조음 기관의 시각화를 이용한 음성 동기화 애니메이션

이성진[1], 김익재[2], 고형석[3]
서울대학교[1 2 3]
{sjlee[1], ijkim[2], ko[3]}@graphics.snu.ac.kr

Speech Animation by Visualizing the Organs of Articulation

Sung Jin Lee[1], Ig Jae Kim[2] , Hyeong Seok Ko [3]
Seoul National University[1 2 3]

요 약

본 논문에서는 음성에 따른 얼굴 애니메이션을 사실적으로 표현하기 위한 조음기관(혀,성대 등)의 움직임을 시각화하는 방법을 제시한다. 이를 위해서, 음성에 따른 얼굴 애니메이션을 위한 말뭉치(Corpus)를 생성하고, 생성된 말뭉치에 대해서 음소 단위의 분석(Phoneme alignment) 처리를 한 후, 각 음소에 따른 조음기관의 움직임을 생성한다. 본 논문에서는 조음기관의 움직임 생성을 위해서 얼굴 애니메이션 처리에서 널리 사용되고 있는 기저 모델 기반 형태 혼합 보간 기법(Blend shape Interpolation)을 사용하였다. 그리고 이를 통하여 프레임/키프레임 기반 움직임 생성 사용자 인터페이스를 구축하였다. 구축된 인터페이스를 통해 언어치료사가 직접 각 음소 별 조음기관의 정확한 모션 데이터를 생성토록 한다. 획득된 모션 데이터를 기반으로 각 음소 별 조음기관의 3 차원 기본 기저를 모델링하고, 새롭게 입력된 음소 시퀀스(phoneme sequence)에 대해서 동기화된 3 차원 조음기관의 움직임을 생성한다. 이를 통해 자연스러운 3 차원 얼굴 애니메이션에 적용하여 얼굴과 동기화된 조음 기관의 움직임을 만들어 낼 수 있다.

Keyword : Lipsynch, organs of articulation, facial animation, blend shape

1. 서론

단순히 감정을 표현하는 얼굴의 표정 움직임과는 달리 음성에 동기화된 얼굴 애니메이션을 좀 더 사실적으로 만들기 위해서는 음성에 따른 입술의 움직임과 입술과 동기화된 조음 기관의 순간적인 움직임을 나타내는 것이 필요 하다. 그러나 캡쳐(capture) 할 수 없을 정도로 빠르게 움직이는 입술과 입 속의 조음 기관을 표현하는데 한계가 있다. 특히, 실제 조음 기관의 움직임은 마커를 붙일 수 없기 때문에, 기존의 모션 캡쳐 기법[4][5]으로는 조음 기관의 움직임 데이터를 만들어 낼 수 없다.

본 논문에서는 이러한 조음 기관의 움직임을 생성해 내는 거에 초점을 맞췄으며, 새로운 음성이 들어 올 때 동기화된 움직임을 만들어 내는 것이 가능하도록 하였다.

먼저, 데이터 수집을 위해, 각 음소 단위로 매개변수 값을 조절하여 저장할 수 있는 'J&T 컨트롤 인터페이스'를 제작하였다. 이 인터페이스는 조음 기관의 기저를 제어할 수 있는 매개변수를 조절함으로써 조음 기관의 변형된 형태를 시각적으로 보고 수정할 수 있도록 구현되었다. 이 인터페이스를 이용하여 실제 음성에 동기화된 초기값(Initial Guess)이 주어졌을 때, 언어치료 전문가들이 각 음소에서의 조음 기관의 변형 형태를 보면서 초기값(Initial Guess)을 수정하여 실제 조음기관의 모양과 비슷한 형태를 나타내는 데이터 값을 저장할 수 있게 하였다.

이렇게 저장된 데이터를 이용하여, 새로운 음성이 들어 올 때 그것과 동기화된 조음 기관의 움직임을 만들어 낼 수 있었다. 본 논문에서는 2 차원 비디오 기반인 MMM[1]을 3 차원으로 확장한 후, 조음 기관 모델에 적용하여, 음성에 동기화된 조음 기관의 움직임을 얻었다. 이를 구현하기 위해, 전체 말뭉치(corpus)에 대해 얻어진 데이터의 평균과 분산 값을 구하였다. 이렇게 구해진 평균과 분산 값을 이용하여 기존의 말뭉치(corpus)에 없는 새로운 입력 음성이 들어 올 때, 그것에 동기화된 음소에 대한 연속된 매개 변수 값을 얻을 수 있었다. 연속된 음소가 있고 그것의 평균 값을 할당하여 이산적으로 표현된 연속된 매개 변수 값들이 존재 할 때, 그 값들 간의 연결을 부드럽게 만들어주기 위해 규칙화 문제(regularization problem)[2][3]을 각 프레임 단위로 풀어내어 매개변수 궤도(parameter trajectory)를 얻었다. 궤도 합성 문제(trajectory synthesis problem)를 통해 얻은 각 프레임에 대한 매개변수 값은 자음을 발음할 때 나타나게 되는 정적인 조음 기관의 형태를 기저로 갖는 형태 혼합 기법의 매개변수에 그대로 적용하여, 새로 입력된 음성에 동기화된 조음 기관의 움직임을 만들어 낼 수 있다.

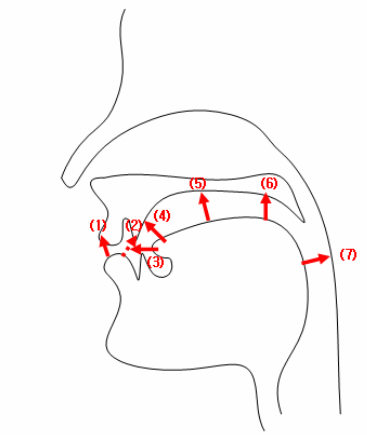
본 연구의 목적은 새로운 음성에 동기화된 조음 기관의 움직임을 생성해 내는 것이다. 실제 촬영하기 어려운 조음 기관의 움직임 데이터를 'J&T 컨트롤 인터페이스'를 통해 모든 말뭉치(corpus)와 동기화된 음성 데이터로 생성해 낸다. 그리고 말뭉치(corpus)에 저장되어 있지 않은 새로운 음성이 들어 올 때, 기존의 말뭉치(corpus) 데이터를 이용하여 새로운 음성에 동기화된 조음 기관의 움직임을 생성해 내는 것이다.

2. J&T 컨트롤 인터페이스

본 논문 에서는 조음 기관의 움직임 데이터를 얻기 위해 'J&T 컨트롤 인터페이스'를 개발하였다[12][13]. 먼저, 조음점(조음 자리)에 따른 음소를 그룹화 시켜 조음기관의 모델을 정의하였다. 그리고 이 모델을 기저모델 기반 선형 합성 방법(Blend shape Interpolation)[9][10][11][17]에 적용시켜 사용자가 원하는 형태로 조음 기관 모양을 제어할 수 있도록 하였다. 조음 기관의 움직임과 연관되는 음소의 키프레임 (keyframe)에 값을 주고 캣물롬 보간 기법(catmull Rom interpolation)을 통해 모든 말뭉치(corpus)에 대한 조음기관의 움직임을 만들어 냈다. 구현된 인터페이스를 언어치료 전문가에게 주고 새롭게 입력된 음소 시퀀스(sequence)에 대해 만들어진 조음 기관의 움직임을 수정함으로써, 조음기관의 3 차원 움직임 데이터를 얻을 수 있었다.

2.1 조음기관 모델

조음할 때의 조음 기관의 막음이나 좁힘이 이루어지는 조음점(조음 자리)에 따라서 자음[14][15][16][18][19]을 나눌 수 있다. [그림 2-1]을 참고 하여, 위 아래 입술 사이에서 나는 양순음(1), 위 앞니와 아랫입술 사이에서 나는 순치음(2), 위 아랫니 사이에서 나는 치(간)음(3), 윗잇몸 치경에서 나는 치경음(4), 윗잇몸과 경구개 접경 지역에서 나는 경구개치음(5), 혀 뒷부분이 연구개에 닿거나 접근하여 나는 연구개음(6), 성문이 폐쇄되거나 마찰되어 나는 성문음(7), 그리고 혀를 구부려 나는 권설음(8)으로 구분되어 진다.



[그림 2-1]

본 논문 에서는 조음점(조음자리)에 따른

분류를 기준으로 조음 기관의 정적인 형태를 정의 하였다. [그림 2-1]의 양순음(1)과 치조음(2)은 입술의 움직임을 통해서 결정되기 때문에, 캐릭터(character)의 외관 입술 움직임을 통해서 표현 될 수 있다. 따라서 조음 기관(혀, 이, 성대)의 형태에는 영향을 미치지 않아 하나의 그룹으로 묶어 형태를 정의 하였다. 이를 바탕으로 따른 8 개의 그룹을 7 개의 조음 기관 형태로 정의 하여 표현하였다[그림 2-2].

(a) 양순음/순치음 (b) 치(간)음

(c) 치경음 (d) 권설음

(e) 경구개치음 (f) 연구개음

(g) 성문음

[그림 2-2]

2.2 형태 혼합 기법

이번 장에서는 형태 혼합 기법(Blend Shape)을 사용하여, 조음 기관의 새로운 움직임을 만들어 내는 방법에 대해서 제안한다.

먼저, 형태 혼합 기법의 기저 모델을 조음 기관 모델을 바탕으로 정의 한다(2.2.1). 그리고 이를 조합하여 새로운 형태의 조음 기관의 모델을 만들어 내고(2.2.2), 캣물롬 보간 기법(catmull Rom interpolation)을 사용하여 새로이 입력된 말에 대한 조음 기관의 초기값(Initial Guess) 움직임을 생성해 낸다(2.2.3).

2.2.1 기저 모델 정의

이번 장에서는 3.1 절에서 정의한 조음 기관 모델을 바탕으로 형태 혼합 기법(Blend Shape)을 수행하기 위한 기저 모델을 정의 한다. 19 개의 자음을 조음점(조음 자리)의 위치에 따라 8 가지로 그룹화 하고, 각 그룹에 해당하는 모델을 만들어 내 가장 기본적인 7 개의 조음 기관 모델을 구축 하였다.

이렇게 생성된 조음 기관 모델을 혀와 혀를 제외한 구강구조(이, 성대)로 나눠 두 개의 기저 집합으로 표현 하였다.

먼저, 혀의 기저들을 만들어 내기 위해 조음 기관 모델을 표현하기 위한 대표 자음을 선택한 후, 명시하도록 하였다. 양순음과 순치음에 대해서 ㅂ-계열 <ㅂ(b), ㅍ(p), ㅃ, ㅁ(m), (f), (v)>, 치(간)음에 대해선 ㅿ-계열<(ㅿ), (ð) >, 그리고, 나머지들에 대해선 다음처럼 분류하여 나타낼 수 있다. ㄷ-계열<ㄷ(d), ㅌ(t), ㄸ, ㅅ(s), ㅆ, ㄴ(n), ㄹ(l)>, r-계열<(r)>, ㅈ-계열<ㅈ(z), ㅊ(tʃ), ㅉ>, ㄱ-계열<ㄱ(g), ㅋ(k), ㄲ, ㅇ(ŋ)>,ㅎ-계열<ㅎ(h)>. 이것을 바탕으로 본 논문 에서는 7 개의 기저들을 다음과 표현할 수 있다.

$$\Im = \{T_{ㅂ}, T_{\theta}, T_{ㄷ}, T_{r}, T_{ㅈ}, T_{ㄱ}, T_{ㅎ}\} \quad (2.1)$$

ㅂ-계열 자음을 발음 할 경우, 발음하지 않을 때의 혀 모양을 그대로 유지 하며 입술을 이용하여 발음 되기 때문에, $T_{ㅂ}$ 을 중립(neutral) 형태로 놓을 수 있다. $T_{ㅂ}$ 를 초기 형태(initial shape)인 $\breve{T}_0$ 과 같이 놓을 수 있다. 그리고 나머지 6 개의 기저에 대해서도 $T_{ㅈ}$ 를 시작으로 시계방향 순으로 다음과 같이 재정의 한다[17].

$$\breve{\Im} = \{\breve{T}_1, \breve{T}_2, \breve{T}_3, \breve{T}_4, \breve{T}_5, \breve{T}_6\} \quad (2.2)$$

혀를 제외한 구강구조, 즉 이와 성대로 구성되어 있는 조음 기관의 기저도 정의 하였다. 턱의 상하 움직임만을 고려 하였기 때문에 1 개의 자유도 만이 존재 한다. 따라서, 턱을 다물고 있는 상태를 중립(neutral) 형태로 놓고, 턱을 벌리고 있는 상태를 한 개의 기저로 표현하였다.

$$\breve{\Re} = \{\breve{J}_1\} \quad (2.3)$$

이렇게 혀와 혀를 제외한 조음 기관(이, 성대)에 의해서 정의된 기저를 이용하여 새로운 형태를 만들어 내는 방법에 대해서 다음 (2.2.2)장에서 좀더 자세히 살펴 보도록 하겠다.

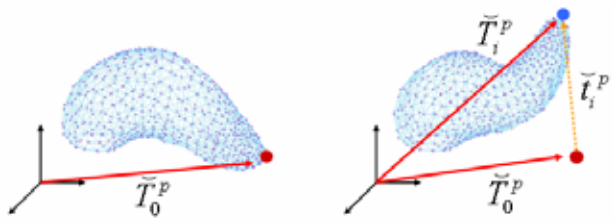
2.2.2 형태 혼합 기법

혀 기저들의 경우 $\breve{\Im}$ 과 $\breve{T}_0$, 모두 619 개의 점(vertex)로 이루어져 있다. 한 점(vertex) P 에 대해서

$$\breve{t}_i^{\,p} = \breve{T}_i^{\,p} - \breve{T}_0^{\,p} \left(i=1, 2, \ldots, 6\right) \qquad (2.4)$$

$$\breve{j}_1^{\,p} = \breve{J}_1^{\,p} - \breve{J}_0^{\,p} \qquad (2.5)$$

로 중립(neutral) 형태와 각 기저 들간의 차이를 이와 같은 방법으로 표현할 수 있다.



변형 전 변형 후

[그림 2-3]

혀의 6 개에 대한 $\breve{t}_i^{\,p}\left(i=1, 2, \ldots, 6\right)$ 가 주어졌을 때, 각각에 0 과 1 사이의 가중치 값을 할당하고 선형 보간(linear combination) 방법에 의해 중립(neutral)형태에서 원하는 혀 모양으로 변형 시킬 수 있다.

실제 혀가 움직일 때의 각 프레임(frame) h 에 대한 가중치 벡터가

$$\alpha^h = [a_1^h, a_2^h, a_3^h, a_4^h, a_5^h, a_6^h]^T \qquad (2.6)$$

일때, 이 식을 일반화한 형태는 다음과 같이 된다. 한 점 p 가 아닌 전체 모델에 대한 식의 일반화 형태는 다음 과 같이 된다.

$$\breve{T}\,\Big|^h = \breve{T}_0 + \sum_{i=1}^{n} a_i^h \breve{t}_i \,\Big|^h \qquad (2.7)$$

혀를 제외한 제외한 구강 구조의 형태 합성 방법도 혀와 마찬 가지 방법으로 나타낼 수 있다.

$$\breve{J}\,\Big|^h = \breve{J}_0 + \beta^h \breve{j}_1 \,\Big|^h \qquad (2.8)$$

(2.7)과 (2.8) 식을 이용하여 각 프레임(frame) h 에 대한 $<\alpha^h, \beta^h>$ 의 값이 주어지면, 조음 기관의 변형된 형태를 얻을 수 있다.

2.2.3 초기값(Initial Guess)의 생성

입력된 음성에 대한 조음 기관의 움직임을 얻기 위해서는 시간에 따른 $<\alpha^h, \beta^h>$ 의 연속된 값을 얻어내는 것이 필요 하다. 본 논문 에서는 이러한 움직임을 만들어 내기 위해, 3 차 보간 기법을 사용하였다.

입력 음성이 있을 때, 시간에 따른 연속된 음성의 흐름을 $\{P_t\}$ 로 정의한다. 예를 들어, '토끼'라는 묵음(s: silence)를 포함하는 입력 음성이 있을 때, $\{P_t\}_{t=1}^{26}$ = <s, s, s, s, s, ㅌ, ㅌ, ㅌ, ㅗ, ㅗ, ㅗ, ㅗ, ㅗ, ㄲ, ㄲ, ㄲ, ㅣ, ㅣ, ㅣ, ㅣ, ㅣ, s, s, s, s, s>로 표현될 수 있다. 여기서 각각의 요소(element)는 형태 혼합 기법을 사용하여 나타내어지는 조음 기관의 변형 형태이다. 이렇게 나열된 음소를 연속된 같은 구간으로 구분한 후, 그 음소 구간 에서의 키프레임(keyframe)을 지정한다.

키프레임(keyframe)은 l_t 를 연속된 같은 음소의 구간 길이라 할 때, $\lceil l_t/2 \rceil$ 과 같이 표현할 수 있다.

키프레임(keyframe)을 보간 하기 위한 데이터 점 $(x_0, x_{s1}, x_{ㅌ}, x_{ㅗ}, x_{ㄲ}, x_{ㅣ}, x_{s2}, x_{26})$ 으로 정한다. 조음기관의 움직임이 묵음(silence)에서 시작해서 묵음(silence)로 끝나기 때문에, x_0 과 x_{26} 을 추가하여 경계점에서 함수의 값이 0 을 갖게 하여, 묵음(silence)을 표현하도록 한다.

이렇게 정의된 데이터 값에 각 자음에 해당하는 값을 할당하여 이 점을 지나는 3 차 스플라인(spline) 곡선이 값의 영향을 받을 때 취하는 모양을 갖도록 한다.

2.3 조음 기관의 움직임 데이터 수정

2.1 장과 2.2 장의 작업을 통해 새로운 음성에 동기화된 조음 기관의 움직임을 얻어낼 수 있었다. 하지만 단지 초기값(initial guess)을 이용하여 예상되는 조음 기관의 움직임을 나타내는 것이기 때문에 데이터의 신뢰도가 상당히 떨어진다.

본 논문 에서는 언어 치료 전문가에게 'J&T 컨트롤 인터페이스'를 제작하여, 이러한 초기값(initial guess)를 쉽게 수정한 후 저장할 수

있도록 하였다. 모든 말뭉치(corpus)에 대해 언어치료 전문가가 데이터를 처리할 수 있도록 하였으며, 이를 이용하여 모든 말뭉치(corpus)에 대한 조음 기관 움직임 데이터를 얻을 수 있었다.

3. 조음 기관의 움직임 생성

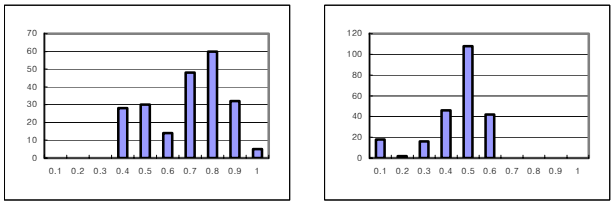
2 장에서 구축한 J&T 컨트롤 인터페이스를 이용하여 모든 말뭉치(corpus)의 연속된 음소에 대한 매개변수 값을 얻었다. 4 장에서는 이런 매개변수 값을 이용하여 새로운 음성이 들어 올 때, 그것에 동기화된 조음 기관의 애니메이션을 만들어 내는 방법에 대해서 설명한다.

먼저, 대한 평균과 분산 값을 구한다(3.1). 그리고 구한 평균과 분산을 이용하여, 새로운 음성에 동기화된 매개변수의 궤도(trajectory)를 생성해 낸다(3.2)[1]. 시간에 동기화된 매개 변수를 기저에 적용시켜 조음기관의 움직임을 만들어 낸다.

3.1 음소의 평균과 분산

2 장을 통해 얻은 모든 말뭉치(corpus)의 각 프레임에 해당하는 데이터는 (2.7)식과 (2.8)식에서 알 수 있듯이 7 개의 매개변수 값으로 이루어져 있다. 이러한 데이터를 같은 음소로 레이블링(labeling)된 값들끼리 40 개의 그룹으로 클래스화 하여 묶었다.

아래 그림에서 보는 것과 같이 $\{P_t\}_{t=1}^{26}$ 각 클래스로 묶인 매개변수 값들은 비슷 한 값을 나타내는 경향이 있다. 아래 그림의 (a)는 'ㅌ' 클래스로 묶었을 때, $\breve{t}_3$ 에 대한 매개 변수 값들의 히스토그램을 나타내고, 그림 (b)는 모음 'ㅗ' 클래스로 묶인, $\breve{j}_1$ 에 대한 매개 변수의 히스토그램을 나타낸다.



(a) (b)

[그림 3-1]

이렇게 각 클래스로 그룹화된 값들은 일정한 중심값에 몰리는 경향이 있음을 알 수 있다. 이를 바탕으로 본 논문에서는 각 클래스에 해당하는 매개변수 값들의 평균(mean)과 공분산(covariance)을 구하였다.

공분산 행렬에서의 요소들은, 각 요소들 사이의 상관관계에 대한 수치를 나타낸다. 그러나, 본 논문에서는 공분산 행렬 요소를 나타내는 평균값 요소들은 각각 서로 다른 기저를 제어하는 매개변수 이므로, 각 요소들 사이의 상관관계는 전혀 고려될 필요가 없다. 따라서, 공분산 행렬의 대각 부분 요소인 분산 값만을 고려하여, 각 클래스에서의 흩어진 정도를 분석 하여 나타낼 수 있다.

3.2 매개변수 궤도 생성

새로 입력된 음성에 대한 매개변수의 궤도를 생성해 내기 위해, 본 논문에서는 [1] 에서 보여진 규칙화 이론(regularization theory)을 사용한다.

함수 f 의 몇 개의 샘플 값들이 산발적으로 흩어져 있을 때, 샘플을 지나는 함수 f 의 곡선은 무수히 많이 그려질 것이다. 이렇게 무수히 많은 해(solution)를 갖는 것은, 수학적으로 잘 정립되지 않은 문제(ill-posed problem)이다. 이러한 문제는 사전지식(prior knowledge)를 통해 하나의 유일한 해를 갖도록 풀 수 있으며, 이 방법을 사용 하는 것이 바로 규칙화 이론(regularization theory)이다.

규칙화 이론(regularization theory)은 함수 f 가 '매끄럽다(smooth)' 하다는 가정하에, 두 개의 유사한 입력 값이 들어 올 때, 그에 해당하는 두 개의 출력 값은 유사 하다는 것이며, 다음과 같은 수식으로 표현할 수 있다.

$$H[f] = E[f] + \lambda\phi[f] \qquad (3.1)$$

$E[f]$ 는 대략화의 신뢰도(fidelity)를 나타내는 부분으로 사전지식(prior knowledge)를 사용하여 값을 실제 값에 가깝게 제약 시키도록 하는 부분이다. $\phi[f]$ 는 함수 f 가 매끄러운(smooth) 형태로 나타낼 수 있도록 하는 매끄러움 함수(smoothness function) 부분으로 λ 값에 따라 함수의 매끄러움(smoothness) 정도를 조절할 수

있다 λ 는 규칙화 매개변수(regularization parameter)라고 불리는 양수이다. λ 값이 작아지면 함수 f 는 더 매끄러운(smooth) 형태의 그래프 모양을 나타낸다.

식(3.1)의 첫째 항은 데이터와의 근접함을 제어하고, 두 번째 항은 그래프의 매끄러운 정도를 제어한다. 여기에서 λ 값을 조절함으로써, 두 항 사이의 교환(trade-off)를 조절할 수 있다.

이렇게 정의된 식(3.1)의 기능성(functional) $H[f]$ 의 극점에서의 함수 $\overline{f}$ 의 값을 얻기 위해, H 의 기능성 미분(functional derivative)을 행한다. 그리고 얻은 미분 수식의 $dH/df = 0$ 을 계산하여, 극점에서의 함수 $\overline{f}$ 을 얻는다.

본 논문에서는 새로운 음성에 대한 매개변수의 궤도를 얻어 내기 위해, 식(3.1)을 식(3.2)와 같이 표현하였다.

$$E = \left(\tilde{y} - \tilde{\mu}\right)^T \tilde{D}^T \tilde{\Sigma}^{-1} \tilde{D} \left(\tilde{y} - \tilde{\mu}\right) + \lambda \tilde{y}^T \tilde{W} \tilde{y} \tag{3.2}$$

$\left(\tilde{y} - \tilde{\mu}\right)^T \tilde{D}^T \tilde{\Sigma}^{-1} \tilde{D} \left(\tilde{y} - \tilde{\mu}\right)$ 이 값을 구하기 위한 항이며, $\lambda \tilde{y}^T \tilde{W} \tilde{y}$ 이 함수의 매끄러움(smoothness)를 결정하는 항이다. 이 두 항으로 구성된 함수 E 을 최소화 시키는 y 을 찾음으로써, 매개변수의 궤도를 생성해 낼 수 있다.

(3.2)식의 $\tilde{y}$ 는 실제 음성과 동기화되어 기저에 할당되게 되는 위 식의 결과 값으로써, 새로 입력된 음소의 전체 프레임을 T 라 하자. 한 개의 y_t 는 혀를 제어하는 6 개의 매개변수와 혀를 제외한 조음기관을 제어하는 1 개의 매개변수로 이루어졌기 때문에 7×1 의 벡터로 이루어져 있다. 따라서 전체 프레임에 해당하는 $\tilde{y}$ 는 $7T\times1$ 차원으로 이루어진 벡터이다.

$\tilde{\mu}$ 과 $\tilde{\Sigma}$ 는 (3.1.1)에서 구한 평균과 분산 값을 이용하여, 각 프레임에 대한 값을 할당한다. 위 식에 보여진 $\tilde{\mu}$ 는 해당 음소를 나타내는 프레임에 3.1 장에서 구한 7 개의 매개변수에 대한 평균 값을 할당하여, 전체 T 시간에 해당하는 $7T\times1$ 크기를 갖는 벡터를 위의 형태로 나열하여 정의 하였다.

전체 프레임에 대한 분산을 나타내는 $\tilde{\Sigma}$ 도 위와 비슷한 방법으로 3.1 장에서 구한 Σ 을 이용하여, 다음 행렬 형태로 나타내었다.

$$\tilde{\mu} = \begin{bmatrix} \mu_{p1} \\ \mu_{p2} \\ \vdots \\ \mu_{pT} \end{bmatrix} \qquad \tilde{\Sigma} = \begin{bmatrix} \Sigma_{p_1} & & & \\ & \Sigma_{p_2} & & \\ & & \ddots & \\ & & & \Sigma_{pT} \end{bmatrix}$$

각 프레임에 해당하는 7×7 행렬을 전체 T 시간에 대해 할당하여, $7T\times7T$ 크기를 갖는 분산 행렬을 나타냈다.

이렇게 구한 분산 값은 동시조음(coarticulation) 현상을 해결하기 위해, 사용되었다. 음소의 분산 값이 작은 경우, 매개변수 데이터들이 평균값에 몰려 있으므로, 이웃하는 음소에 동시조음(coarticulation) 현상을 거의 미치지 못한다. 반면에 분산 값이 큰 경우, 매개변수 데이터들이 흩어져 있으므로, 이웃하는 음소에 따라 그것에 유연하게 동시조음(coarticulation) 현상을 잘 반영할 수 있도록 원하는 음소를 선택하여 동시조음 효과를 잘 표현한다. 본 논문 에서는 모든 음소에 같은 동시 조음 효과를 적용 하기 위해, 위의 분산 행렬의 역행렬을 구하여 모든 프레임에 있는 음소에 적용하여, 모든 프레임에 대해 동시조음(coarticulation) 현상에 대해 같은 가중치를 할당하였다.

식(3.2)의 $\tilde{D}$ 는 지속기간 가중치 행렬로써, 같은 음소가 연속적으로 나타나는 지속기간이 긴 경우, 가중치를 적게 주고 짧은 경우 가중치를 많이 주는 부분이다. 이를 통해 모든 음소가 전체 프레임에 대해서 동등한 영향을 미치도록 한다. 연속적으로 나타나는 음소의 지속기간이 긴 경우, 전체 그래프를 생성하는 과정에서 같은 평균 값이 전체 프레임에 많은 부분이 할당 되어, 전체 궤도(trajectory)의 결정에 영향을 많이 미친다. 이럴 경우, 상대적으로 적은 지속기간을 갖는 음소는 긴 지속기간을 갖는 음소의 궤적에 묻히게 되어, 거의 표현되지 않는다. 이와 같은 것을 지속기간 가중치 행렬 $\tilde{D}$ 를 사용하여 해결할 수 있다.

$$\tilde{D} = \begin{bmatrix} D_{P_1} & & & \\ & D_{P_2} & & \\ & & \ddots & \\ & & & D_{P_T} \end{bmatrix}$$

전체 프레임 T 에 대한 하나의 음소(element)의 지속기간이 t_P 일 때, 그 때의 그 음소에 대한 가중치는 $\sqrt{1-(t_P/T)}$ 로 나타낼 수 있다. 그리고 이 값을 하나의 요소(element)로 사용하여, 각 음소에 대한 행렬 D_P 는 다음과 같이 표현된다.

$$D_P = \begin{bmatrix} \sqrt{1-\frac{t_P}{T}} & & & \\ & \sqrt{1-\frac{t_P}{T}} & & \\ & & \ddots & \\ & & & \sqrt{1-\frac{t_P}{T}} \end{bmatrix}$$

이 대각 행렬은 $t_P \times t_P$ 이다. 이와 같은 형태의 행렬들을 이용하여 $\tilde{D}$ 형태의 전체 프레임에 대한 대각 행렬을 만들어 낼 수 있다.

이렇게 평균과 분산, 그리고 지속기간 가중치 행렬을 이용하여, 식 (3.2)의 첫 번째 항을 만들 수 있다. 이를 바탕으로 식(3.2)식을 완성하기 위해, 두 번째 항의 매끄러움 행렬 $\tilde{W}$ 를 살펴보자. 일반적으로 매끄러움(smoothness)를 나타내기 위해 사용되는 행렬은 (4.10)과 같다.

$$W = \begin{bmatrix} -I & I & & \\ & -I & I & \\ & & \ddots & \\ & & -I & I \end{bmatrix}$$

I 행렬은 7×7 크기이고, W 행렬의 가로축은 전체 프레임이 T 일 때 $7T$ 이고, 세로 축은 $7(T-1)$ 이다.

본 논문에서는 [1] 에서 사용한 1 차 매끄러움(smoothness)행렬 W^TW 를 사용하지 않고, 더 매끄러움(smoothness)를 만들어 내기 위해, W^TWW^TW 과 같은 2 차 매끄러움 행렬을 사용하였다.

식 (3.2)에 나타난 각 항의 벡터와 행렬을 정의한 후, 이 식을 함수 E 을 최소화 시키는 $\tilde{y}$ 을 찾음으로써, 매개변수의 궤도(trajectory)를 생성해 낼 수 있다. 이 값을 찾기 위해 함수 E 를 미분하여 이 식이 0 이 되는 점, 즉 극점에서의 $\tilde{y}$ 를 구할 수 있다. $dE/dy = 0$ 을 구하기 위해, 식(4.4)를 미분하였다.

$$2\tilde{D}^T\tilde{\Sigma}^{-1}\tilde{D}\,(\tilde{y}-\tilde{\mu}) + 2\lambda\tilde{W}\tilde{y} = 0$$

$$\tilde{D}^T\tilde{\Sigma}^{-1}\tilde{D}\,\tilde{y} - \tilde{D}^T\tilde{\Sigma}^{-1}\tilde{D}\tilde{\mu} + \tilde{D}\lambda\tilde{W}\tilde{y} = 0$$

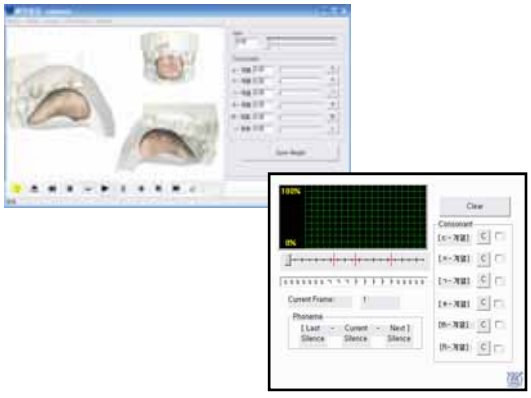
$$(\tilde{D}^T\tilde{\Sigma}^{-1}\tilde{D} + \lambda\tilde{W})\tilde{y} = \tilde{D}^T\tilde{\Sigma}^{-1}\tilde{D}\mu \qquad (3.3)$$

식 (3.3)을 CG(Conjugate Gradient) 을 사용하여, 벡터 $\tilde{y}$ 을 구할 수 있었다. 구한 벡터 $\tilde{y}$ 의 값들을 각 기저에 해당하는 7 개의 매개변수에 프레임 단위로 할당함으로써, 전체 프레임에 대한 매개변수의 궤적(trajectory)을 구할 수 있었다. 그리고 이 궤적의 값들을 식(2.7)과 (2.8)에 할당하여 조음 기관의 움직임을 생성해 낼 수 있다.

4. 결과

J&T 컨트롤 인터페이스

본 논문 에서는 조음 기관에 대한 3 차원 모델을 사용자가 쉽게 제어할 수 있는 인터페이스를 구현하였다.



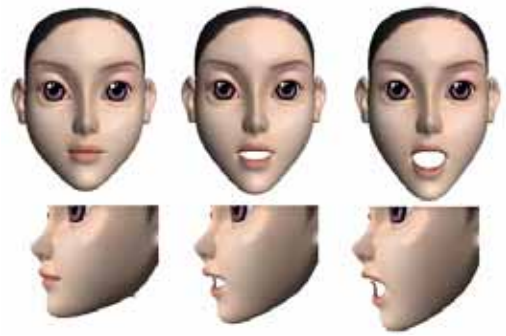
[그림 4-1]

'J&T 컨트롤 인터페이스'는 자음에 대한 6 개와 턱에 해당하는 1 개 총 7 개의 클래스에 대한 슬라이더를 움직임으로써, 중립(neutral) 형태의 조음 기관 모양에서 원하는 클래스의 모양으로 독립적으로 변형시킬 수 있다.

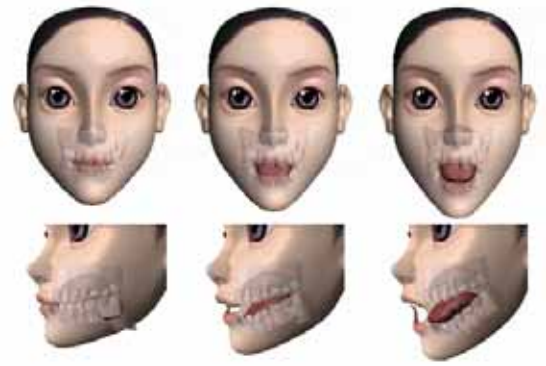
이 인터페이스를 이용하여, 새로 입력된 음성과 동기화된 조음 기관의 움직임을 나타내는 연속된 매개변수의 값을 생성해 낼 수 있다. 입력으로

프레임 단위의 연속된 음소가 들어 오고, 그것에 동기화된 초기값(initial guess)이 만들어 진다. 그리고, 수정하길 원하는 프레임으로 시간 슬라이더를 옮긴 후, 값을 변화 시켜 저장할 수 있다. 이 방법을 통해 모든 말뭉치(corpus)에 대해서 연속된 매개변수 값을 얻을 수 있다.

J&T 컨트롤 인터페이스를 통해 얻은 매개 변수 값들을 이용하여, 새로운 입력 음성이 들어 올 때, 그것에 동기화된 조음 기관의 움직임을 MMM(Multidimensional Morphable Model) 모델을 통하여 생성해 냈다. 본 논문에서는 모든 말뭉치(corpus)에 대한 조음 기관의 매개변수 값을 얻지 못하였다. 대신에 이를 수행하기 위한 실험 값들을 추출하여, 평균과 분산 값들을 구하였고, 식(3.3)을 사용하여 새로운 입력 음성에 대한



(a) 모션 캡쳐를 통해 얻은 얼굴 애니메이션



(b) 얼굴의 움직임과 동기화된 조음 기관 애니메이션

[그림 4-2]

연속된 매개변수 값들을 생성해 냈다. 이렇게 생성된 매개 변수 값을 2.2.2 절에서 정의한 기저 모델에 적용하여, 음성에 동기화된 조음기관의 움직임을 얻을 수 있었다. 가장 먼저 조음 기관의 움직임을 생성해 냈고, 이를 바탕으로 실제 모션 캡쳐 방법을 사용하여 3.2 절을 통해 재생성 해낸 얼굴 움직임과 동기화 하여 애니메이션을 만들어 냈다[그림 4.2].

5. 결론

본 연구에서는 조음 기관의 움직임 데이터를 획득하기 위한 J&T 컨트롤 인터페이스를 구현하였고, 인터페이스를 통해 생성된 데이터로부터 [1]에 기반한 매개변수 궤도(trajectory)를 만들어 냈다. 그리고 이를 통해 새로 입력된 음성에 동기화된 조음 기관의 움직임 생성해 낼 수 있었다.

일반적으로 얼굴이나 몸(body)의 움직임 데이터(motion data)를 획득하기 위해, 마커 기반 모션 캡쳐 방법을 이용하거나 카메라 기반 3 차원 좌표 생성 방법을 사용한다. 하지만, 이러한 방법으론 얼굴 외관이 아닌, 얼굴 내부(혀와 성대를 포함하는 조음 기관)의 움직임 데이터를 얻을 수 없다. 본 실험에서는 'J&T 컨트롤 인터페이스'를 구현하여, 음성에 동기화된 조음 기관의 움직임 데이터를 생성해 냈다.

이를 위해, 혀의 움직임을 결정하기 위해 23 개(19 개의 한국어 자음+f, v, Θ, ð)의 자음[14][15][16][18][19]을 6 개의 클래스로 구분하여, 형태 혼합기법을 수행하기 위한 6 개의 정적인 형태를 기저로 정의 하였다. 그리고 혀를 제외한 조음 기관의 움직임을 생성하기 위해 1 개의 DOF 로 기저를 정의 하였다. 따라서 총 7 개의 매개변수를 제어 함으로써, 원하는 형태의 조음 기관의 모양을 변형 시킬 수 있었고, ' J&T 컨트롤 인터페이스' 를 통해 음성에 대한 움직임 데이터 즉, 음성과 동기화된 연속된 매개변수를 얻을 수 있었다.

이렇게 생성된 데이터를 [1]에 기반하여 새로운 음성이 입력 될 때, 그것에 동기화된 조음 기관의 움직임을 생성해 냈다. 동시 조음(coarticulation) 문제를 하기 위해, 분산 행렬을 고려 하였다. 그리고 두 음소에 기반된

매개변수 값 사이의 매끄러움(smoothness)을 추가 하기 위해 2 차 매끄러움(smoothness) 행렬을 사용하였다. 이를 이용하여, 입력 음성에 동기화된 조음 기관의 움직임을 생성해 냈다.

향후 과제로는, ‘J&T 컨트롤 인터페이스’에서 입력 음성에 동기화된 초기값(Initial Guess)을 생성해 내는 부분이다. 같은 클래스에 포함되는 자음들 사이의 차이를 명확히 하지 않았다. 예를 들어, 예사소리 ‘ㅂ’과 된소리 ‘ㅃ’, 그리고 거센소리 ‘ㅍ’ 사이에는 같은 정적인 조음 기관의 모양을 나타낼지라도, 발음 할 때 공기의 압력이나 발음 지속 시간에 따라 확연히 다른 자음이 된다. 이를 제어할 수 있는 컨트롤 인터페이스를 추가한다면, 좀더 명확한 초기값(Initial Guess) 측정이 가능할 것이고, 언어 치료 전문가 들이 이 인터페이스를 통해 더 많은 데이터 처리를 좀 더 명확하게 처리 할 수 있을 것이다.

또한, 이 논문을 이용하여 3 차원 캐릭터를 이용한 장애 아동 재활 치료 프로그램으로 개발이나, 영화 캐릭터에서의 조음 기관의 움직임 생성에 응용할 수 있을 것이다.

참고문헌

[1] T. Ezzat, G. Geiger, T. Poggio. In Proceedings of SIGGRAPH 2002, Trainable Videorealistic Speech Animation.

[2] Girosi F., Jones M., and Poggio T. 1993 Priors, stabilizers, and basis functions: From regularization to radial, tensor, and additive splines. Tech. Rep. 1430.

[3] Wahba G. 1990 Splines Models for Observational Data. Series in Applied Mathematics, Vol 59.

[4] B. deGraf, In Proceedings of SIGGRAPH 1989, Course Notes22, 'Performance' Facial Animation.

[5] L.Willams. In Proceedings of SIGGRAPH 1990, Performance - Driven Facial Animation.

[6] C. Bregler, M. Covell, and M. Slaney. In Proceedings of SIGGRAPH 1997, Video Rewrite: Driving Visual Speech with Audio.

[7] M. Brand, In Proceedings of SIGGRAPH 1999, Voice Puppetry.

[8] F.I. Parke. Proceedings ACM annual conference 1972, Computer generated animation of faces.

[9] F.I. Parke. PHD thesis, University of Utah 1974, A parametric model for human faces.

[10] J. Kleiser. In Proceedings of SIGGRAPH 1989, A fast, efficient, accurate way to represent the human face.

[11] T. Blanz and T. Vetter, In Proceedings of SIGGRAPH 1999, A morphable model for the synthesis of 3d faces.

[12] F. Pighin, R. Szeliski, and D.H. Salesin, In Proceedings of International Conference on Computer Vision 1999, Resynthesizing facial animation through 3d model-based tracking.

[13] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D.H. Salesin, In Proceedings of SIGGRAPH 1998, Synthesizing realistic facial expressions from photographs.

[14] 오정란, 현대 국어음운론. 형설 출판사, 서울, 1993.

[15] 이기문, 김진우, 이상억. 국어음운론. 學研社, 서울, 2000

[16] 이호영. 국어 음성학. 태학사, 서울, 1996.

[17] B. Choe and H. Ko. Analysis and synthesis of facial expressions based on hand-generated muscle actuation basis. In Proceedings of Computer Animation 2001 Conference, pages 12-19, Nov.2001

[18] 강진철, 조선어 실험 음성학 연구, 한국문화사, 서울, 1996

[19] 원경식, 영어 음성학, 탑출판사, 서울, 2001