Proceedings of KFIS Spring Conference 2006 Vol. 16, No. 1.

발화속도를 고려한 3차원 얼굴 모형의 퍼지 모델 기반 립싱크 구현

Human-like Fuzzy Lip Synchronization of 3D Facial Model Based on Speech Speed

박종률[1], 최철완[2], 박민용[3]

[1] 서울시 서대문구 연세대학교 전기전자공학부
E-mail: beluga21@yeics.yonsei.ac.kr
[2] 서울시 서대문구 연세대학교 전기전자공학부
E-mail: cwchoi55@empal.com
[3] 서울시 서대문구 연세대학교 전기전자공학부
E-mail: mignpark@yonsei.ac.kr

요 약

본 논문에서는 음성 속도를 고려한 새로운 립싱크 방법에 대해서 제안한다. 실험을 통해 구축한 데이터베이스로부터 음성속도와 입모양 및 크기와의 관계를 퍼지 알고리즘을 이용하여 정립하였다. 기존 립싱크 방법은 음성 속도를 고려하지 않기 때문에 말의 속도와 상관없이 일정한 입술의 모양과 크기를 보여준다. 본 논문에서 제안한 방법은 음성 속도와 입술 모양의 관계를 적용하여 보다 인간에 근접한 립싱크의 구현이 가능하다. 또한 퍼지 이론을 사용함으로써 수치적으로 정확하게 표현할 수 없는 애매한 입 크기와 모양의 변화를 모델링 할 수 있다. 이를 증명하기 위해 제안된 립싱크 알고리즘과 기존의 방법을 비교하고 3차원 그래픽 플랫폼을 제작하여 실제 응용 프로그램에 적용한다.

Key Words : Lip-synch, Fuzzy modeling, Speech speed, Table look-up, 3D modeling

1. 서 론

얼굴을 통한 의사소통은 인간의 감정과 의도를 표현하는 매우 효과적이고 자연스러우면서도 즉각적인 방법 중 하나이다. 표정과 머리의 움직임 등의 다양한 의사 표현 중에서도 의미 전달에 있어서 가장 명확한 방법은 음성 언어를 이용한 대화이다. 컴퓨터의 처리능력이 빠른 속도로 향상되면서 영상 출력장치를 이용하여 대화를 이용하는 인터페이스의 구현이 가능해졌고 최근에는 실시간 3차원 그래픽을 통하여 인간에 가까운 아바타를 재현한 인터페이스가 개발되고 있다.

3차원 아바타의 표정과 립싱크를 모델링하기 위해서는 보간법(interpolation), 매개변수법(parameterization), 의사 근육 모델링(pseudo-muscle modeling) 등의 여러 가지 방법들이 이용된다[3]. 보간법은 두 키 프레임(key frame) 사이의 프레임을 생성하는 가장 빠르고 상대적으로 간편한 방법이기 때문에 다양한 분야에서 사용되고 있다. 정교하게 재현된 모형을 사용하여 사실적인 영상을 얻을 수 있는 방법이지만 미리 정해진 프레임 사이의 보간을 통해서 중간 과정의 영상을 만들어내기 때문에 얼굴이 표현할 수 있는 다양한 요소들의 조합을 통한 동적인 변화를 표현하기 어렵다는 단점이 있다. 그러나 실제 인간이 표현 가능한 표정과 입주위의 움직임 중 의사소통에 명확한 의미를 가지는 것은 제한되어 있으며 이들의 부분적인 조합을 통해 빠르면서도 세부적인 묘사가 동시에 가능하기 때문에 보다 실제에 가까운 형태의 모델을 구현할 수 있다.

일반적으로 적용되고 있는 립싱크 기술은 주요 모음을 발음할 때의 입 주의의 형태를 모음 음소와 일대일 대응시키는 방법을 사용한다. 일부 애니메이션 등에서는 모션 캡처(motion capture)를 통한 자연스러운 동작 구현이 가능

하지만 모든 발음의 전환에 대응하는 방대한 양의 자료가 필요하기 때문에 아직까지는 실시간 처리에 적합하지 않다. 각 발음과 그에 해당하는 입의 형태를 단순 대응 하는 것은 처리 속도와 구현의 용이성에 커다란 이점을 제공하지만 단순화된 움직임으로 부자연스러운 동작을 표현하게 된다.

본 논문에서는 특징점(feature point) 검출을 통해 화자의 발화 속도와 그에 따른 개구부(mouth opening)의 크기 변화의 관계를 퍼지 이론을 사용하여 밝히고 이를 TTS (text-to-speech) 기반의 실시간 립싱크 구현에 적용하는 방법을 제안한다. 입 주변에 부착한 표식의 위치 변화를 검출하여 화자의 말하는 속도에 따른 개구부 크기 변화를 자료화하여 테이블룩업(table look-up) 방법을 사용한 퍼지 모델링을 통해 발화 속도와 개구부 크기와의 관계를 정립한다. 세밀한 3차원 모델을 간편하고 빠르게 사용할 수 있는 보간법을 기반으로, 보간법의 취약점인 제한된 움직임을 통한 부자연스러운 립싱크를 인간의 움직임에 가깝도록 발화 속도마다 다른 개구부 크기를 사용하여 보완한다. 마지막으로 본 논문에서 제안한 방법을 증명하기 위하여 기존의 방법과 비교하고 3차원 그래픽 플랫폼을 제작하여 실제로 립싱크를 구현한다.

1. 특징점의 추출 및 레이블링

발화 속도에 따른 개구부의 크기 변화를 측정하기 위하여 입 주변의 4개의 특징점을 선정하여 그 위치에 녹색과 파란색의 원형 표식을 부착한다. 피실험자의 발화 모습을 동영상으로 촬영한 후 각 프레임마다의 영상을 처리하여 표지의 좌표를 얻는다. 얻고자 하는 표식을 쉽게 검출하기 위하여 다음과 같이 영상에서 각 화소의 색상을 정규화한다.

$$Normal\_R(x,y) = 255 \times \frac{Img\_R(x,y)}{Img\_R(x,y)+Img\_G(x,y)+Img\_B(x,y)}$$

$$Normal\_G(x,y) = 255 \times \frac{Img\_G(x,y)}{Img\_R(x,y)+Img\_G(x,y)+Img\_B(x,y)}$$

$$Normal\_B(x,y) = 255 \times \frac{Img\_B(x,y)}{Img\_R(x,y)+Img\_G(x,y)+Img\_B(x,y)}$$

정규화 후 녹색과 파란색 성분을 추출하기 위하여 다음의 과정을 적용한다.

IF ($Normal\_G(x,y) > Normal\_R(x,y)$ and $Normal\_G(x,y) > Normal\_B(x,y)$)
$Img\_G(x,y) = 255$

ELSE
$Img\_G(x,y) = 0$

IF ($Normal\_B(x,y) > Normal\_R(x,y)$ and $Normal\_B(x,y) > Normal\_G(x,y)$)
$Img\_B(x,y) = 255$

ELSE
$Img\_B(x,y) = 0$

이 과정을 거치면 표지가 위치한 곳은 255, 아닌 곳은 0의 값을 갖는 영상을 얻게 된다.

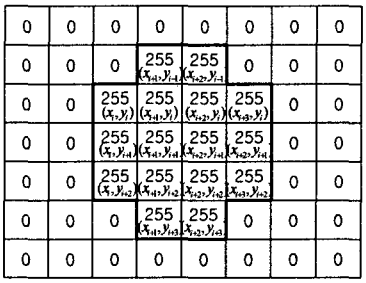


그림 1. 색상 추출 후 두 단계의 값만을 가지도록 처리된 영상의 구조

이 영상에서 표지의 중심점을 계산하기 위하여 혼합 인접(m-adjacency)법을 이용하여 인근의 녹색 혹은 파란색 점들을 레이블링(labeling)한 후 각 그룹에 소속된 점들의 좌표를 평균하여 중심점의 위치를 검출한다.

$$x_{center} = \frac{x_1 + \cdots + x_n}{n}$$

$$y_{center} = \frac{y_1 + \cdots + y_n}{n}$$

위와 같이 얻은 중심점이 정수가 아닐 경우 가장 가까운 정수로 변환하여 각 표식의 최종적인 위치를 얻는다.

2. 퍼지 모델링

2.1 퍼지 모델

퍼지 시스템의 구조는 퍼지화(fuzzification), 비퍼지화(defuzzification), 퍼지 제어 규칙(fuzzy control rule) 및 퍼지 추론(fuzzy inference)의 세 부분으로 크게 나눌 수 있다. 다음과 같은 퍼지 규칙을 고려 할 때,

$$Rule\ l:\ IF\ x_1 is\ A_1^l,\ \ldots,\ \text{and}\ x_n is\ A_n^l$$
$$THEN\ y^l is\ B^l$$

퍼지 추론 작업은 입력 공간에서의 퍼지 집합 A′을 출력 공간에서의 퍼지 집합 B′에 대응시키는 작업이다. 프로덕트 퍼지 추론기는 다음과 같다.

$$\mu_{B'}(y) = \max_l \left[\sup_{x \in U} \left(\mu_{A'}(x) \prod_i \mu_{A_i^l}(x_i) \mu_{B^l}(y) \right) \right]$$

실변수를 퍼지 집합으로 바꾸는 퍼지화기와 퍼지 추론의 결과인 퍼지 집합을 실변수로 바꾸는 비 퍼지화기는 각각 싱글톤 퍼지화기와 중심 평균법에 의한 비퍼지화 방법을 쓴다.

$$\mu_{A'}(x) = \begin{cases} 1 & \text{if } x = x^* \\ 0 & \text{otherwise} \end{cases}, \quad y^* = \frac{\sum_{l=1}^{M} y^l w_l}{\sum_{l=1}^{M} w_l}$$

이러한 구조로 이루어진 퍼지 시스템은 다음과 같이 실수를 입출력으로 하는 비선형 함수가 된다.

$$f(x) = \frac{\sum_{l=1}^{M} y^l (\Pi_i \mu_{A_i^l}(x_i))}{\sum_{l=1}^{M} (\Pi_i \mu_{A_i^l}(x_i))}$$

여기서 y^l은 퍼지 제어 규칙에서 후건부 퍼지 집합 B^l의 중심값이다.

2.2 테이블룩업(table look-up) 방법을 이용한 퍼지 시스템 설계

실험적으로 얻은 입출력 쌍으로부터 퍼지 시스템을 설계하기 위하여 테이블룩업 방법을 사용한다. 다음과 같이 주어진 N개의 입출력 쌍

$$(x_0^p y_0^p),\ p = 1, 2, \ldots, N$$
$$x_o^p \in U = [\alpha_1, \beta_1] \times \cdots \times \subset R^n$$
$$y_0^n \in V = [\alpha_y, \beta_y] \subset R$$

로부터 입출력 공간을 모두 포함하는 N^i개의 퍼지 집합 $A_i^j (j = 1, 2, \ldots, N_i)$을 정의한다. 정의된 퍼지 집합으로부터 각 입출력 쌍에 대하여 입력과 출력의 소속값 $\mu_{A_i^j}(x_{0i}^p)$와 $\mu_{B^l}(y_0^p)$를 얻어 그 결과로 다음과 같은 퍼지 IF-THEN 규칙을 생성한다.

$$IF\ x_1 is\ A_i^{j*}\ \text{and}\ \cdots\ \text{and}\ x_n is\ A_n^{j*},\ THEN\ y is\ B^{l*}$$

이때 얻은 규칙들은 동일한 IF 부분을 갖더라도 일치하지 않는 THEN 파트를 가질 수 있으므로 이러한 경우에 발생하는 충돌을 제거하기 위하여 다음의 방법으로 각 규칙마다 차수를 부여한다.

$$D(rule) = \prod_{i=1}^{n} \mu_{A_i^{j*}}(x_{0i}^p) \mu_{B^{l*}}(y_0^p)$$

서로 충돌이 없는 모든 규칙들과, 충돌이 발생하는 규칙 중 가장 차수가 높은 규칙들로 최종적인 퍼지 기반 규칙을 구성한다. 이 규칙을 2.1에서 정의한 퍼지 시스템에 사용하여 퍼지 모델을 생성한다.

3. 실험 및 결과 고찰

얼굴 표면에 부착한 표식의 좌표를 검출하여 개구부 크기를 측정한다.

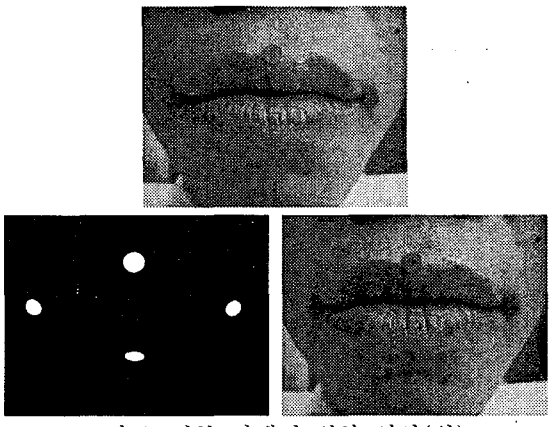
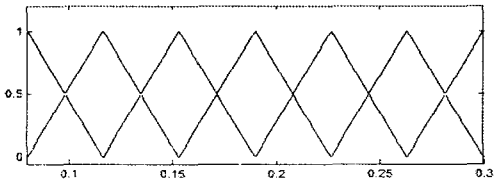


그림 2. 기본 상태의 원본 영상(위),
색상 추출 결과(좌)와 검출된 표식(우)

평상 대화 수준의 성량으로 주요 모음인 'ㅏ', 'ㅔ', 'ㅓ', 'ㅗ', 'ㅜ', 'ㅣ'를 정확히 발음할 때의 입모양을 기준으로 발화시의 개구부의 크기와의 비율을 결정한다. 이때 하나의 음절을 발음하는데 걸리는 시간의 평균값과 입술이 열린 폭의 비율을 퍼지 시스템의 입력과 출력으로 정의한다. 정의된 입력과 출력을 만족하는 시스템을 모델링하기 위해 테이블룩업 방법을 사용하여 실험을 통해 얻은 입출력 쌍으로부터 퍼지 기반 규칙을 생성한다. 소속 함수는 입력과 출력 각각 7개의 삼각형 함수로 정의한다.



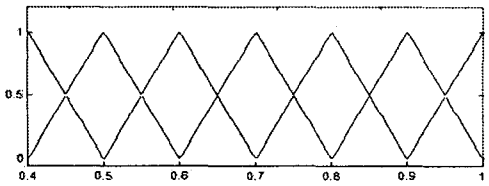


그림 3. 입력(위)과 출력(아래)에 대한 소속 함수

실험은 총 12명을 대상으로 이루어졌으며 피실험자가 '천천히', '보통', '빠르게'의 세 단계의 발화를 세 번씩 수행하여 그 평균값으로 획득한 자료를 사용하였다. 설계된 퍼지 시스템의 실제 입출력 그래프는 다음과 같다.

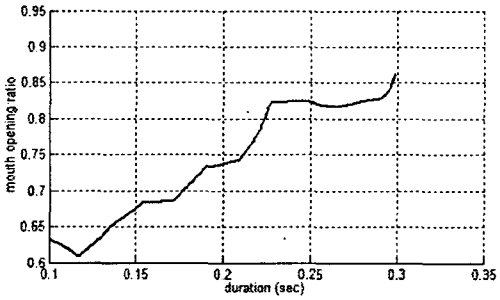


그림 4. 시스템 입출력 그래프

다음 그래프는 TTS 소프트웨어의 출력 자료를 제안한 시스템에 적용한 결과를 나타낸다. 범례에서 Line1은 구현한 퍼지 시스템을 적용하지 않은 기존 시스템의 것이며 Line3은 본 시스템을 적용한 결과이다. Line2는 TTS로 합성한 음성을 피실험자가 발화할 때의 자료로 본 논문에서 제안한 방법을 적용한 결과와의 차이를 관찰할 수 있다. 발화 시작부분의 오차가 크고 제안한 시스템과 다른 모습을 보여주는 부분들이 보이지만 기존의 방법에 비하여 실제의 입 움직임에 가까운 결과를 나타낸다.

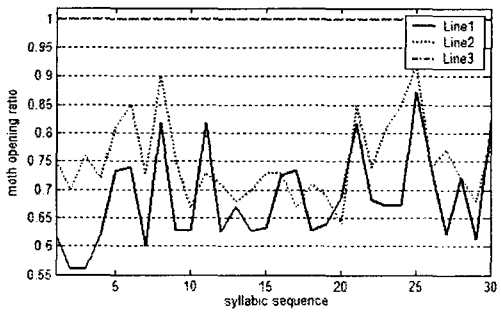


그림 5. 성능 비교

다음 그림은 실제 구현한 3차원 아바타 립싱크에 제안한 방법의 결과를 적용한 것이다.

그림 5. 기존의 방법(좌)과
발화 속도가 고려된 립싱크 방법(우)의 구현 비교

4. 결론 및 향후 과제

본 논문에서는 발화 속도와 발화자의 입 크기 변화와의 관계를 자료화하여 테이블룩업 방법을 사용해 퍼지 시스템을 설계하여, 보간법인 모핑(morphing) 기법을 이용해 온 기존의 립싱크 방법을 개선할 수 있는 방법을 제안하였다. 실험과 실제 립싱크 응용 프로그램에 적용한 결과 기존의 방법에 비하여 실제 인간의 발화 모습에 가까워졌음을 확인하였다.

더욱 인간과 흡사한 립싱크 방법을 개발하기 위해서는 개구부의 움직임을 결정하는 요소인 음량의 크기와 받침에 따른 혀의 위치에 대한 고려와 이러한 요소들을 립싱크 방법에 적용하는 연구가 필요할 것이다.

참 고 문 헌

[1] Sy-sen Tang, "Lip-Sync in Human face Animation Based on Video Analysis and Spline models", International Multimedia Modeling Conference, pp. 102-108, 2004.

[2] 경규민, "Automatic 3D Facial Movement detection from Mirror-reflected Multi-Image for Facial Expression Modeling", 정보 및 제어 심포지엄, pp. 113-115, 2005.

[3] J. Noh, "A survey of Facial modeling and Animation Techniques", USC Technical Report, pp. 99-105, 1998

[4] Li-Xin Wang, "A Course in Fuzzy Systems and Control", Prentice Hall, 1997.

[5] R. C. Gonzalez and R. E. Woods, "Digital Image Processing," Second Edition. Prentice Hall, 2002.