FAES : 감성 표현 기법을 이용한 얼굴 애니메이션 구현

On the Implementation of a Facial Animation Using the Emotional Expression Techniques

민용식*, 김상길**
호서대학교 컴퓨터공학부*, 한영신학대학교**

Yong-Sik Min(ysmin@office.hoseo.ac.kr)*, Sang-Kil Kim(ksknet2000@hytu.ac.kr)**

요약

본 논문은 여러 가지 감정들 중에서 4가지 감정의 범주 즉, 중성, 두려움, 싫증 및 놀람을 포함한 음성과 감성이 결합되어진 얼굴의 표정을 좀 더 정확하고 자연스러운 3차원 모델로 만들 수 있는 FAES(a Facial Animation with Emotion and Speech) 시스템을 구축하는데 그 주된 목적이 있다. 이를 위해서 먼저 사용할 훈련자료를 추출하고 난후에 감성을 처리한 얼굴 애니메이션에서는 SVM(Support vector machine)[11]을 사용하여 4개의 감정을 수반한 얼굴 표정을 데이터베이스로 구축한다. 마지막으로 얼굴 표정에 감정과 음성이 표현되는 시스템을 개발하는 것이다. 얼굴 표정을 위해서 본 논문에서는 한국인 청년을 대상으로 이루어졌다. 이런 시스템을 통한 결과가 기존에 제시된 방법에 비해서 감정의 영역을 확대시킴은 물론이고 감정인지의 정확도가 약 7%, 어휘의 연속 음성인지가 약 5%의 향상을 시켰다.

■ 중심어 : | 애니메이션 | 멀티미디어 HCI 응용 | 음성 감성 인식 |

Abstract

In this paper, we present a FAES(a Facial Animation with Emotion and Speech) system for speech-driven face animation with emotions. We animate face cartoons not only from input speech, but also based on emotions derived from speech signal. And also our system can ensure smooth transitions and exact representation in animation. To do this, after collecting the training data, we have made the database using SVM(Support Vector Machine) to recognize four different categories of emotions: neutral, dislike, fear and surprise. So that, we can make the system for speech-driven animation with emotions. Also, we trained on Korean young person and focused on only Korean emotional face expressions. Experimental results of our system demonstrate that more emotional areas expanded and the accuracies of the emotional recognition and the continuous speech recognition are respectively increased 7% and 5% more compared with the previous method.

■Keyword : | Animation | Multimedia HCI Applications | Speech Emotion Recognition |

* 본 연구는 2004년도 호서대학교 교내 특별연구과제로 수행되었습니다.

접수번호 : #041116-002
접수일자 : 2004년 11월 16일
심사완료일 : 2004년 12월 10일
교신저자 : 민용식 e-mail : ysmin@office.hoseo.ac.kr

I. 서 론

사람과 컴퓨터간의 의사 전달 수단으로 특정 입출력 장치를 사용하였지만, 최근에는 인간의 음성이나 얼굴 등을 이용하여 자연스럽고 지능적인 인터페이스를 구현하고자 하는 연구들이 진행되고 있다.

인간의 얼굴을 직접적으로 애니메이션 하는 것은 사실적인 얼굴 표현을 위해 통제되어야만 하는 많은 매개변수를 사용하므로 해서 사용자들로부터 많은 항의를 받고 있다. 만화제작자들의 그와 같은 어려움을 경감하기 위해 말을 하는 애니메이션 기술[4, 5]은 목소리와 얼굴 움직임 사이를 mapping 하여, 그 다음 음성 시그널에서 얼굴 애니메이션을 조정하고 있다. 그러나 대부분의 이전 얼굴 애니메이션 시스템들은 음성 시그널에 존재하고 있는 감정들을 뚜렷하게 염두에 두지 않았다. 같은 내용을 말할 동안 사람들은 그들이 행복 한가 또는 슬프냐에 따라 상당히 다른 얼굴 표현을 가질 수 있다. 감정은 얼굴 애니메이션 시스템에서 반드시 고려되어야만 한다[6].

이러한 얼굴 애니메이션에 대한 연구는 1974년 Parke의 파라메트릭 모델(parametric model)을 시초로 시작되었으며, 얼굴 애니메이션 연구의 주된 방향은 감정과 입술 움직임 등을 처리하기 위한 정확하고 효율적인 방법을 찾는 것으로 얼굴 표현을 계층적 구조를 갖는 영역들로 구조화한 Platt and Badler의 연구와 얼굴의 변형을 안면 근육들의 움직임의 결과로 해석하여 얼굴 변형에 영향을 미치는 근육들을 시뮬레이션 한 Waters의 연구 등이 대표적인연구로 볼 수 있다[3].

Hong Chen은 비모수적 샘플링 기법을 이용한 얼굴 스케치 생성방식을 제한하였다. 제한된 방식은 학습과 비모수적인 샘플링 기법을 이용해서 스케치를 자동으로 생성해 준다. 예술가에 의해서 그려진 스케치 이미지들을 데이터 베이스화 해 놓고, 그 데이터 베이스의 샘플링 이미지들을 변형하여 합성하는 방식을 사용한 방법이다 [2].

Li에 의해서 제시된 방법은 기존의 단순히 얼굴 표현에서 벗어나 음성과 더불어서 감성을 지닌 얼굴 애니메이션을 구현한 방법이다. 이 방법에서의 문제점은 첫째로 감정의 범주 가운데서 단지 4개의 부분 즉, 노여움, 행복과 슬픔 그리고 중립에 대한 것만 다루었을 뿐 그 이외 감정의 범주인 두려움, 싫증과 놀람에 대해서는 언급을 하고 있지를 있다. 둘째로는 사람을 대상으로 4가지의 다른 감정(보통, 행복, 노여움, 슬픔)을 분류하는데 그 정확도가 63.5%이며, 음성 인지 분야에서 어휘의 연속 음성 인지 작업에 대한 캐릭터 정확도는 90% 이상이다[8, 11].

본 논문은 이러한 문제점을 해결하기 위해서 얼굴의 특징점을 정확하게 추출한 후에 얼굴에 감정을 지닌 음성과 더불어서 감정추출을 함에 있어서 4개의 감정의 범주를 더 확장을 시키는 FAES(감성을 지닌 얼굴 애니메이션 구현)를 제안 한다. 이들 4가지 감정의 범주로는 중성, 두려움, 싫증 및 놀람을 의미한다. 또한 4가지의 감정을 포함한 음성과 감성이 결합되어진 얼굴의 표현을 정확하고 자연스러운 3차원 모델로 만들 수 있는 FAES(A Facial Animation with Emotion and Speech) 시스템을 구축한다. 이 시스템을 통해서 화상회의, 가상현실, 교육, 영화 등에서 활용될 수 있도록 하는데 그 주된 목적이 있다. 이런 시스템을 구성함에 있어서 감정을 나타내는 얼굴 애니메이션과 대화시 입술 모양의 변화를 중심으로 하는 대화 애니메이션 부분을 포함하여 설계한다. 얼굴 애니메이션은 약간의 문화적 차이를 제외하면 거의 세계 공통적인 보편적 요소로 이루어진 반면, 대화 애니메이션은 언어에 따른 차이를 충분히 고려하여 구성을 하여야 하므로 본 논문에서는 그 대상을 한국인 청년으로 하여 시스템을 구성한다.

본 논문의 구성은 제 2장에서는 MPEG-4에서 표준화한 얼굴 정의 파라미터(FDP)와 얼굴 애니메이션 파라미터(FAP)를 정의함과 동시에 음성의 특징 점을 추출에 대해서 알아보고, 제 3장에서는 FAES 시스템 설계에 대하여 기술한다. 제 4장에서는 실험결과 및 분석해 대하여 언급하고, 제 5장에서 결론을 맺는다.

II. 얼굴 정의와 애니메이션 파라미터

음성을 포함한 감정을 지닌 얼굴 애니메이션의 시스

템을 구성하기 위해서 우선 먼저 얼굴과 음성의 특징 점을 분석을 하여야만 한다. 따라서 본 절에서는 MPEG-4에서 표준화한 얼굴 정의 파라미터(FDP)와 얼굴 애니메이션 파라미터(FAP)를 정의함과 동시에 음성의 특징 점 추출에 대해 알아보고자 한다[1].

1. FDP(Facial Definition Parameter)의 정의

FDP는 총 84개의 파라미터로 구성되는데 얼굴의 형태를 나타내는 중요한 특징 점들의 3차원 좌표 값으로 정의된다. FDP는 눈, 눈썹, 코, 입, 턱, 뺨, 혀, 치아, 귀, 머리 그리고 얼굴의 회전 등을 나타내는 총 9개의 그룹으로 나뉘어 정의되고 있다. MPEG-4에서 정의한 FDP는 얼굴의 뒷모습이나 목 부분에 대한 정보는 다루지 않고 [그림 1]과 같이 측면에 대한 정보까지만 다루고 있다. FDP에서 정의된 적은 수의 특징 점만으로는 자연스러운 얼굴을 구현할 수 없고, 보조 특징 점들을 추가하거나 일반모델을 이용하여 얼굴을 구현하여야 한다[8].

2. FAP(Facial Animation Parameter)의 정의

FAP는 얼굴 근육의 움직임과 관련하여 정의되어 자연스러운 표정 합성과 입술 동기화 등을 표현할 수 있도록 한다. FAP는 크게 상위 파라미터와 하위 파라미터로 구성되는데, 상위 파라미터는 표정을 정의하는 표정 파라미터와 음운의 발음에 따른 얼굴의 움직임을 정의하는 파라미터로 구성된다. FAP는 이러한 2개의 상위 파라미터와 함께 FDP의 움직임에 따라 정의된 66개의 하위 파라미터를 포함하여 총 68개의 파라미터들로 구성되어 있다.

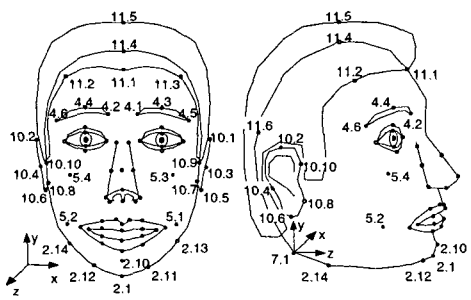

그림 1. MPEG-4에서 정의된 FDP

[그림 1]에서 표시된 검은 점들은 FAP에 의해 움직이는 특징 점들을 나타내고 있다[10].

3. 음성 특징 점 추출

[10]에서 사용한 것처럼 음향의 두드러진 점을 발췌하여 감정에 음성을 삽입시키도록 하고자 한다. 각각의 어조에 대해 다음과 같은 요소를 가지고 16 차원의 벡터를 계산했다.

- 리듬과 관련된 통계: 말하는 속도, 소리가 나는 지역 사이의 평균 길이, 최대 횟수/(최소 + 최대)횟수, 위경사의 횟수/경사의 횟수
- 평온한 음높이 시그널에 대한 통계 : 최소, 최대, 중앙, 표준 편차
- 평온한 음 높이의 미분계수에 대한 통계 : 최소, 최대, 중앙, 표준 편차
- 개별 소리 부분에 대한 통계 : 최소 평균, 최대 평균
- 개별 경사에 대한 통계 : 평균 양수 편차, 평균 음수 편차

III. FAES 시스템 설계

1. 시스템 개요

본 논문은 음성을 컴퓨터에 입력하여 음성 시그널이 파생되어 감정으로 나타나는 얼굴 애니메이션 구현 시스템(FAES)을 구축한다. 입력된 신호에 따라 많은 얼굴 표정이 나타날 것인데 그 중에서 4개의 감정 카테고리 즉 중성, 두려움, 싫증 및 놀람의 감정표현이 얼굴에 나타나는 얼굴 애니메이션 시스템을 목표로 한다.

[그림 2]는 FAES 시스템의 개요를 보여 주고 있다. 시스템은 크게 음성을 처리한 얼굴애니메이션과 감성을 처리한 얼굴 애니메이션으로 구성된다. 첫째 사용할 훈련자료를 추출하고, 둘째 감성을 처리한 얼굴 애니메이션에서는 SVM(Support vector machine)[11]을 사용하여 4개의 감정(중립, 놀람, 싫증, 두려움)을 수반한 얼굴 표정을 데이터베이스로 구축한다. 셋째 현실감 있고 자연스러운 얼굴 애니메이션이 이루어질 수 있도록 음

성을 구현하여서 이것들을 음성 데이터베이스화한다. 넷째 음성과 음성 시그널에서 파생된 감정이 합하여 나타나는 얼굴의 모습을 보여 주고 있다. 마지막단계에서는 우리가 얻고자 하는 얼굴 표정에 감정과 음성이 표현되는 시스템을 개발[알고리즘 FAES_main_routine 참조]하여 실제 얼굴 애니메이션을 구현한다.

```
procedure FAES_main_routine
 // 감정을 동반한 음성을 이용한 얼굴 애니
    메이션 생성 방법 //

{
 1. training_data( );
   // 훈련자료 추출 //
 2. Recognize_continuous_speech( );
   //연속적인 음성 인식과 데이터베이스 관리//
 3. Recognize_emotional_data( );
   //감정 얼굴 표정과 데이터베이스 관리//
 4. mixture_with_emotion_and_speech( );
   //감정이 수반된 얼굴에 음성 동기화 구성//
 5. Output
   // 원하는 결과를 원하는 위치에 표현 //
}
```

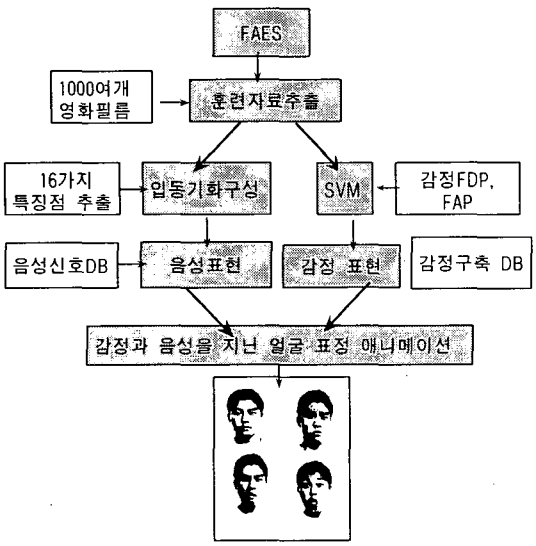

그림 2. FAES 시스템 구성도

표 1. 각 감정에 대한 샘플을 다룬 횟수

중립	두려움	놀람	싫증
311	384	205	197

2. 훈련자료 추출

본 논문에서 제시한 시스템을 구성하기 위해서 일차적인 목표로 실험 장치에서 어조를 4가지(중립, 놀람, 싫증, 두려움)로 분류하기위해 평균길이가 약 10초인 필름을 1,097개 모았다. 어조를 발췌하기위해 필름에 나오는 사람들 중에서 성별과 나이가 다른 사람들을 구분했다. 어조를 잘 구별할 수 있는 5명의 얼굴 애니메이션 전문가를 선정하여 구분해 놓은 사람들의 말을 각각 듣게 하고 감정을 4가지(중립, 놀람, 싫증, 두려움)로 분류하게 하였다. 분류한 자료를 다시 5명이 토의하여 만장일치로 동의하였을 때만 훈련자료로 사용하였다. [표 1]은 실험 대상이었던 각각의 분류에 대한 샘플처리 횟수이다. 이것들을 DB로 저장하였다.

분류된 자료들이 한국인 청년에게 나타나는 결과를 확인하기 위하여 분류된 자료를 모델에게 들려주었다. 그리고 모델에게 느낀 감정을 표정 짓게 하고, 그 얼굴 표정을 디지털 카메라로 1,097번 촬영하였다(표 1 참조). 촬영한 사진을 다시 선별하여 0%, 25%, 50%, 75%, 100%로 구분하여 그 사진의 얼굴 표정부분을 일러스트레이터를 이용해서 그려진 그림을 분류[알고리즘 training_data참조]하여 정리한 것이 [그림 3]이다.

```
procedure training_data
{  // 훈련 데이터를 추출하여 각 감정별로
      데이터베이스 구축//
 1 단계 : // 영화필름 capture //
    Capture movie film for 10 seconds
    and then insert
 2 단계 : // 화자들의 어조 발췌 //
   2.1 Retrieve the speech according to the
       sex
   2.2 insert data into speech_database on the
       age
```

```
3 단계 : // 얼굴 촬영 //
  3.1 capture directly from a digital camera(450
      만 화소)
  3.2 Process the emotional data depending
      on(Neutral, Surprise, Dislike, Fear)
      and then classifying each emotional
      data from 1000 movie pictures
4단계://5명의 감정별 데이터베이스 구축//
  4.1 select 5 persons for the emotional aspect
  4.2 detect and determine each person's
      speech
  4.3 divide each emotional data according to
      the speech
  4.4 if training data are OK then insert then
      into database
}
```

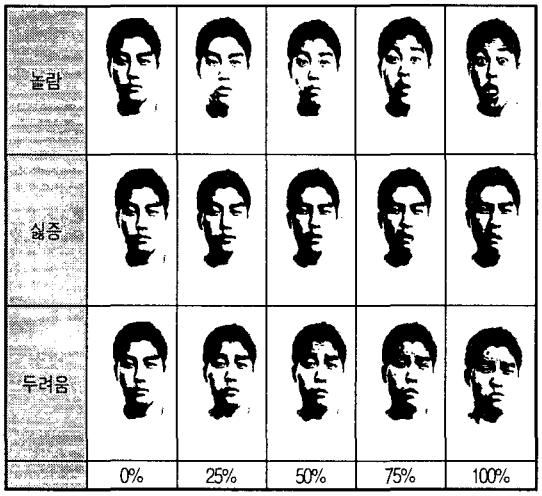

그림 3. 중립, 놀람, 싫증 및 두려움

3. 연속 음성 감정 인지

본 시스템에서 얼굴 모델을 작동시키기 위해 연속 입력 음성을 사용했다. 이들 각각의 음성을 적당한 범주로 분류하기 위해 여러 차원의 판별 장치를 설치했다. 감정 인지 자 부분은 음성 세트로부터 훈련을 통해서 얻었다. 또한 감정이란 부분은 짧은 기간 내에 안정적일 수 있기 때문에 연속적인 음성을 구로 분절시키고, 각각의 구에 대해 그 구에서 각 감정의 비율을 나타내는 인지 연산을 적용했다. [알고리즘 Recognize_continuous_speech]. 분절 연산은 다음과 같은 가정에 근거한 것이다.

- 감정 강도는 주어진 구 내 에서는 불변하다.
- 감정은 구가 침묵 또는 분류기가 구의 감정 분류를 정할 수 없을 때는 중립이다.

```
procedure Recognize_continuous_speech
{ //연속 음성 감정을 인식하기 위한 순서 //
//1. 판별장치 설치 //
 Classifying proper ranges depending on each
 person's speech
//2. 감정인지자 훈련 //
Training emotional data from speaking set
//3. 연속 음성 구로 분절 //
 Divide the sentence(or phrase) from continuous
 speech because a short period will provide a
 stable position
//4. 감정 인지 비율 연산적용 //
 Calculate and represent the percentage of each
 sentence(or phrase)
}
```

음향 특징이 음의 높이에서 발췌되기 때문에 구는 음의 높이 수치가 T 기간 동안 zero로 나타난 영역에서 분리될 수 있다. T는 감정의 평온을 조정하기 위한 매개변수로서 사용될 수 있다. 예로 T는 음성 감정이 엄청나게 바뀔 때는 작아야만 한다. 한편 큰 T는 평온한 감정 이동을 생기게 할 때 사용된다.

분석을 위해 각각의 구를 전체로 처리했다. 오디오 자료에서 발췌한 다른 특징들은 상호 관련이 되기 때문에 비선형의 분류자가 설계되어야만 한다. 현존하는 감정 인지 연산은 대개 K에 가장 가까운 이웃 또는 중립 네트워크에 근거를 두고 있다. 본 논문의 시스템에서는 다른 특징 사이의 상호 관계를 고려하지 않고 빨리 다룰

수 있는 SVM를 사용했다. 더구나 훈련 자료도 분류 자를 얻은 후, 인지 과정에서 분리시켰으며, 이를 위해서 Gaussian kernel[11]을 이용하는 SVM을 도구를 사용했으며 이때 사용한 식은 다음과 같다.

$$K(xi,xj)=e^{-\|x_i-x_j\|^2/2\sigma^2} \quad (1)$$

4개의 1-υ-r(1: 나머지) SVM들은 나머지 3개에서 한 개를 구별하게 훈련을 받았다. 훈련 과정 이후 우리는 두 범주의 4개 분류자 Si(υ)(i=1....4)를 얻었는데, 여기서 υ는 feature vector이다. 인지 과정은 다음과 같이 나타낼 수가 있다.

$$\begin{cases} v \in class & (if\ S_i(v) > 0) \\ v\ is\ rejected\ by\ classifier\ i & (if\ S_i(v) \le 0) \end{cases} \quad (2)$$

식별력이 있는 단계가 좀 더 많은 훈련 샘플을 가지고 CLASS를 접근하는 경향이 있기 때문에 SVM의 성능은 플러스 및 마이너스 훈련 샘플 수에 의해 영향을 받을 것이다. 각각의 분별기를 위해서 자료 세트에서 무작위로 훈련 샘플을 뽑아 거의 같은 플러스 및 마이너스 훈련 샘플 개수를 뽑았다. 예를 들면 감정 "두려움"을 인식하는 SVM을 훈련하기 위해 플러스 샘플로서 150개의 싫증으로 인식된 어조를 선택했다. 마이너스 샘플은 놀람으로 인식된 50개의 샘플, 중립으로 인식된 50개의 샘플, 두려움으로 인식된 50개의 샘플로 구성했다.

표 2에선 SVM의 성능과 SVM 이 훈련 샘플과의 관계를 보여 주고 있다. 분류자의 성능이 자동 분류가 주관적 라벨과 같은 라벨을 낳을 수 있는가 없는가의 판단에 의해 평가 된다. 또한 실험에서 가장 분명하게 인지될 수 있는 감정은 중립이며, 놀람 및 두려움의 인지 수행은 상대적으로 낮다. 그러나 여기서 발견한 것은 인간 청취자조차도 놀람과 두려움을 혼돈하기 쉽다는 것을 알 수가 있다[알고리즘 face_animation_with_emotion].

표 2. SVM 성능과 훈련 샘플 개수

	훈련 샘플 숫자				정확도
	D	S	N	F	
싫증	162	46	49	52	77.16%
놀람	31	102	31	32	65.64%
중립	61	68	194	64	83.73%
두려움	31	34	31	96	70.59%

(D: Dislike (싫증), S: Suprise(놀람), F: Fear(두려움), N: Neutral (중립))

```
procedure face_animation_with_emotion
// 감정을 이용한 얼굴 애니메이션 생성//
// 얼굴 특징 제어선(Control Line) 표시//
{
  1. FDA와 FAP를 이용한 주요 얼굴 특징(눈, 입, 턱, 코)점을 이용한 감정 찾기
  2. Gaussian 필터를 이용한 곡선추출
}
```

4. 입 동기화 구성

얼굴 애니메이션의 사실적인 표현을 높이기 위해, 음성[12, 13]로부터 입 동기화 구성을 종합적으로 제어하는 연산을 개발하였다. 입 동기화 구성은 아래의 3가지 문제점을 고려하여 구성한 것이다.

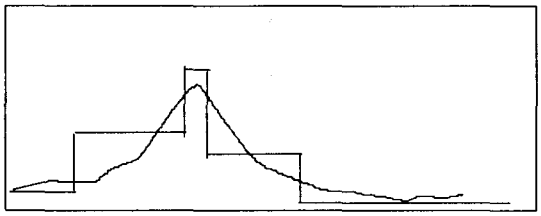

그림 4. 감정여과 곡선(실선은 감정인지를, 횡성은 감정 곡선을 의미한다)

- 시청각 신호를 어떻게 표현할 것인가
- 시청각 매핑을 어떻게 정의할 것인가
- 최적의 모델 매개변수를 어떻게 훈련할 것인가

일반적으로 말해, 음성 인지 시스템에서 음성 신호들

은 3가지 다른 단계, 즉 앞 끝(또는 신호 단계), 음향 모델(또는 음소 단계), 언어 모델(단어 단계)로 나타낼 수가 있다. 3단계 각각이 비록 입 동기화 구성 시스템 안에서 적용될 수 있을지라도, 특정하게 적용시키기 위해 좀 더 생각을 해 봐야만 한다. 높은 단계의 신호 표현은 좀 더 많은 상태 암시와 연관이 되기 때문에 후자의 두 가지 방법을 이용해 보다 좋은 결과를 달성할 수가 있다. 동시에 좀 더 높은 입력 신호 단계는 좀 더 복잡한 시스템을 필요로 한다. [그림 5]는 1,097개의 훈련자료들을 4개의 감정, 중립, 놀람, 싫증 및 두려움의 정도에 따라 분류한 입 모양이다.

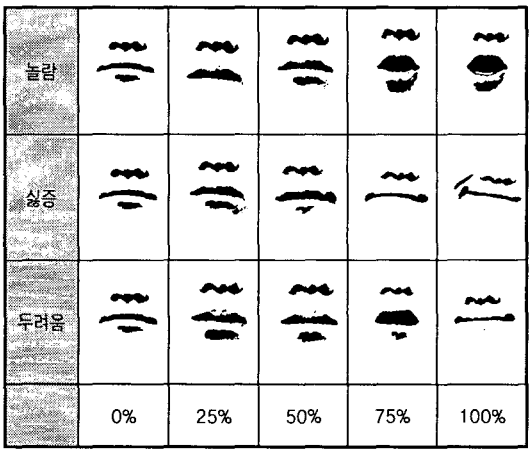

그림 5. 감정, 중립, 놀람, 싫증 및 두려움으로 분류한 입 모양

Ⅳ. 실험결과

실험 장치에서 어조를 4가지(중립, 놀람, 싫증, 두려움)로 구분하여 평균길이가 약 10초인 필름을 1097개를 추출을 했다. 이들 필름에서 어조를 성별과 나이별로 구분하였고, 감정은 4가지(중립, 놀람, 싫증, 두려움)로 구분을 하여서 훈련자료로 사용기 위해서 DB로 저장하였다.

이와 같이 분류된 자료들이 한국인 청년에게 나타나는 결과를 확인하기위해 분류된 자료를 모델에게 들려주었다. 그리고 모델에게 느낀 감정을 표정짓게 하여, 그 얼굴 표정을 디지털 카메라로 촬영하였다. 촬영한 사진을 다시 선별하여 얼굴 표정에 따라서 0%, 25%, 50%, 100%로 구분하여 사용을 하였다.

또한, 입 동기화 구성에 있어 3가지 문제점인 시청각 표현, 시청각 매핑 정의, 최적의 모델 매개변수 훈련을 고려해서 1097개의 훈련 데이터를 만들었다. 이렇게 만들어진 자료들을 4개의 감정, 중립, 놀람, 싫증 및 두려움의 정도(0%, 25%, 50%, 75%, 100%)에 따라 입 모양을 구분하였다.

이렇게 얻어진 얼굴 표현과 입 동기화 구성 결과들을 얻은 후, 최종 애니메이션 [알고리즘 mixture_with_emotion_and_speech 참조]를 만들기 위해 두 가지 요소를 결합하였다. [그림 6]은 FAES 시스템이 만든 3가지 애니메이션 순서, 즉 (A)는 감정만을 처리한 얼굴 애니메이션, (B)는 음성 즉 입 모양만을 처리한 얼굴 애니메이션, (C)는 두 가지 다 결합(A+B)시킨 것을 보여 주고 있다. 감정 인지 결과에 따라 설득력 있는 얼굴 표현이 만들어질 수 있음을 보여주고 있다. 또한 입 구성은 캐릭터의 음성과 함께 동시성을 가진다.

```
procedure mixture_with_emotion_and_ speech
 //감정과 음성을 합성한 얼굴 애니메이션//
{
  face_animation_with_emotion( );
  recognize_speech_from_emotional_db( );
  recognize_speech_from_speech_db( );
  output; // 입을 동기화 구성으로 얼굴
            애니메이션//
 }
```

이들을 통한 실험결과의 비교분석은 표 2에 제시된 것과 같다. Hong Chen이 제시한 방법[2]는 단순히 얼굴 애니메이션만 다루는 반면에 본 논문에서는 감성과 더불어서 음성을 지닌 얼굴 애니메이션이 가능하도록 하였다. 그리고 Li 등에 의해서 제시한 방법[11]에 비해서는 감정의 영역을 4개 더 확대를 시킴과 동시에 감정을 분류하는 정확도가 거의 7%이상의 향상을 가짐은 물론 어휘의 연속 음성인지 작업에 대한 정확도 역시 5% 이상의 향상을 지님을 볼 수가 있었다.(표 3 참조)

표 3. 기존의 방법과의 비교

	기본개념	문제점	개선점
Hong Chen의 방법[2]	-얼굴 스케치만을 생성	-다양한 얼굴 특징점 추출 부족 -입력된 윤곽선만을 변형해서 생성 -웹상에서 사용불가능	
Li 외 방법[11]	-감성과 더불어서 음성을 지닌 얼굴 애니메이션	-감성 분류를 단지 4개로 제한(노여움, 행복, 슬픔 과 중립) -감정 인식 정확도 63.5%, 연속 음성인지도 90%	-웹상에서 사용가능 -감정과 음성을 얼굴에 표현
본 논문에서 제시한 FAES 방법	-감성과 더불어서 음성을 지닌 얼굴 애니메이션	-단순히 얼굴만이 아닌 몸동작도 같이 구현을 해야만 함	-감정의 범위를 4개 더 확대 -감정 인식 정확도 70%, 연속 음성인지도 95%

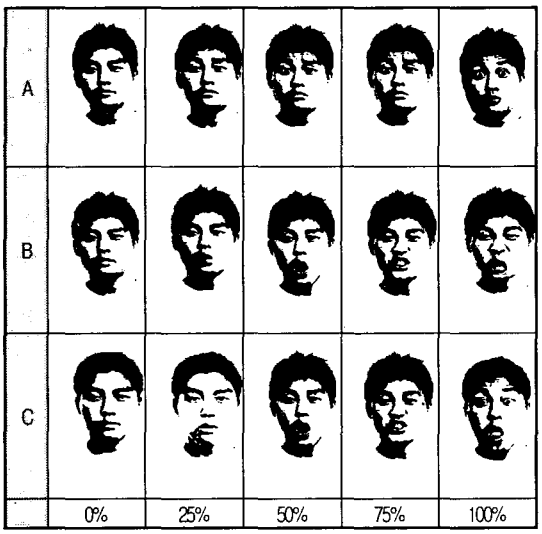

그림 6. A는 감정, B는 입 모양, C는 감정과 입 모양 모두 나타남

V. 결 론

본 논문은 감정과 더불어 음성으로 얼굴을 애니메이션 하는 FAES 시스템 구성을 제안함과 동시에 이들 시스템을 개발한 것이다.

이런 시스템 구성을 위해서, 다른 화자들로부터 천 개 이상의 어조를 이용해서(본 시스템에서는 영화 필름을 사용함) 이들을 4가지 감정, 즉 중립, 싫증, 놀람 및 두려움을 인지하기 위해 SVM을 사용했다. 이를 위해 입력 음성이 주어지면, 감정 인지 자는 모든 감정에 대해 완만한 이동곡선을 만들면서 이것이 얼굴 모델을 움직이게 하는데 사용된다. 얼굴 모델은 컴퓨터로 그린 형판들로 구성된다. 얼굴 표현은 이들 형판들 사이를 이동함으로서 애니메이션 된다.

또한 각각의 얼굴 프레임의 입 모양은 입력 오디어로부터 일치하게 된다. 입 동기화 구성 연산은 전통적인 음소 매핑보다는 음향 신호를 사용하였으며, 언어로부터 독립적으로 구성을 하고 실시간으로 처리할 수 있게 하였다. 그런 다음 입 모양이 최종 얼굴 애니메이션을 만들 수 있게 이동된 얼굴 표현 이미지와 함께 조립된다. 감정 및 음성에 중점을 둔 얼굴 애니메이션 시스템이 PC 및 인터넷상에서 매우 유용할 것이다. 이런 방법으로 구성된 시스템을 실제 만화나 애니메이션 영화에 이용을 하면 얼굴의 표정을 좀 더 정확하면서도 자연스러운 3차원 모델로 구성을 할 수가 있다. 그리고 본 논문에서 제시된 어조에 따른 감성을 지닌 얼굴 애니메이션의 실험 결과가 믿음만하고 우수한 성능을 가지고 있음을 입증할 수가 있었다.

향후 계획은 감정 및 음성에 중점을 둔 얼굴 애니메이션 시스템에 감정과 음성에 따라 움직임이 다르게 나타나는 몸동작에 대한 연구로 얼굴과 몸 전체가 연결되어 변화되고 움직이는 시스템을 구축할 계획이다.

참 고 문 헌

[1] 심연숙, "자연스러운 표정 합성을 위한 3차원 얼굴 모델링 및 합성 시스템", 한국 인공지능학회 논문지, 제11권 제2호, pp.1~4, 2000.

[2] 박연출, "음영합성 기법을 이용한 실사형 아바타 얼굴 생성", 한국인터넷정보학회 논문지, 제5권 제5호, pp.79~91, 2004.

[3] http://kidbs.itfind.or.kr/KIDBS/ITStrategy/30Summary/Animation.pdf

[4] M. Brand. "Voice puppetry," In Proc. ACM Siggraph99, pp.21~28, 1999.

[5] C. Bregler, M. Covell, and M. Slaney, "Video Rewrite: driving visual speech with audio," In Proc. ACM Siggraph97, pp.353~360, 1997.

[6] J. Cassell, C. Pelachaud, N. I. Badler, M. Steedman, B. Achorn, T. Beckett, B. Douville, S. Prevost, and M. Stone, "Animated conversation: rule-based generation of facial display, gesture and spoken intonation for multiple conversational agents," In Proc. ACM Siggraph94, pp.413~420, 1994.

[7] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," In Proc. ICSLP 1996, pp.855~860, 1996.

[8] E. Chang, J. L. Zhou, S. Di. C. Huang, and K. F. Lee, "Large vocabulary mandarin speech recognition with different approaches in modeling tones," In Proc. ICSLP 2000, pp.983~986, 2000.

[9] M. Escher, I. Pandzic, and N.M. Thalmann, "Facial Deformation for MPEG-4," SIGGRAPH 1998.

[10] ISO/IEC JTC1/SC29/WG11/MPEG97 N1820, "SNHC Verification Model 5.0," 1997.

[11] Y. Li, F. Yu, Y.Q. Xu, E. Chang, and H. Y. Shum, "Speech-Driven Cartoon Animation with Emotions," IEEE Computer Animation 2002, pp.365~371, 2002.

[12] T. Chen and R. R. Rao, "Audio-visual integration in multimodel Communication," IEEE Proceedings, pp.837~852, 1998.

[13] Y. Li and H. Y. Shum, "Animating cartoon face from video," In 6th International Conference on Control, Automation, Robotics and vision, pp.840~855, 2000.

저 자 소 개

민 용 식(Yong-Sik, Min) 정회원

- 1981년 : 광운대학교 전자계산학과 학사
- 1983년 : 광운대학교 전자계산학과 석사
- 1991년 : 광운대학교 전자계산학과 박사
- 1987년 3월~현재 : 호서대학교 컴퓨터공학부 교수

<관심분야 > : 병렬 알고리즘, 컴퓨터 애니메이션 알고리즘, 컴퓨터그래픽스

김 상 길(Sang-Kil, Kim) 정회원

- 1990년 : 경희 대학교 석사
- 2003년~현재 : 호서대학교 박사과정 재학 중
- 1998년 3월~현재 한영신학대학교 부교수

<관심분야> : 컴퓨터 애니메이션, 컴퓨터 그래픽스