표정짓고 말하는 가상 얼굴의 실시간 합성

Realtime Synthesis of Virtual Faces with Facial Expressions and Speech

송 경 준*, 이 기 영**, 최 창 석***, 민 병 의*
(Kyung Joon Song*, Ki Young Lee**, Chang Seok Choi***, Byung Eui Min*)

요 약

본 논문에서는 고품질의 얼굴 동영상과 운율이 첨가된 음성을 통합하여 자연스런 가상얼굴을 실시간으로 합성하는 방법을 제안한다. 이 방법에서는 한글 텍스트를 입력하여, 텍스트에 따라 입모양과 음성을 합성하고, 얼굴 동영상과 음성의 동기를 맞추고 있다. 먼저, 텍스트를 음운 변화한 후, 문장을 분석하고 자모음사이의 지속시간을 부여한다. 자모음과 지속시간에 따라 입모양을 변화시켜 얼굴 동영상을 생성하고 있다. 이때, 텍스트에 부합한 입모양 변화뿐만 아니라, 두부의 3차원 동작과 다양한 표정변화를 통하여 자연스런 가상얼굴을 실시간으로 합성하고 있다. 한편, 음성합성에서는 문장분석 결과에 따라 강세구와 억양구를 정하고 있다. 강세구와 억양구를 이용하여 생성된 운율모델이 고품질의 음성합성에 필요한 지속시간, 억양 및 휴지기를 제어한다. 합성단위는 무제한 어휘가 가능한 반음절과 triphone(VCV)의 조합이며, 합성방식은 TD-PSOLA를 사용한다.

ABSTRACT

This paper proposes a real time method for synthesizing the natural virtual faces by integrating the high-quality facial image sequences and the prosodic speech. The method synthesizes the lip shapes and the speech according to the Korean text, and synchronizes with the lip shapes and the speech. After changing phonology of the input text and analyzing the sentences, the method gives the duration time between a consonant and a vowel each syllable of the sentences. The facial image sequences are synthesized by changing the lip shapes according to the consonants, the vowels and the duration times. The real time synthesis of the natural virtual face also includes the three dimensional head motion and the various facial expressions as well as the lip shape changes. In speech synthesis process, accentual phrases and intonational phrases is determined according to the sentence analysis results. The prosodic model for speech synthesis is composed of accentual phrases and intonational phrases. This prosodic model controls duration time, intonation and pause of synthesized speech. Synthesis units constitute of demi-syllables and VCV-triphones which can make unlimited vocabularies, and TD-PSOLA is used as the synthesis method.

I. 서 론

인간이 사회생활을 영위하는데 있어서 얼굴은 중요한 역할을 한다.[1][2] 손이나 발을 보고 개인을 식별하기는 어려울 때가 많지만, 얼굴을 보고 식별하지 못하는 경우는 거의 없다. 즉, 개인식별기능이 있기 때문이다. 상대와 대화를 할 때도 얼굴을 마주하고 말과 얼굴표정을 이용하여 자기의 의도와 감정을 표현하는 개인간 통신 기능이 있다. 이외에도 청각, 시각, 촉각, 음식물섭취 및 미각기능 등 다양한 기능을 갖추고 있어, 인간에 있어서 정보 입출력은 거의 얼굴을 통해서 이루어진다고 해도 과언이 아니다. 숫자와 문자중심의 맨·머신 인터페이스에 있어서도, 얼굴의 대화기능에 착안하여 얼굴표정과 말을 이용한 인터페이스가 궁극적인 휴먼인터페이스로 인식되어 최근 많은 관심의 대상이 되고 있다. 이러한 연구는 80년도 초반 Lippman, Lewis, Parke등에 의해 산발적으로 제안되어 오다가[3][4], 80년도 중·후반에 Alan Kay, Harashima(原島), Morishima(森島)등에 의해 본격적으로 연구되어 왔다.[5]-[8] 90년도 초반부터는 MPEG4에 얼굴합성기술과 TTS(Text To Speech)가 함께 포함되어 세계각국에서 활발히 연구되고 있다.[9]

초기에는 얼굴 표정의 합성과 음성의 합성이 개별적으로 연구되어 왔다. 얼굴 표정의 합성은 K.Waters등에 의해 얼굴 근육의 움직임에 따라 형상모델을 변형하여, 그 모델에 음영(shading)을 부여하는 렌더링 방법을 이용하였다.[10] 한편, Harashima그룹은 차세대 영상통신 방식으로서 모델

* 전자통신연구원
** 관동대학교 정보통신공학과
*** 명지대학교 전자정보통신공학부
접수일자 : 1998년 4월 22일

기반 영상부호화(Model-Based Image Coding)방식을 제안하고, 이 방식의 핵심기술로서 얼굴표정을 합성하게 되었다.[11]-[15] 이 방법에서는 K. Waters의 방법과는 달리 실제 얼굴 사진을 이용하여 고품질의 표정합성을 시도하였다. 동 그룹의 Morishima는 실세 음성과 합성영상을 통합하여 그래픽엔진 상에서 가상의 얼굴을 만들었고,[16] Kaneko (金子)는 합성음성과 합성얼굴을 통합한 시스템을 보드로 제작하였다.[17] MPEG4에 TTS가 채택되자 각국에서는 자국어와 동기된 합성음성과 합성영상을 통합하는 연구가 활발하게 되었다.[9]

한편, 국내에서는 ETRI의 Lee등과 필자들에 의해 한글에 부합한 연구가 이루어지고 있다. Lee등은 MPEG4표준안에 따라 MPEG4 TTS Interface를 제안하고 있다.[18] 그러나, Lee등의 인터페이스에서는 얼굴에서 입모양에 주목하고 있기 때문에, 두부의 3차원 동작과 얼굴 전체의 표정은 표현하기가 어렵다. 나아가서, 얼굴 동영상을 합성하기보다는 입모양 패턴을 음절에 대응시키고 있기 때문에, 자연스런 애니메이션을 얻기가 곤란할 것으로 보인다. 이에 대해, 필자등은 두부의 3차원 동작, 얼굴 전체의 표정과 함께 입모양을 프레임별로 파라미터를 조직하여 동영상을 생성하고 있기 때문에, 자연스런 애니메이션이 가능한 동영상을 얻을 수 있다. 또한, 합성 음성과 음절별 지속시간에 따라 동기를 맞추고 있기 때문에, 합성음성과 합성얼굴 동영상이 일체화된 시스템을 구축 할 수 있게 된다. 이러한 시스템의 구축을 위해 필자등은 Harashima 그룹에서 고품질 표정합성에 대하여 연구를 한 후,[14][15] 한글에 부합한 가상얼굴의 실현을 목표로 하고 있다.[19]-[24] 먼저, 한글자모음을 분석한 후, 한글발음에 필요한 기본형 입모양을 11개 패턴으로 분류하였다. 그리고, 합성얼굴동영상과 합성음성의 동기구현 방법에 대해서 검토하고 있다.

음성합성은 기계가 인간의 음성을 합성하는 기술로서 인간의 발성모델을 토대로 연구되고 있으며 성도의 전달함수와 성대의 진동특성을 모델링하여 구현되어 왔다.[22][24] 이러한 모델링에 의한 합성방식의 대표적인 것으로 LPC 계열의 파라메타 합성방식[25]-[27]이 주류를 이루어 오고 있었다. 그러나, 프랑스에서 non-parametric 합성방식으로 pitch-synchronous하게 분석하여 운율조절이 용이한 PSOLA 합성방식[28]-[30]을 제안한 이후, 합성음성의 명료도와 자연성이 대폭 향상된 방식으로 TD-PSOLA와 LP-PSOLA 등이 개발되었다. TD-PSOLA는 시간축에서 원음성 파형을 그대로 pitch-synchronous하게 분석한 non-parametric 합성방식이며, LP-PSOLA는 parametric 합성방식이다. 여기서 TD-PSOLA 방식이 LPC분석을 이용하는 LP-PSOLA 합성방식보다 명료성, 자연성이 향상될 뿐만아니라, 실시간으로 합성음성을 얻을 수 있으므로 TD-PSOLA 방식[24]을 이용하여 운율을 첨가한 음성을 합성한다. 또한 고품질의 한국어 음성합성을 위하여 필자등은 언어학적으로 제시된 억양구(intonational phrase)와 강세구(accentual phrase)[31][32]를 이용하여 운율모델을 생성하고 한국어 음성합성시스템[33]에 적용하고 있다.

본 연구에서는 얼굴표정과 운율이 첨가된 음성을 통합하여 가상얼굴을 실시간으로 합성하는 방법을 제안한다. 이 방법에서는 텍스트에 부합한 입모양의 변화뿐만 아니라, 두부의 3차원 움직임과 얼굴의 다양한 표정변화를 실시간으로 합성할 수 있다. 또한, 음성합성방법에서는 한글문장으로 된 텍스트를 억양구와 강세구로 분석한 후, 운율모델을 생성하고 TD-PSOLA 방식을 이용하여 실시간 합성한다. 이와 같이 다양한 얼굴표정변화와 더불어 두부의 3차원동작이 가능한 얼굴과 운율제어가 가능한 무제한 음성을 실시간으로 합성함으로써 생동감 있는 가상얼굴을 실현할 수 있게 된다.

II. 가상얼굴 시스템의 개요

고품질의 얼굴동영상을 실시간으로 합성하여, 입모양이 자연스럽게 변화하는 애니메이션을 얻기 위해서, 본 연구에서는, 얼굴의 3차원 형상모델을 구성하고, 실제 사진을 텍스처메핑 하는 방법을 이용한다. 이러한 가상얼굴을 실현하기 위해서는 합성음성에 동기된 입모양 변화가 가장 중요한 요소이지만, 가상얼굴에 자연스러움을 향상시키기 위해서는 두부의 3차원 동작과 얼굴전체의 표정변화 또한 빼놓을 수 없는 요소이다. 표정 짓고 말하는 가상얼굴 시스템의 개요를 그림 1에 나타낸다.

먼저 문장해석부에서는 한글로 텍스트를 입력하면, 텍스트를 음운 변화한 후, 문장 및 어절을 검출하고, 자모음을 코드로 변환한다. 변환된 자모음코드에 따라, 음성지속시간 DB로부터, 초성자음과 중성모음사이, 중성모음과 중성자음사이의 지속시간을 부여한다. 음절별 자모음과 지속시간에 따라 얼굴동영상과 음성을 각각 합성한다.

얼굴동영상 합성부에서는 얼굴합성을 시작하기 전에 가상 얼굴의 대상이 되는 얼굴정면상의 실세 사진을 수집하여 DB를 구축한다. 구축된 DB의 사진은 얼굴의 3차원 형상모델을 정합하여 개인 얼굴의 3차원 형상 모델을 얻는다. 입력된 텍스트의 자음과 모음에 따라 입모양을 결정하고, 지속시간에 따라 자음과 모음사이에 생성해야 할 프레임수를 산출한다. 이 과정에서, 두부의 3차원 동작과 표정변화에 대한 정보를 부여한다. 이 정보들은 파라미터로 표현되어, 얼굴동영상의 프레임별로 입모양, 두부동작, 표정변화의 파라미터가 결정된다. 이 파라미터로 얼굴의 3차원 형상모델을 변형한 후, 그래픽 가속기로 실시간 텍스처 매핑을 하면 얼굴동영상을 생성할 수 있게 된다.

음성합성부에서는 운율모델을 생성하기 위하여 운율분석을 수행하지만 이미 얼굴동영상 합성을 위해 입력된 한글 텍스트의 분석 결과를 그대로 이용하므로 분석시간을 최소화하고 있다. 이 분석과정에서 생성된 운율모델은 고품질의 음성합성에 필요한 운율을 제어한다. 합성방식은 합성단위 데이터베이스가 pitch-synchronous한 원음성 파형인 TD-PSOLA를 사용한다.

제 III장에서 얼굴동영상의 합성에 필요한 두부동작, 표정변화, 입모양 파라미터와 실시간합성에 대하여, 제 IV장에서는 고품질 음성합성을 위한 운율분석과 실시간 합성에 대하여 기술한다.

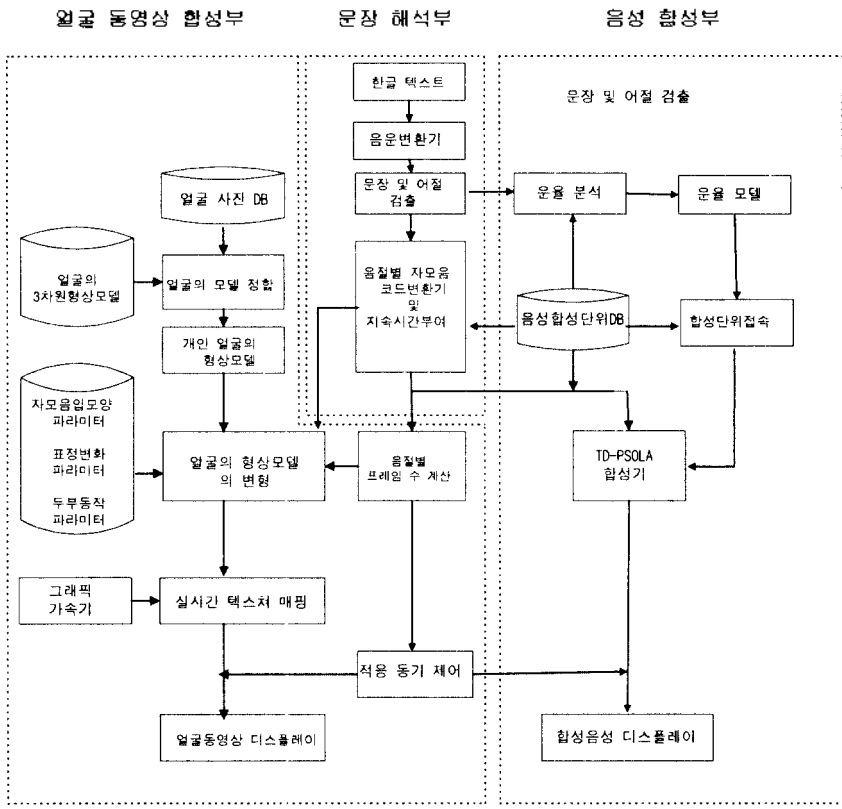


그림 1. 가상 얼굴의 실시간 합성의 개요
Fig. 1. Configuration for real time synthesis of the virtual faces.

III. 얼굴동영상의 실시간 합성

1. 얼굴의 3차원 모델

고품질의 얼굴동영상을 합성하기 위해서는 얼굴의 3차원 형상모델이 근간이 된다. 본 연구에서는 그림 2와 같은 형상모델을 준비하고 있다. 나아가서, 그림 3과 같은 실제 얼굴사진을 이용하여 합성동영상의 현실감을 높이기 위해서, 형상모델을 그림 3에 정합한다.(그림 4)[13][14][15] 이 과정은 음영(shading)을 이용하는 방법의 경우에는 불필요하다. 그러나, 음영을 이용하여 현실감 넘치는 특정개인의 얼굴을 합성하기가 곤란하기 때문에, 본 연구에서는 실제 사진을 이용하여 현실감 넘치는 얼굴동영상을 얻고 있다.

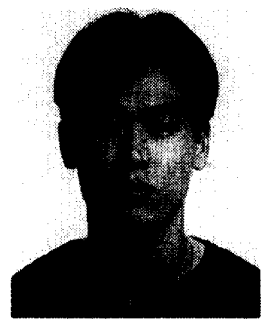


그림 3. 실제 얼굴 사진
Fig. 3. A real facial image.

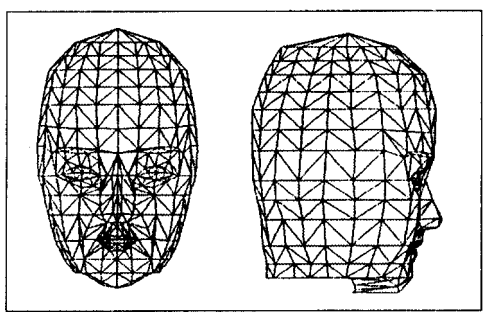


그림 2. 얼굴의 3차원 형상모델
Fig. 2. A 3-D facial shape model.

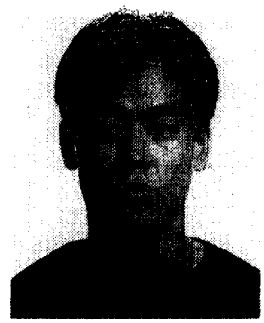


그림 4. 실제 사진과 형상모델의 정합
Fig. 4. Adjustment of the real image and the shape model.

2. 두부의 3차원 동작파라미터

두부의 3차원 동작은 회전운동과 병진운동으로 나눌 수 있으며, 카메라의 원근에 대응하는 스케일링(scaling)을 생각할 수 있다. 이러한 운동은 다음 어파인 변환으로서 실현할 수 있다.[34]

$$P' = SR(P - P_C) + P_C + T \tag{1}$$

여기서 $P = (x, y, z)^T$, $P' = (x', y', z')^T$, $P_c = (x_c, y_c, z_c)^T$는 각각 운동전, 운동후, 운동중심의 점이 된다. $T = (t_x, t_y, t_z)^T$는 x, y, z축방향의 병진량을 나타낸다. 또 S는 원근을 나타내는 행렬

$$S = \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & s_z \end{bmatrix} \tag{2}$$

이고, s_x는 x축 방향의 스케일을 나타낸다.

$$R = R_x R_y R_z \tag{3}$$

이다. R_x, R_y, R_z는 각각 x, y, z축을 중심으로 회전운동을 나타내는 행렬이다.

예를들면,

$$R_x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & -\sin\theta_x \\ 0 & \sin\theta_x & \cos\theta_x \end{bmatrix} \tag{4}$$

이다. θ_x는 x축을 중심으로 한 회전각이다. 이러한 행렬변환을 통해 얼굴의 3차원동작을 실현할 수 있다. 이와 같이 스케일링(s_x, s_y, s_z), 회전운동(θ_x, θ_y, θ_z), 병진운동(t_x, t_y, t_z)의 9개 파라미터로 두부의 3차원 동작을 기술할 수 있다.

3. 표정변화 파라미터

얼굴의 표정변화는 안면근육의 수축이완으로 이루어진다. Ekman은 복잡하게 분포되어있는 안면근육을 표정변화의 관점에서 독립적으로 움직이는 근육을 분류하고, 독립적인 근육의 움직임에 대한 표정변화를 Action Unit(AU)라고 정의해 놓았다.[35] 필자 등은 AU를 얼굴의 3차원 형상모델과 결부시켜 표정 애니메이션에 대한 연구를 지속해 오고있다.[14][15] 본 연구에서도 AU를 채택하여 표정변화를 표현하는 파라미터로 삼는다. AU는 44개가 정의되어 있으나, 필자들은 이들 중 34개의 AU를 실현하여, AU를 조합하고, AU강도를 변화시켜 여러 가지 표정을 고품질로 합성하고 있다. 실현된 AU리스트를 표 1에 나타낸다. AU중에서 빠져있는 번호는 문헌[35]에 정의되어 있지 않거나, 필자들이 아직 실현하지 않은 번호이다. 안구의 회전은 AU로 정의되어 있지 않으나, 필자들이 표정합성에 필요하다고 판단하여 AU에 포함시켜 놓고 있기 때문에, AU의 번호는 부여하지 않았다.

표 1. 안면 근육의 실현된 기본 동작(AU : Action Unit)의 일람
Table 1. List of Implemented Action Units of the facial muscles.

AU no.	AU 명	AU no.	AU 명
1	눈썹 내측을 올린다	20	입술 양단을 옆으로 당긴다
2	눈썹 외측을 올린다	23	입술을 강하게 다문다
4	눈썹을 내린다	24	입술을 상하로 누른다
5	윗눈꺼풀을 올린다	25	턱을 내리지 않고 아래 입술을 내린다
6	뺨을 올린다	26	턱을 내리면서 아래 입술을 내린다
7	눈꺼풀을 긴장시킨다	27	입을 크게 벌린다
8	입술을 서로 접근시킨다	28	입술을 빨아들인다
9	코에 주름을 잡는다	29	아래 턱을 내민다
10	윗입술을 올린다	30	턱을 좌우로 이동시킨다
11	팔자주름을 깊게 한다	32	입술을 깨문다
12	입술 양단을 끌어올린다	35	볼을 빨아들인다
13	입술 양단을 예리하게 끌어올린다	41	윗눈꺼풀을 내린다
14	보조개를 만든다	42	눈을 가늘게 뜬다
15	입술 양단을 내린다	43	눈을 감는다
16	아랫입술을 내린다	44	눈을 작게 뜬다
17	턱을 올린다	45	눈을 깜빡인다
18	입술을 좁힌다	46	윙크
			안구의 회전

4. 입모양 파라미터

텍스트에 대응한 입모양을 변화시키기 위해서는, 한글의 자모음에 대한 입모양을 정의해 놓을 필요가 있다. 필자 등은 한글의 자모음의 입모양에 대하여 조사 분류한 결과, 초기의 입모양은 초성자음에 의해서 이루어지고 중성모음에 따라 변화하여 종성으로 끝난다는 것을 알 수 있었다. 그 결과, 자음2개, 모음8개, 묵음1개, 도합11개에 대한 기본형 입모양 패턴을 준비하여, 모든 한글 발음에 대한 입모양을 합성하고 있다.[19] 즉, 자음은 입술소리(ㅁ, ㅂ, ㅃ, ㅍ)와 그 외의 소리의 2종류로 분류하고, 모음은 'ㅏ', 'ㅓ', 'ㅗ', 'ㅜ', 'ㅡ', 'ㅣ', 'ㅐ', 'ㅔ'의 8 종류로 분류하고 있다. 나아가서, 자모음에 대응한 입모양으로는 볼 수 없으나, 입모양 변화에는 필요한 묵음에 대한 입모양을 1개 추가한다. 이것은 입모양의 변화가 없는 무표정으로 한다. 'ㅐ', 'ㅔ'의 모음을 기본형 입모양으로 분류한 것은 'ㅏ'와 'ㅣ'의 입모양이 연속적으로 변화하는 것이 아니라, 입모양이 일관되게 변화하기 때문이다. 'ㅗ'와

표 2. 기본형 입모양에 대한 AU의 종류와 강도
Table 2. AUs and their intensities for the fundermental lip shapes.

입모양	AU의 조합과 강도
'ㅏ'	12(0.1), 26(0.45)
'ㅓ'	26(0.3)
'ㅗ'	18(0.3), 26(0.3)
'ㅜ'	18(0.5), 25(0.2)
'ㅐ'	12(0.1), 26(0.4)
'ㅔ'	26(0.3)
'ㅣ'	15(0.2), 20(0.2), 25(0.4)
'ㅡ'	20(0.3), 25(0.1)
입술소리	23(0.3)
그외소리	26(0.2)
묵음	입모양의 변화가 없는 무표정

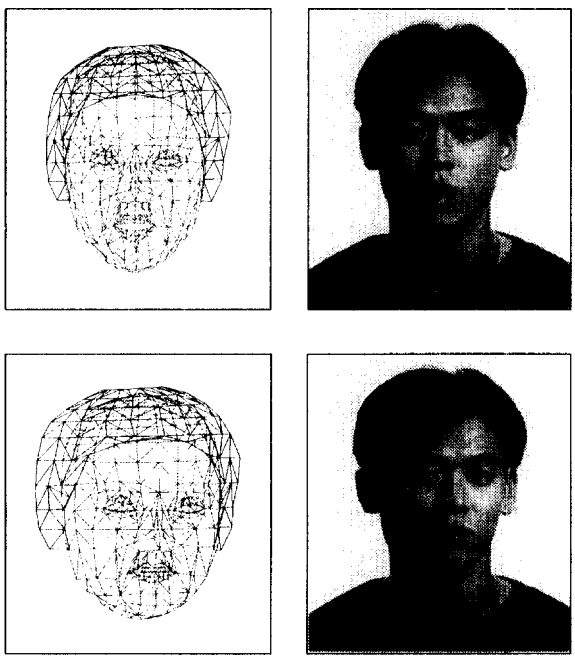
같은 복모음은 'ㅗ'와 'ㅣ'의 입모양이 연속적으로 변화하고 있기 때문에, 단모음의 연속으로 취급할 수 있기 때문에 기본형에 포함시키지 않고 있다. 표 2에 기본형 입모양에 대한 AU의 종류와 강도를 나타낸다.

5. 얼굴동영상의 프레임 생성

합성얼굴 동영상과 합성음성의 동기를 맞추기 위해서는 음절별 지속시간에 따라, 생성되는 얼굴동영상의 프레임 수를 조절해야한다. 프레임수는 얼굴동영상의 프레임당 합성속도를 지속시간으로 나눈 것이다. 즉, 텍스트 「합성」을 예를 들면, 초성「ㅎ」과 중성「ㅏ」의 입모양을 키프레임으로 하여, 이들 사이의 지속시간을 프레임 합성시간으로 나누면 프레임 수를 결정할 수 있다.[20-21] 프레임수가 구해지면, 초성과 중성의 입모양 파라미터를 프레임별로 보간하여, 자연스럽게 변화하는 입모양을 애니메이션으로 합성할 수 있게 된다. 또한, 중성「ㅏ」와 종성「ㅂ」 사이의 변화 또한 동일한 방법으로 합성할 수 있다.

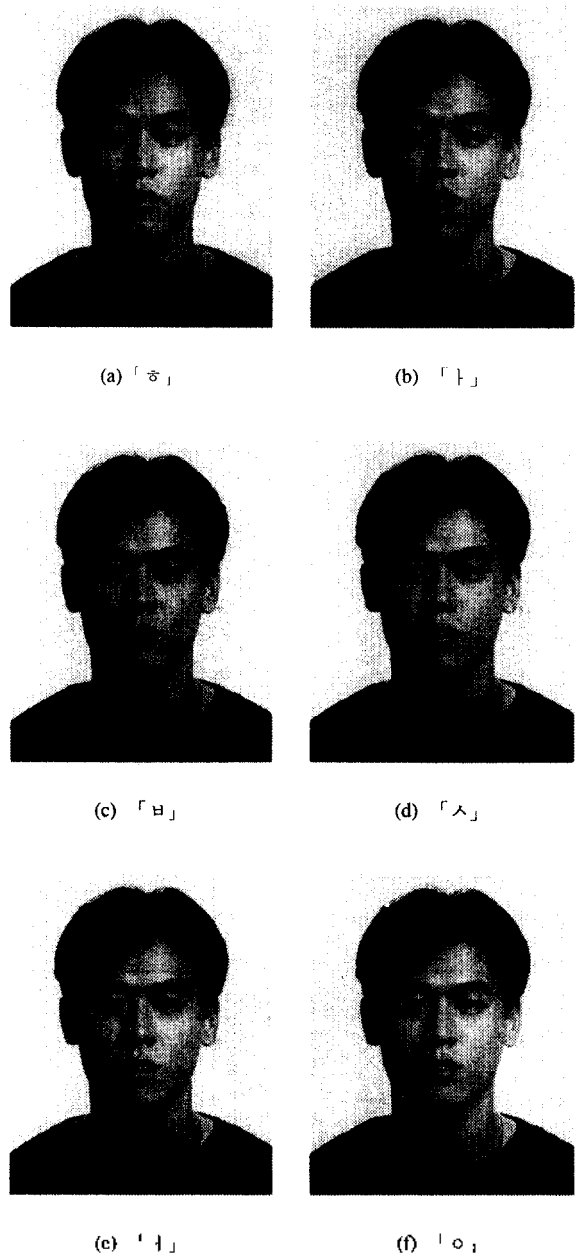
6. 실시간 합성

얼굴동영상 합성은 얼굴의 3차원 형상모델의 변형과 변형된 모델에 텍스처 매핑의 2단계로 나누어 진다. 이것을 그림 5에 나타낸다. 즉, 그림 4의 정합된 모델을 그림 5(a)와 같이 변형하여 텍스처 매핑하면, 그림 5(b)와 같은 영상을 얻을 수 있다. 이하의 합성영상은 동일한 방법으로 얻어진다. 형상모델의 변형은 펜티엄 PC에서도 실시간으로 변형이 가능하지만, 텍스처 매칭은 초당 2~3프레임 정도이다. 이 정도의 속도로는 동영상의 실시간 합성의 효과를 얻기가 곤란하므로, 텍스처 매핑에서는 그래픽 가속기를 사용하고 있다. 그래픽 가속기를 사용하면, 초당 40 프레임정도의 텍스처 매핑이 이루어지므로, 얼굴동영상 실시간 합성이 가능하다. 그림 6은「합성」의 발음시 자모음에 대한 키프레임을 나타낸다. 이 키프레임은 표 2의 AU 파라미터에 따라 합성한 것으로, 키프레임 사이의 중간 프레임은 AU파라미터를 지속시간에 따라 보간하여 합성하고 있다. 그림 7은 표정변화의 합성예를, 그림 8은 두부의 3차원 동작의 합성예를 나타내고 있다. 이 그림들로부터 한 장의 사진을 가지고 현실감 넘치는 고품질의 가상 얼굴을 합성할 수 있음을 알 수 있다.



(a) 형상 모델 변형 (b) 텍스처 매핑

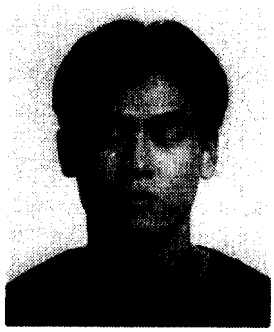
그림 5. 그림 4의 형상모델을 변형하여 합성한 영상
Fig. 5. Images synthesized by deforming the shape model in Fig. 4.



(a)「ㅎ」 (b)「ㅏ」 (c)「ㅂ」 (d)「ㅅ」 (e)「ㅓ」 (f)「ㅇ」

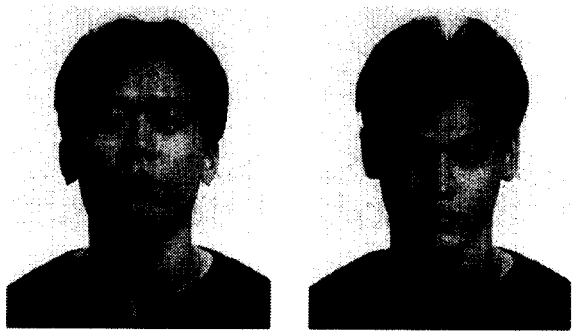
그림 6.「합성」의 자모음 입모양 변화에 대한 키프레임의 합성예
Fig. 6. Synthesis of the key frames for the consonants and the vowels of the korean syllable 「Hap sung」.

(a) AU 1 (b) AU 1+2

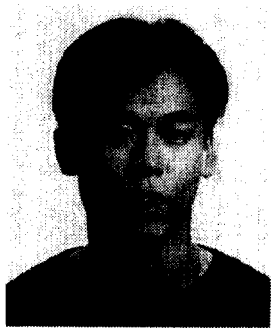


(c) AU 43

그림 7. 얼굴 표정의 변화의 합성 예
Fig. 7. Synthesis for the facial expression changes.



(a) 「상」 (b) 「하」



(c) 「우」

그림 8. 두부 3차원 동작에 대한 합성 예
Fig. 8. Synthesis for 3-D head motions.

IV. 운율분석과 실시간 음성합성

음성합성과정에서는 먼저 입력된 텍스트를 분석하여 얼굴동영상과 동기구현이 가능하도록 음절별 코드를 생성하고 지속시간을 부여한다. 본 연구의 음성합성방법은 TD-PSOLA를 사용한다. 이 합성방식은 음성파형을 pitch-synchronous 하게 데이터베이스에 축적한 후 합성시 다시 접속해 주는 방식으로 실시간 처리가 가능하다. 또한 합성된 음성의 자연성 향상을 위하여 텍스트로부터 운율을 분석한다.

1. 운율분석

언어학적으로 제시된 억양구(intonational phrase)와 강세구(accentual phrase)를 이용하여 한국어 발화문장(utterance)에 대한 운율구조는 그림 9와 같다.

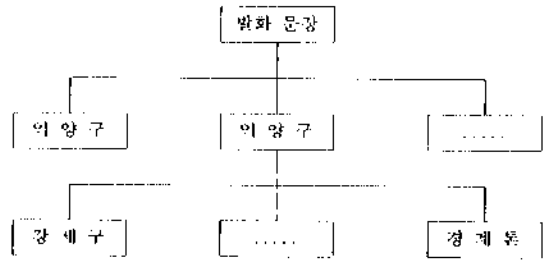


그림 9. 한국어 운율구조
Fig. 9. The prosodic structure of the Korean language.

이러한 한국어 운율구조를 음성합성에 적용하기 위하여 입력된 텍스트로부터 지속시간, gkqtj억양 및 휴지기를 구하기 위한 운율의 분석과정은 다음과 같다.

첫째, 지속시간은 텍스트 분석과정에서 부여된 음절별 지속시간을 그대로 이용한다.

둘째, 텍스트로부터 억양을 분석해 내기 위해 한 문장의 각 어절을 강세구로 하되, 쉼표(,) 또는 마침표(.)로 끝나는 마지막 서술어절을 경계톤으로 하며, 억양구는 시작하는 강세구부터 경계톤까지로 한다. 여기서 강세구 및 경계톤은 서울말씨를 기준으로 음절수에 따라 각각 LHLH 형태 및 감최형태의 피치주파수로 하고 있으며 각 문장 단위의 지속시간 동안 설정된 피치기준선과 강세구를 중첩하여 운율모델을 생성한다.

세째, 휴지기는 이미 분석된 억양구마다 인정시간을 삽입한다.

그림 10은 입력된 텍스트 "지연이는 학교에 갑니다."를 운율분석하여 생성된 운율모델이다.

2. 실시간 음성합성

그림 11은 TD-PSOLA 방식을 이용하여 데이터베이스를 구축하는 분석과정 및 합성과정을 보이고 있다. 이 데이터베이스는 음성파형을 pitch-synchronous 하게 축적하였으므로 실시간 합성이 가능하다.

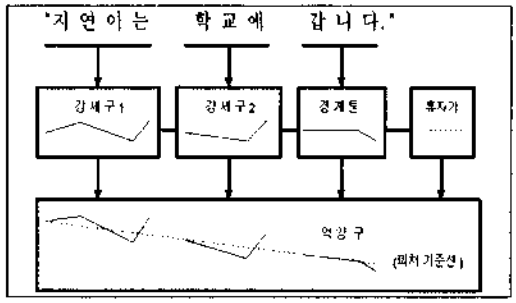


그림 10. 한국어 운율모델
Fig. 10. The prosodic model of the korean language.

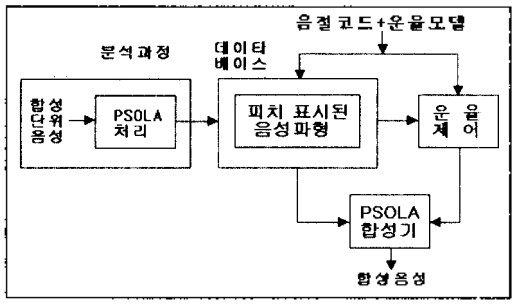
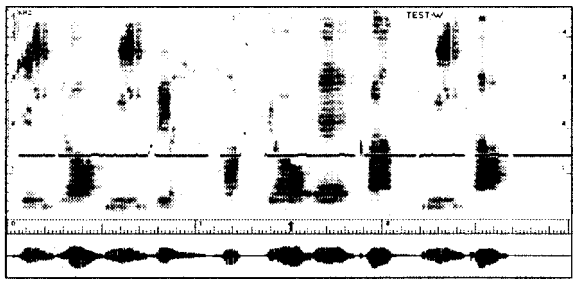
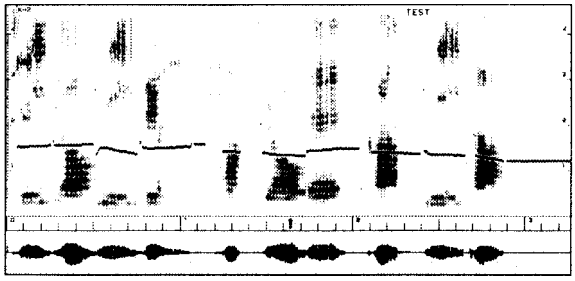


그림 11. TD-PSOLA 합성방식
Fig. 11. TD-PSOLA Synthesis method.

그림 12는 "지연이는 학교에 갑니다."라는 입력문장에 대해 (a)는 운율모델을 적용하지 않고 합성한 결과이며, (b)는 한국형 운율모델을 적용하여 합성한 결과이다.



(a) 운율모델을 적용하지 않은 경우



(b) 운율모델을 적용한 경우

그림 12. 합성음성의 비교("지연이는 학교에 갑니다.")
Fig. 12. Comparison of the synthesized speech.

V. 시스템 구현

표정 짓고 말하는 가상얼굴의 실시간 합성시스템은 펜티엄PC에 그래픽 가속기를 삽입하여 구현하고 있다. 이 시스템의 유저 인터페이스를 그림 13에 나타낸다. 문장분석과 음성 합성은 펜티엄CPU가 담당하고 있고, 동영상합성에 있어서는 형상모델의 변형은 펜티엄CPU가, 실시간 텍스쳐 매핑은 그래픽 가속기가 담당하고 있다. 가상 얼굴은 고품질 동영상 합성을 위해 실제 사진을 이용하고 있다. 그림 13의 유저 인터페이스에 나타나 있는 대상인물은 얼굴 3차원 형상모델만 정합되어 있으면 자유롭게 변경이 가능하다. 즉, 현재의 인물이나 과거의 인물이라 할지라도 한장의 정면 사진만 있으면, 자유롭게 표정짓고 말을 할 수가 있게 된다. 텍스트는 키보드 또는 텍스트파일로 입력할 수도 있다.

고품질의 무제한 어휘에 대한 음성합성 시스템을 구현하기 위하여 이용하는 합성방식은 TD-PSOLA 방식이며 데이타베이스의 합성단위는 반음절과 triphone (VCV)의 조합이다. 데이터베이스의 용량은 약 7Mbyte 이지만 pitch-synchronous한 원음성의 파형이므로 실시간 합성에 용이하다.

이와 같은 시스템을 구성하여 신문기사등을 입력으로 하여 시스템을 시험해본 결과, 고품질의 동영상과 음성을 합성하여, 자연스러운 가상얼굴을 실현할 수 있음을 알 수 있었다.

그림 13. 시스템의 유저 인터페이스
Fig. 13. User Interface of this system.

VI. 결 론

본 논문에서는 표정 짓고 말하는 가상얼굴의 실시간 합

성방법에 대해서 제안했다. 고품질의 가상얼굴의 동영상을 실시간으로 합성하기 위해서, 얼굴의 3차원 형상모델을 구성하여, 실제 얼굴사진의 텍스처 매핑을 이용하고 있다. 한글 텍스트에 따라 운율이 첨가된 음성과 얼굴동영상의 입모양을 합성하고, 음절별 지속시간에 따라 합성음성과 합성동영상의 동기를 맞추고있다. 나아가서, 가상얼굴의 합성에서 중요한 정보인 두부의 3차원 동작과 표정변화를 실현하는 방법도 제시했다. 텍스처 매핑은 그래픽 가속기를 이용하여 초당 40프레임정도를 합성하여, 현실감 있는 가상얼굴을 구현할 수 있었다. 또한 한글문장을 억양구와 강세구로 분석하여 운율모델을 생성하는 과정을 제시하였으며, 이 모델을 이용하여 고품질의 음성을 합성할 수 있었다. 얼굴동영상의 합성에 있어서, 각프레임을 실시간으로 합성하는 것도 중요하지만, 자연스러운 동작의 실현과 얼굴주름 등의 표현도 필요하다고 생각된다.

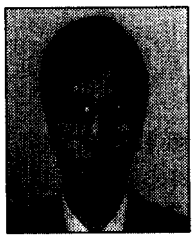
참 고 문 헌

1. V. Bruce, "Recognizing Faces", Lawrence Erlbaum Associates, 1988.
2. P. Ekman and W. V. Friesen, "Unmasking Face", Prentice-Hall, 1975.
3. J. P. Lewis and F. Parke, "Automated Lip_Synch and Speech Synthesis for Character Animation", Chi+Gi 1987 conf. Proc., pp.143-147, 1987.
4. A.Lippman, "Semantic Bandwidth Compression : Speech-maker", Picture Coding Symposium, pp.29-30, 1981.
5. A. Kay, "Computer Software", Sci. America, vol.251, no.3, pp.191-207, 1984.
6. 森島, 岡田, 原島, "知的インタフェースのため表情合成法の 検討", 日本電子情報通信學會論文誌, vol.J73-D-II, no.3, pp.351-359, 1990.
7. S. Morishima, K. Aizawa and H. Harashima, "An Intelligent Facial Image Coding Dirven by Speech and Phones", IEEE ICASSP, 39M8.7, pp.1795-1798, 1989.
8. S. Morishima and H. Harashima, "A media Conversion from speech to facial image for intelligent man-machine interface", IEEE JSAC, vol.9, no.4, pp.594-600, 1991.
9. ISO/IEC/JTC1/SC29/WG11 N1666Pub, April 1997.
10. K. Waters : "A Muscle Model for Animating Three-Dimensional Facial Expression", Computer Graph. vol.21, no.4, pp.17-24, 1987.
11. 原島 博, "知的映像符號化と知的通信", 日本テレビジョン學會誌, vol.42, no.6, pp.519-525, 1988.
12. H. Harashima, K. Aizawa and T. Saito, "Model-Based Analysis Synthesis coding of Videotelephone Images", Trans. IEICE Japan, vol.E72, no.5, pp.452-459(May 1989).
13. K. Aizawa and H. Harashima, "Model-Based Analysis Synthesis Image Coding(MBASIC) System for a person's Face", Signal Process. Image Com., vol.1, no.2, pp.139-152, 1989.
14. 崔昌石, 原島 博, 武部 幹, "顔の3次元モデルに基づく表情の記述と合成", 日本 電子情報通信學會論文誌, vol.J73-A, no.7, pp.1270-1280, 1990.
15. C.S.Choi, K.Aizawa, H. Harashima, T.Takebe, "Analysis and Synthesis of Facial Image Sequences in Model-Based Coding", IEEE Trans. Circuit. Sys. Video Tech., vol.4, no.3, pp.257-275, 1994.
16. S. Morishima, "Better face Communication", ACM SIGGRAPH'95, Visual Proceedings, p.117, 1995.
17. 金子 正秀, 小池, 淳, "テキスト情報に對應した口形形象變化する顔動畵像の合成", 日本 電子情報通信學會論文誌 D-II, vol.J75, no.2, pp.203-215, 1992.
18. H. S. Lee, M. S. Hahn and J. C. Lee, "MPEG-4 TTS Interface", Proc. ICSP, pp.177-181, 1997.
19. 이용동, 최 창석, 최 갑석, "휴먼인터페이스를 위한 한글 음절의 입모양 합성", 통신학회논문지, vol.19, no.4, pp.614-623, 1994.
20. K. Y. Lee and C. S. Choi, "Synchronized Realization of Synthetic Speech and Synthetic Facial Image Sequences for Virtual Reality", Proc. ICSP, pp.187-190, 1997.
21. 송경준, 이기영, 최창석, 양광호, "가상현실을 위한 합성 얼굴동영상과 합성 음성의 동기구현", 음향학회 학술대회 논문집, vol.15, no.1(s), pp.107-112, 1996.
22. J.L.Flanagan, *Speech analysis, Synthesis and Perception*, 2nd. Ed., Springer-Verlag, Berlin, 1972.
23. S.Furui, M.M.Sondhi, *Advances in Speech Signal Processing*, Marcel Dekker, Inc., New York, Basel.Hong Kong, pp.741-853, 1992.
24. Eric Keller, *Fundamentals of Speech Synthesis and Speech Recognition*, John Wiley & Sons, pp.69-127, 1995.
25. B.E.Caspers, B.S.Atal, "Changing pitch and duration in LPC synthesized speech using multipulse excitaion," J.Acoust. Soc. Amer., suppl., Vol.73, No.1, pp.S5, Spring, 1983.
26. T. Takagi, T. Umeda, "Voice quality conversion with correction of spectral distortion by pitch manipulation, and its subjective evaluation," the Transactions of the Institute of Electronics, Information and Communication Engineers A Vol.J73-A, No.3, pp.387-396, Mar. 1990.
27. T. Takagi, "Voice quality conversion," Trans.Televison, Vol. 47, No.12, pp.28-32, 1993.
28. F. Charpentier, M. G. Stella, "Diphone SynthesisUsing Overlap-add Technique for Speech Waveforms Concatination," ICASSP 86, pp.2015-2018, 1986.
29. E. Moulines, F. Charpentier, "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones," Speech Communication 9, pp.453-467, 1990.
30. H. Valbret, E. Moulines, J. P. Tubach, "Voice transformation using PSOLA Technique," EUROSPEECH 91, pp.345-348, 1991.
31. Sun-Ah Jun, *The Phonetics and Phonology of Korean Prosody*, Doctoral Dissertation, The Ohio State University, 1993.
32. Kiyoung Lee, Minsuck Song, "Automatic segmentatin of Korean prosodic phrases," ICSP 97, pp.1407-1410, 1997.
33. 이기영, 송민석, "악센트구와 억양구의 운율패턴을 이용한 음성합성시스템," 1997년도 한국음향학회 정기총회 및 학술논문 발표대회, pp. 753-756, 1997.
34. Edward Angel, "Interactive Computer Graphics", Addison Wesley, pp.144-152, 1997.
35. P. Ekman and W. V. Friesen, "Facial Action Coding System", Consulting Psychologist Press, 1977.

▲송 경 준(KyungJoon Song)

1982년 : 명지대학교 전자공학과(학사)
1984년 : 명지대학교 전자공학과(석사)
1982년～1984년 : 명지대학교 전자공학과 실험조교
1985～현재 : 한국전자통신연구원 입체정보연구팀 선임연구원
※주관심분야 : 가상현실, HCI, 멀티미디어

▲이 기 영(Kiyoung Lee)



1984년 : 명지대학교 전자공학과 졸업.
1986년 : 명지대학교 전자공학과 석사과정 졸업
1992년 : 명지대학교 전자공학과 박사과정 졸업
1993～현재 : 관동대학교 전자정보통신공학부 부교수

▲최 창 석(Changseok Choi)

1978년 : 홍익대학교 전자공학과 졸업
1988년 : 일본 가나자와 대학원 전기정보공학과 석사과정 졸업
1991년 : 일본 가나자와 대학원 전기정보공학과 박사과정 졸업
1984년～1992년 : 산업기술 정보원 책임 연구원
1993년～현재 : 명지대학교 정보통신공학과 부교수

▲민 병 의(Byungeui Min)

1982년 : 한양대학교 전자공학과(학사)
1984년 : 한국과학기술원 전기 및 전자학과(석사)
1992년 : 한국과학기술원 전기 및 전자공학과(박사)
1984년～1987년 : 대림산업기술연구소
1987년～현재 : 한국전자통신연구원 입체정보연구팀 팀장
※주관심분야 : 가상현실, 멀티미디어시스템, 에이전트