論文2000-37CI-5-6

음성/영상의 인식 및 합성 기능을 갖는 가상캐릭터 구현

(Cyber Character Implementation with Recognition and Synthesis of Speech/Image)

崔 光 杓 *, 李 斗 誠 *, 洪 光 錫 *

(KwangPyo Choi, DooSung Lee, and KwangSeok Hong)

요 약

본 논문에서는 음성인식, 음성합성, Motion Tracking, 3D Animation이 가능한 가상캐릭터를 구현하였다. 음성인식으로는 K-means 128 Level VQ와 MFCC의 특징패턴을 바탕으로 Discrete-HMM 알고리즘을 이용하였다. 음성합성에는 반음절 단위의 TD-PSOLA를 이용하였으며, Motion Tracking에서는 계산량을 줄이기 위해 Fast Optical Flow Like Method를 제안하고, 3D Animation 시스템은 Vertex Interpolation방법으로 Animation을 하고 Direct3D를 이용하여 Rendering을 하였다.

최종적으로 위에 나열된 시스템들을 통합하여 사용자를 계속적으로 주시하면서 사용자와 함께 구구단 게임을 할 수 있는 가상캐릭터를 구현하였다.

Abstract

In this paper, we implemented cyber character that can do speech recognition, speech synthesis, Motion tracking and 3D animation.

For speech recognition, we used Discrete-HMM algorithm with K-means 128 level vector quantization and MFCC feature vector. For speech synthesis, we used demi-syllables TD-PSOLA algorithm. For PC based Motion tracking, we present Fast Optical Flow like Method. And for animating 3D model, we used vertex interpolation with Direct3D retained mode.

Finally, we implemented cyber character integrated above systems, which game calculating by the multiplication table with user and the cyber character always look at user using of Motion tracking system.

Ⅰ. 서 론

현재 많은 분야에서 음성인식, 음성합성, 화상인식, 3차원 Animation 기술이 연구되어 왔고 각각의 분야에서 적지 않은 시스템들이 개발되고 있다. 음성 분야는 컴퓨터 뿐만 아니라 통신을 이용한 금융서비스가 확대됨에 따라 전화상의 음성인식 시스템으로, 음성합성 시스템은 기존의 ARS을 대신하여 좀 더 동적인 역할을 하는 시스템으로서 자리를 잡고 있다. 그리고 화상인식은 제품의 생산분야에서, 3차원 Animation은 방송, 영화 분야에서 좋은 결과를 나타내고 있다.

그러나 Human Computer Interface 분야에서는 위에 열거한 시스템들이 각각의 독립적인 장치로 운영될 것이 아니라, 횡단적인 접근 방법을 시도하여 각 시스템을 통합하여 좀 더 사용자에게 친숙한 형태로 시스템을 구성하는 것이 중요한 문제로 대두되고 있다.

본 논문에서는 사실성 있는 3차원 캐릭터 Model의

* 正會員, 成均館大學校 電氣電子 및 컴퓨터工學部
(Department of Electrical & Computer Engineering, Sungkyunkwan University)

接受日字:2000年1月24日, 수정완료일:2000年7月26日

얼굴변화와 TTS 합성음에 따른 입 모양 변화의 실시간 합성 방법[16]을 제안하고, DHMM 기반의 음성인식 시스템과 TD-PSOLA 방식의 음성합성을 구현하였다. 그리고 Motion Tracking은 계산량을 줄이고 배경 잡음에 덜 민감하게 하기 위하여 Fast Optical Flow Like Method를 제안하고, 이들을 통합한 응용 예로서 PC상에서 실시간으로 사용자와 말로써 구구단 게임이 가능한 가상캐릭터를 구현하였다.

이 통합 예는 각 시스템들의 독립적인 실행을 위해서 서로의 Process를 침범하지 않고 계산량과 CPU 점유율을 최소화하여 설계하였고 시스템들간의 통신은 Event 방식을 취함으로써 서로의 현재 상태를 투명하게 하면서 즉각적인 반응을 하도록 설계하였다.

Ⅱ. 3차원 캐릭터 Animation 시스템

인간의 얼굴을 Modeling하는 것은 가장 난해한 Modeling 분야 중 하나로써 이는 인간의 얼굴이 다른 신체 부위에 비해 매우 복잡할 뿐 아니라 사람을 볼 때 가장 먼저, 그리고 자세히 보게 되는 부위가 얼굴이기 때문에 미묘한 차이도 쉽게 부각 될 수 있기 때문이다.

1. 3차원 캐릭터 Model의 구성

얼굴의 Animation을 구현하기 위한 방법들은 실제로 얼굴 모습을 표현하는 피부 Model을 어떤 방법으로 만들었는가에 따라 좌우될 수 밖에 없다. 3차원 피부 Model 생성에 가장 흔하게 이용되는 Polygon Mesh의 경우 상대적으로 조작이 쉽지만 부드러운 곡면으로 이루어지는 얼굴을 사실적으로 표현하기 위해서는 Model을 구성하는 다각형의 수가 많아져야 하는 단점이 있다. 1990년 Waite는 B-Spline 곡면을 이용하여 얼굴 Model을 표현하였는데[1], 이 Model은 얼굴 표정들을 나타낼 때 비교적 적은 수의 Data로 사실적인 표현을 할 수 있었으나 치아 및 눈썹 등과 같이 세밀한 부분의 처리가 불가능하였다.[2]

본 논문에서는 3차원 캐릭터의 Model을 구성함에 있어서 PC환경에서의 실시간 계산을 위하여 많은 계산을 필요로 하는 캐릭터의 정밀도를 높이기 보다는 눈에 자연스러울 정도의 사실성을 중심으로 한 Polygon Mesh방식으로 얼굴을 Modeling하였다. 그리고 하나의 Frame으로 얼굴 Mesh를 구성하면 Frame의 색상 정보를 한 가지만 가지게 된다. 그러므로 각각의 Frame으로 각 얼굴의 부위를 구분하여 각각의 색상정보로 구성하였다.

Modeling Tool은 일반적인 3차원 Modeling Tool을 사용하였으며 DirectX의 Direct3D Rendering Engine[3]을 이용하여 Windows Platform에서의 3차원 Rendering 및 Animation을 수행하였다. 그림 1은 위와 같은 방법으로 구현한 3차원 가상캐릭터의 얼굴 모습이다.

(c) Wire Frame (d) Solid Fill

그림 1. Modeling된 가상캐릭터의 3차원 얼굴
Fig. 1. 3D face of cyber character.

2. 3차원 캐릭터의 기본 동작 및 표정 변화

(1) 3차원 캐릭터의 기본 동작

캐릭터 Model의 Animation 작업은 크게 나누어 감정 표현, 즉 표정의 Animation과 대화 시 입술 모양의 변화(Lip Synchronization)를 중심으로 하는 대화 Animation 부분으로 구분할 수 있다. 표정 Animation일 경우 약간의 문화적 차이를 제외한다면 거의 세계 공통의 보편적인 요소들로 이루어지는 반면, 대화 Animation의 경우는 언어에 따른 차이를 고려해야 한다.

얼굴 Model에 대한 동작 제어 Animation 기법은 주로 Parametric Model에 의한 접근 방법이라고 할 수 있다. 즉, 표정 변화의 기본 요소들의 조합에 의해 임의의 표정이나 발음 모양을 조합하는 방식이다. 이렇게 함으로써 적은 수의 Parameter의 제어만으로 다양한 표정 및 대화 Animation의 생성이 가능해 진다. 이 기법은 Key Frame Animation[2]에 비해 수작업이 줄어드는 장점이 있다.

Parametric Model의 근간이 되는 것이 바로 1970년대에 Ekman과 Friesen에 의해 제시된 Facial Action Coding System(FACS)[4]으로 이것은 각 안면 근육들의 위치와 운동 형태 및 이들이 얼굴 표면에 미치는 영향을 조사하고 이로부터 43개의 대표적 Action Unit(AU)를 포함한 총66개의 AU들을 정의한 것이다. Action Unit (AU)이란 안면 근육의 변화에 의해 얼굴에 나타

날 수 있는 표현들의 기본 요소로, 거의 모든 표정이 하나 또는 그 이상의 Action Unit들의 조합에 의해 표현될 수 있다.

본 논문에서는 이러한 여러 가지 Animation 기법 중에서 Parametric Model에 의한 접근 방법을 이용하여 모두 20개의 Action Unit을 선정하였다. 표 1에 본 논문에서 선정한 Action Unit을 정의하였고, Action Unit에 정의 되지 않은 표정은 Action Unit의 조합으로 생성해 낸다.

표 1. 선정된 Action Unit
Table 1. Selected action unit.

번호	Action Unit	번호	Action Unit
1	'아' 발음의 입모양	11	눈 상하 회전
2	'어' 발음의 입모양	12	얼굴 좌우 회전
3	'오' 발음의 입모양	13	얼굴 상하 회전
4	'우' 발음의 입모양	14	얼굴 좌우 갸우뚱
5	'으' 발음의 입모양	15	눈썹 슬픔
6	'이' 발음의 입모양	16	눈썹 화남
7	'에' 발음의 입모양	17	눈썹 웃음
8	좌측 눈꺼풀 내림	18	얼굴 화남
9	우측 눈꺼풀 내림	19	얼굴 웃음
10	눈 좌우 회전	20	아무런 표정 없음

Vertex Interpolation

Vertex Interpolation은 Lip Synchronization나 얼굴 표정의 변화에 적용하였다.

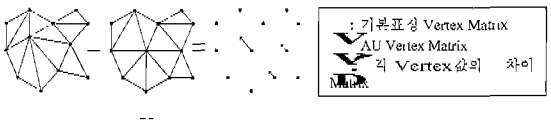


그림 2. Vertex Interpolation의 기본 형태
Fig. 2. Basic shape of vertex interpolation.

그림 2와 같이 아무런 표정이 없는 얼굴을 기본 표정으로 가정하고 기본 표정 얼굴의 i번째 Frame에 대한 Vertex 집합을 $\mathbf{V}_i$, j번째 Action Unit의 i번째 Frame에 대한 Vertex 집합을 $\mathbf{V}_{ji}$'라고 하면, 두 i번째 Frame의 Vertex 집합의 차이 $\mathbf{D}_{ji}$는 (2.1)식과 같이 정의 할 수 있다.

$$\mathbf{D}_{ji} = \mathbf{V}_{ji}' - \mathbf{V}_i \tag{2.1}$$

$\mathbf{D}_{ji}$는 기본 표정 얼굴과 각 Action Unit에 대한 변화량으로 정의될 수 있고, 이러한 변화량에 각 Action Unit에 따른 Weight w_j를 준 것에 기본 표정 얼굴의 Vertex 집합을 더함으로서 (2.2)식과 같이 Vertex Interpolation된 특정한 얼굴 표정의 각 Frame에 대한 Vertex 집합 $\mathbf{V}_i''$을 구할 수 있다.

$$\mathbf{V}_i'' = \left(\sum_j w_j \mathbf{D}_{ji}\right) + \mathbf{V}_i \tag{2.2}$$

Frame Interpolation

Frame Interpolation은 그림 3과 같은 3가지의 회전 Weight를 동시에 처리함으로써 이루어진다.

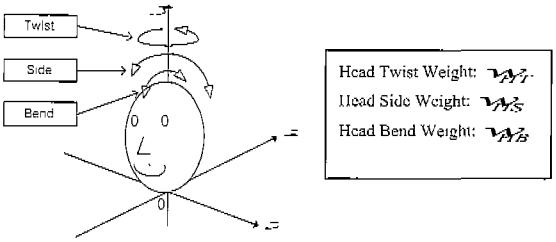


그림 3. 얼굴의 회전 방향과 그에 따른 Weight
Fig. 3. Rotational direction of face and its weight.

2. 3차원 캐릭터의 표정 변화

가상캐릭터와 사용자 간의 의사 통신에 단순히 TTS를 통한 음성 출력만으로는 전달하려는 뜻을 정확히 표현하기 어렵기 때문에 정보에 대한 명료도를 높이기 위한 감정에 따른 얼굴 표정 변화가 필요하다. 3차원 캐릭터의 얼굴 표정은 Vertex Interpolation과 Frame Interpolation을 이용하여 여러 가지의 Action Unit에 대한 Weight를 주어 합성하는 방법으로 생성해낸다. 그림 4는 여러 가지의 Action Unit의 합성을 통한 몇 가지의 복합 표정변화를 나타냈다.

(a) 웃음 :
AU#17 w_j=1.0
AU#19 w_j=0.7

(b) 화남 :
AU#16 w_j=0.7,
AU#18 w_j=0.7

그림 4. 복합 표정 변화 및 그에 따른 Action Unit Weight
Fig. 4. Variation of compound facial expression and its weight.

Ⅲ. 음성합성 시스템

음성합성 시스템은 반음절(Demi-syllable) 단위 TD-PSOLA (Time Domain-Pitch Synchronous Overlap-and Add) 방식[5]을 여성화자 Data Basc에 적용하여 TTS 방식으로 구현하였다. 한국어 표준 음운규칙 적용하여 자연스러운 음운 변화를 구현하였다.

전체 음성합성 시스템의 구조는 그림 5와 같다.

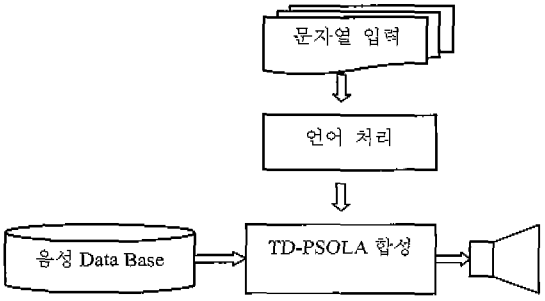


그림 5. 음성합성 시스템의 구조
Fig. 5. Structure of speech synthesis system.

1. 음성 Data Base

합성음의 Data Base는 반음절단위로 하였다. 한국어의 한 음절이 기본적으로 CVC형태의 초성, 중성, 종성을 가지므로 두 가지 형태의 반음절 즉, CV형의 초성-중성과 VC형의 중성-종성의 음성 Data Base만 작성하면 452개의 반음절로 한국어 음성을 무제한 합성할 수 있다.

반음절 단위의 Data Base를 작성하기위해 음절을 두 개의 반음절로 나누는데 그렇게 함으로써 초성과 중성, 중성과 종성의 연결부위인 천이구간 정보를 그대로 저장할 수 있었고, 접합부위의 부자연스러운 현상은 선형적 보간에 의해 어느 정도 해결할 수 있었다.

2. 언어 처리

본 논문의 음성합성 시스템은 한글, 영문, 숫자 등이 문서에 혼용되어 있을 경우 문장 분석을 통하여 한국어 텍스트로 변환하여 합성하였다.

하나의 음성 출력은 문장 단위로 처리를 하고 한 문장의 끝은 마침표로 구분하였다. 입력된 문장을 최종적으로 한국어 음성으로 합성하여 출력하려면 입력문장이 한국어의 발음상의 문장으로 다시 변환되어야 하는데 이를 위해 여러 가지 한국어 문법이 적용되어야 한다. 기존의 완성형에서 입력될 수 없는 몇 가지 음절을 모두 수용하기 위하여 한글의 경우 Uni-Code로 처리하였다.

영문의 경우는 이미 작성해 놓은 12만개 정도의 단어에 대한 국제음성기호와 한글 대조표를 이용하여 해당단어가 발음될 수 있도록 하였다. 또한 ODBC Data Base System을 이용하여 알파벳 순으로 탐색을 하여 상용 단어는 물론 약어까지 처리가 가능하게 하였다.

숫자는 문장 내에 사용되는 형태에 알맞게 변환 되도록 하였다. 숫자와 연결되는 앞 뒤 음절을 분석하여 해당 숫자의 성격을 파악한 다음 그 성격에 따른 자릿수에 맞도록 정확하게 한국어로 변환하였다.

이러한 방법으로 최초 입력문장에 대해 각각 처리를 하게 되면 음절단위의 초성,중성,종성의 세가지 구성요소가 결정이 된다. 이와 같이 입력문장 자체에 대한 일차적인 변환이 되면 다음에 발음상의 규칙을 적용하여 초성, 중성, 종성의 배열을 실제 발음상의 구조로 다시 변환하였다[6]

표 2에 본 합성시스템에 적용된 음운변동규칙을 나타내었다.

표 2. 한국어 자,모음의 변동 규칙
Table 2. Change rule for Korean consonant and vowel.

<자음의 변동 규칙>

대표 변동 규칙	내 용
대치	평폐쇄음화, 비음화, 유음화, 조음위치동화, 경음화
탈락	자음군단순화, 용어간말, ㅎ탈락, 중복자음감축, ㄷ탈락, ㄹ탈락, 비음탈락
첨가	중복자음화, ㄴ 첨가, 동음첨가
축약	유기음화(순행적 유기음화, 역행적 유기음화)

<모음의 변동 규칙>

대표 변동 규칙	내 용
대치	모음조화, 움라우트, y반모음화, w반모음화, 전설모음화
탈락	'으'탈락, 동모음 탈락, y 탈락, w 탈락
첨가	y 첨가
축약	모음축약
이음과정	불파음화, 유성음화, 설측음화, 구개음화, w조음위치동화, y탈락

본 논문에서는 적용된 규칙에 맞지않는 몇 가지 경우를 예외로 분류하여 처리하였다.[7]

3. TD-PSOLA 합성

본 논문에서의 음성합성 방식은 TD-PSOLA 방식을 이용하였다. 이 방식은 이미 작성해 놓은 음성 Data Base를 시간영역에서 합성하는 방식으로 다른 방식에 비해 합성음의 명료도나 자연성이 뛰어나다. 미리 구성해 놓은 반음절 단위 Data Base로부터 원하는 음절을 생성하기 위해서는 음절의 앞부분과 뒷부분의 반음절을 시간 축에서 붙여서 만들어 낸다. 그리고 음성에서 각 Pitch를 측정하고 이를 기준으로 Windowing을 한 후 Windowing된 데이터를 중첩하여 TD-PSOLA합성을 한다. 하지만 단순히 반음절을 붙이면 전위반음절과 후위반음절의 접합 부분에서 음성이 자연스럽지 못하게 된다. 이러한 문제를 해결하는 방법은 전위반음절과 후위반음절을 중첩하여 접합 시키고, 접합부분을 보간하여 자연스럽게 하도록 해야 한다.

Ⅳ. 음성인식 시스템

본 논문에서는 사용자와 가상캐릭터와의 구구단 게임을 할 수 있도록 DHMM (Discrete Hidden Markov Model)을 이용한 음성인식 시스템을 표 3과 같이 구성하였다[8]

표 3. 음성인식 시스템의 구성 및 Algorithm
Table 3. Construction and algorithm of speech recognition system.

인식 대상	구구단 게임
인식 Algorithm	DHMM(Discrete Hidden Markov Model), 10 State
VQ	K-Means 128 level
특징 Feature	16차 Weighted MFCC(Mel Frequency Cepstrum Coefficient)
음성 Data Base	구구단 게임을 하기 위한 숫자 단어와 게임 운영시 필요한 단어(일일~구구,일~팔십일,사이버맨,예,아니오,남자,여자 등) 사무실환경, 남자 100명 여자 30명, 11025Hz, 16bit Sampling

HMM은 음성 인식 시스템 전체 구성 중에서 훈련 과정 및 인식 과정을 수행하는 Algorithm으로서 1970년대 말부터 음성 인식 Algorithm으로 많이 사용되어 왔다. 최근에는 높은 인식률과 빠른 인식 시간 때문에 대용량 음성 인식 시스템에 많이 사용되고 있다.[8]

본 논문에서는 이러한 HMM Algorithm 중 Symbol의 개수가 유한한 DHMM Algorithm을 이용하여 화자 독립 음성인식 시스템을 구현하였다. 일반적으로 DHMM은 인식단어의 수가 1000단어 이하 일 경우와 Training용 음성 Data의 수가 적을 때는 CHMM (Continuous Hidden Markov Model)보다는 인식 성능이 좋은 것으로 알려져 있다.[9]

Ⅴ. Motion Tracking 시스템

1. Motion Tracking

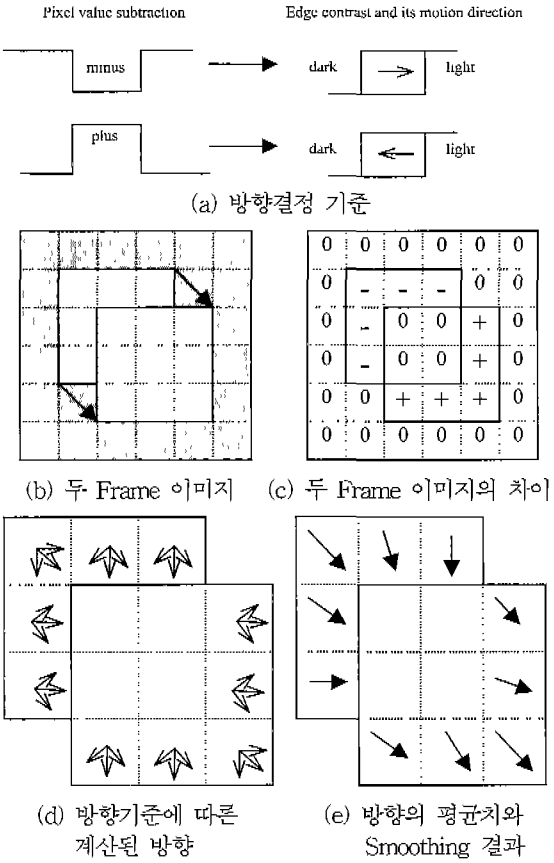
임의의 카메라로 받은 이미지로부터 움직임을 추적하는 것을 Motion Tracking이라고 하는데, 이러한 방법은 일반적으로 이미지의 작은 Segment단위로 Correlation을 계산하는 방법과 시간상의 이미지 차로 구하는 방법 등 여러 가지가 있다.[10] Correlation을 이용하는 방법은 계산량이 많을 뿐 아니라, 배경이 복잡하거나 물체가 회전을 할 경우에는 움직임을 잘 못 Detection하는 경우가 많기 때문에 본 논문에서는 복잡한 배경에서도 적용이 가능하면서 PC상에서의 실시간 계산을 위하여 시간상의 이미지 차로 구하였다. 그러나 단순히 시간상의 이미지 차로 구현한 것은 물체의 움직임의 방향 요소를 측정할 수 없기 때문에 보완적인 요소로 Optical Flow Method[11]를 사용하였다. 그렇지만 전형적인 Optical Flow Method는 현재 PC상에서 실시간 계산이 불가능하므로 본 논문에서는 Fast Optical Flow like Method를 제안하고 3차원 가상캐릭터 Model의 눈과 얼굴의 회전에 적용을 하였다.

2. Fast Optical Flow Like Method

전형적인 Optical Flow Method는 수식이 복잡하고 모든 Pixel의 계산을 요구함으로 Hardware로서 구현하기 전에는 실시간 계산이 불가능하다. 그러나 Optical Flow Method의 기본 아이디어가 앞 뒤 이미지 Pixel의 차이가 빠르게 변화하는 쪽을 움직임의 방향으로 간주하는 것을 생각한다면 그림 6과 같이 Fast Optical Flow Method를 만들어 낼 수 있다.[12]

그림 6과 같은 방법은 전형적인 Optical Flow보다 빠른 계산을 할 수 있다. 그러나 배경이 하나의 색으로 되어 있지 않으면 제대로 된 값을 얻을 수 없으므로 본 논문에서는 Fast Optical Flow의 단점을 보완하기

위해 그림 7과 같은 Fast Optical Flow Like Method을 사용하였다. 이 방법은 카메라로부터 받은 이미지의 시간상의 차를 이용하여 배경 이미지를 제거하는 것이다.

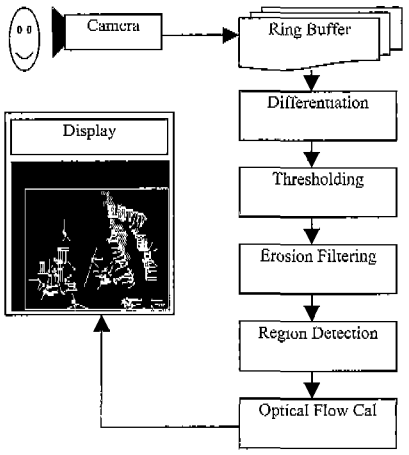


(a) 방향결정 기준

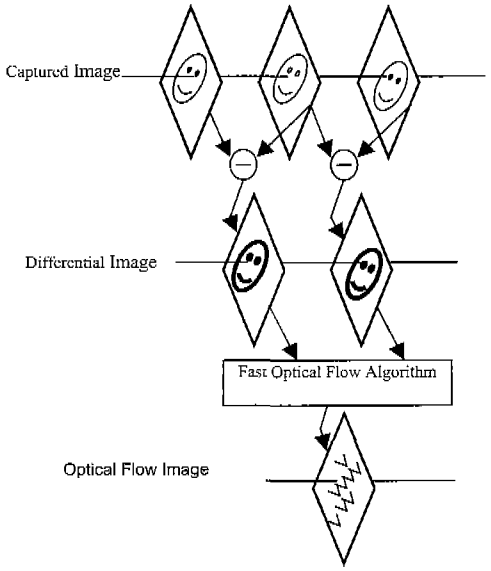
(b) 두 Frame 이미지 (c) 두 Frame 이미지의 차이

(d) 방향기준에 따른 계산된 방향 (e) 방향의 평균치와 Smoothing 결과

그림 6. Fast Optical Flow Method.
Fig. 6. Fast optical flow method.



(a) 전체 시스템 구성도



(b) Fast Optical Flow Like Method Algorithm

그림 7. Fast Optical Flow Like Method.
Fig. 7. Fast Optical Flow Like Method.

VI. 시스템 통합

1. Lip Synchronization

Lip Synchronization는 반음절 단위의 TD-PSOLA 방식으로 구현한 TTS와 연동하여 이루어진다. 음절이 발음될 때 대부분의 입 모양은 주로 모음이 결정하지만 순음(입술의 마찰에 의한 소리, 'ㅂ','ㅁ','ㅍ' 등)과 같은 자음성분에 의해서도 모양이 결정된다. 그리고 'ㅖ'와 'ㅐ'는 한국어의 표준 단모음[13]에는 포함 되지 않지만 이중 모음의 음운 천이 특성을 지녔다기보다는 단모음의 특성을 많이 따르므로 주요 모음으로 간주했다. 그러나 'ㅢ'와 같은 이중 모음은 'ㅡ'에서 'ㅣ'로의 천이 특성을 가지기 때문에 위에 정의한 주요 단모음으로 분해, 표현 가능하다. 초성이 순음일 경우 입술이 순간적으로 붙으므로 그에 따라 표현을 해 주지 않으면 입의 모양이 어색하게 된다.

음성의 출력과 3차원 캐릭터 입 모양과의 동기화는 일반적으로 PC의 Timer를 이용하여 구현하지만 그러한 방식은 PC의 Timer가 부정확하다는 이유와 사실상 많은 계산을 하고 있을 때는 Timer가 작동을 하지 않기 때문에 적절한 방법이 아니다.

좀 더 정확한 방법은 음성이 실제적으로 출력 될 때 한 음절이 출력 시작, 안정구간의 시작 그리고 안정

구간의 끝 지점이 되면 특정한 Event를 발생시키는 방법이다. 만약 "가곡"이라는 음성을 발음하기 위해서는 그림 8과 같다.

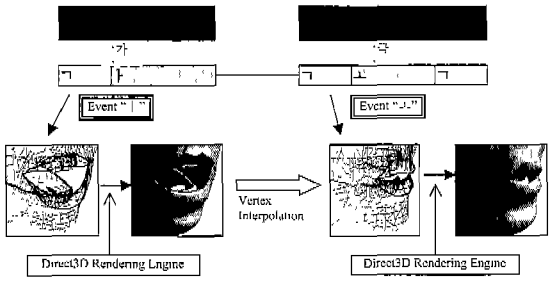


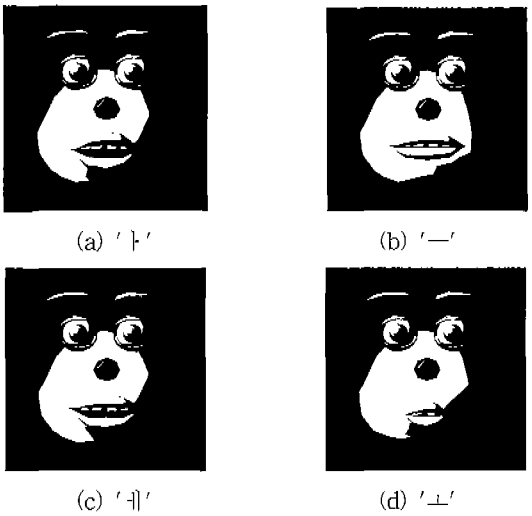
그림 8. Lip Synchronization의 예 ("가곡"을 발음할 때)

Fig. 8. An example of the lip synchronization,

그리고 출력할 음성의 세기(P)에 따라 입의 움직임 강도를 바꿔줘야 하므로 출력하기 전 미리 음성의 세기 정보를 순서대로 TTS에서 지정해 주고 Weight를 (6.1)식과 같이 다시 계산하여 보간을 할 필요가 있다.

$$\mathbf{v}_I'' = \left(\sum_J PW_J \mathbf{d}_{JI}\right) + \mathbf{V}_I \qquad (6.1)$$

위와 같은 방법으로 자모음을 표현 한 것을 그림 9에 나타냈다.



(a) 'ㅏ' (b) 'ㅡ'

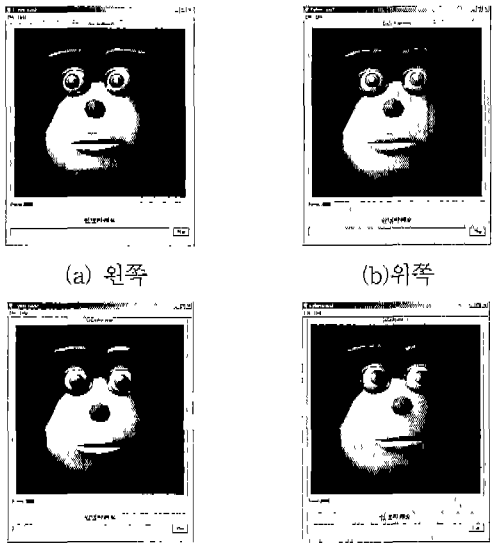
(c) 'ㅔ' (d) 'ㅗ'

그림 9. 발음에 따른 입술 변화

Fig. 9. Lip variation for pronunciation.

2. Motion Tracking과 3D Animation 통합

본 논문에서는 Motion Tracking을 이용하여 가상캐릭터의 눈 및 얼굴을 회전 시킴으로써 사용자를 계속적으로 주시하도록 하였다. 그림 10은 사용자의 위치에 따른 가상캐릭터의 얼굴과 눈의 회전을 나타낸 것이다. Motion Tracking 시스템은 많은 계산량을 요구한다. 그러므로 상대적으로 3D Animation 시스템이나 음성인식 시스템에 자원이 부족하게 할당됨으로써 부자연스러운 3D Animation이나 오인식을 나타내는 음성인식 시스템을 설계하게 된다. 원활한 연동을 위해서는 CPU 자원의 적절한 할당이 중요하다.



(a) 왼쪽 (b)위쪽

(c) 아래쪽 (d) 오른쪽

그림 10. 사용자의 위치에 따른 가상캐릭터의 눈,얼굴 회전

Fig. 10. Eye and face rotation of cyber character for user's position.

3. 시스템 통합 원칙

3차원 캐릭터 Animation과 음성합성, 음성인식, Motion Tracking의 통합은 구구단 게임을 설계하는 것으로 통합되었다. 이 4가지 Engine의 통합의 원칙은 다음과 같다.

(ㄱ) 모든 각 Engine이 독립적으로 설계 되어 서로의 Process를 침범해서는 안 된다.

(ㄴ) 하나의 Process가 실행될 경우에도 다른 Process가 실행 권한을 잃지 않아야 한다.

(ㄷ) 각 Engine들은 모두 Event방식으로 서로간에 통신을 한다.

(ㄹ) 각 Engine들은 현재 동작 상태에 대해서 다른

Engine에게 투입해야 한다.

(ㅁ) Timer의 부정확성으로 인해 PC의 Timer를 사용하지 않는다.

(ㅂ) 메모리와 계산량 최소를 위해 동적인 메모리 할당과 Loop를 가능한 짧게 해야 한다.

이러한 방식으로 설계한 시스템의 전체 종속 관계와 실행 Flow를 그림 11에 나타내었다.

캐릭터의 사실성을 부각시키기 위하여 아무런 Event가 발생되지 않을 경우에는 눈의 깜박임과 얼굴, 눈에 약간의 움직임을 두었다. 그리고 사용자가 음성명령을 내리고 있는 상황이나, 합성을 하고 있는 동안에도 가상캐릭터의 눈과 얼굴의 방향은 사용자를 볼 수 있도록 계속적으로 Motion Tracking을 하였고 사용자로의 명령은 음성인식 시스템에서 분석을 하고 그에 따라 TTS 음성합성 시스템과 3차원 캐릭터 Animation 시스템이 연동을 하여 명령에 대한 결과 물을 사용자에게 제시하였다.

Ⅶ. 성능 평가

본 논문에서 구성한 통합 시스템의 성능평가 대상 PC는 Pentium Ⅱ 400MHz의 CPU에 Frame grabber가 없는 일반적인 USB port 카메라, DirectX 가속이 가능한 비디오카드와 Full Duplex가 가능한 사운드카드를 장착하여 실험하였다.

음성합성 시스템 성능 평가를 위해서 합성하기 전의 원음과 합성 후 음성의 LPC 12차 분석을 통한 Smoothing된 Spectrum 차이를 보면, 평균 약 9dB정도의 차이를 나타냈다.

음성인식 시스템에서 총 인식 대상 단어의 수는 구구단 게임을 진행 할 때 필요한 한자리 숫자음과 두자리 숫자음, 진행상 필요한 단어 등 총 135개 이며 화자종속 발성일 경우에는 평균92%, 화자독립 발성일 경우 평균 88.3%의 성능을 나타내었다..

Motion Tracking 시스템 평가에서 중요한 것은 정확

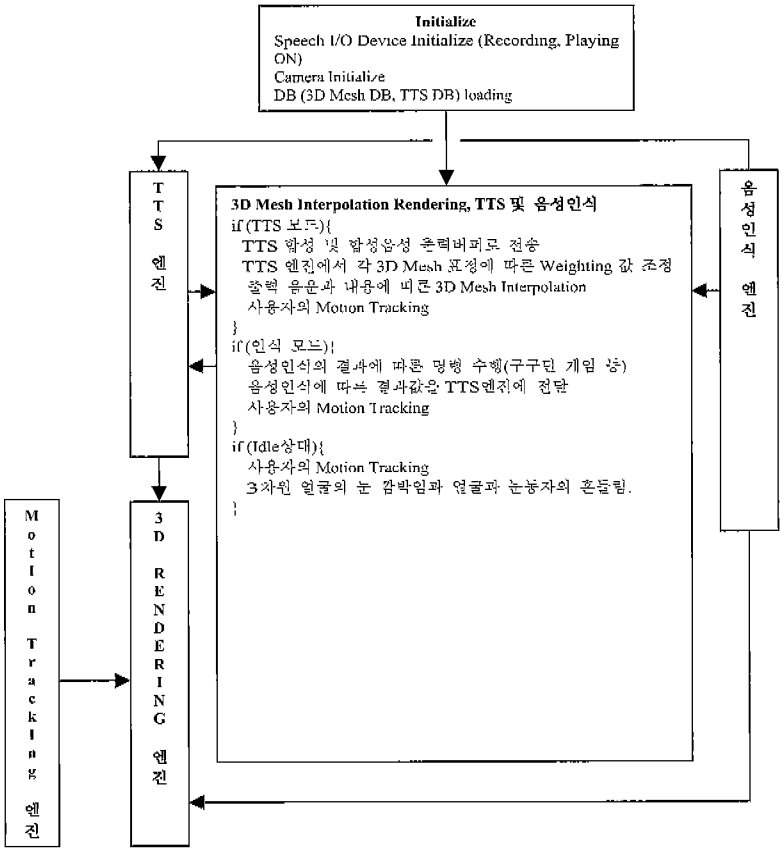


그림 11. 통합 시스템(구구단 게임)의 기본 구조
Fig. 11. Basic structure of integrated system.

Motion Tracking 시스템 평가에서 중요한 것은 정확성과 초당 처리할 수 있는 Frame수 이다. 본 논문의 Motion Tracking 시스템은 계산량을 단순화 시킴으로써 초당 3~4Frame의 속도를 나타냈다. 즉 다시 말해서 카메라 앞을 약 250msec 보다 천천히 움직이는 것은 Tracking이 가능하고, 그 보다 빨리 움직이는 것은 Tracking이 되지 않는 다는 것이다. 대부분의 사람의 움직임은 250msec보다 카메라 앞을 빨리 움직이는 일이 적으므로 Motion Tracking이 가능하였다.

Ⅷ. 결 론

본 논문에서는 새로운 Human Interface인 3차원 가상캐릭터를 구현하기 위하여, TD-PSOLA방식의 음성합성 시스템과 DHMM 음성인식 시스템, Fast Optical Flow Like Method를 이용한 Motion Tracking, Vertex Interpolation을 이용한 3차원 캐릭터 Animation를 구현하였고 이들을 통합하여 PC상에서 실시간으로 사용자와 구구단 게임을 할 수 있는 통합 시스템을 구현하였다.

3D 캐릭터 Animation은 Parametric 합성 방식을 이용하여 20개의 Action Unit을 선정하고 Vertex Interpolation과 Frame Interpolation 방법을 통하여 거의 모든 감정 표현과 Lip Synchronization 모양을 조합해 낼 수 있었다.

음성합성 시스템은 음운 변환 규칙을 적용하여 자연스러운 음질을 나타냈고, 3D 캐릭터의 입 모양과의 동기화는 Timer를 사용하는 것이 아니라 Event방식으로 구현 함으로써 시간에 따라 변화하는 입 모양을 정확한 시점에 표현하였다.

Motion Tracking 시스템에는 배경 이미지를 제거하는 방법을 채택함으로써 복잡한 배경 환경에서도 Tracking이 가능하였고 Fast Optical Flow Like Method를 이용하여 계산량을 최소화 시켰다. 그리고 이를 통합 시스템에 적용하여 사용자의 위치를 추적하면서 계속적으로 사용자를 주시하도록 하였다.

음성인식 시스템에서는 DHMM 인식 Algorithm을 이용하여 통합시스템에서 사용자와 가상캐릭터가 구구단 게임을 할 수 있는 인식 시스템을 구현하였다.

선정된 Algorithm들은 서로의 Process를 침범하지 않고 PC에서도 실시간으로 각각의 시스템들이 독립적으로 작동하도록 조합 계산량과 CPU 사용량을 최소화하여 통합하였다.

이와 같이 통합하여 구현된 가상 캐릭터는 Human Computer Interface 분야에서 사용자에게 친숙한 가상 도우미의 역할과 방송산업에서의 가상 나레이터, 교육분야에서 원격교육등에 응용될 것으로 기대된다.

참 고 문 헌

[1] C.T. Waite, "The Facial Action Control Editor, Face: a Parametric Facial Expression Editor for Computer Generated Animation", *Master Thesis*, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1990.

[2] 김웅순, 김영수, "3차원 캐릭터 애니메이션 기술동향", *정보과학회지*, 제17권2호, pp 48-99, 1999

[3] Microsoft DirectX 6.1 SDK Direct3DRM Document, 1998.

[4] P. Ekman and W. V. Friensen, "Facial Action Coding System", *Consulting Psychologist Press*, 1977.

[5] F. Charpentier, E. Moulines, "Pitch-synchronous wave-form processing techniques for text-to-speech synthesis using diphones" Proc. Eurospeech, pp. 2:13-19, 1989.

[6] 배주채, "국어음운론 개설", *신구문화사*, 1996

[7] 김종우 외 3, "지능적 휴먼-컴퓨터 인터페이스를 위한 무제한 음성합성 시스템 구현", *대한전자공학회* 멀티미디어 연구회 창립학술발표, pp209-212, 1999

[8] Lawrence Rabiner, Biing-Hwang Juang, "Fundamentals of Speech Recognition", *Prentice Hall*, 1993.

[9] Chin-Hui Lee, Frank K.Soong, "Automatic Speech and Speaker Recognition", *Kluwer Academic Publishers*, 1996.

[10] Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, "MACHINE VISION", *McGraw-Hill* 1995

[11] Berthold Klaus Paul Horn, "Robot Vision", *MIT Press*, 1986.

[12] W.T. Freeman, "Computer Vision for Interactive Computer Graphics", *IEEE CGA*, pp 42-53 May-June 1998.

[13] "한국어 맞춤법 통일안", *한글학회*, http://www.hangeul.or.kr

[14] 이주상, 유지상 "MPEG-4 SNHC 기반 얼굴 객체의 구현" *Telecommunication Review* ,제8권 3호, pp 400-409, 1998

[15] 박재용, 박승수, "실시간 얼굴 애니메이션에서 효율적인 표정관리와 한글 립싱크", *한국정보과학회*, HCI '99 학술대회, pp 675-686, 1999

[16] 최광표 외 3, "사이버 에이전트를 위한 3D얼굴 애니메이션", 한국정보처리학회, 제13회 산학연멀티미디어학술대회, pp 204-207, 1999

[17] Nadia Magnenat Thalmann, "Interactive Computer Animation", *Prentice Hall*, 1996.

저 자 소 개

崔 光 杓(正會員)

1973년생. 1998년 2월 성균관대학교 전자공학과 졸업(공학사). 2000년 2월 성균관대학교 전기전자 및 컴퓨터공학부 졸업(공학석사). 현재 동 대학원 박사과정. 주관심분야는 신호처리, 데이터통신

李 斗 誠(正會員)

1952년 12월 22일생. 1979년 2월 성균관대학교 전자공학과 졸업(공학사). 1981년 2월 성균관대학교 대학원 전자공학과 졸업(공학석사). 2000년 2월 성균관대학교 대학원 전기전자 및 컴퓨터공학과 박사과정 수료. 1983년 9월 - 현재 서일대학 전자과 부교수. 주관심분야는 음성인식, 디지털신호처리

洪 光 錫(正會員)

1985년 성균관대학교 전자공학과 공학사. 1988년 성균관대학교 전자공학과 공학석사. 1992년 성균관대학교 전자공학과 공학박사. 1990-1993 서울보건전문대학 전산정보처리과 전임강사. 1993-1995 제주대학교 정보공학과 전임강사. 1995-현재 성균관대학교 전기전자 및 컴퓨터공학부 부교수. 관심분야 : 휴먼-컴퓨터 인터페이스, 음성인식 및 합성