신기술 논고

MPEG-4 오디오 기술 동향

한민수*, 강경옥**, 변경진*
(*한국정보통신대학원 대학교, **한국전자통신연구원)

ABSTRACT

In this survey paper, the emerging MPEG-4 audio technology is described. In the previous MPEG-1 and the MPEG-4 audio works, only the natural audio and the speech coding techniques were the standardization objects. But in the MPEG-4 audio standardization, not only the natural audio and the speech coding, but also the structured audio and the synthetic speech techniques are included. The purpose of this expansion can be summarized as the preparation for the versatile high-quality multimedia services supposed to emerge in the 21st century.

I. MPEG-4 오디오 개요

MPEG-4 오디오는 다양한 오디오 부호화 방법을 통합한 새로운 형태의 오디오 부호화 표준이라고 할 수 있다. 즉, 합성음이 동반된 자연음, 고품질 전송에서 저비트율의 전송에 이르기까지, 음악이 동반된 음성, 간단한 음향부터 복잡한 사운드 트랙 음향까지, 또한 양방향성(interactive)이 있는 콘텐트까지를 모두 다룬다고 할 수 있다. MPEG-4 오디오의 새로운 개념을 요약하면 다음과 같다.

- MPEG-4는 전송 표준을 정의하고 있지 않다.
- MPEG-4 오디오는 저비트율 부호화를 사용하고 있다.
- MPEG-4 오디오는 다양한 도구들을 가진 객체 기반 부호화 표준이다.
- MPEG-4 오디오는 합성음(synthetic sound)에 대한 부호화 방법도 제공한다.

한편, MPEG-4 오디오는 서로 간에 특별한 관련이 없고 각각 독립적인 목표 기능을 갖는 다양한 도구 셋을 제공하고 있으며 이러한 도구들을 다양한 응용목적에 적용하기 위하여 프로파일(profile)을 정의하여 사용하고 있다. MPEG-4 오디오의 도구들은 다음과 같이 분류될 수 있다.

- 합성음성과 자연음성의 전송과 복호를 위한 음성 부호화 도구(speech tool)
- 레코딩된 음악 및 기타 사운드 트랙의 전송과 복호를 위한 오디오 부호화 도구(audio tool)
- 합성 음악과 기타 사운드의 극저 비트율 description과 전송, 단말기에서의 합성을 위한 합성 도구(synthesis tool)
- 객체 기반 부호화, 대화형(interactive) 기능, AV 동기를 위한 composition tool
- Recording 없이 여러가지 다른 비트율로 전송할 수 있는 bit stream을 생성하기 위한 scalability tool

다음은 MPEG-4오디오에서 다루는 각각의 도구에 대하여 간략히 설명한다.

• MPEG-4 speech coding tools

자연 음성과 합성 음성에 대한 2가지 형태의 부호화 도구를 제공하고 있다. 자연음성 부호화 도구는 2~24 kbps 사이의 자연 음성에 대한 압축 및 복원을 할 수 있다. 파라메트릭 음성 부호화 기술인 HVXC(harmonic and vector excitation coding) 부호화 기술과 선형 예측 부호화 방법인 CELP(code excited linear prediction) 부호화 기술의 2가지 기술이 사용된다.

HVXC는 bitrate scalability 기술을 사용하여 2.0 kbps에서 4.0 kbps 사이의 고정 비트율로 동작하며, 가변 비트율 모드에서는 1.2 ~ 1.7 kbps 사이의 저비트율로 동작할 수 있다. HVXC는 100~3800 Hz 대역 신호에 대해 8 kHz 표본화 주파수에서 communication quality에서 거의 toll quality에 가까운 음성을 제공한다. 또한, 복호화 과정에서 음질에 영향없이 재생 속도와 피치를 가변할 수 있다.

기존의 CELP는 단일 비트율로 압축하며 특정 목적에 최적화 되어 있는 반면에, MPEG-4 CELP는 압축도 하나의 기능으로 제공되며 AAC Scalable 도구 등과 결

합되어 scalability 기능을 제공할 수 있는 기본 부호기(basic coder)로 사용될 수 있다. MPEG-4 CELP는 bitrate 및 bandwidth scalability 기능도 제공하고 있으며, 100~3800 Hz 대역 신호에 대해서는8 kHz의 표본화 주파수를, 50~7000 Hz 대역 신호에 대해서는 16 kHz의 표본화 주파수를 제공하고 있다

합성음성 부호화 도구는 TTS(Text-to-Speech) 합성 시스템에 대한 인터페이스를 제공하여 합성음성에 의한 저비트율 동작이 가능한 한편 facial animation 또는 동영상과 결합하여 저비트율의 회의통신(videoconferencing) 등에 사용할 수 있다.

• MPEG-4 general audio coding tools

MPEG-4는 모노, 스테레오 또는 멀티채널 오디오 객체로 구성된 오디오 채널당 6 kbps에서 수 백 kbps까지의 비트율로 natural audio를 부호화할 수 있다. 또한 MPEG-4 도구 셋 내에서 MPEG-2 AAC에 추가적인 도구를 사용함으로써 고품질 압축 기능을 제공한다. MPEG-4 오디오는 고품질 오디오를 부호화하는 도구 셋으로 구성된 general audio(GA) 부호화기를 정의하고 있으며, 이 부호화 기술에서는 지각 필터뱅크, 복잡한 마스킹 모델, noise-shaping 기술, 채널 결합 및 가능한 한 고품질을 유지하면서 최대의 압축률을 제공할 수 있는 noiseless coding 및 비트할당 방법을 사용한다.

채널 당 6 kbps에서 64 kbps 까지의 비트율에 대해, MPEG-4 표준은 AAC에 대한 확장(extension)과 콘텐트 저작자가 비트율에 따라 사용할 도구를 가변함으로써 고품질을 얻을 수 있는 TwinVQ 도구를 제공한다.

• MPEG-4 audio synthesis tools

오디오(general audio)에 대한 합성 방법을 제공하는 MPEG-4 도구 셋을 MPEG-4 SA (structured audio) coder라고 한다. SA 부호기는 합성음을 기술하는 매우 일반적인 방법과 복호 단말에서 합성음을 생성하는 방법을 제공한다. 고품질 스테레오를 0 kbps에서 2~3 kbps 사이의 비트율로 전송할 수 있다.

SA는 합성방법을 표준화하는 것이 아니라 합성방법을 기술하는 언어를 표준화 하는 것이다. 합성 오디오(synthetic audio)는 악보(score)의 제어 하에 오디오 신호를 생성할 수 있는 악기 모듈에 의하여 전송되며, 여러 가지 다른 악기가 하나의 SA bit stream으로 전송되고 사용될 수 있다. 악보는 특정한 시간에 다양한 악기가 전체 음악 연주를 구성하도록 다양한 악기를 호출하는 순차적 명령어의 집합이다. 악기를 기술하는 포맷을 SAOL(Structured Audio Orchestra Language)이라고 하며, 악보를 기술하는 포맷을 SASL(Structured Audio Score Language)라고 한다.

• MPEG-4 audio composition tools

오디오 composition 도구는 개별적인 다수의 오디오 객체를 하나의 사운드 트랙으로 믹싱하는 기술이다. MPEG-4에서는 다채널 믹스 자체가 다른 부호화 도구를 사용한 각각의 오디오 객체와 함께 전송될 수 있으며, 또한 다운믹스를 하기 위한 일련의 정보(instruction)를 bit stream에 포함할 수 있다. 다수의 객체가 수신됨에 따라 이들을 개별적으로 복호화한 후 이를 청취자에게 바로 재생하는 것이 아니라 다운믹스를 하기 위한 일련의 정보(instruction)를 사용하여 하나의 사운드 트랙을 만든 후 이를 청취자에게 재생하는 것이다. 이와 같이 객체 기반 부호화 방법을 사용함으로써 하나의 사운드 트랙을 부호화 하는 기존의 부호화 방법에 비하여 상당한 정보량을 절감할 수 있으며, 또한 객체와 청취자 사이의 대화성을 제공할 수 있다는 장점이 있다.

• MPEG-4 audio scalability tools

MPEG-4의 많은 bit stream은 한가지 방법 또는 다른 방법으로 scalability를 가지고 있다.

가변등급(scalability)이란 bit stream의 어떤 부분만 사용하여 복호화가 가능하며, 더 낮은 품질과 대역 또는 선택된 내용으로 의미있는 오디오 신호를 만드는 것을 의미한다.

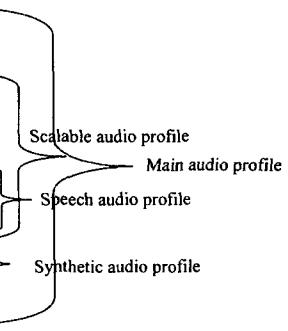
Bitrate scalability란 하나의 bit stream이 더 낮은 비트율의 bit stream으로 해석되어 이들의 조합이 의미있는 신호로 부호화될 수 있는 것을 의미한다. bit stream의 해석은 전송 중에나 디코더에서 수행되며, 각각의 natural audio 부호화 도구 내에서나 서로 다른 natural audio 부호화 도구 간의 조합에 의해 가능하다.

Bitrate scalability의 특수한 경우로서 bandwidth scalability가 있으며, 이는 주파수 스펙트럼의 일부를 표현하는 bit stream의 일부를 전송 중에나 디코딩할 때 버릴 수 있는 것을 의미한다. 이는 CELP부호화기에서 사용되며 확장 계층(extension layer)이 협대역 기본 계층(base layer) 인코더를 광대역 부호기로 변환한다. 또한 GA 부호기 도구들도 서로 다른 부호화 계층에 대한 매우 유연한 대역폭 제어를 가능하게 한다. 이외에 서로 다른 복잡도를 가진 인코더 간에 의미있는 bit stream을 생성할 수 있는 encoder complexity scalability와 주어진 bit stream을 서로 다른 레벨의 복잡도를 가진 디코더에 의하여 복호할 수 있는 decoder complexity scalability를 제공하고 있다.

한편, MPEG-4 오디오는 프로파일(profile)과 이와 관련된 도구(tool)를 표 1과 같이 정의하고 있다. 각 프로파일은 객체(object)들로 구성되며 표 1에 MPEG-4 오디오에서 다루는 모든 객체를 기술하고 있다. 또한,

표1. MPEG-4 오디오의 객체

도구(Tools) / 객체(Objects)	13818-7 main	13818-7 LC	13818-7 SSR	PNS	LTP	TLSS	Twin VQ	CELP	HVXC	TTSI	SA tools	SABSF	MIDI	hierarchy
Null														
AAC main	X			X										contains AAC LC
AAC SSR			X	X										
AAC LC		X		X										
AAC LTP		X		X	X									
AAC scalable		X		X	X	X								
TwinVQ					X		X							
CELP								X						
HVXC									X					
TTSI										X				
Main synthetic											X	X	X	contains W/T &Algor. synthesis
Wavetable synthesis												X	X	Contains General MIDI
General MIDI													X	
Algorithmic Synthesis and Audio FX											X			



MPEG-4 오디오의 4개의 프로파일과 해당 객체들의 관계를 함께 나타내었다. 표에서 'X' 는 해당 객체를 기술하는데 필요한 도구를 나타내는 것이다.

각 객체의 정의는 다음과 같다.

- AAC main object: ISO/IEC 13818-7(MPEG-2 AAC)에서 정의한 AAC main 프로파일과 상당히 유사하나, PNS(perceptual noise shaping) 도구를 추가적으로 필요로 한다. Bit stream syntax가 MPEG-2 AAC와 호환적이다. 이bit stream을 디코딩할 수 있는 디코더는 MPEG-2 AAC의 bit stream도 디코딩할 수 있다.
- AAC LC(low complexity) object: ISO/IEC 13818-7(MPEG-2 AAC)의 AAC LC 프로파일의 MPEG-4 버전으로, PNS(perceptual noise shaping) 도구를 추가적으로 필요로 한다.
- AAC SSR(scalable sampling rate) object: ISO/IEC 13818-7(MPEG-2 AAC)의 AAC SSR 프로파일의 MPEG-4 버전으로, PNS(perceptual noise shaping) 도구를 추가적으로 필요로 한다.
- AAC Scalable object: 이 객체는 비트율 및 대역폭과 관련된 scalability 기능을 제공하기 위한 bit stream syntax를 사용하며, TwinVQ 및 CELP 부호기 등과의 조합을 포함하여 많은 scalable 조합이 가능하다. 그러나 모노 또는 스테레오 객체만 지원한다.
- Twin VQ object: MDCT 계수를 양자화하는 GA 부호화 scheme에 속하며, MPEG-2 AAC의 허프만(Huffman)부호화 대신에 고정 비트율의 벡터 양자화(VQ)를 사용한다. 저비트율의 모노 및 스테레오 오디오 부호화가 가능하며, AAC scalable object와의 조합에 의한 scalable 오디오 부호화 scheme을 사용할 수 있다.
- CELP object: CELP 음성 부호화 도구를 사용하며, 8kHz와 16kHz의 표본화 주파수에서 4~24kbps의 비트율을 제공한다. CELP bit stream의 scalable decoding을 위하여 비트율 및 대역폭과 관련된 scalability 기능을 제공한다. 항상 하나의 모노 신호만을 포함한다.
- HVXC object: HVXC 부호화 도구를 사용하며, scalable 및 non-scalable 모드에서 2.0~4.0 kbps의 고정 비트율을 제공하고, 가변 비트율 모드에서는 2 kbps 미만의 비트율을 제공할 수 있다. 피치와 속도를 가변할 수 있으며, 8 kHz 표본화 주파수와 모노 채널만을 지원한다.

본 원고의 구성은 2장에서 MPEG-4 General Audio 부호화 기술을 다루고 3장에서 MPEG-4 음성 부호화 기술에 대하여 설명한 후4장에서 MPEG-4 TTS에 대하여 설명하고 마지막으로5장에서 간단한 결론을 내렸다. MPEG-4오디오 기술 중Structured Audio등 몇몇 중요성이 떨어진다고 판단되는 분야는 임의로 자세한 언급을 생략하였다.

II. MPEG-4General Audio 부호화 기술

GA(general audio) 부호화는 음악신호에 대해 모노의 경우에는 채널 당 6 kbps부터, 스테레오의 경우에는 채널 당(스테레오 신호당) 12 kbps부터 방송용 오디오 품질인 64 kbps 또는 그 이상의 비트율로 부호화가 가능하다.

MPEG-2 AAC의 syntax가 MPEG-4 GA 부호화에 의해 완전히 지원되며, AAC의 모든 특징이 MPEG-4에도 적용된다. AAC에 기반을 둔 MPEG-4 도구들은 bitrate scalability나 극저 비트율에서 개선된 부호화 효율의 제공 등의 부가 기능을 제공하는 다른 MPEG-4 GA 부호화 도구들과 함께 사용할 수 있다. Bitrate scalability는 GA 부호화 도구 만으로 또는 GA 부호화 도구가 아닌 외부 코어 부호화 도구(예를 들어 CELP)와의 조합에 의해 얻을 수 있다.

MPEG-4 GA 부호화 도구에는 MPEG-2 AAC와 관련된 도구와 Twin-VQ양자화 및 부호화 모듈과 관계되는 MPEG-4 add-on 도구가 있다. Twin-VQ는 AAC 양자화 모듈의 대안으로 사용될 수 있으며, 채널 당 6 kbps 이상의 비트율을 제공하나, 16 kbps 미만의 고정 비트율에서 사용하는 것이 좋다.

GA부호화 기술은 시간/주파수 변환 부호화와 지각 부호화의 방식을 사용하며, 인간의 청각 현상을 부호화에 이용하기 위한 심리음향 모델링(Psychoacoustic modelling), 시간영역의 신호를 주파수 영역으로 변환하여 부호화의 효율을 높이는 시간/주파수 변환, 주파수 영역의 다채널 오디오 신호를 압축 부호화하기 위한 오디오 데이터 압축 부호화, 압축된 오디오 데이터를 실제 양자화하여 bit stream을 구성하는 양자화 및 bit stream 형성으로 크게 구분할 수 있다. 압축 처리부의 블록 단위의 접속은 그림 1과 같다.

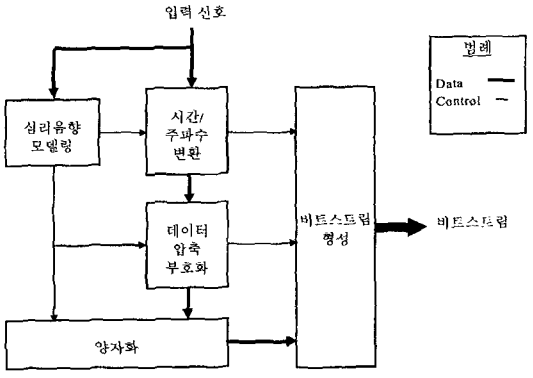


그림 1. GA 부호화 처리의 블록도

• 심리음향 모델링

심리음향 모델링은 다채널 오디오의 지각 부호화를 위한 인간의 청각 특성을 모델링하기 위한 기능으로서 입력 오디오의 특성을 추출하며, 대역별로 인간의 청각에 의해 감지되지 않는 양자화 잡음의 정도를 계산하여 부호화에 필요한 비트의 할당 시 반영으로써 최적의 부호화를 달성하도록 한다.

심리음향 모델링은 이 외에도 오디오의 부호화를 위하여 필요로 하는 청각 모델링 파라미터들을 계산하고 출력한다

• 시간/주파수 변환

시간/주파수 변환에서는 일반적으로 시간영역의 신호보다 주파수 영역의 신호를 부호화 하기 용이한 특성을 이용하기 위하여 시간영역의 오디오 신호를 주파수 영역의 오디오 신호로 변환하는 부분이다. 시간/주파수 변환 방식으로는 MDCT(Modified Discrete Cosine Transform)를 사용한다.

• 데이터 압축 부호화

오디오의 데이터 압축 부호화를 위해서는 시간/주파수 변환부에서 출력되는 주파수 스펙트럼의 진폭을 줄이거나 예측할 수 있는 방법들이 필요하다. 이러한 기능들은 포락선을 이용하기도 하며, 채널 간의 관계에 착안하기도 하고, 시간적인 신호의 변화를 이용하기도 한다.

시간영역에서의 데이터 압축 방법으로서는 이전 프레임의 스펙트럼으로부터 현재 프레임의 스펙트럼을 예측하는 프레임간 예측 및 이전의 상관성이 높은 신호에 대하여 예측을 행하는 피치 성분 예측을 사용할 수 있는데, 예측 파라미터와 예측 오차 만을 전송 함으로서 전송 데이터의 양을 감소 시킬 수 있다.

채널 간의 관계에 의한 데이터의 압축 방법으로는 좌, 우 채널로서 구분되는 각 채널 쌍에 대해서 하나의 채널에 대해서 다른 채널의 레벨 차이 만을 전송 함으로서 실제 전송되는 데이터의 양을 줄이는 세기(intensity)/결합(joint) 부호화, 좌, 우 채널의 신호를 M/S(Middle/Side) 채널로서 변환하여 데이터를 줄이는 M/S 변환을 사용할 수 있다.

포락선을 이용하는 방법으로서는 포락선을 생성할 수 있는 LPC(Linear Prediction Coefficient)에 의해 포락선이 제거된 신호를 생성 함으로서 데이터의 변화 폭을 줄여 신호 대 잡음 비를 향상시킬 수 있는 포락선 제거 필터를 사용할 수 있다.

• 양자화

양자화에서는 데이터 압축 부호화에 의해 압축된 주

파수 스펙트럼을 심리음향 모델링을 이용하여 주어진 비트율에 대해서 최적의 양자화 레벨을 할당하는 방법을 사용한다.

양자화된 주파수 스펙트럼들은 할당된 비트에 의해 표현되는 값들로 구성이 되는데 이들을 보다 적은 비트로 표현하기 위해 디코더에서 원래의 값들을 복원할 수 있는 상태로 부호화하는 방법, 예를 들면 허프만 부호화, 벡터 양자화(vector quantization: VQ)를 사용하여 보다 감소된 비트 수로 부호화할 수도 있다.

• MPEG-4 GA 인코더 및 디코더의 도구 셋

그림 1의 GA 부호화 처리 블록도에 따른 MPEG-4 GA의 인코더 및 디코더의 도구들을 각각 그림 2와 3에 나타낸다. 아래에 GA 디코더에 사용되는 각 도구에 대해 간략히 설명한다.

Bit stream 역다중화 도구(bitstream demultiplexer tool)는 MPEG-4 GA bit stream을 해석하여 bit stream을 각 도구에 해당하는 부분으로 분류하고 각 도구에 해당 정보를 전달한다. 이 도구의 출력은 다음과 같다.

- 다음의 2가지 방법 중 하나에 해당하는 양자화된 스펙트럼: 섹션 정보 및 무잡음 부호화 스펙트럼(noiselessly coded spectra)(AAC) 또는 코드 벡터 인덱스 셋(TwinVQ)
- M/S 정보(선택사항)
- 예측기 부가정보(선택사항)
- PNS (perceptual noise substitution) 정보(선택사항)
- 인텐시티 스테레오(intensity stereo) 및 결합 채널(coupling channel) 제어 정보(2가지 모두 선택사항)
- TNS (temporal noise shaping) 정보(선택사항)
- 필터뱅크 제어 정보
- 이득 제어 정보(선택사항)
- Bitrate scalability 관련 부가 정보(선택사항)

무잡음 복호화 도구(noiseless decoding tool)는 bit stream 역다중화 도구로부터 정보를 받아 그 정보를 해석(parsing)하고 Huffman 부호화 데이터를 복호하고, 양자화된 스펙트럼과 Huffman 및 DPCM 부호화 스케일팩터를 복원한다. 이 도구의 입력은 무잡음 부호화 스펙트럼과 이 스펙트럼에 대한 섹션 정보이며, 출력은 복호화된 스케일팩터의 정수값과 스펙트럼에 대한 양자화 값이다.

역양자화 도구(inverse quantizer tool)는 스펙트럼에 대한 양자화 값을 받아 정수값을 non-scaled의 역양자화 스펙트럼으로 복원한다. 이 양자화기는 비균일(non-uniform) 양자화기이다.

스케일팩터 도구(scalefactor tool)는 정수값의 scalefactors를 실제값으로 복원하며, 스케일팩터 값을 사용하여 non-scaled의 역양자화 스펙트럼에 곱한다. 이 도구의 출력은 스케일 팩터가 곱해진(scaled) 역양자화 스펙트럼이다.

M/S 도구(M/S tool)는 결정 정보의 제어 하에 Middle/Side 스펙트럼 쌍(pair)을 좌/우(Left/Right) 스펙트럼 쌍으로 변환하여, 결과적으로 스테레오 이미지 quality를 개선하며 때로는 부호화 효율을 증가시킨다. 이 도구의 입력은 M/S 결정정보와 채널 쌍과 관련된 스케일팩터가 곱해진 역양자화 스펙트럼이며, 출력은 M/S 복호화된 채널 쌍과 관련된 스케일팩터가 곱해진 역양자화 스펙트럼이다.

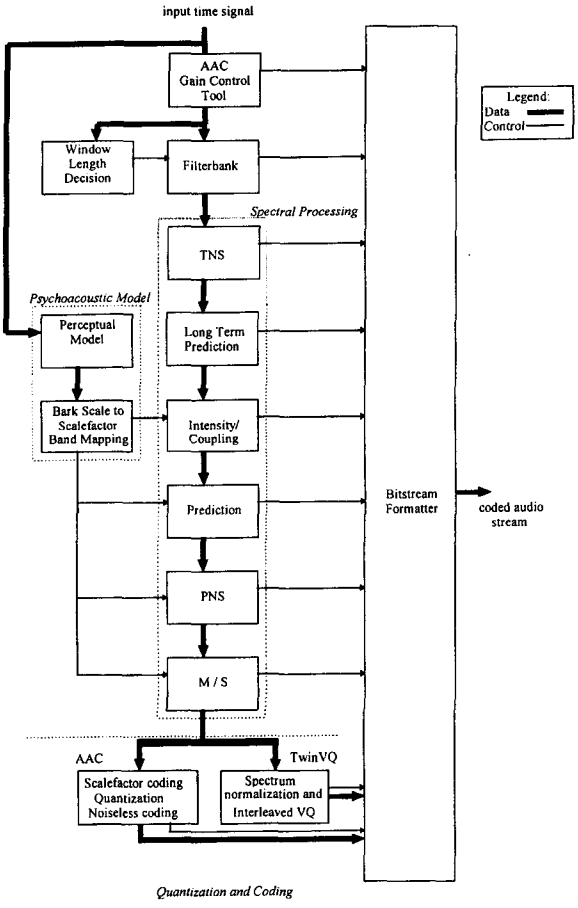


그림 2. MPEG-4 GA 인코더의 구조(non-scalable)

예측 도구(prediction tool)는 인코더에서 수행한 예측 처리 과정을 역으로 수행한다. 예측기 상태 정보의 제어하에 인코더에 추출한 redundancy를 재삽입하며, 2계 역방향 적응 예측기(second order backward adaptive

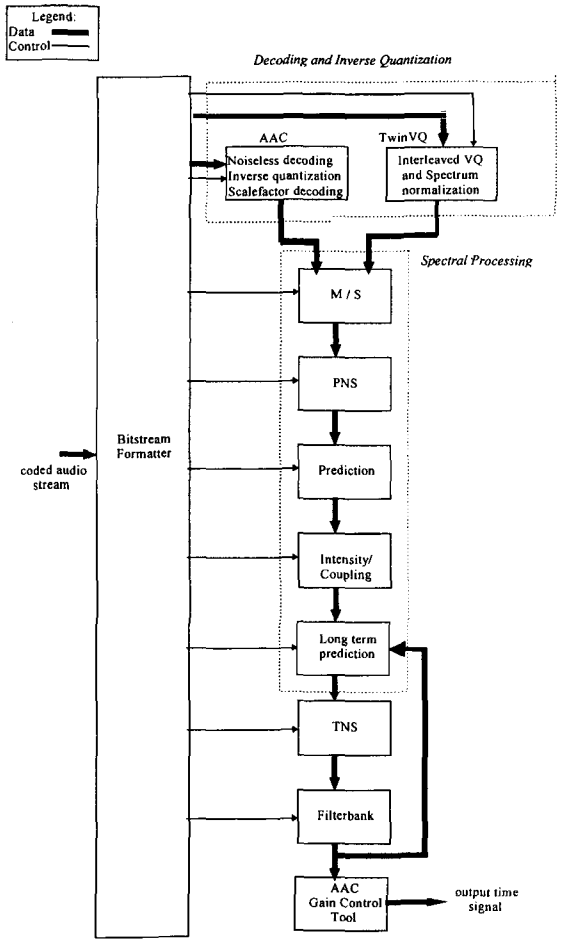


그림 3. MPEG-4 GA 디코더의 구조(non-scalable)

predictor)로 구현된다. 입력은 예측기 상태정보 및 부가정보와 스케일팩터가 곱해진 역양자화 스펙트럼이며, 출력은 예측이 적용된 스케일팩터가 곱해진 역양자화 스펙트럼이다. 만약 예측 도구가 사용되지 않으면, 스케일팩터가 곱해진 역양자화 스펙트럼이 수정없이 전달된다.

예측도구에 대한 대안으로 순방향 적응 LTP (forward adaptive long term prediction) 도구가 사용될 수 있다. 이 도구의 입력으로는 복원된 시간영역의 디코더 출력과 스케일팩터가 곱해진 역양자화 스펙트럼이며, 출력은 예측이 적용된 스케일팩터가 곱해진 역양자화 스펙트럼이다.

PNS(perceptual noise substitution) 도구는 잡음성(noise-like) 신호성분에 대한 효율적인 표현수단을 제공하여 채널 스펙트럼에 잡음 대치(noise substitution) 복호화를 수행한다. 이 도구에의 입력은 역양자화된 스펙트럼과 PNS 제어 정보이며, 출력은 역양자화된 스펙트럼이다.

세기 스테레오/결합(intensity stereo / coupling) 도구는 스펙트럼 쌍에 대한 세기 스테레오 디코딩을 수행한다. 부가적으로, 결합제어 정보에 따라 종속적으로 스위칭되는 결합채널로부터 관련 데이터를 스펙트럼에 더한다. 입력은 역양자화된 스펙트럼과 세기 스테레오 제어 정보 및 결합 제어 정보이며, 출력은 세기 및 결합 채널 복호화가 수행된 후의 역양자된 스펙트럼이다. 세기 스테레오 도구와 M/S 도구는 주어진 스케일팩터 밴드 및 스펙트럼 쌍에 대해 서로 배타적으로 동작된다.

TNS(temporal noise shaping) 도구는 부호화 잡음의 미세 시간 구조(fine time structure)를 제어한다. 인코더에서, TNS는 신호의 시간 진폭 포락선을 평탄화하며, 디코더에서는 TNS 정보의 제어하에 인코더의 역과정을 수행하여 실제 시간 진폭 포락선을 복원한다. 이 과정은 스펙트럼의 일부에 필터링을 적용하여 수행한다. TNS 도구의 입력은 TNS 정보와 역양자화된 스펙트럼이고, 출력은 역양자화된 스펙트럼이다.

필터뱅크(filterbank) 도구는 필터뱅크 제어 정보와 이득 제어 정보의 유무에 의하여 지시되는 것에 의해 인코더에서 수행된 주파수 매핑의 역과정을 적용한다. 필터뱅크 도구로는 IMDCT (inverse modified discrete cosine transform)가 사용된다. 이득 제어 도구가 사용되지 않으면, IMDCT 입력은 적용 창의 sequence에 따라 1024 또는 128개(또는 960 또는 120개)의 스펙트럼 계수로 구성된다. 이득 제어 도구가 사용되면, 창의 sequence 값에 따라 256 또는 36개의 계수로 구성된 4개의 셋을 사용하는 것으로 필터뱅크가 구성된다. 입력은 역양자화된 스펙트럼과 필터뱅크 제어 정보이며, 출력은 시간영역의 복원된 오디오 신호이다.

필터뱅크의 대안으로 한 프레임이 960개의 샘플로 구성되며 따라서 프레임 길이가 정수값(시간 단위로)을 갖는 것을 생각할 수 있다. 즉, 48 kHz 표본화 주파수에 대하여 정확히 20 msec의 프레임 크기를 갖는다. 이는 CELP/AAC bitrate scalability 조합의 경우에 유용하다. 이 경우에는 10 msec의 정수배의 프레임 크기를 갖는 CELP 계층(layer) 프레임과 AAC enhancement 계층 프레임이 결합된 구성이 가능해 진다.

이득 제어(gain control) 도구는 인코더의 PQF (polyphase quadrature filter) 필터뱅크의 이득 제어에 의해 생성된 4개의 주파수 대역 각각에 대해 분리된 시간 영역에서의 이득 제어를 적용하여, 4개의 주파수 대역을 모아 이득 제어 도구의 필터뱅크를 사용하여 시

간 영역의 파형을 복원한다. 이 도구의 입력은 시간영역의 복원된 오디오 신호와 이득 제어 정보이며, 출력은 시간영역의 복원된 오디오 신호이다. 이득 제어 도구가 사용되지 않으면 시간영역의 복원된 오디오 신호가 바로 디코더 출력으로 전달되며, 이 도구는 SSR(scalable sampling rate) 객체에만 사용한다.

스펙트럼 정규화(spectrum normalization) 도구는 복원된 평탄 스펙트럼을 디코더에서 실제의 값으로 변환한다. 스펙트럼 포락선은 LPC 계수, Bark 스케일 포락선, 주기적인 피크 성분 및 이득에 의해 규정된다. 이 도구의 입력은 복원된 평탄 스펙트럼과 LPC 계수, Bark 스케일 포락선, 주기적인 피크 성분 및 이득 정보이며, 출력은 복원된 실제 스펙트럼이다.

Interleaved VQ 도구는 벡터 인덱스를 코드북 테이블 look-up과 스펙트럼의 inverse interleaving 을 사용하여 TwinVQ 디코더에서의 평탄 스텍트럼으로 변환한다. 적응 비트 할당 대신에 인코더에서의 가중 왜곡 척도를 사용하여 양자화 잡음을 최소로 하며, AAC 양자화 도구의 대안으로 사용한다. 이 도구의 입력은 코드 벡터 인덱스 셋이며, 출력은 복원된 평탄 스펙트럼이다.

주파수 선택 스위치(Frequency Selective Switch: FSS) 도구는 AAC 부호화 계층과 TwinVQ 및 CELP 부호화 계층과의 결합(combination)을 제어하기 위하여 사용된다. 이때, TwinVQ 및 CELP 부호화 계층은 scalable 구성에서 기본 계층 부호기(base layer coder)로 사용된다.

업 샘플링(up-sampling) 필터 도구는 scalable 구성에서 기본 계층 부호기로 사용되는 CELP 코어 부호기(core coder)를 AAC 확장 계층(extension layer)의 표본화 주파수로 적응시킨다. 이 도구의 입력은 AAC 확장 계층의 표본화 주파수보다 낮은 표본화 주파수를 갖는 CELP 코어 부호기의 출력이며, 출력은 AAC 확장 계층의 표본화 주파수에 매칭된 up-sampled CELP 코어 부호기 출력이다. 이 때 이 출력은 AAC 확장 계층과 동일한 주파수 및 시간 분해능을 가지며 주파수 영역으로 변환된 것이다.

III. MPEG-4음성 부호화 기술

MPEG-4 audio coding에서는 2 kbps에서 64 kbps까지의 넓은 bit rate를 지원하고 있는데 원하는 bit rate에서 가장 좋은 음질을 얻기 위하여 다음과 같은 3가지 형태의 codec을 제공하고 있다. 낮은 쪽의 bit rate (2 - 6 kbps)를 지원하는 parametric codec, 중간의 bit rate (6 - 24 kbps)를 위한 Code Excited Linear Predictive (CELP) codec, 그리고 높은 쪽의 bit rate (16 kbps 이상)를 위한 Time-to-Frequency (TF) codec이 있다.

ITU-T에서 표준으로 제정된 speech coder로는 6.3/5.3 kbps의 G.723, 8 kbps의 G.729, 16 kbps의 G.728, 32 kbps의 G.721등이 있다. 그러나 MPEG4 speech coder는 8 KHz mode에서는 2-24 kbps, 16 KHz mode에서는 16-64 kbps까지의 다양한 bit rate를 지원할 수 있다. 즉 ITU_T의 speech coder는 8 KHz mode에서 2 kbps와 같은 낮은 bit rate를 지원하지 못하고 16 kHz mode에서 높은 bit rate를 지원하지 못하는 단점이 있다. 더욱이 MPEG-4 speech coder에서는 bit rate scalability, complexity scalability 등의 기능이 제공되는 장점이 있다. 특히 2.0 kbps의 bit rate는 국제표준 중에서 가장 낮은 bit rate이며, 음질 측면에서는 일상적인 대화를 하는데 지장이 없을 정도이고 4.8 kbps의 FS1016표준 보다 더 나은 성능을 가지고 있다.

본 절에서는 이러한 MPEG-4 speech codec 중에서 인터넷 폰이나 디지털 이동통신에서와 같이 낮은 bit rate가 요구되는 응용분야에서 사용될 수 있는 parametric codec의 알고리즘 동작원리에 대하여 상세히 설명한다. Parametric codec은 크게 두개의 블록으로 구성되어 있다. 그 중 하나는 HVXC(Harmonic Vector eXcitation Coding) codec으로써 음성신호를 2 kbps-4 kbps사이로 압축 복원하는 블록이고, 또 하나는 HILN(Harmonic and Individual Line plus Noise) codec으로써 4 kbps이상의 bit rate로 음악과 같은 음성 이외의 신호를 압축 복원하는 블록이다. 이러한 두 가지의 codec이 서로 조합되어 넓은 영역의 신호에 대하여 낮은 bit rate로 신호 압축 복원하는 parametric codec을 구성하고 있다. 그러나 여기서는 음성신호를 압축 복원하는 HVXC codec에 대해서만 encoder 부분과 decoder 부분으로 나누어 설명하고자 한다.

- HVXC Encoder

HVXC codec에서는 LPC 잔여신호를 harmonics와 stochastic vector를 사용하여 코딩하고 있다. 즉 신호가 유성음일 때는 LPC 잔여신호의 spectral envelope를 벡터 양자화하여 코딩하고, 무성음일 때는 Vector excitation coding(VXC) 기법을 사용하여 코딩하기 때문에 HVXC codec이라 불리운다. 아래의 그림 4에 이러한 HVXC 인코더의 블록도를 나타내었다.

HVXC 인코더는 8 kHz로 샘플링된 음성신호를 입력으로 받아서 32 msec (256 샘플)의 프레임 길이로 LPC 분석을 수행하며, 20 msec(160 샘플) 길이의 프레임 간격을 가지고 있으므로 20 msec 마다 한번씩 LPC 분석을 수행하게 된다. LPC 잔여신호는 양자화된 LPC 파라미터를 사용하는 역필터에 입력 음성신호를 통과시켜 얻어지며, pitch와 spectral magnitude를 구하는데 사

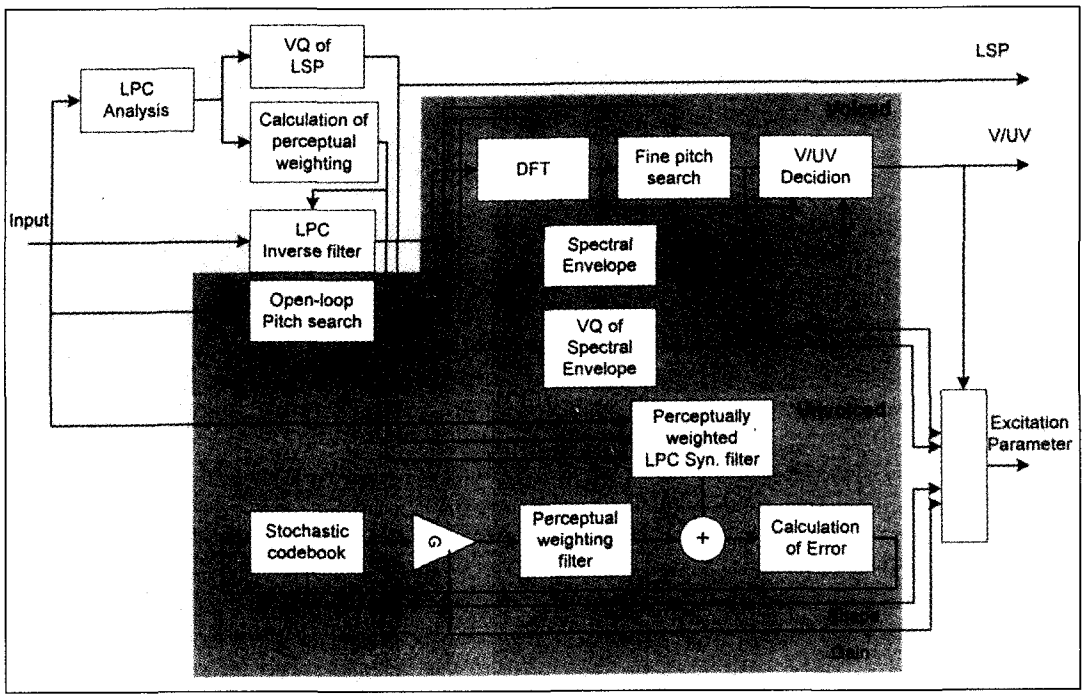


그림 4. HVXC 인코더 블록도

용된다. 여기서 LPC 잔여신호에 대한 spectral envelope는 MBE coder와 거의 유사한 과정으로 계산된다. 유성음 구간에 대해서는 spectral envelope의 가중화 왜곡치를 구하여 벡터 양자화되고, 무성음 구간에 대해서는 vector excitation coding을 위하여 폐루프 검색 과정이 수행된다.

HVXC 인코더의 bit rate는 정상모드 일 때 2.0 kbps이고 enhanced 모드에서는 4.0 kbps 그리고 가변 비트율에서는 평균적으로 1.2-1.7 kbps의 비트율을 가지고 있다. 아래의 표 2와 표 3에 각 비트율에 따른 파라미터의 비트 할당표를 나타내었다.

표 2. HVXC codec의 비트 할당

Parameter	Voiced	Common	Unvoiced
LSF 1		18 bits / 20ms	
LSF2		8 bits / 20ms	
V/UV		2 bits / 20ms	
Pitch	7bits / 20ms		
harmonic 1 shape	4+4bits / 20ms		
harmonic 1 gain	5bits / 20ms		
harmonic 2 split	32bits / 20ms		
VXC 1 shape			6 bits/10ms
VXC 1 gain			4 bits/10ms
VXC 2 shape			5 bits/5ms
VXC 2 gain			5 bits/5ms
Total (1) 2Kbps	40 bits/20msec		40 bits/20msec
Total (1 & 2) 4Kbps	80 bits/20msec		80 bits/20msec

표 3. 가변 비트율일 때의 비트 할당

Mode (V/UV)	Background Noise	UV	MV, V
V/UV LSP Excitation	2 bit 0 0	2 bit 18 bit 8 bit (gain only)	2 bit 18 bit 20 bit (pitch & harmonic spectral parameter)
Total	2bit/20msec 0.1 kbps	28bit/20msec 1.4 kbps	40bit/20msec 2.0 kbps

그리고 알고리즘 지연은 정상 지연모드일 때 56 msec (인코더 46 msec, 디코더 10msec)이고, 낮은 지연 모드일 때 33.5 msec (인코더 26msec, 디코더 7.5msec)이다. 인코더에서 알고리즘 지연은 26 msec와 46 msec 중에서 선택 될 수 있는데 46 msec가 선택되면 피치검출 시 한 프레임의 look-ahead를 하게 된다. 그리고 26 msec가 선택되면 현재 프레임만 가지고 피치 검출을 수행하게 된다. 다음의 그림 5에 지연 모드를 설명하기 위한 프레임 구조를 나타내었다.

• LPC 분석 및 LSP 양자화

매 프레임마다 해밍 창함수가 가해진 신호에 대해 autocorrelation 방법을 사용하여 10차의 LPC 계수가 구해지게 된다. 이렇게 구해진 LPC 계수는 LSP 파라미터로 변환된 후 벡터 양자화 방법을 사용하여 양자화 된다. 양자화 과정에서는 interframe prediction 없는 two-stage VQ 방법과 VQ 와 interframe predictive VQ 결합한 방법의 두가지 방법이 사용되는데 두가지 방법 중 양자화 에러가 적은 쪽의 결과를 취하게 된다. 첫번째 단에서는 아래의 식과 같이 10차의 LSP 계수를 5 bit로 벡터 양자화한다.

$$err1[n] = \sum_{i=0}^{dim-1} \{(lsp[sp+i] - lsp\_tbl[n][m][i])^2 \cdot w[sp+i]\}$$

두번째 단에서는 5차씩 두개로 쪼개서 양자화 하는데 앞의 5차에 대해서는 7 bit, 뒤의 5차에 대해서는 5 bit로 양자화하고 interframe prediction에 대하여 1 bit를 할당하게 된다. Interframe prediction이 없는 경우는 다음

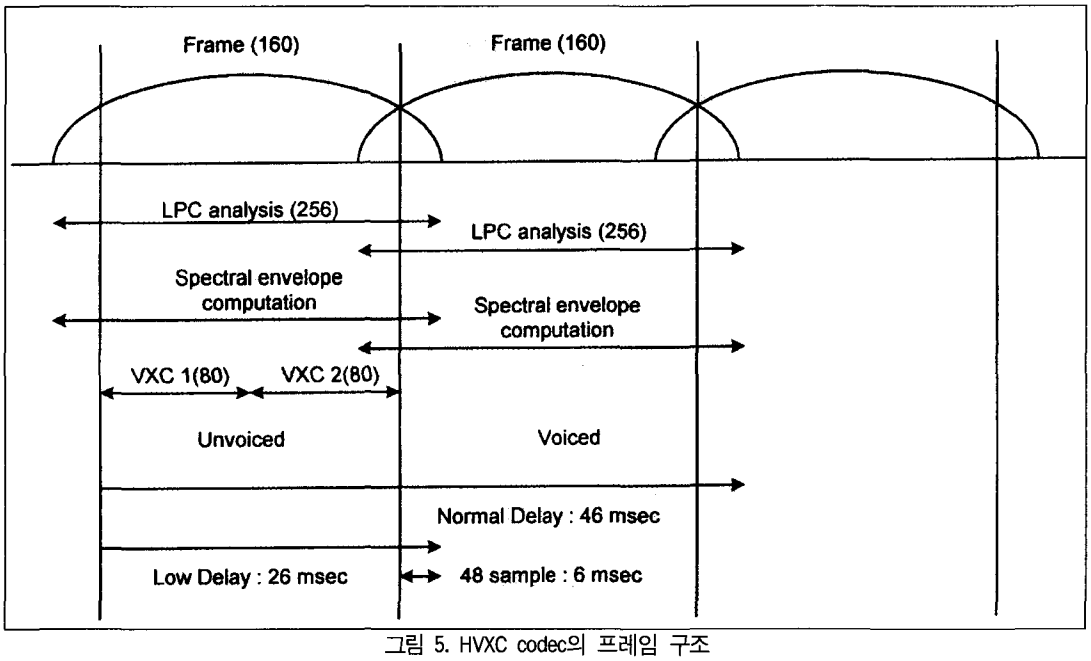


그림 5. HVXC codec의 프레임 구조

과 같은 식을 이용하여 양자화 한다.

$$err2\_total = err2[0] + err2[1]$$

$$err2[n] = \sum_{i=0}^{dim-1} \{(lsp\_res[sp+i] - sign[n] \cdot d\_tbl[n][m][i])^2 \cdot w[sp+i]\}$$

$$lsp\_res[sp+i] = lsp[sp+i] - lsp\_first[sp+i]$$

그리고 interframe prediction을 수행하는 경우는 다음과 같은 식을 이용한다.

$$err2\_total = err2[0] + err2[1]$$

$$err2[n] = \sum_{i=0}^{dim-1} \{(lsp\_pres[sp+i] - sign[n] \cdot pd\_tbl[n][m][i])^2 \cdot w[sp+i]\}$$

$$lsp\_pres[sp+i] = lsp[sp+i] - \{(1 - ratio\_predict) \cdot lsp\_first[sp+i] + ratio\_predict \cdot lsp\_previous[sp+i]\}$$

그리고 enhanced mode일 때는 세번째 단을 동작시켜 앞의 두 단에서 발생된 양자화 에러에 대하여 추가적으로 8 bit를 더 할당하여 10차의 벡터 양자화를 수행하게 된다.

• Pitch Estimation

초기의 피치 값은 LPC 잔여신호의 autocorrelation 값을 계산하여 최대값을 가질 때를 open-loop 피치 값으로 결정한다. 그리고 좀더 정확한 피치 값을 얻기 위하여 피치 트래킹을 수행하게 된다. 인코더에서 낮은 지연모드일 때는 지연시간을 26 msec이내로 맞추기 위하여 현재 프레임과 과거의 프레임 만을 사용하여 피치 트래킹을 수행하고 정상 지연모드일 때는 피치 트래킹에 미래의 한 프레임을 더 사용하게 되어 총 46msec의 지연시간이 발생된다.

다음 과정으로는 유성음, 무성음 구간에 따라 잔여신호를 코딩하기 위하여 유무성음 구간을 결정하게 된다. 유무성음의 판별은 20 msec 단위로 하게 되며 유성음, 무성음, 혼합음 구간으로 나누어진다. 그리고 이러한 판단은 합성된 스펙트럼과 원래 스펙트럼 사이의 유사성, signal power, maximum autocorrelation (LPC residual/residual signal power), zero crossing 등의 파라미터를 이용하여 결정하게 된다.

• Harmonic 크기 추출

Harmonic 크기 추출 과정은 fine pitch 검색 과정과 spectral envelope 검출 과정으로 구성되어 있다. Fine pitch 검색 과정은 앞에서 구해진 open-loop 정수 피치 지연값을 이용하여 원래 신호의 스펙트럼과 합성된 스펙트럼 간의 에러가 최소화되도록 0.25 샘플의 fractional

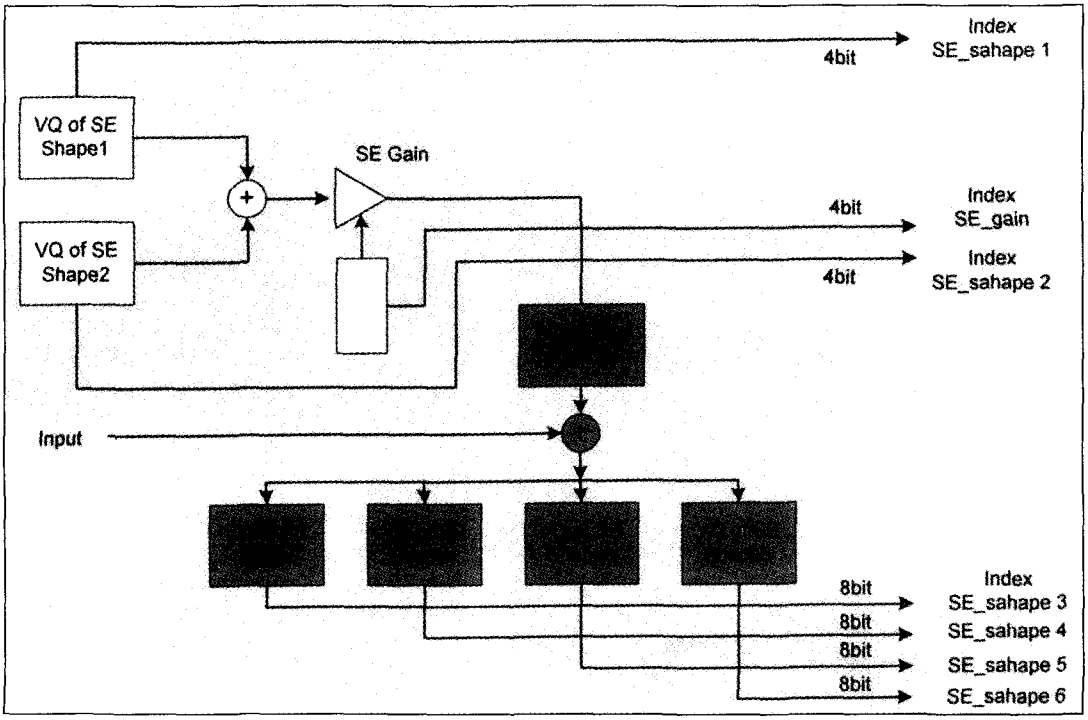


그림 6. Spectral envelope의 벡터 양자화

피치 지연값을 결정하게 된다. 이 과정에서 스펙트럼의 크기도 동시에 구해지게 된다. 원래 신호의 스펙트럼을 구하기 위해서는 LPC 잔여신호의 256 point DFT를 하게 된다. Spectral envelope는 각 harmonic에서의 스펙트럼 크기의 집합으로 볼 수 있다. 이러한 스펙트럼의 크기는 아래의 식과 같이 원래 신호의 스펙트럼 X(j)와 E(j)를 이용하여 최적의 크기 Am을 구하면 얻어지게 된다.

$$\varepsilon_m = \sum_{j=a_m}^{b_m} (|X(j)| - |A_m||E(j)|)^2$$

$$|A_m| = \frac{\sum_{j=a_m}^{b_m} |X(j)||E(j)|}{\sum_{j=a_m}^{b_m} |E(j)|^2}$$

이러한 harmonic vector는 44차원을 갖는 spectral vector를 사용하여 양자화되며 다음의 그림6와 같이 Multi-Stage Vector Quantization (MSVQ) 형태로 양자화 된다.

• 시간영역 인코딩

무성음 구간인 경우에는 다음의 그림7에 보인 것과 같은 VXC (Vector Excitation Coding)알고리즘이 사용된다.

VXC 알고리즘의 과정은 입력 음성신호를 인지가중화한 후 영입력 응답을 빼서 기준이 되는 신호 r(n)을 만든 후, 코드북 벡터 s(n)을 여기 신호로 하는 합성된 신호 syn(n)을 만들어 다음의 식과 같이 두 신호 사이의 에러를 최소화 하는 코드북 벡터와 이득을 찾게 된다.

위와 같은 검색과정은 실제로 다음과 같은 2 단계의 과정을 거쳐서 이루어진다.

- Search s(n) that maximize
- Search g that minimize ,

위에서2 kbps 모드일 때는 매10 msec마다 shape vector s(n)은6 bit으로 양자화되고, 이득은4 bit으로 양자화 된다. 그리고4 kbps 모드일 때는 매5 msec 마다5 bit의 shape과3 bit의 이득 코드북이 추가된다.

• HVXC Decoder

HVXC 디코더는2 kbps와4 kbps의 비트율로 음성신호를 디코딩 한다. 그리고 가변 비트율 모드에서는 평균 비트율이 1.2 - 1.7 kbps이 되도록 디코딩한다. 아래의 그림 8에 디코더의 블록도를 나타내었다.

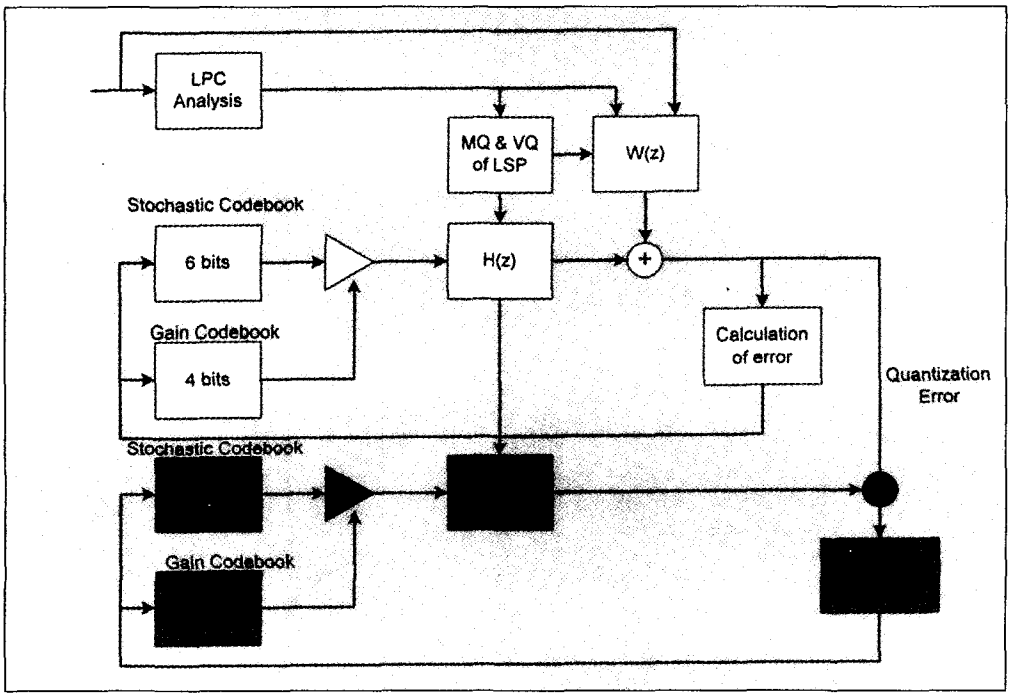


그림 7. 무성음에 대한 Vector excitation coding

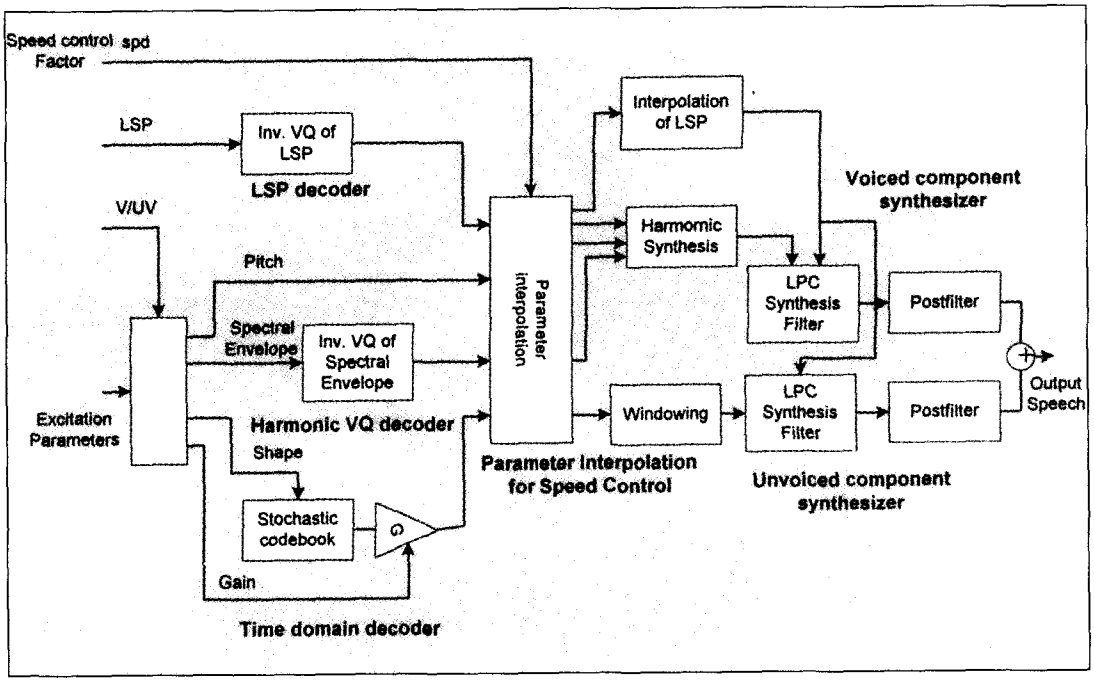


그림 8. HVXC 디코더의 블록도

HVXC 디코더의 기본적인 동작은 우선 수신된 양자화되어 있는 파라미터들을 디코딩한 후 유성음의 프레임인 경우는 sinusoidal 합성(harmonic 합성)에 의해 여기신호를 생성하고, 무성음의 프레임인 경우는 코드북을 참조하여 여기신호를 생성한 후, LPC 합성을 수행하여 음성신호를 합성한다. 마지막으로 합성된 음성신호의 음질을 향상시키기 위하여 후처리 필터링을 수행한다.

• LSP 디코딩

LSP 파라미터의 디코딩 과정은2 kbps의 비트율을 갖는 경우는 2단의 디코딩 과정을 거치게 되고4 kbps의 경우는8 bit의 코드북이 더 추가된다. 여기서 얻어진 LSP 계수는 선형보간 과정을 거쳐서 매 2.5 msec 마다 갱신된다. 그리고 보간된 LSP 값은 합성필터에서 사용하기 위하여 LPC 계수로 변환된다.

• Harmonic VQ 디코딩

2 kbps 모드의 경우는 2단계의 과정을 거쳐서 잔여신호 벡터의 역 양자화 과정을 수행하고, 4 kbps의 경우는 추가적인 역 양자화기가 사용된다. 아래의 그림 9에 spectral envelope의 양자화기를 나타내었다.

Harmonic spectral의 크기에 대한 디코딩은2 kbps의 경우 2단의 shape벡터 양자화기와 스칼라 이득 양자화기를 결합하여 수행한다. 각 shape벡터에 대한 코드북에는4 bit가 할당되어 있고, 이득의 양자화에는5 bit가 할당되어 있다. 그리고 shape 벡터의 차원은 44로 고정되어 있으므로 원래 신호의 spectral 벡터의 차원을 복구하기 위해서는 차원 변환과정을 거쳐야 된다. 그리고 4 kbps의 경우에는 위의 그림 6에서와 같이 shape 3,4,5,6에 대해 split VQ 블록이 더 추가된다.

• 시간영역 디코딩

무성음의 경우에는 수신한 코드북의 인덱스를 이용하여 여기신호를 생성하게 된다. Shape 벡터와 이득은 매10 msec 마다 갱신되며, 2 kbps의 경우 첫째단에서 생성된 출력만 사용하게 되고4 kbps의 경우는 첫째단의 출력에 둘째단의 출력이 더해진다. 여기서 둘째단의 shape과 이득은 매 5msec마다 갱신된다. 그리고 첫째단의 shape의 차원은 80이고6 bit의 shape 코드북과4 bit의 이득 코드북을 사용하고 있으며, 둘째단의 차원은 40이고5 bit의 shape 코드북과3 bit의 이득 코드북을 사용하고 있다.

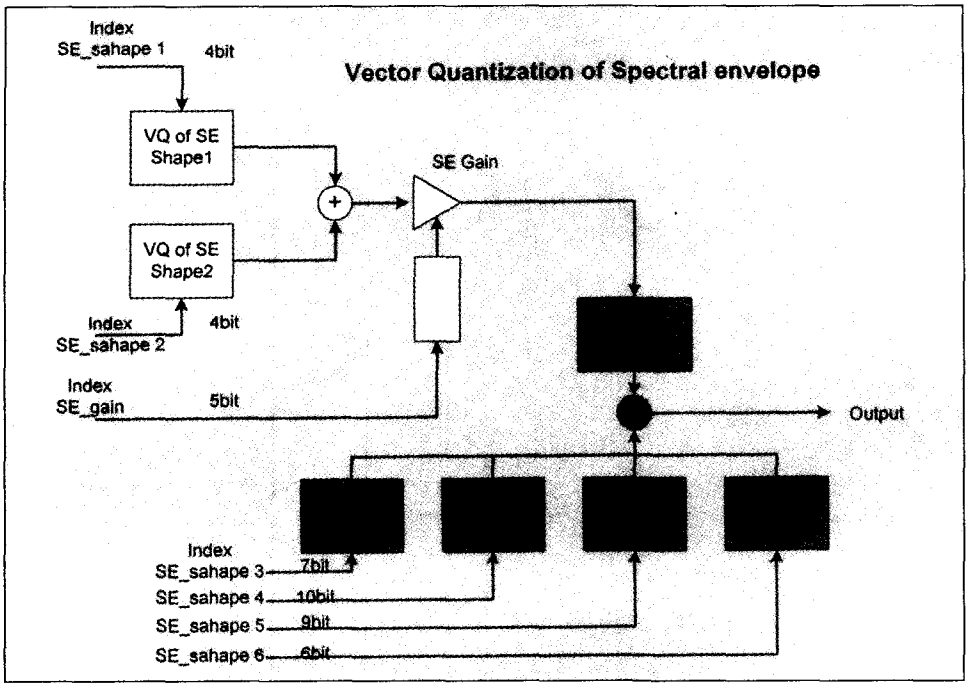


그림 9. Spectral envelope의 벡터 양자화

• 유성음 성분의 합성

유성음 성분의 합성기는 아래의 그림10과 같다.

유성음 성분의 합성기는 harmonic 여기신호 합성, 잡음 성분의 추가, LPC 합성, 그리고 후처리 필터로 구성되어 있다. Harmonic 여기신호를 효율적으로 합성하기 위해서는 harmonic 크기로부터 주기적인 여기신호를 얻으면 된다. 그리고 여기서 얻어진 주기신호에 노이즈 성분을 더해주면 유성음에 대한 여기신호를 얻을 수 있다. 이렇게 얻어진 여기신호는 LPC 합성필터와 후처리 필터링을 거쳐 최종적인 유성음 신호를 얻게된다.

• 무성음 성분의 합성

무성음 성분의 합성기는 LPC 합성필터와 후처리 필터로 구성되어 있다. 무성음 성분의 여기신호는 VXC

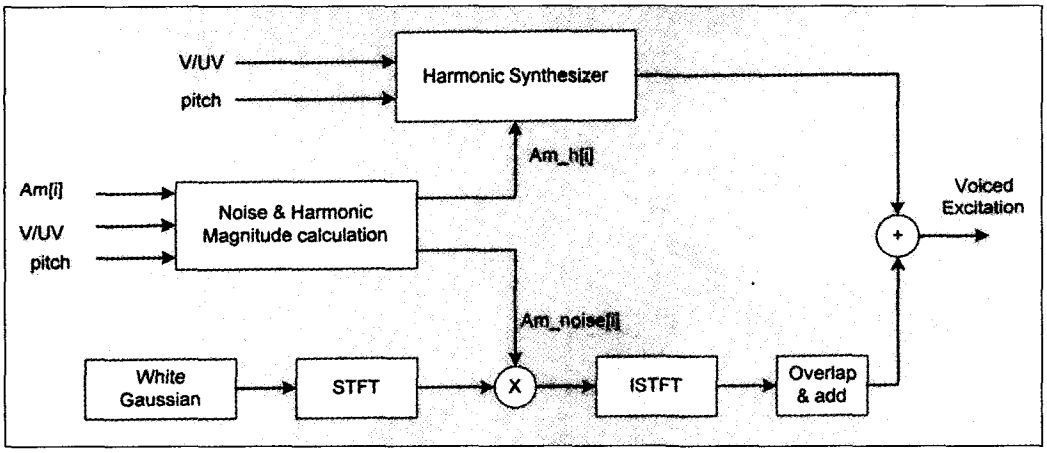


그림 10. 유성음 성분의 합성기

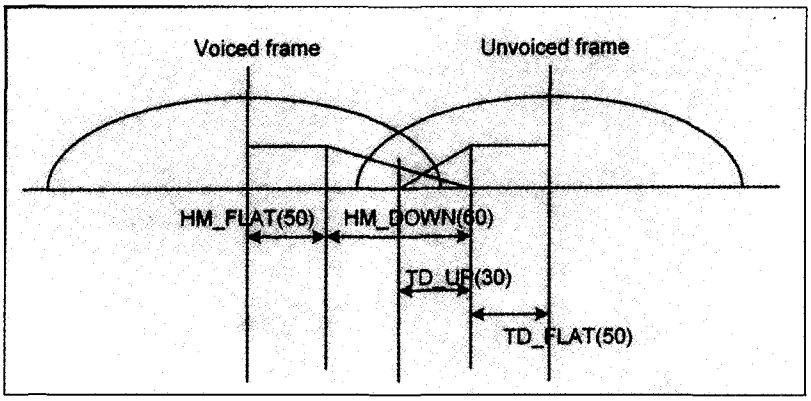


그림 11. 유성음에서 무성음으로 이어지는 경우의 윈도우

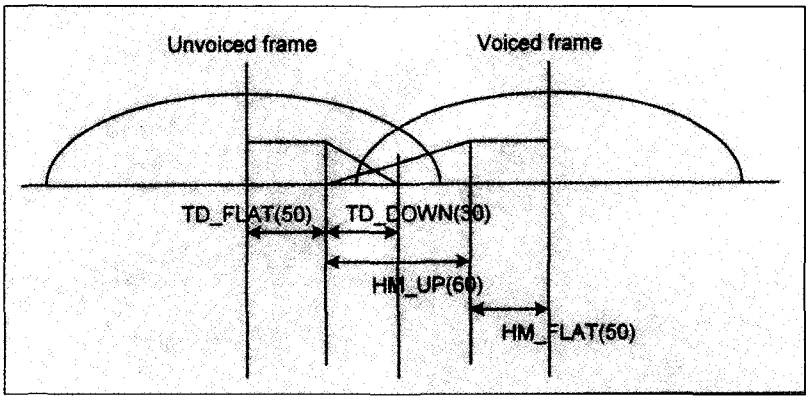


그림 12. 무성음에서 유성음으로 이어지는 경우의 윈도우

방법으로 얻어지고, LPC 합성필터에서 사용되는 계수는 선형보간없이 현재의 프레임에 대한 것만 사용하게 된다. 그리고 무성음 성분의 여기신호와 유성음 성분의 여기신호를 부드럽게 연결하기 위하여 윈도우를 사용한다. 아래의 그림 11에 유성음에서 무성음으로 이어지는 경우와 그림 12에 무성음에서 유성음으로 이어지는 경우를 나타내었다.

위의 그림9와 10에서 TD_UP=30, TD_FLAT=50, HM_DOWN=60, HM_FLAT=50, TD_DOWN=30, HM_UP=60 이다.

이러한 윈도우는 무성음의 프레임인 경우만 사용된다. 즉 무성음 프레임이 유성음 프레임이나, 유무성 혼합음 프레임에 인접한 경우에만 사용된다. 그리고 LPC 합성 필터링은 유성음, 무성음에 상관없이 수행된다. 마지막으로 LPC 합성필터의 출력이 후처리 필터링 과정을 거치게 되면 최종적인 디코더의 출력인 음성신호가 얻어지게 된다

Ⅳ. MPEG-4 TTS 기술

그림 13에 MPEG-4 TTS의 전체 구조를 보였다. 이 그림에서 FAP는 Facial Animation Parameter를 의미한다. FAP에 대해 부언하자면 합성될 음의 종류에 따라 FA 도구에서 적절한 입 모양을 합성하기 위해 필요한 변수들로서 구체적인 정의는MPEG-4 Video Committee Draft에서 찾을 수 있다. 한편 Compositor에서 speech synthesizer로 향하는 화살표로 표시된 User event는 사용자가 정의할 수 있는 기능들로서 발성 속도, 발성자의 성과 나이, 합성음의 크기, trick mode 등의 제어가 가능하다.

현재 이 세상에는 다양한 종류의, 또 여러 품질의

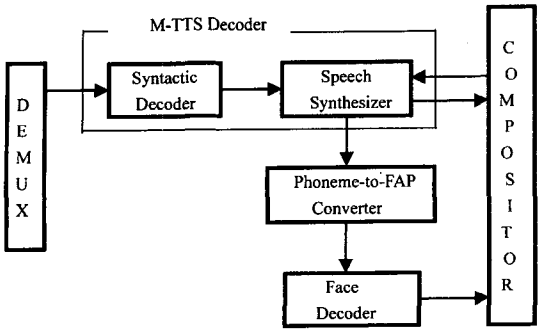


그림13. MPEG-4 TTS 전체 구조

TTS가 존재하고 있으므로 TTS 자체를 표준화 한다는 것은 불가능한 일이다. 따라서 MPEG-4 TTS의 표준화 대상은 그림 13에서 상자로 둘러싼 부분, 즉 디코더 만이다. 바꿔 말하자면 Demux로부터 MPEG-4 TTS 디코더로 전달되는bit stream을 표준화함으로써 기존의 TTS들이 약간의 보완 및 수정 만으로도 MPEG-4 TTS가 제공하는 다양한 기능들을 구현할 수 있도록 하는 것이 MPEG-4 TTS 표준화의 목적이다.

그림 13에서 볼 수 있듯이 MPEG-4 TTS는 기존의 일반적인 TTS가 갖는 문자열로부터 음성을 합성해 내는 기능 외에 부가 기능들이 요구되므로 기본적으로 다음과 같은 인터페이스들이 정의되어야만 기존의 TTS를 MPEG-4 TTS구조로 변경이 가능하다.

- Demux와 TTS 디코더 간의 인터페이스
- TTS 디코더와 TTS간의 인터페이스
- 사용자와 TTS간의 인터페이스
- TTS와 Phoneme-to-FAP converter 간의 인터페이스
- TTS와 오디오 디코더 간의 인터페이스

• MPEG-4 TTS 기능

그림 13에서 알 수 있듯이 MPEG-4 TTS는 문자열로부터 음성을 합성해 내는 기능 외에 원래 발성 문장의 운율을 재현할 수 있어야 하며 합성되는 음성에 적합한 입 모양을FA도구가 생성할 수 있는 정보를 제공하여야 한다. 한편 입술 모양 패턴을 이용하여 Moving Picture 나 Animated Picture를 dubbing할 수 있어야 하며 Animated Picture의 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 "시작", "정지", "재시작", "앞으로", "뒤로" 등의 trick mode기능들도 제공되어야 한다. 따라서 이러한 모든 기능을 제공하기 위하여 MPEG-4 TTS 각각의 인터페이스는 다음과 같이 정의된다.

- Demux와 TTS 디코더 간의 인터페이스: Demux는 입력으로 들어온 MPEG-4 방식의 audiovisual 정보로부터 MPEG-4 TTS에 해당되는 정보의 bit stream을 MPEG-4 TTS 디코더로 보내 준다.
- TTS 디코더와 TTS간의 인터페이스: Demux로부터 입력된 TTS용 bit stream을 받아 TTS 디코더는 다음과 같은 정보를 해당 TTS로 보내 준다.
  1) TTS가 문장 만으로, 또는 FA나 동영상과 연동되어 구동되는가에 대한 정보
  2) 합성되어야 할 문장 내용
  3) 운율정보가 있는 경우 피치, 에너지, 지속시간 등의 운율정보
  4) 입술 모양 패턴 정보가 있는 경우 입술 모양 패턴 정보
  5) Trick mode 가능 여부에 대한 정보
- 사용자와 TTS간의 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 mode에 대한 정보이다. Trick mode 정보는 "시작", "정지", "재시작", "앞으로", "뒤로"에 대한 정보들을 의미한다. MPEG-4 TTS는 발성속도 및 음의 크기를 16단계로 변화 시킬 수 있으며 합성음의 피치 변화폭도 조정이 가능하다. 한편 입술 모양 패턴은 현재 256종류가 사용 가능하며 합성음 화자의 나이도 4가지 중 하나를 선택할 수 있다.
- TTS와 Phoneme-to-FAP converter 간의 인터페이스: MPEG-4 TTS는 Phoneme-to-FAP converter로 최소한 합성되는 음소에 대한 정보를 전달해 주어야 한다. 이 경우 음소보다는 변이음 정보가 보다 정확한 입술 모양의 생성에 유리하므로 MPEG-4 TTS는 일반적으로 사용되는 발음 기호 대신 International Phonetic Alphabet(IPA)를 전달할 수 있도록 디자인 되었다.
- TTS와 오디오 디코더 간의 인터페이스: TTS는 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.

• MPEG-4 TTS의 Bit Stream Syntax

표 4과 표 5에 MPEG-4 TTS 인코딩 및 디코딩에 필요한 bit stream정보를 보였다.

표 4에 나타난 변수들에 대한 설명은 다음과 같다.

표 4. 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)*	
}	

- TTS_Sequence_Start_Code: 'XXXXX'와 같은 16진수로서 TTS Sequence의 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 36개국에 대해 정의되어 있음
- Prosody_Enable: 운율정보가 있으면 '1' 아니면 '0'
- Video_Enable: TTS가 동영상과 연동되면 '1' 아니면 '0'
- Lip_Shape_Enable: 입술 모양 패턴 정보가 있으면 '1' 아니면 '0'
- Trick_Mode_Enable: 사용자가 trick mode를 이용할 수 있으면 '1' 아니면 '0'

표 5에 표시된 변수들에 대한 설명은 다음과 같다.

- TTS_Sentence_start_Code: 'XXXXX'와 같은 16진수로서 합성문장의 시작을 알려줌.
- 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: 16단계의 발성속도 정보
- Length_of_Text: 합성해야 할 전체 텍스트의 길이 정보
- TTS_Text: 임의의 길이를 갖는 입력 텍스트 스트링
- 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 IPA를 표준으로 함)
- Dur_each_Phoneme: 음소의 지속 시간 정보(msec)
- F0_Contour_each_Phoneme: 음소의 피치 contour 정보(음소의 0%, 50%, 100% 에서의 3값으로 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-frame과 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

표5. 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
}	
}	
}	
}	

V. 결론

이상에서 살펴본 바와 같이 현재의 MPEG-4오디오는 자연 음향에서 합성 음향까지, 자연 음성에서 합성 음성까지, 또 사용자가 사용할 수 있는 품질의 음성 및 음향을 선택하여 들을 수 있도록 다양한 음성 및 음향에 대한 압축 복원 기술을 제공한다. 이는 다가올 20세기의 고품질 멀티미디어 통신 및 방송 시대에서 출현이 예측되는 다양한 서비스들을 가능하게 해 주는 첨단 기술들의 집합이라 할 수 있다. 즉, 고품질 멀티미디어 통신 뿐만 아니라 다양한 멀티미디어 컨텐츠를 MPEG-4 기술을 사용하여 보다 쉽게 제작할 수 있게 해주는 것이다. 물론 MPEG-4기술들이 과연 MPEG-1이나 MPEG-2기술처럼 널리 쓰일 것이냐는 아직 논란의 여지는 있다. 그러나 MPEG-4기술의 표준화에 참여하고 기여한 기관들이 AT&T, NTT, BT, FT, NHK, Apple, Sony, Ericsson, Philips, Sony, Yamaha, Matsushita 등의 해외 선진 연구 기관 및 업체 뿐만 아니라 국내의 ETRI, KAIST, 삼성, LG, 현대, 대우 등이었다는 것을 고려한다면 MPEG-4기술의 앞날에 대하여 낙관적인 결론을 내린다 해도 큰 무리는 아니라 판단된다.

마지막으로 첨언하고 싶은 것은 지난 MPEG-1이나 MPEG-2의 표준화 때와는 달리 MPEG-4의 경우, 외국 기관에서 놀랄 정도로 한국의 기여가 무척 컸으며 그 중 많은 부분이 표준화되었다는 점이다. 이는 다가오는 지식 정보화 사회에서의 지적 재산권의 중요성을 새삼 거론치 않더라도 기술 측면에서의 국위 선양에도 크게 공헌하였다고 말할 수 있다. 끝으로 총알없는 전쟁터에서 우리 기술을 표준화하기 위해 많은 고민과 충고를 함께 했던 MPEG Korea 회원들께 심심한 감사를 표하는 동시에 MPEG-7에서도 계속하여 분발해 주시길 부탁드리는 바이다.

참 고 문 헌

1. ISO/IEC 14496-3 (MPEG-4 Audio CD),"edited by MPEG-4 Audio Editing Group, Oct. 1997.

필 자 소 개

한 민 수

- *1979년 2월: 서울대학교 전기공학과 졸업 (학사)*
- *1981년 2월: 서울대학교 대학원 전기공학과 졸업 (석사)*
- *1989년 12월: Univ. of Florida 전기 및 전자공학과 졸업 (박사)*
- *1990. 2월 - 1997년 12월: 한국전자통신연구원 책임연구원*
- *1998년 1월 - 현재: 한국정보통신대학원대학교 공학부 부교수*
- *주관심 분야: 음성 분석, 합성, 인식 및 음성/음향 코딩, 입체음향*

강 경 옥

- *1962년 11월 17일생*
- *1985년 2월: 부산대학교 물리학과 졸업 (이학사)*
- *1988년 2월: 부산대학교 대학원 물리학과 졸업 (이학석사)*
- *1991년 2월 - 현재: 한국전자통신연구원 무선방송기술연구소 선임연구원*
- *주관심 분야: 오디오 부호화 알고리즘, 음향신호처리, 3-D오디오*

변 경 진

- *1962년 2월11일생*
- *1987년 2월: 국민대학교 전자공학과 졸업 (학사)*
- *1998년 3월 - 현재: 한국정보통신대학원대학교 석사과정*
- *1987년 3월 - 현재: 한국전자통신연구원 통신회로연구실 선임연구원*
- *주관심 분야: 음성 부호화 알고리즘, DSP 설계*