MPEG-4 TTS (Text-to-Speech)

한민수
한국정보통신대학원대학교
e-mail: mshahn@icu.ac.kr

Abstract

It cannot be argued that speech is the most natural interfacing tool between men and machines. In order to realize acceptable speech interfaces, highly advanced speech recognizers and synthesizers are inevitable. Text-to-Speech(TTS) technology has been attracting a lot of interest among speech engineers because of its own benefits. Namely, the possible application areas of talking computers, emergency alarming systems in speech, speech output devices for speech-impaired, and so on. Hence, many researchers have made significant progresses in the speech synthesis techniques in the sense of their own languages and as a result, the quality of currently available speech synthesizers are believed to be acceptable to normal users. These are partly why the MPEG group had decided to include the TTS technology as one of its MPEG-4 functionalities. ETRI has made major contributions to the current MPEG-4 TTS among various MPEG-4 functionalities. They are; 1) use of original prosody for synthesized speech output, 2) trick mode functions for general users without breaking synthesized speech prosody, 3) interoperability with Facial Animation(FA) tools, and 4) dubbing a moving/animated picture with lib-shape pattern information.

I. 머리말

인간과 기계, 또는 컴퓨터와의 의사 소통 수단으로 음성이 이용될 수 있다면 가장 자연스런 방법이라는 것은 누구나 동의하는 바이다. 그러나 이를 위해선 고품질의 음성인식기와 음성합성기가 필요한 것이 사실이며 이는 현재의 기술과는 격차가 있는 것 또한 사실이다. 문자/음성 변환기(TTS)는 그 자체가 갖는 많은 장점 때문에 1950 년대부터 활발히 연구되어 왔다. 즉 성능이 좋은 문자/음성 변환기가 구현될 경우 그 기술은 말하는 컴퓨터, 발성 장애자용 발성 도구, 음성을 이용한 시스템 및 망의 비상사태 안내, 전자 우편 음성 낭독기, 가상현실 기술 분야에서의 avatar 의 음성을 이용한 의사표현 수단 등 그 응용 분야가 무척 다양하다. 따라서 많은 음성 공학자들이 음성합성 기술을 평생의 연구과제로 연구하기 시작하였으며 그 결과로써 포만트 합성기, LPC 합성기, LSP 합성기, PSOLA 합성기 등을 선보였으며 합성음의 품질도 크게 개선되었다..

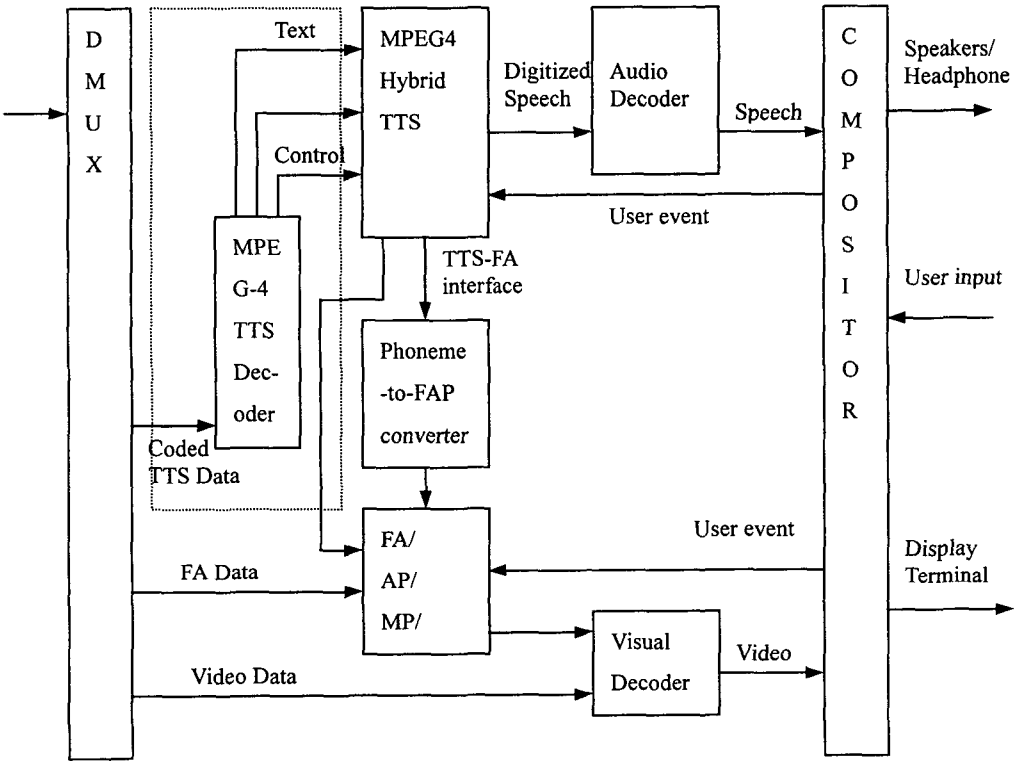
이와 같은 노력에 힘입어 발전된 TTS 기술은 현재 그 합성음의 품질이 낭독체 정형문의 경우 일반 사용자들이 알아 듣는데, 즉 합성된 문장의 의미를 이해하는 데는 큰 불편함이 없는 수준까지 와있으며 그 결과로 주변에서 손쉽게 상용화된 TTS 제품을 접할 수 있

게 되었다. 물론 아직은 최고 수준의 합성기라 할지라도 합성음의 자연성이 만족할 만한 수준은 아니나 이도 적용 분야를 한정시킬 경우 조만간 사용자에게 불편하지 않은 수준의 합성음을 제공할 수 있을 것이라는 것이 전문가들의 일반적인 견해이다.

한편 영상/음향과 관련된 첨단 기술에 대한 국제표준화를 담당하고 있는 ISO 산하 Moving Picture Expert Group(MPEG)에서는 자연영상/음향의 효율적인 고품질 압축/복원 기술을 중심으로 이미 표준화가 완료된 MPEG-1 이나 MPEG-2 의 경우와는 달리 현재 표준화가 완료되어가는 MPEG-4 의 경우 자연영상/음향 뿐만이 아니라 합성영상/음향도 그 대상으로 하고 있으며 압축/복원 방식도 구성요소 별로 최적의 기술을 적용할 수 있도록 지향하고 있다. 따라서 MPEG-4 의 표준화 대상에는 향후 가상현실 등에 필수적인 FA 도구, Body Animation(BA) 도구, 합성 입체 영상 코딩 도구, 합성음향, TTS 등이 포함된 것이다.

본 논문의 구성은 2 장에서는 제안된 MPEG-4 TTS 의 구조를 간략히 설명하고 3 장에서는 MPEG-4 TTS 가 갖는 기능들에 대하여 소개한다. MPEG-4 TTS 에 대한 인터페이스 bitstream 을 4 장에서 소개한 후 5 장에서는 이를 이용한 MPEG-4 TTS 시연시스템 및 표준화 경과에 대하여 간략히 설명하고 마지막으로 6 장에 결론을 맺었다.

II. MPEG-4 TTS 구조



<그림 1> MPEG-4 TTS 전체 구조

그림 1 에 MPEG-4 TTS 의 전체 구조를 보였다. 이 그림에서 FAP, AP, MP 는 각각 Facial Animation Parameter, Animated Picture, Moving Picture 를 의미한다. FAP 에 대해 부언하자면 합성될 음의 종류에 따라 FA 도구에서 적절한 입 모양을 합성하기 위해 필요한 변수들로서 구체적인 정의는 MPEG-4 Video Draft of International Standard 에서 찾을 수 있다. 한편 User event 는 사용자가 편의에 따라 조정할 수 있는 기능들로서 합성음의 발성 속도, 발성자의 성과 나이, 합성음의 크기, trick mode 등의 제어가 가능하다.

현재 이 세상에는 다양한 종류의, 또 여러 품질의 TTS 가 존재하고 있으므로 TTS 자체, 즉 합성방식이나 사용 데이터베이스 등을 표준화 한다는 것은 불가능한 일이다. 따라서 MPEG-4 TTS 의 표준화 대상은 그림 1 에서 점선으로 둘러싼 부분, 즉 디코더 만이다. 바꿔 말하자면 Demux 로부터 MPEG-4 TTS 디코더로 전달되는 bitstream 을 표준화함으로써 기존의 TTS 들이 약간의 보완 및 수정 만으로도 자신이 가지고 있는 품질의 합성음으로 MPEG-4 TTS 가 제공하는 다양한 기능들을 구현할 수 있도록 하는 것이 MPEG-4 TTS 표준화의 목적이다.

그림 1 에서 볼 수 있듯이 MPEG-4 TTS 는 기존의 일반적인 TTS 가 갖는 문자열로부터 음성을 합성해 내는 기능 외에 4 장에서 기술될 부가 기능들이 요구되므로 기본적으로 다음과 같은 인터페이스들이 주의 깊게 정의되어야만 기존의 다양한 TTS 를 MPEG-4 TTS 구조로 변경하는 것이 가능하다.

1) Demux 와 TTS 디코더 간의 인터페이스
2) TTS 디코더와 TTS 간의 인터페이스
3) 사용자와 TTS 간의 인터페이스
4) TTS 와 Phoneme-to-FAP converter 간의 인터페이스
5) TTS 와 오디오 디코더 간의 인터페이스

III. MPEG-4 TTS 기능

그림 1 에서 알 수 있듯이 MPEG-4 TTS 는 문자열로부터 음성을 합성해 내는 기능 외에 원래 발성 문장의 운율을 재현할 수 있어야 하며 합성되는 음성에 적합한 입 모양을 FA 도구가 생성할 수 있는 정보를 제공하여야 한다. 한편 입술 모양 패턴을 이용하여 Moving Picture 나 Animated Picture 를 dubbing 할 수 있어야 하며 Animated Picture 의 경우 발성음에 따라 입술 모양 패턴 정보를 활용하여 화면에서의 입술 모양을 제어할 수 있어야 한다. 뿐만 아니라 사용자의 편의를 위하여 발성음의 발성 속도, 크기 및 화자의 성과 나이를 선택할 수 있어야 하며 저장된 매체를 이용하는 경우를 위하여 start, stop, pause, replay, forward, backward 등의 trick mode 기능들도 제공되어야 한다. 따라서 이러한 모든 기능을 제공하기 위하여 MPEG-4 TTS 각각의 인터페이스는 다음과 같이 정의된다.

1. Demux 와 TTS 디코더 간의 인터페이스

   Demux 는 입력으로 들어온 MPEG-4 방식의 audiovisual 정보로부터 MPEG-4 TTS 에 해당되는 정보의 bit stream 을 MPEG-4 TTS 디코더로 보내 준다.

2. TTS 디코더와 TTS 간의 인터페이스

   Demux 로부터 입력된 TTS 용 bit stream 을 받아 TTS 디코더는 다음과 같은 정보를 해당 TTS 로 보내 준다.

   1) TTS 가 문장 만으로, 또는 FA 나 동영상과 연동되어 구동 되는가에 대한 정보
   2) 합성되어야 할 문장 내용
   3) 운율정보가 있는 경우 피치, 에너지, 지속시간 등의 운율정보
   4) 입술 모양 패턴 정보가 있는 경우 입술 모양패턴 정보
   5) Trick mode 가능 여부에 대한 정보

3. 사용자와 TTS 간의 인터페이스

   사용자가 이용할 수 있는 기능에 대한 정보가 이 인터페이스에 대하여 정의되어야 한다. 이들 정보는 사용자가 이용하고자 하는 발성 속도, 피치 및 음의 크기에 대한 정보, 선호하는 화자의 성과 나이에 대한 정보 및 trick mode 에 대한 정보이다. Trick mode 정보는 start, stop, pause, replay, forward, backward 에 대한 정보들을 의미한다. MPEG-4 TTS 는 방성속도 및 음의 크기를 16 단계로 변화 시킬 수 있으며 합성음의 피치 변화폭도 조정이 가능하다. 한편 입술 모양 패턴은 현재 256 종류가 사용 가능하며 합성음 화자의 나이도 4 가지 중 하나를 선택할 수 있다.

4. TTS 와 Phoneme-to-FAP converter 간의 인터페이스

   MPEG-4 TTS 는 Phoneme-to-FAP converter 로 최소한 합성되는 음소에 대한 정보를 전달해 주어야 한다. 이 경우 음소보다는 변이음 정보가 보다 정확한 입술 모양의 생성에 유리하고 또 다양한 여러가지 언어를 지원하기 위하여 MPEG-4 TTS 는 일반적으로 사용되는 발음 기호 대신 International Phonetic Alphabet(IPA)를 전달할 수 있도록 디자인되었다.

5. TTS 와 오디오 디코더 간의 인터페이스

   TTS 는 합성된 음성신호에 대한 16-bit 디지털 신호를 오디오 디코더로 보내주며 이 경우 오디오 디코더는 이 디지털 신호의 샘플링 주파수 및 디지털 신호 방식을 알 수 있어야 한다.

IV. MPEG-4 TTS 의 Bit Stream Syntax

<표 1> MPEG-4 TTS bit stream syntax(1)

Syntax	No. of bits
TTS_Sequence() {	
TTS_Sequence_Start_Code	Sc+8=32
TTS_Sequence_ID	10
Language_Code	8
Prosody_Enable	1
Video_Enable	1
Lib_Shape_Enable	1
Trick_Mode_Enable	1
Do{	
TTS_Sentence()	
}while(next-bits()==TTS-Sentence-Start-Code)	
}	

표 1 과 표 2 에 MPEG-4 TTS 인코딩 및 디코딩에 필요한 bit stream 정보를 보였다.

표 1 에 나타난 변수들에 대한 설명은 다음과 같다.

- TTS_Sequence_Start_Code: 'XXXXX'와 같은 16 진수로서 TTS Sequence 의 시작을 의미함
- TTS_Sequence_ID: 합성될 문장이 어떤 객체와 관련되는지에 대한 정보
- Language_Code: 합성될 언어의 종류, 즉 영어, 일어, 한국어, 독어 등의 정보로서 현재 ITU 회원국인 36 개국에 대해 정의되어 있음
- Prosody_Enable: 운율정보가 있으면 '1' 아니면 '0'
- Video_Enable: TTS 가 동영상과 연동되면 '1' 아니면 '0'
- Lip_Shape_Enable: 입술 모양 패턴 정보가 있으면 '1' 아니면 '0'
- Trick_Mode_Enable: 사용자가 trick mode 를 이용할 수 있으면 '1' 아니면 '0'

표 2 에 표시된 변수들에 대한 설명은 다음과 같다.

- TTS_Sentence_start_Code: 'XXXXX'와 같은 16 진수로서 합성문장의 시작을 알려 줌.
- TTS_Sentence_ID: 합성될 문장이 합성해야 할 전체 문장의 몇 번째인가의 정보
- Silence: 현재의 위치가 묵음이면 '1',아니면 '0'
- Silence_Duration: 묵음 지속시간 정보(msec)
- Gender: 남성 화자면 '1', 여성 화자면 '0'
- Age: 합성음 화자의 나이 정보로서 '0'면 어린이, '1'이면 젊은 사람, '2'면 중년, '3'이면 노인을 의미함
- Speech_Rate: 16 단계의 발성속도 정보
- Length_of_Text: 합성해야 할 전체 텍스트의 길이 정보
- TTS_Text: 임의의 길이를 갖는 입력 텍스트 스트링

<표 2> MPEG-4 TTS bit stream syntax(2)

Syntax	No. of bits
TTS_Sequence() {	
TTS_Sentence_Start_Code	Sc+8=32
TTS_Sentence_ID	10
Silence	1
If(Silence)	
{	12
Silence-Duration	
}	
Else	
{	1
Gender	2
Age	
If(!Video-Enable)	
{	
Speech-Rate	
}	
Length-of-Text	
TTS-Text()	
If(Prosody-Enable)	
{	
Dur-enable	1
F0-contour-enable	1
Number-of-Phonemes	1
For(j=0;j<Number-of-phonemes;j++)	10
{	
Symbol-each-Phoneme	8
If(Dur-enable) {	
Dur-each-Phoneme	12
}	
If(F0-Contour-enable) {	
F0-Contour-each-Phoneme	8*3=24
}	
If(Energy-Contour-enable) {	
Energy-Contour-each-Phoneme {	8*3=24
}	
}	
}	
If(Video-Enable)	
{	
Sentence-Duration	16
Position-in-Sentence	16
Offset	10
}	
If(Lip-Shape-Enable)	
{	
Number-of-Lip-Event	10
For(j=0;j<Number-of-Lip-Event;j++)	
{	
Lip-in-Sentence	16
Lip-shape	8
}	
}	
}	
}	

- Dur_enable:음소 지속시간 정보가 있으면 '1', 아니면 '0'
- Energy_Contour_enable: 에너지 contour 정보가 있으면 '1', 아니면 '0'
- Number_of_Phonemes: 전체 입력 텍스트를 합성하는데 필요한 음소 개수 정보
- Symbol_each_Phoneme: 각각의 음소정보(현재는 IPA 를 표준으로 함)
- Dur_each_Phoneme: 음소의 지속 시간 정보(msec)
- F0_Contour_each_Phoneme: 음소의 피치 contour 정보(음소의 0%, 50%, 100% 에서의 3 값으로 정의됨)
- Sentence_Duration: 문장 지속 시간(msec)
- Position_in_Sentence: 현 지점이 합성될 문장의 시작부터 얼마나 지난 지점인가의 정보(msec)
- Offset: 현재 위치가 소속된 관련 동영상의 GOP(Group of Pictures)의 시작점인 I-frame 과 얼마나 떨어져 있는가에 대한 정보(msec)
- Number_of_Lip_Event: 처리해야 할 입술 모양 패턴의 개수
- Lip_in_Sentence: 입술 모양 패턴의 지속 시간 정보(msec)
- Lip_Shape: 입술 모양 패턴 정보

V. MPEG-4 TTS 표준화 경과 및 시연 시스템

MPEG-4 TTS 의 표준화 경과를 간단히 살펴보면 다음과 같다. MPEG 에서 MPEG-4 TTS 에 대한 요구 사항을 처음으로 발표한 것은 NTT 가 처음으로 MPEG-4 TTS 에 대한 기고서를 제안한 1996 년 7 월 핀란드의 Tempere 에서 열린 제 35 차 회의에서였다[1]. 이 때 발표된 MPEG-4 TTS 에 대한 요구 사항은 크게 3 가지로서 발성자의 원래 운율을 합성시 재현할 수 있을 것, FA 도구와 연동이 가능할 것, 운율을 깨지 않으면서 사용자에게 trick mode 를 제공할 수 있을 것의 3 가지였다. 1997 년 9 월 시카고에서 열린 제 36 차 회의에서 한국의 ETRI 가 원래 운율 재현 및 FA 와의 연동 방법에 대한 기고서를 발표하였다[2]. 브라질의 Maceio 에서 같은 해 11 월에 개최된 제 37 차 회의에서 ETRI 에서 운율을 유지하면서도 사용자에게 trick mode 를 제공할 수 있는 기술과 입술 모양 패턴을 이용한 입술 영상제어 기술 및 입술 모양 정보를 이용하기 위한 MPEG-4 TTS 전체 구조를 제안하고 시연하여 MPEG-4 SNHC Verification Model 에 포함시켰다[3]. 1997 년 2 월 스페인의 세비야에서 열린 제 38 차 MPEG 회의에서는 ETRI 에서 MPEG-4 TTS 전체 인터페이스에 대한 bit stream 을 제안하여 수정 없이 MPEG-4 SNHC Verification Model 에 포함되었다[4].

1997 년 4 월 영국의 브리스톨에서 열린 제 39 차 MPEG 회의에서 ETRI 는 MPEG-4 TTS 가 실제 입술 영상과 연동되어 구동 되는 것을 시연하였으며 세비야에서 확정된 MPEG-4 기능을 이미 구현하고 MPEG ftp site 에 해당 software 를 reference software 로 공개한 것이 인정되어 무리 없이 MPEG-4 Audio Working Draft 3.0 에 포함되었다[5]. 같은 해 7 월 말, 스웨덴의 스톡홀름에서 열렸던 제 40 차 MPEG 회의에서는 MPEG 회의 참가자를 대상

으로 공개된 MoMuSys 사의 MPEG-4 S/W Video encoder 와 MPEG-4 Audio Working Draft 3.0 규격을 따르는 MPEG-4 TTS S/W 를 이용하여 동영상과 TTS 가 입술 모양 패턴을 이용하여 동기되어 구동될 수 있음을 시연한 결과 같은 해 10 월 말에 작성된 MPEG-4 Committee Draft Version 1.0 에 제안된 상태대로 MPEG-4 TTS 를 포함시키기로 결정되었다[6]. 그 후 큰 무리 없이 Draft of International Standard 를 거쳐 현재 MPEG-4 표준의 하나로 인정되었다.

< 그림 2. MPEG-4 TTS 시연 시스템의 화면 구성도>

그림 2 에 MPEG-4 TTS 시연 시스템의 화면 구성을 보였다. 이 그림에서 볼 수 있듯이 시연 시스템의 화면은 FA 도구 화면, 실제 입술 모양 화면, 텍스트(Caption) 창 화면, Encoder 화면 및 제어 패널 화면으로 이루어져 있다. 즉 입력 문장에 따라 FA 도구나 실제 입술 모양 표시 도구가 사용자의 선택에 의하여 작동될 수 있도록 구성되었으며 현재 사용 가능 언어는 영어, 일어, 한국어의 3 개이다. 한국어와 일어는 ETRI 에서 개발한 한국어 합성기를 활용하고 있으며 영어의 경우에는 영국의 Cambridge 대학에서 개발하여 공개한 Festiva system 을 이용하여 구현하였다.

한편 사용자는 이 그림에서 알 수 있듯이 발성 속도, 발성자의 성, 발성자의 나이 등을 조절할 수 있으며 trick mode 기능의 도움을 받아 운률을 깨지 않으면서도 pause, stop, replay, forward, backward 등의 기능을 마치 cassette recorder 를 사용하듯이 이용할 수 있다.

VI. 결론

이상에서 살펴본 바와 같이 현재의 MPEG-4 TTS는 그 기능의 대부분과 전체 구조 및 인터페이스 bit stream을 ETRI에서 제안한 것으로 주요 기능은 크게 발성시의 원래 운율 부가 기능, FA 도구와의 연동 기능, 입술 모양 패턴을 이용한 동영상 dubbing 기능, 입술 모양 패턴을 이용한 animated picture의 입술 모양 제어 기능, 사용자용 trick mode 기능의 다섯 가지이다. 이러한 MPEG-4 TTS 기능들은 향후 story-teller on demand, 동영상 dubbing 도구, 가상현실 분야에서의 음성을 이용한 avatar의 의사 표현 도구, 발성 장애자용 음성 발생기, 전자우편 음성 낭독기 등 그 응용 분야가 무척 다양하다.

마지막으로 첨언하고 싶은 것은 아직도 MPEG-4 TTS에는 개발할 기술들, 예를 들면 입술 모양 패턴의 표준화 및 입술 모양 이벤트 사이의 인터폴레이션 기술 뿐만 아니라 합성음의 명료도 및 자연성 개선, 음색 변환 기술 등의 중요한 문제가 많이 남아 있으므로 국내 학계나 산업체에서도 이들을 개발하고 지적재산권을 확보한다면 향후 좋은 결과를 기대할 수 있을 것이라는 점이다.

참고문헌

[1] S. Nakajima, "A Hybrid Scalable Text to Speech Synthesis," Tampere meeting, document ISO/IEC/JTC1/SC29/WG11 M1157, July 1996.

[2] J.C. Lee, S.H. Kim, "Multilevel Scalable TTS Synthesis," Chicago meeting document ISO/IEC/JTC1/SC29/WG11 M1244, September 1996.

[3] M. Hahn, J.C. Lee, "MPEG4 TTS Interface," Maceio meeting, document ISO/IEC/JTC1/SC29/WG11 M1524, November 1996.

[4] M. Hahn, H.S. Lee, J.W. Yang, "Revision of MPEG-4 TTS Interface," Sevilla meeting, document ISO/IEC/JTC1/SC29/WG11 M1739, February 1997.

[5] MPEG-4 Audio Group, "MPEG-4 Audio Working Draft 3.0." Bristol meeting, document ISO/IEC/JTC1/SC29/WG11 N1631, April 1997.

[6] Y.K. Lim, M. Hahn, J.C. Lee, H.S. Lee, "Synchronization of MPEG-4 TTS with Moving Picture," Stockholm meeting, document ISO/IEC/JTC1/SC29/WG11 M2450, July 1997.