멀티미디어 환경을 위한 정서음성의 모델링 및 합성에 관한 연구

Modelling and Synthesis of Emotional Speech on Multimedia Environment

조 철 우 · 김 대 현*
(Cheol-Woo Jo · Dae-Hyun Kim)

ABSTRACT

This paper describes procedures to model and synthesize emotional speech in a multimedia environment. At first, procedures to model the visual representation of emotional speech are proposed. To display the sequences of the images in synchronized form with speech, MSF(Multimedia Speech File) format is proposed and the display software is implemented. Then the emotional speech sinal is collected and analysed to obtain the prosodic characteristics of the emotional speech in limited domain. Multi-emotional sentences are spoken by actors. From the emotional speech signals, prosodic structures are compared in terms of the pseudo-syntactic structure. Based on the analyzed result, neutral speech is transformed into a specific emotinal state by modifying the prosodic structures.

Keywords: emotion, trasformation, multimedia

1. 서 론

정서는 기쁨, 슬픔, 두려움, 화냄 등 인간의 감정상태이다. 이러한 정서는 실제 생활에서 아주 다양하게 나타나며 동일한 정서 상태일지라도 상황에 따라서, 사람에 따라서 전혀 다른 형태로 나타난다.

최근 음성합성 등에서 자연성 향상 등의 목적으로 정서를 표현하려는 시도가 여러 연구자들에 의해 시도되고 있다. 정서정보는 음성에서 여러 가지 형태로 나타나고 있으며 이는 합성음의 자연성에서 중요한 요소의 하나이다. 또한 최근의 멀티미디어 기술의 발달로 정보의 멀티미디어화가 가속화되고 있는데 이와 함께 음성에 영상정보를 부가하여 멀티미디어 형태로 표시하려는 시도도 곳곳에서 이루어지고 있다.

정서적 음성은 음향신호든, 영상신호든 간에 실제 신호를 얻는 것부터가 어려운 일이

* 창원대학교 공과대학 제어계측공학과

므로 분석된 자료를 가지고 일정한 법칙에 의해서 합성해 내는 것은 더욱 어렵다. 그러나 정서의 종류와 화자를 제한하고 값을 변화시킬 파라미터의 수를 제한한다면 가능한 일이라고 생각되어 본 연구를 수행하게 되었다.

본 논문에서는 정서적 음성을 멀티미디어 형태로 표현하는 한 가지 방법을 제안하고, 그것을 실제로 구현하는 과정과 결과에 대해 논의하고자 한다.

2. 정서의 표현의 멀티미디어적 요소

인간의 정서, 또는 감정상태는 매우 다양하다. 보통 기본 정서라고 일컬어지는 기쁨, 슬픔, 화냄, 두려움 등의 네 가지 정서 외에도 여러 가지의 다른 정서로 구분될 수 있으며, 동일한 단어로 표시되는 정서라고 할지라도 말하는 사람의 상황에 따라서 전혀 다른 표현으로 나타나게 된다. 이러한 정서의 다양성 및 복잡성 때문에 아직도 정서를 일관되게 정의하는 것은 매우 어려운 문제이다. 대개는 기본 정서라고 하는 네 가지 범주를 연구대상으로 삼고 있다. 정서를 멀티미디어 형태로 표현하고자 할 경우 영상표현과 음성표현으로 나누어 볼 수 있다. 영상표현은 사람의 얼굴표정이 되며, 음성표현은 정서의 종류에 따른 음성의 특징의 변화가 된다. 그리고 이 두 가지 표현이 서로 동기되어 나타내어져야 한다. 영상표현을 위해서는 우선 입모양이 말소리와 동기되어야 하며 표정이 적절한 시간에 정서상태를 나타내도록 표시되어야 한다.

3. 정서적 음성의 영상표현

음성을 출력할 경우 소리만으로 출력하기보다는 소리와 영상을 함께 출력할 경우 인간의 인지율이 높아진다는 보고가 있다. 컴퓨터에서 음성이 출력될 경우도 영상정보와 같이 출력된다면 효과적인 정보전달이 가능하기 때문에 최근 영상과 소리를 결합시키는 연구가 많이 시도되고 있다. 이와 같은 시도는 영상인식 또는 컴퓨터 그래픽 기술과 음성인식, 합성 등 음성처리기술을 결합하여 의사전달과정에 도움을 주고자 하는 시도의 일환이다. 기존의 연구 중에도 2차원 또는 3차원 영상을 이용하여 영상정보가 부가된 음성출력을 구현한 사례가 많이 있다. 그러나 그러한 시도들에서는 얼굴영상을 3차원 메쉬구조를 이용하여 변화하는 영상을 구현함으로써 많은 계산시간이 필요하였으며, 취급해야 할 데이터의 양도 많아서 범용 PC시스템이 아닌 빠른 속도를 갖는 그래픽 워크스테이션급의 컴퓨터가 필요하였다.[1][2][3][4] 이를 보완하기 위하여 새로운 멀티미디어 화일형식 MSF가 제안되었다.[5,6] 본 논문에서는 MSF 형식화일을 생성해 주는 부호화 프로그램을 작성하는 과정에 관하여 기술하고 그 결과에 대하여 논의한다.

3.1 MSF 멀티미디어 화일형식

영상 및 소리를 포함하는 멀티미디어 화일형식으로는 AVI, MPEG등이 있다. 이러한 화일형식들은 소리와 함께 동영상 데이터를 압축된 형태로 내장하고 있기 때문에 크기가 커지게 된다. 그러나 응용분야에 따라 화일 크기가 크면 전송이나 보관 등에 시간이 걸리

기 때문에 좋지 않은 경우가 있다. 이와 같은 단점을 보완하기 위하여 고안된 것이 MSF 형식이다.

MSF형식은 화일의 내부에 크기가 큰 영상데이터를 직접 갖고 다니지 않고 정해진 영상데이타의 인덱스만을 소리화일에 부가된 형태로 갖고 있기 때문에 크기가 기존의 동영상화일에 비해서 아주 작아진다는 것이 장점이다. 이러한 장점 때문에 인터넷에서 플러그인으로 동영상 애니메이션을 구현한다던가 TTS의 출력을 애니메이션 형태로 내보낸다던지 할 경우 유용하다. 또 한가지 장점은 소리에 포함된 영상데이타를 선택하는 영상데이타베이스에 따라서 언제든지 바꿀 수 있다는 것이다. 예를 들어 사람의 음성에 따라 변화하는 입모양 영상을 부가할 수도 있고, 또는 관련된 다른 영상을 부가하여 볼 수도 있다. 그림 1은 MSF 화일형식의 구조이다.

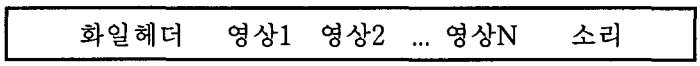


그림 1. MSF화일형식

3.2 화일작성기의 구성

MSF화일작성기는 음성신호 또는 소리신호를 원하는 부분으로 잘라서 시간축정보를 얻는 부분과 구해진 시간정보를 이용하고 필요한 영상데이터베이스를 선택하여 MSF형식으로 만들어 주는 부분의 두 부분으로 나눌 수 있다. 그림 2 는 MSF부호화기의 전체 구성을 나타낸 것이다.

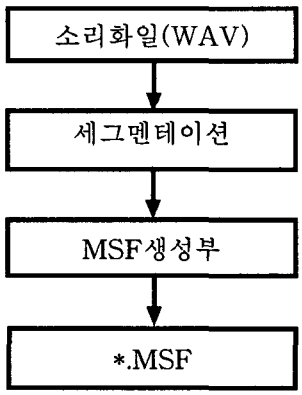


그림 2. MSF부호화기의 구성도

다음으로 음성화일을 MSF형식으로 변환하는 과정을 설명한다. 음성화일을 각 음소로 분할하기 위하여 그래픽 음성 편집기를 구성하였다. 편집기는 MATLAB으로 구현하였다. 그림 3은 그래픽 음성편집기의 화면을 보여준다.

편집기에는 신호의 시간축 파형, 스펙트로그램, 에너지, ZCR(Zero Crossing Rate) 등이 표시되며, 시각적 편집에 의해 각 음소의 시작시간, 지속시간을 기록한다. 이러한 작업의

결과로 시간정보 화일이 만들어진다. 표 1은 시간정보화일의 예이다.

표 1. 시간정보화일의 형식

시간(2바이트)	사건코드(2바이트)
10	s
25	a
35	z
54	a

표 1에서 시간정보화일은 각 사건에 대한 시간정보와 사건코드를 2바이트씩으로 기록한 것이다. 사건코드는 음성의 경우는 해당음소기호를 기타 동영상의 경우는 영상의 번호가 부여되게 된다. 영상의 번호는 이미 만들어져 공유하고 있는 영상데이타베이스에 의해 결정된다. 시간정보화일의 확장자는 tdd 로 한다.

영상정보를 주어진 음성 화일에 부가하는 과정은 다음과 같다.

(1) 시간정보 화일을 읽어 들인다.
(2) 음성(또는 소리) 화일을 읽어 들인다.
(3) MSF헤더부를 기록하여 새로운 화일로 저장한다.

이 과정에서는 tdd 화일과 wav형태로 기록된 소리화일을 읽어들여 각 시간구분에 따른 사건코드를 추출하고, 이에 따른 영상데이타베이스를 선택한 뒤 tdd 정보를 wav 화일의 앞에 부가하여 MSF형식으로 기록한다.

이와 같은 방법에 의해 구성된 MSF화일은 적은 용량을 가지면서도 동영상을 효과적으로 표시해 줄 수 있다. 본 실험에서는 음성에 따른 입술영상 데이타베이스와 음악소리에 따른 동영상애니메이션의 두 가지에 대하여 구현해 보았는데 효과적인 동영상 구현이 가능하였다. 보다 자연스러운 영상의 구현은 세그멘테이션 과정에서의 섬세함과 그래픽구현의 자연성을 통해 얻어질 수 있다. 세그멘테이션 과정을 위하여 세그멘테이션 도구를 개발하였다. 그림 3에 동작화면을 보인다. 앞으로 세그멘테이션 과정에서 신호의 특성을 측정하여 자동으로 음소를 추출한다던지 음의 종류에 따라 감성적인 방법으로 적합한 영상을 찾아내어 부가하는 형태로 영상을 부가하는 지능형 부호화기를 개발할 예정이다.

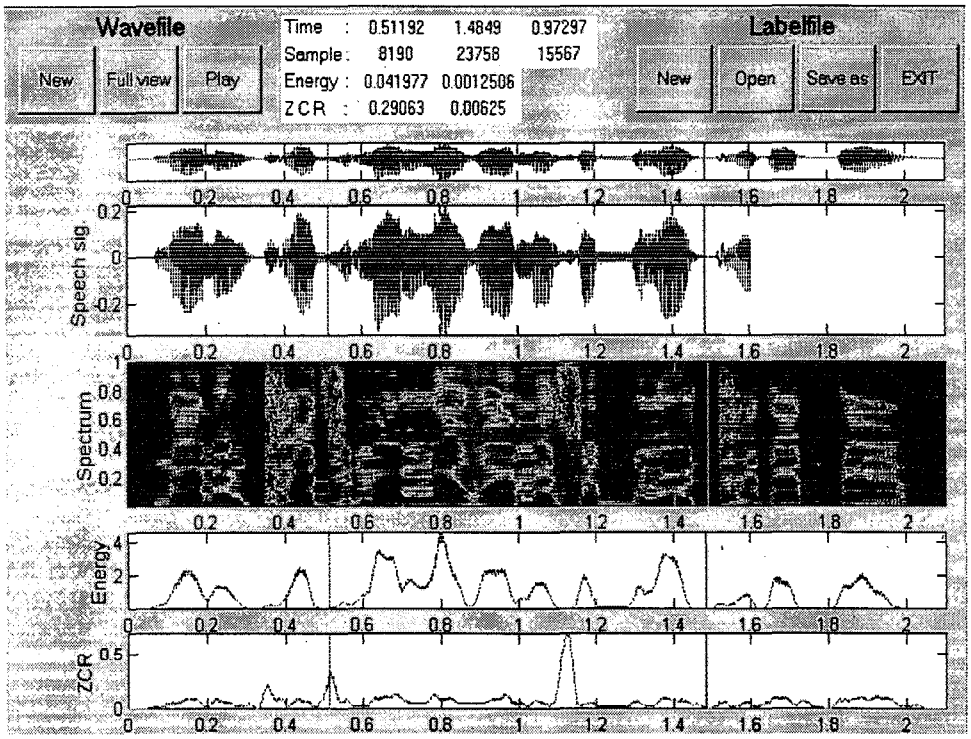


그림 3. 세그멘테이션 도구

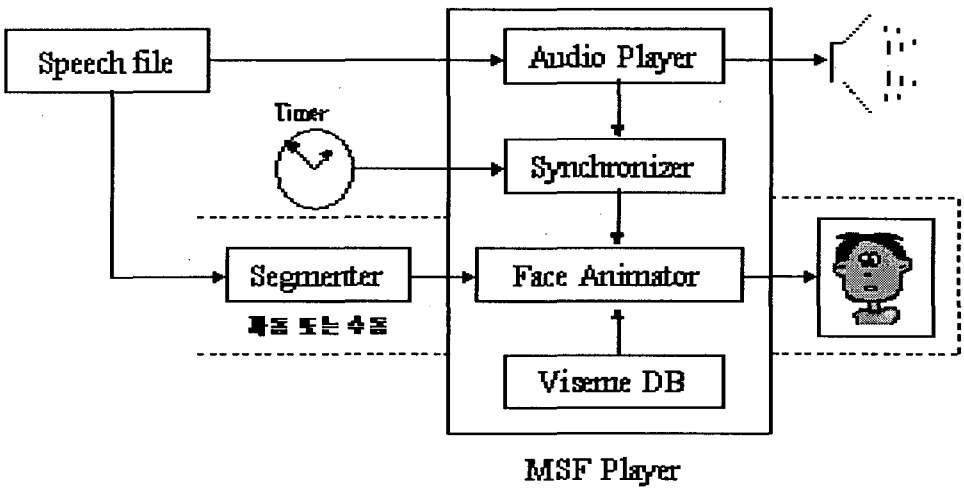
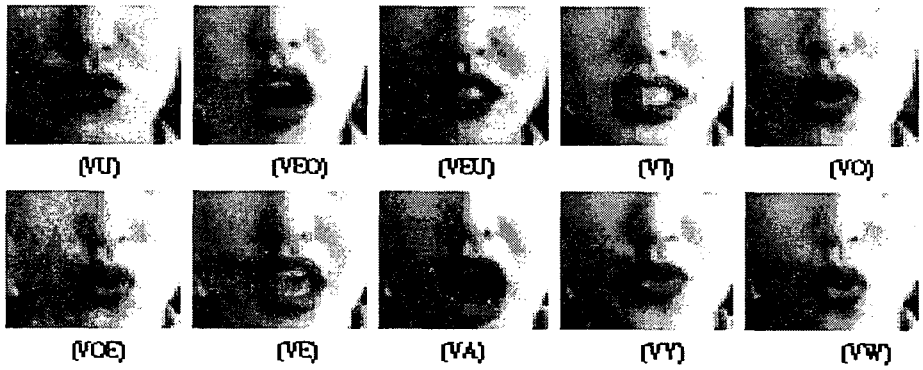


그림 4. MSF 플레이어의 구성

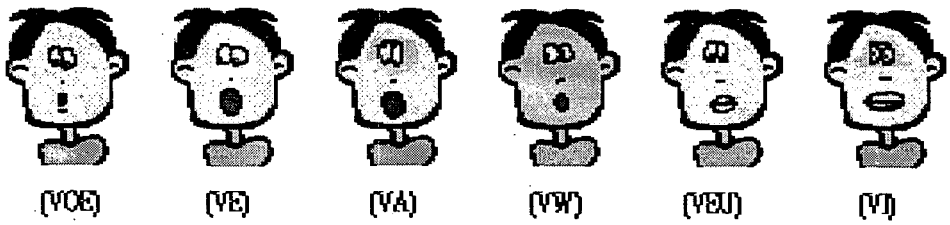
그림 4는 MSF플레이어의 구성도를 보인 것이다. 음성화일은 세그멘터에 의해 부가된 시간영역 정보와 영상데이타를 합해서 소리와 동기된 형태로 출력하게 된다.[6]

3.3 영상정보데이타의 구성

영상정보데이타는 기본적인 입모양의 영상과 표정영상을 포함한다. 그림 5는 이들 중 모음에 대한 입모양과 애니메이션 영상의 예이다.



(a) 측정된 실제입모양



(b) 간소화된 애니메이션 형태의 입모양

그림 5. 측정된 입모양의 형태와 간소화된 형태

정서의 영상화는 별도의 측정없이 다소 과장된 형태의 애니메이션 이미지를 사용하였다. 이렇게 과장된 이미지를 사용함으로써 실제 영상보다 간단하면서도 의도를 더 잘 전달할 수 있다.

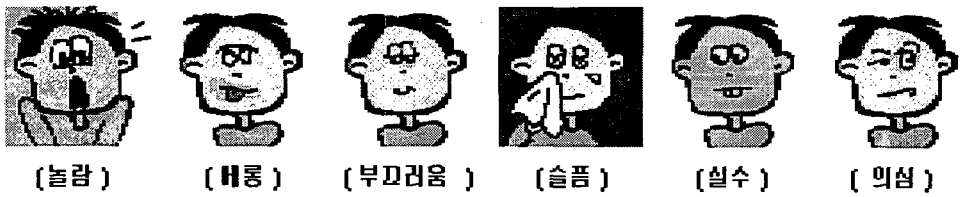


그림 6. 애니메이션화된 표정

4. 정서적 음성의 표현

음성을 원하는 정서상태를 갖도록 합성해 내는 방법은 여러 가지가 있겠지만 본 연구에서는 낭독체 음성의 내용을 구성요소에 따라 정서적 음성의 형태로 변환시키는 방법을

사용하였다. 기존의 문장음성합성을 위한 TTS를 사용한다면 보다 다양한 음성을 합성하는 것이 가능하겠으나 충분한 고음질을 갖는 TTS를 구성하는 것 자체가 별도의 큰 과제가 되기 때문에 여기서는 미리 녹음된 낭독체 문장음성을 변환하는 방법을 사용하였다.

정서적 음성의 분석은 앞에서 수집된 4가지 종류의 정서상태를 갖는 5가지 문장음성에 대하여 행해졌다. 정서적 음성의 특징을 나타내는 인자는 여러 가지가 있으나[7,8,9] 본 연구에서는 피치의 변동폭, 평균피치의 변화, 지속시간의 변화(발성속도의 변화)를 중심으로 분석하였다.

화자간 서로 다른 정서상태에 따른 상이한 정서음성이 유발될 수 있기 때문에 서로 다른 화자간의 정서음성을 비교한다는 것은 의미가 없다고 판단된다. 따라서 본 실험에서는 동일한 화자의 경우에 대해 서로 다른 정서상태의 음성을 비교하는 방법으로 분석하였다. 사용된 문장은 4명의 화자중 정서상태가 뚜렷이 드러난다고 보아지는 화자 1의 음성을 택하여 4개의 문장에 대해 분석하였다.[10] 각각의 문장의 종류는 다음과 같다.

(1) 나는 가지말라고 하면서 문을 닫았다.
(2) 이제 그만하자.
(3) 나는 ____입니다.
(4) 바람과 햇님이 서로 힘이 더 세다고 다투고 있을때

문장을 분석하기에 앞서 각 문장의 요소들을 주어부, 서술부, 목적어부, 각 부분의 끝음절 및 휴지부로 나뉘어 진다. 이러한 분석은 수작업에 의해 수행되었다. 이렇게 나눈 이유는 분석결과 정서적 음성의 차이에서 오는 가장 큰 변화가 피치와 지속시간이었는데 이들의 변화가 가장 크게 관찰된 부분이 이와 같은 요소들의 경계부분이었다.

표 2. 문장의 요소별 기호

구분	기호
주어부	SUB
서술부	DES
목적어부	OBJ
각 부의 끝음절	DEE
휴지부	PAU

입력된 음성은 수작업으로 각 부분으로 나뉘어 레이블링 된 다음 각 부분의 특성변화가 그림 7은 문장(1)의 각 정서별 피치패턴의 변화는 보인 것이다. 4가지 다른 정서에 대해 동일한 문장을 발성길이를 맞추어 그려서 피치패턴을 비교할 수 있게 그린 것이다. 그림의 가로축에서 볼 수 있듯이 각각의 지속시간은 서로 다르다.

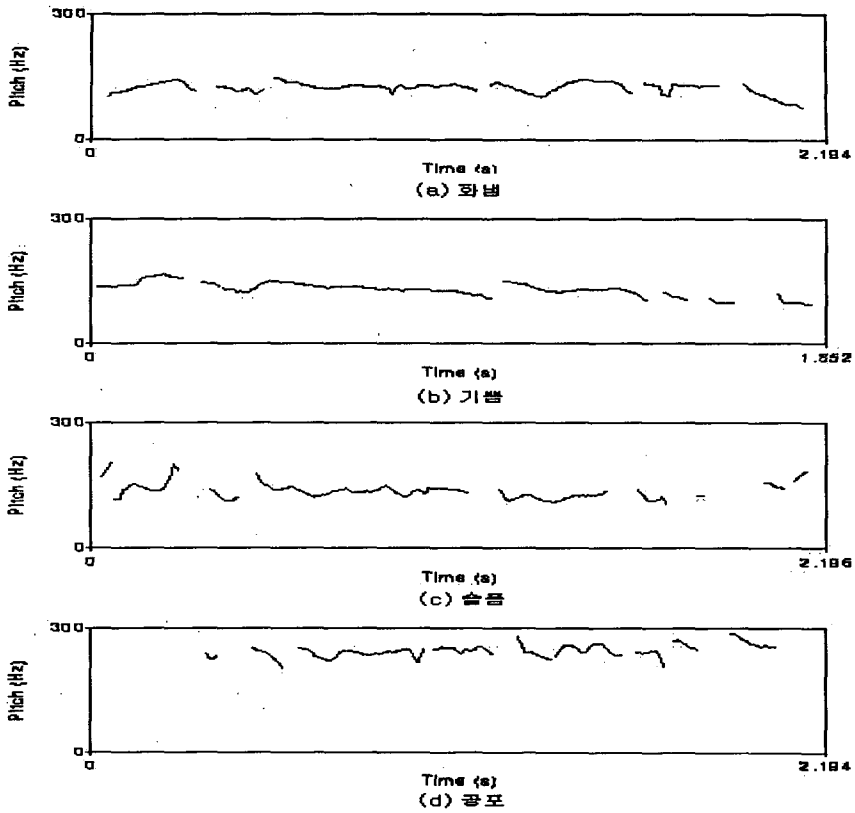


그림 7. 정서별 피치패턴의 변화

표 3. 각 문장별 지속시간의 측정치

	문장1	문장2	문장3
화냄	2131	677	874
기쁨	1768	566	2145
슬픔	2595	821	1753
공포	1959	683	1972
정상	2970	1050	1760

표 4는 문장(1)에 대한 각 부분의 레이블링된 결과예이다.

표 4. 문장(1)에 대한 레이블링 결과

난		가지말라고하면	서	문을	닫았	다
SUB	PAU	DES	DEE	OBJ	DES	DEE

이와 같은 레이블링 자료를 기준으로 측정된 정서별 지속시간과 피치변화를 고려하여 각 정서별 피치변화규칙을 표 5와 같이 만들었다.

표 5. 정서별 피치변화 규칙

	피치	지속시간
화냄	전체 +40 %	전체 -30 %
기쁨	폭 : +10 % +20%~-20 %	전체 +20 %
슬픔	Random 변화	전체 +30 %
공포	전체 +100%	전체 -10 %

이와 같은 피치변화규칙은 표 6과 같이 코우딩된다.(화냄의 경우)

표 6. 피치변화의 규칙 코딩

```
1 1368.5359 PAU P +1 PW +1 PR 0 D -30
1368.5359 7805.1807 SUB P +40 PW +1 PR 0 D -30
7805.1807 9248.8537 PAU P +1 PW +1 PR 0 D -30
9248.8537 26605.5226 DES P +40 PW +1 PR 0 D -30
26605.5226 31100.2836 DEE P +40 PW +1 PR 0 D -30
31100.2836 36820.8259 OBJ P +40 PW +1 PR 0 D -30
36820.8259 42292.2629 DES P +40 PW +1 PR 0 D -30
42292.2629 45222.5491 DEE P +40 PW +1 PR 0 D -30
```

여기서 P는 피치변화율(%), PW는 피치의 폭변화율(%), PR은 피치에 랜덤변화를 추가 여부(0/1), 그리고 D는 지속시간의 변화율(%)을 나타낸다.

코우딩된 자료에 의해 PSOLA분석 및 합성법에 의해 정서적 중성 문장의 변환이 이루어지게 된다.

4. 실험 및 검토

정서적 음성의 분석에는 연구실에서 개발한 분석 및 합성용 프로그램을 이용하여 PSOLA 방법으로 분석 및 합성을 행하였다. PSOLA 방법은 Pitch Synchronous OverLap and Add method의 약자로 다른 방법보다 비교적 쉽게 고품질의 음성을 합성할 수 있기 때문에 TTS 등에서 많이 사용되고 있는 방법이다. 그림 3은 PSOLA 방법의 수행과정과 PSOLA 방법에 의한 분석도구의 화면을 나타낸 것이다.

우선 정서적 음성을 발성하지 않은 제 3의 화자가 발성한 음성으로부터 각 피치의 위

치를 검출하고 각 요소들로 세그멘테이션한다. 세그멘테이션 작업이 끝나면 레이블 화일로 저장된다. 이 레이블 화일을 기준으로 각 요소에 대하여 지정된 변환작업을 거쳐 정서적 음성으로 합성해 내게 된다.

합성된 음성들은 20명의 화자에게 들려주고 인지되는 정서적 상태를 표시하도록 하였다. 표 7은 각각의 문장에 대해 입력된 정서상태와 측정된 정서상태의 분포를 나타내 준다. 여기서 문장(4)는 분석에 사용되지 않는 제 3의 화자의 음성을 제안된 규칙에 의해 변환한 것이다.

표 7. 정서상태 측정표

I \ O	기쁨	슬픔	화냄	공포	보통
기쁨	1	3	1	1	14
슬픔	1	16	0	2	1
화냄	0	1	10	4	5
공포	0	4	6	8	2
보통	0	5	1	0	14

(a) 문장 (1)

I \ O	기쁨	슬픔	화냄	공포	보통
기쁨	2	1	1	0	16
슬픔	0	13	4	0	3
화냄	5	3	8	2	2
공포	0	10	4	6	0
보통	0	2	10	0	8

(b) 문장 (2)

I \ O	기쁨	슬픔	화냄	공포	보통
기쁨	0	5	0	1	14
슬픔	1	17	1	1	0
화냄	4	1	1	13	1
공포	0	2	0	13	5
보통	2	1	2	3	12

(c) 문장 (3)

I \ O	기쁨	슬픔	화냄	공포	보통
기쁨	8	1	0	0	11
슬픔	0	18	0	2	0
화냄	1	0	8	5	6
공포	2	4	3	4	7
보통	3	0	0	1	16

(d) 문장 (4)

그림 8에서 두 번째 윈도우는 PSOLA분석을 위한 것이다. 세 번째 윈도우에서는 피치 값의 변화를 나타내고 있으며, 네 번째 윈도우에서는 그래픽 인터페이스를 이용하여 지속시간 조절을 행할 수 있게 하였다. 다섯 번째 윈도우에는 변경된 파라미터로 합성된 음을 그렸다.

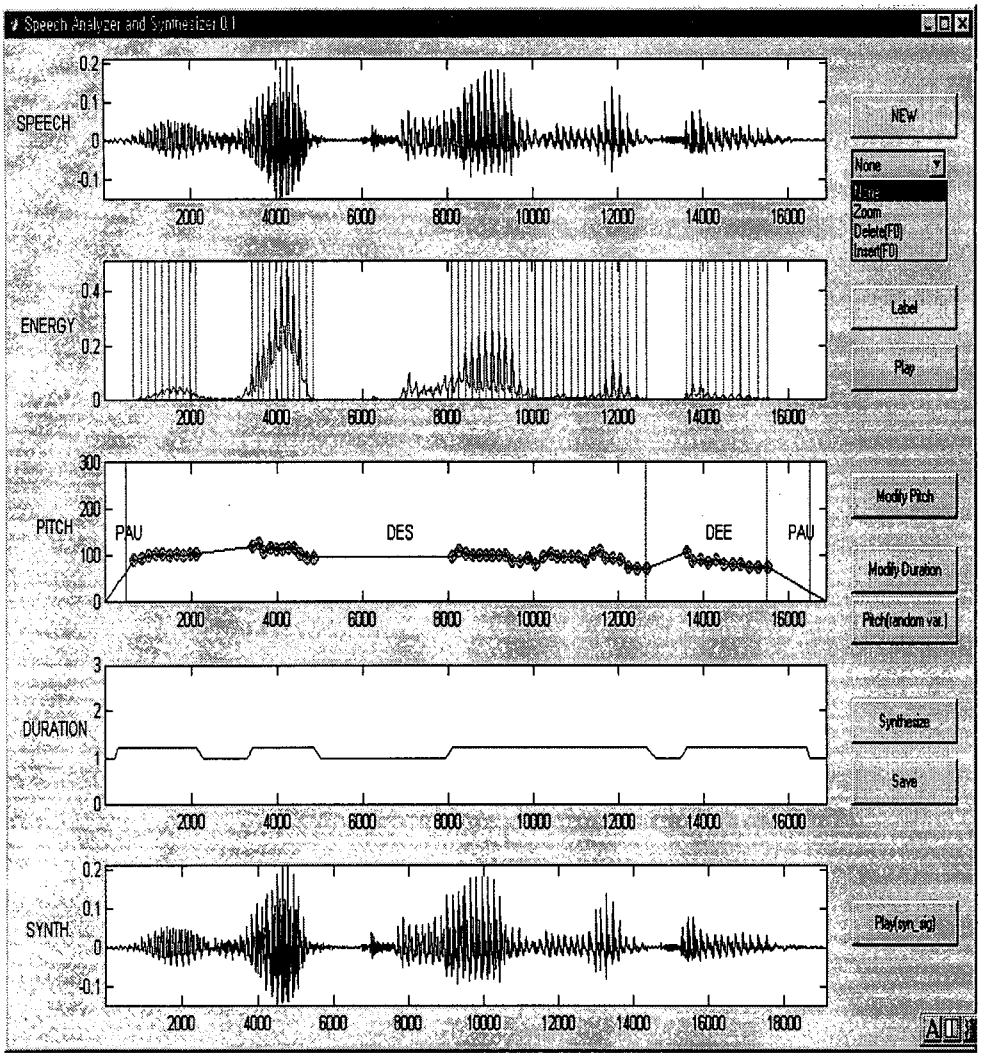


그림 8. PSOLA 분석, 레이블링 및 합성도구

표 8. 정서별 식별율(%)

	기쁨	슬픔	화냄	공포	보통
문장(1)	5	80	50	40	70
문장(2)	10	65	40	30	40
문장(3)	0	85	5	65	60
문장(4)	40	90	40	20	80

표 8은 정서별, 문장별 식별율을 나타낸다. 이 표에서 볼 수 있듯이 변환된 음성이 원

하는 정서상태로 바르게 인지된 경우는 그리 높지 않다. 그러나 특이한 점은 슬픔의 경우 65-90 %의 식별율을 갖고 있다는 점이다. 이는 보통의 낭독체 음성의 식별율이 그다지 높지 않는데 비해 주목할만한 일인데. 이것은 슬픔의 경우 구현해 준 피치패턴의 변환형태가 특징으로 적합했음을 보여준다. 그러나 기쁨(Happy)의 경우는 피치패턴이나 지속시간으로 볼 때 보통음성과 실제로도 구분하기 어려웠으며, 따라서 규칙화하기도 어려움을 알 수 있다. 화냄과 공포의 경우 50 % 근방의 인지율을 나타내었는데 이는 이 두 가지 정서가 유사한 특성을 가지는 때문이라고 볼 수 있다. (즉, 피치가 올라가며, 지속시간이 길어진다.)

5. 결 론

본 논문에서는 정서적 음성을 멀티미디어 형태로 표시할 수 있게 모델링하는 방법을 제안하고 제안된 방법에 의해 합성된 음성을 청취실험에 의해 평가하였다. 영상정보의 면에서는 MSF 동영상 파일 형식을 제안하고, 음성과 동기된 영상제시시스템을 구현하여 입모양과 표정을 음성의 내용에 맞추어 표시할 수 있게 하였다. 음성은 보통의 낭독체 문장을 세그멘테이션하여 정서상태의 문장으로 변환하였다.

측정결과 영상정보의 경우는 과장된 애니메이션을 이용하여 효과적으로 정서정보의 전달이 가능하였으나 음성정보의 경우는 정서상태의 정규화, 정서적 데이터베이스의 부족으로 슬픔상태를 제외하고는 효과적인 정서전달이 구현되지 않았다. 그러나 제시된 방법은 정서적 음성을 멀티미디어 형태로 전달할 수 있게 하는 효과적인 모델링 방법을 구현하여 제시하였다고 본다. 앞으로 이러한 모델에 적용할 파라미터와 규칙을 보완한다면 여러 가지 정서음성을 멀티미디어 형태로 구현할 수 있을 것으로 생각된다.

감사의 글

이 연구는 1997년도 한국과학재단 핵심전문연구과제 "멀티미디어 환경을 위한 정서음성의 모델링 및 합성에 관한 연구" (과제번호: 971-0917-104-2)의 연구결과입니다. 연구비를 지원해 주신 한국과학재단에 감사드립니다.

참 고 문 헌

[1] Shigeo Morishima, Hiroshi Harashima. 1991. "A Media Conversion from Speech to Facial Image for Intelligent Man-Machine Interface." IEEE Journal on Selected Areas in Communications 9(4), 594-600.

[2] W. Goldenthal, K. Waters, J-M Van Thong, O. Glickman. 1997. "Driving Synthetic Mouth Gestures: Phonetic Recognition for FaceMe!." Proc. Eurospeech 4, 1995-1998.

[3] F. Lavagetto, P. Lavagetto. 1995. "A New Algorithm for Visual Synthesis of Speech." Proc. Eurospeech, 303-306.

[4] Jonas Beskow. 1995. "Rule-based Visual Speech Synthesis." Proc. Eurospeech, 299-302.

[5] 조철우, 정인화. 1998. "애니메이션 영상을 이용한 멀티미디어 음성출력기의 구현." 창원대학교 정보통신연구소 논문집 제2집, 147-150.

[6] Cheol-Woo Jo. 1998. "MSF Format for the Representation of Speech Synchronized Moving Image." Proc. of ICSLP 4, 1631-1635.

[7] Marray, I. R. and Arnott, J. L. 1993. "Toward the Simulation of emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion." JASA, 93(2), 1097-1108.

[8] Klasmeyer, G, Sendmeier, W. F. 1995. "Objective Voice Parameters to Characterize the Emotional Speech." ICPhS'95-Stockholm 1, 182-185.

[9] Mozziconaxxi, S. 1995. "Pitch Variations and Emotions in Speech." ICPhS'95-Stockholm 1, 178-181.

[10] 조철우, 조은경, 민경환. 1997. "정서정보의 변화에 따른 음성신호의 특성분석에 관한 연구." 한국음향학회지 16(3), 33-37.

접수일자 : '99. 2. 17.
게재결정 : '99. 3. 12.

▲ 조 철 우
경남 창원시 사림동9
창원대학교 제어계측공학과(우: 641-773)
Tel: (0551) 279-7552, Fax: (0551) 262-5064
e-mail: cwjo@sarim.changwon.ac.kr

▲ 김 대 현
경남 창원시 사림동9
창원대학교 제어계측공학과(우: 641-773)
Tel: (0551) 279-7559
e-mail: midas03@mail.taegu.net