의 조합으로 이중모음을 처리한 결과 좀 더 세밀한 립싱크를 구현하였다.

4.3 영어기반의 립싱크 시스템과의 비교

영어 기반 립싱크 발음의 구분은 한국어 발음 체계와는 다른 구분법으로 나뉘어 있으며, 모음 위주로 발음 현상이 일어나는 한국어 구조와는 다른 아예 발음 할 수 없거나 표현 방식이 어려운 /f/,/th/,/sh/,/v/,/l/ 등과 같은 발음이 추가되어 있다.

상용화된 영어 위주의 립싱크 툴과 본 논문에서 개발된 한국어 전용 립싱크 툴을 간단한 문구의 한국어 음성 및 문장을 이용하여 비교, 검증하였다. 영어 립싱크 툴에서는 'ㅏ'발음은 표현될 수 있지만, 한국어 발음 'ㅕ', 'ㅖ', 'ㅗ' 등은 발화 타임만 찾아 낼 뿐 비슷한 발음 Morph 얼굴 모델로 연결시키질 못한다. 오히려 '하', '세' 발음에서 'ㅎ, ㅅ'의 영향으로 [sh], [s] 등이나 기타 엉뚱한 발음으로 구분해서 입술 모양을 만들어 낸다.

개발된 한국어 립싱크 시스템을 이용하면 "안녕하세요"를 "아 + ㄴ + 여 + 아 + ㅅ + 에 + 오 "와 같이 모음 위주의 발화 타임과 Morph 데이터를 얻을 수 있다.

영어 위주 립싱크 툴이나 한국어 위주 립싱크 툴 모두 시스템에 적용 후 3D 툴에서 2차, 3차 편집이 이루어져야 하지만, 모음 위주로 입술 움직임이 대부분 결정되는 한국어 말하기 특성상 한국어 전용 립싱크 툴의 사용이 제작시간을 현저히 줄일 수 있고 제작 방법에 더 효과적이라 할 수 있겠다.

5. 결론 및 향후 과제

본 논문에서는 한국어를 기반으로 발화된 음성 데이터와 이에 해당하는 텍스트 정보에서 한국어 음소를 분할하고 동기화 정보를 추출하여 보다 정확하고 자연스러운 입술 애니메이션 키 데이터를 생성하는 3D 모델링 데이터와 각종 3D 툴 간에 호환 가능한 립싱크 애니메이션 프로그램을 구현하였다.

3D 립싱크 애니메이션 시스템을 구현하는데 있어서 기존 영어 중심으로 되어 있는 음소 분할 방법의 문제점을 제시하고 한국어 기반의 음성 인식과 자동 음소 분할 알고리즘을 제안하였

다. 또한 분할된 음소를 립싱크에 적당하도록 입모양을 중심으로 한 음소 셋으로 재구성하였으며, 포먼트 주파수를 이용한 입술모양 대체 알고리즘을 제안하여 립싱크에 사용될 모든 얼굴 모델이 없어도 자연스러운 애니메이션이 구성되도록 하였다. 그러나 애니메이션의 보간 방법이 선형보간법으로 한정되어 있어 2차식의 보간 방법에 비하여 연속성이 떨어져 경우에 따라 딱딱한 느낌이나 애니메이션이 끊기는 듯 한 현상이 나타났다.

본 논문의 향후 연구 과제로는 애니메이션 키 데이터를 생성하는 방법을 현재의 1차식의 선형 보간법에서 2, 3차식으로 다양화하여 연속성을 향상시키는 방법이 연구되어야 할 것이다.

참 고 문 헌

[1] 정문열, “3D 얼굴 애니메이션의 기술 동향”, 정보과학회지, 제21권, 제7호, pp.18-23, 2003.

[2] 이윤근, 박준, 김상훈, “음성인터페이스 기술”, 전자통신동향분석 제20권, 제5호, 2005.

[3] 정차균, “한국어 음성인식”, 전자공학회지, 제24권 제9호, pp.1071-1078, 1997.

[4] 홍광석, “음성정보처리 기술”, 성균관대학교 휴먼컴퓨터연구실.

[5] http://www.garycmartin.com/mouth_shapes.html “Preston Blair phoneme series”.

[6] http://www.thirdwishsoftware.com/magpiepro.html “Third Wish Software & Animation”.

[7] Brugnara, F. and et al., “Automatic segmentation and labeling of speech based on hIDden Markov model”, Speech communication, vol.12, pp.357-370, 1993.

[8] Svendsen, T. and Siong, F.K., “On the Automatic segmentation of speech signal”, Proc. of IEEE ICASSP 87, pp.77-80, Apr., 1987.

[9] Rabiner, L.R. and Sambur, M.R., “An algorithm for determining the endpoints of isolated utterances”, The Bell system technical journal, vol.54, No.2, pp.297-315, Feb., 1995.

[10] Sarikaya, R. and John, H.L., “Robust speech activity detection in the presence of noise”, Proc. of ICSLP 98, vol.4, pp.1455-1458, 1998.

[11] 김무중, 권철홍, “자동 음성 분할 시스템의 성능 향상”, 한국음향학회:학술발표대회 논문집 제21권, 제1(s)호, 2002.

[12] 김무중, 권철홍, “HMM 및 보정 알고리즘을 이용한 자동 음성 분할 시스템”, 음성과학, vol.9, no.4, pp.265-274, 2002.

[13] 조성문, “현대 국어의 모음 체계에 대한 음향음성학적인 연구”, 한국언어문화학회:한국언어문화, 제24집, pp.427-441, 2003.

[14] 김지성, “한글에 부합한 실시간 얼굴 애니메이션”, 명지대학교 석사학위논문, 1999.

[15] 강용성, “실시간 Lip-synch 시스템을 위한 음소 인식, 음소 합병 및 입술모양 보정 알고리즘”, 고려대학교 석사학위논문, 2002.

[16] 이형호, “음성인식을 이용한 실시간 립싱크 애니메이션 시스템 개발”, 공주대학교 석사학위논문, 2001.

[17] 김명곤, “RBF를 이용한 자연스러운 립싱크 구현”, 전남대학교 석사학위논문, 2003.

[18] 고락신, “3차원 모델의 움직임 보정을 통한 Speech Animation에 관한 연구”, 중앙대학교 석사학위논문, 2004.

정 일 홍

1993년: 애리조나 주립대학 컴퓨터공학과 졸업 (공학석사)

1998년: 애리조나 주립대학 컴퓨터공학과 졸업 (공학박사)

1998년～현 재 : 대전대학교 컴퓨터공학과 부교수

관심분야 : 컴퓨터 그래픽스, 멀티미디어, 애니메이션, 가상현실, 영상처리, 디지털 콘텐츠 등

김 은 지

1999년: 대전대학교 대학원 컴퓨터공학과 졸업 (공학석사)

2006년: 대전대학교 대학원 컴퓨터공학과 졸업 (공학박사)

2002년～2005년 : 정보통신진흥원 지원 프로그래밍 전문 교수

관심분야 : 컴퓨터 그래픽스, 애니메이션, 디지털 콘텐츠, 가상현실 등