저 전송율 음성 부호화 연구 동향
Trends of Low Bit-Rate Speech Coding

최용수*, 강홍구+, 윤대희*
*연세대학교 신호처리 연구센터

Yong-Soo Choi*, Hong-Goo Kang+ and Dae-Hee Youn*
*Center for Signal Processing Research, Yonsei Univ.
+AT&T-Labs Research, Murray Hill, NJ07974, USA
E-mail : *cando@lethe.yonsei.ac.kr, +goo@research.att.com

요 약 문

정보화 시대가 발전함에 따라 음성 통신 및 저장 시스템은 점점 더 우리 생활 깊숙이 자리 잡아 가고 있다. 따라서 급증하는 수요에 보다 더 효과적으로 대처하기 위한 연구가 진행되어 왔다. 그 한가지 예가 기존의 음성 부호화 시스템의 음질을 유지하면서 압축율을 크게 높일 수 있는 부호화 방법에 대한 연구 및 표준화 작업이다. 본 논문에서는 최근 확정된 음성 부호화기 표준안인 US DoD 2.4 kbps MELP, MPEG-4 HVXC, CDMA용 IS-127 EVRC 음성 부호화기에 대해 비교적 자세히 설명하고, 현재 진행 중인 ITU-T 4 kbps 표준안으로 제안된 부호화 방법들의 경향을 살펴본다. 또한 새로운 연구 분야인 인터넷 전화기와 인식-합성 기법을 이용한 아주 낮은 전송율 음성 부호화기에 대한 연구 동향을 소개한다.

1. 서 론

약 20 년간에 걸쳐 음성 부호화 연구는 급격한 발전을 해오고 있다. 이 연구는 현대의 인류 사회에서 필수적인 음성 통신을 효과적으로 수행하려고 하는 노력으로부터 출발되어왔다고 할 수 있다. 단지 4 kHz 에 해당하는 대역에서 '유선전화 음질'(toll quality)라고 부르는 음질을 얻기 위해 128 kbps를 필요했던 음성 부호화 알고리즘은 이제 그 때의 32 분의 1 에 해당하는 4 kbps에서도 같은 성능을 얻을 수 있는 부호화 알고리즘이 제안되는 시기에 이르고 있다. 또한, 정보를 보내는 매개체도 단지 전화선을 이용한 유선 통신에서부터 무선 통신으로 확장되어 왔고, 이제는 음성뿐만 아니라 오디오, 영상 및 문자 등을 통합한 멀티미디어 정보를 전달할 수 있는 네트워크 통합 시대로 확장되고 있는 실정이다. 따라서, 음성 부호화 알고리즘도 이제는 단지 음성만을 위한 응용 분야에서 탈피하여 모든 정보의 통합 환경에도 쉽게 적용 가능하도록 하는 변화 가능한 방법으로의 다양화가 필요한 실정이다. MPEG-4 표준안은 이러한 요구를 충족시키기 위한 중요한 출발점이 되고 있다.

현재 4 kbps 이상의 저 전송율에서 확정된 음성 부호화 표준안은 CELP(Code Excited-Linear Prediction)[1]가 주류를 이루고 있으며, 일명 분석-합성(Analysis-By-Synthesis:ABS) 부호화의 대표적인 방법이다. 최근에는 진행 중인 ITU-T 4 kbps 음성 부호화기 표준안으로 제안된 부호화기의 동향을 살펴보면 CELP 계열과 STC(Sinusoidal Transform Coding)[2]나 MBE(MultiBand Excitation)[3]과 같은 하모닉(harmonic) 계열이 혼재해 있다. 그러나, 4 kbps 이하에서는 CELP 방법은 유성음에 존재하는 주기 성분을 정확히 모델링하지 못하게 됨에 따라 성능이 급격히 저하된다. 이러한 전송율에서 확정된 표준안이나 제안되고 있는 방법들은 LPC(Linear Predictive Coding) 보코더 계열이나 하모닉 계열이 주류를 이루고 있다. MPEG-4 표준안 HVXC(Harmonic Vector eXcitation Coding)[4]는 MBE에 근간을 두고 있으며, 미국방성 2.4 kbps 표준 부호화기 MELP(Mixed Excitation Linear Predictive Vocoder)[5]는 LPC 보코더에 기본을 두고 있다. 또한 WI(Waveform Interpolation)[6] 같이 보다 개선된 하모닉 부호화기도 제안되고 있다.

본 논문은 이러한 시대적 변화의 흐름을 이해하고 앞으로의 연구 방향을 설정하기 위해, [7]에 이어서, 최근에 이루어진 음성 부호화 표준화 안을 검토하고, 현재 이루어지고 있는 표준화 동향을 정리한다. 또한, 국내외에서 이루어지고 있는 새로운 음성 부호화 연구 동향에 대해 간단히 살펴보기로 한다.

본 논문의 구성은 다음과 같다. 2 장에서는 최근의 국제 표준인 MPEG-4 HVXC와 미국방성의 새로운 2.4 kbps 표준안인 MELP와 그리고 IS-127 EVR

C(Enhanced Variable-Rate Codec)[8]에 대해 살펴본다. 3 장에서는 현재 4 kbps ITU-T 표준안에 제안되고 있는 부호화기들의 특징을 살펴보고, 4 장에서는 최근에 새롭게 연구되고 있는 인터넷 전화기의 연구 동향과 음성 인식과 합서 기법을 이용한 아주 낮은 전송율 음성 부호화 경향을 살펴봄으로써 앞으로의 연구 방향을 가늠해 본다. 마지막으로 5 장에서 결론을 맺는다.

2. 최근 국제 부호화 표준안

2.1 US DoD 2.4 kbps MELP Vocoder

1996년 5월, 애틀랜타에서 열린 ICASSP(International Conference on Acoustics, Speech, and Signal Processing)에서 새로운 미국방성 2.4 kbps 부호화기 표준안으로 MELP를 확정 발표하였다. AT&T의 WI, MIT의 AMBE(Advanced MBE) 등을 포함하여 총 7 개의 후보들 중 최종적으로 Texas Instrument의 MELP가 선정되었다. 기존의 표준안 부호화기인 2.4 kbps LPC-10e(FS1015)와 4.8 kbps DoD CELP(FS1016)에 비해 음질 면에서 보다 우수하고 무엇보다도 복잡도가 낮다는 것이 가장 큰 장점이라고 할 수 있다. 그림 1은 2.4 kbps MELP 부호화기이고 표 1은 비트 할당을 보여준다.

2.4 kbps MELP의 주요한 특징은 다음과 같다.

1. 8 kHz sampling, 16bit PCM input signal.
2. Frequency band: 100 ~ 3800 Hz
3. Frame: 22.5 ms (180 samples) + 0.01 %
4. Multiband mixing model:
   Mixed pulse & noise excitation:
5. Transition regions & erratic glottal pulses:
   Periodic or aperiodic pulses:
6. Adaptive spectral enhancement:
   Postfiltering
7. Harsh quality reduction:
   Pulse dispersion filter

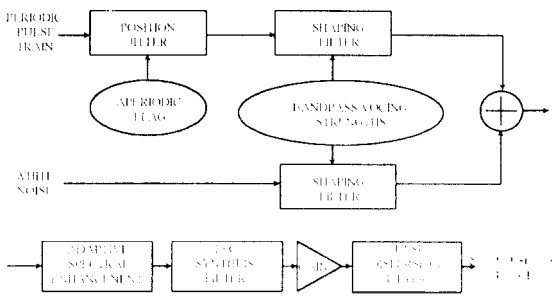

그림 1. 2.4 kbps MELP 부호화기.

표 1. 2.4 kbps MELP 비트 할당.

Parameters	Voiced	Unvoiced
LSF	30	30
Fourier Magnitudes	8	
Gain(2 per frame)	8	8
Pitch, Overall Voicing	7	7
Bandpass Voicing	4	
Aperiodic Flag	1	
Error Protection		13
Sync. bit	1	1
Total Bits / 22.5 ms frame	54	54

2.2 MPEG-4 HVXC

MPEG-4[9][10]는 멀티미디어 분야에서 차세대 주력 표준안을 목표로 하여 이전의 오디오/비디오 부호화 표준과는 구분될 수 있는 새로운 기능들을 정의하고 있으며, 주요 응용 분야로는 고정 혹은 이동 단말기, 데이터 베이스 검색, 통신과 새로운 형태의 대화형 서비스 등이 있다. MPEG-4는 상호작용과 압축율, 그리고 범용 액세스를 허용하는 오디오/비디오 부호화 표준을 제공한다.

MPEG-4 오디오는 새로운 기능의 추가와 압축율의 향상에 목표를 두고 있으며, 지금까지 분리되었던 고품질 오디오 부호화, 음성 부호화와 컴퓨터 음악 등을 하나의 표준에 융합시키려 하고 있다. MPEG-4 오디오 부호화에 의해 저장되거나 전송될 수 있는 신호는 고품질 오디오신호(모노, 스테레오, 멀티채널), 중간급 품질의 오디오신호, 광대역 음성신호, 협대역 음성신호, 이해할 수 있는 음성신호, 음소나 다른 표현에 의해 문자를 음성으로 변환한 합성 음성, 음악 기술 언어(music description language)에 의한 합성 음향 등이다. 또한 MPEG-4 오디오 부호화는 포함하는 기능들은 가변성(scalability), 제한된 시간의 오디오 비트 스트림(limited time bit stream), 피치 조절/시간축 조절(pitch change/time scale change), 편집성(editability), 지연(delay)이다. 따라서, 요구 조건들을 만족시키기 위해서 MPEG-4 오디오는 고도의 구성 능력을 갖는 도구상자과 알고리듬들의 집합체로 구성된다.

MPEG-4 오디오 부호화 표준안은 전송 비트별로 각기 다른 방식들을 선정하였으며 본 논문에서 다루고자하는 저전송율에서는 Sony의 HVXC[4]가 선정되었으며 1998년 11월 국제표준안 완성을 목표로 최종 작업이 진행되고 있다.

HVXC는 하모닉 부호화와 CELP 부호화에 기반을 둔 효율적인 신형예측 분석신호의 부호화 알고리듬이다. 유성음일 경우에는 LPC 잔차신호의 스펙트럼 특성을 백터 양자화하고 무성음일 경우에는 백색 여기신호 부호화를 수행한다.

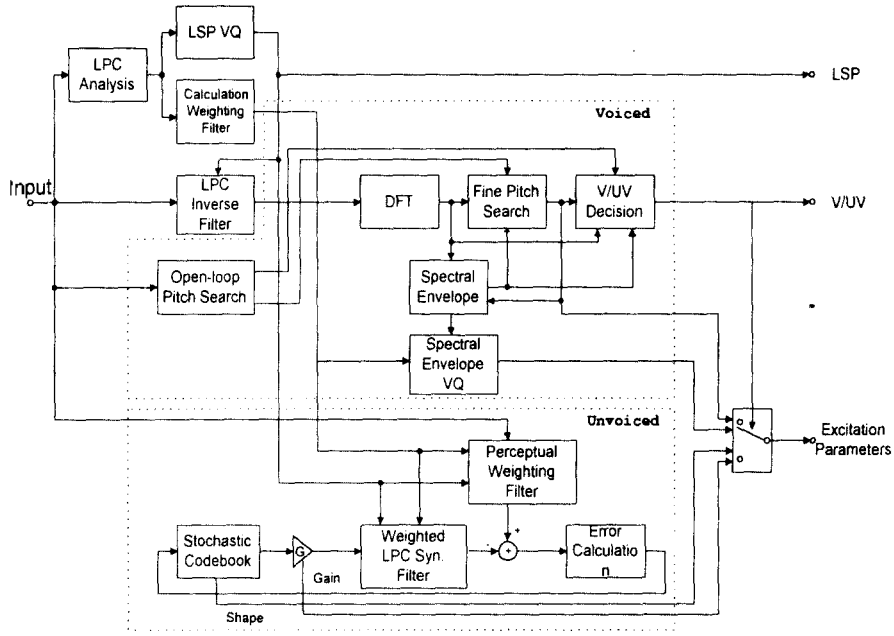

그림 2. HVXC 부호화기.

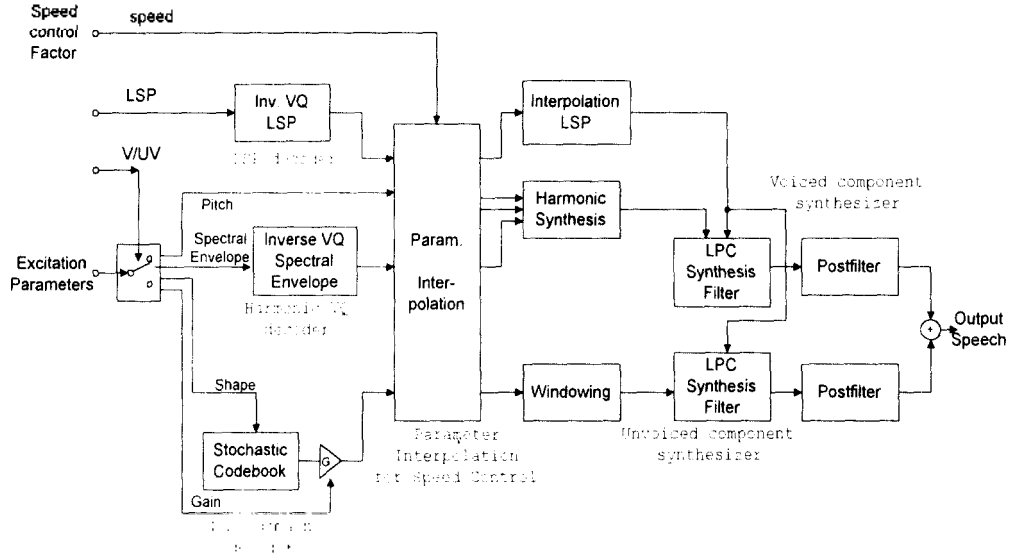

그림 3. HVXC 복호화기.

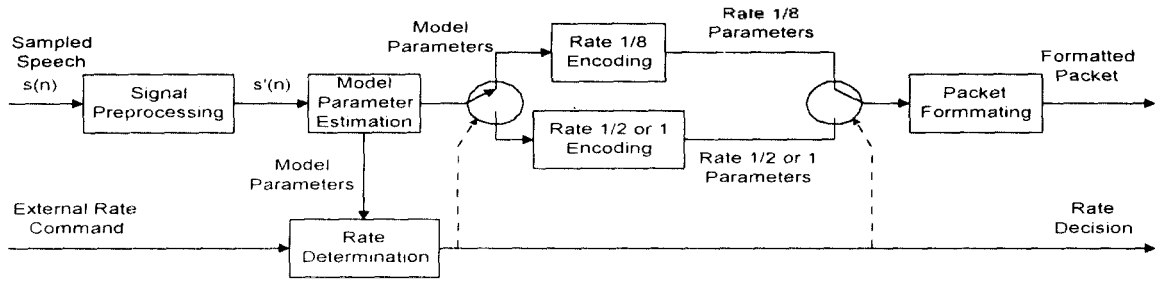

그림 4. EVRC 개요도.

그림 2와 3은 각각 HVXC 부호화기, 복호화기이며 표 2는 2 kbps 비트 할당을 보여준다. HVXC는 다음과 같은 주요한 특징을 갖는다.

1. Frame: 20 ms, 10 ms subframe(unvoiced)
2. Weighted VQ of variable dimension spectral vector
3. Fast harmonic synthesis algorithm by FFT
4. Interpolative coder parameters for speed/pitch control
5. As low as 33.5 ms total algorithmic delay is supported
6. 2~4 kbps scalable mode is supported
7. variable rate coding for less than 2 kbps is supported

표 2. 2 kbps HVXC 비트 할당.

	Voiced	Common	Unvoiced
LSF1		18	
LSF2		8	
V/UV		2	
Pitch	7		
Harmonic1 shape	4+4		
Harmonic1 gain	5		
Harmonic2 split	32		
VXC1 shape			6×2
VXC1 gain			4×2
VXC2 shape			5×4
VXC2 gain			3×4
Total bits/20ms	40		40

2.3 IS-127 EVRC

우리 나라를 비롯해서 CDMA(Code Division Multiple Access) 디지털 셀룰러용 음성 부호화기로 Qualcomm사의 IS-95 가변 전송율(8,4,2,1 kbps) QCELP(Qualcomm Code Excited Linear Prediction)[11]가 상용되고 있다. 그러나 QCELP는 비교적 구현이 간단한 반면에, 음질은 그다지 만족스럽지 못하였으므로 이를 대체해 제안된 것이 EVRC[8]이다. 이 방법의 가장 큰 특징은 선형예측 여기신호의 비선형적인 피치 주기를 선형식으로 변형하여 부호화를 효율적으로 하는 것이다. 이러한 음성의 피치의 선형적 변형이 주관적 음질에는 큰 영향이 없음에 착안한 것이다. EVRC는 QCELP 보다 복잡한 알고리듬을 갖고 있지만 음질 면에서는 상당히 우수하다. 그림 4는 EVRC의 개요도이며, 그 특징은 다음과 같이 요약될 수 있다.

1. Noise suppression preprocessing
2. Generalized analysis-by-synthesis coding
3. Open-loop pitch search & interpolation
4. Time-warping of the original residual
   ⇒ linear pitch contour
   ⇒ easy pitch prediction (adaptive codebook search)
   ⇒ bit-rate reduction
5. Algebraic codebook
6. Adaptive postfiltering
7. High quality speech but high complexity

EVRC는 RCELP(Relaxation CELP)[12][13] 알고리듬에 기반하고 가변 전송율에 따라 적절히 수정되며 CDMA 환경에서 강인한 부호화기이다. RCELP는 일반화된 분석-합성 부호화(generalized analysis-by-synthesis coding)로서, 전통적인 CELP 부호화기와는 달리 목적 신호가 원신호가 아니라 선형적인 피치 컨투어(contour)를 갖도록 원래의 잔차 신호를 시간 축에서 워핑(warping)하여 얻은 수정된 잔차 신호이다. 피치 컨투어는 현재 프레임에서 개회로 검색으로 찾은 피치와 이전 프레임에서 찾은 피치를 선형 보간을 하여 얻는다. 이와 같은 단순화된 피치 컨투어를 사용하면 고정 코드북의 여기 신호와 채널 손상 방지(channel impairment protection)에 비트를 더 할당할 수 있다는 장점이 있다. 이 결과 에러가 없는 채널 환경에서 인지되는 음질의 손상없이 성능을 향상시킨다. 그림 5는 EVRC 부호화기를 모듈별로 상세히 보여준다.

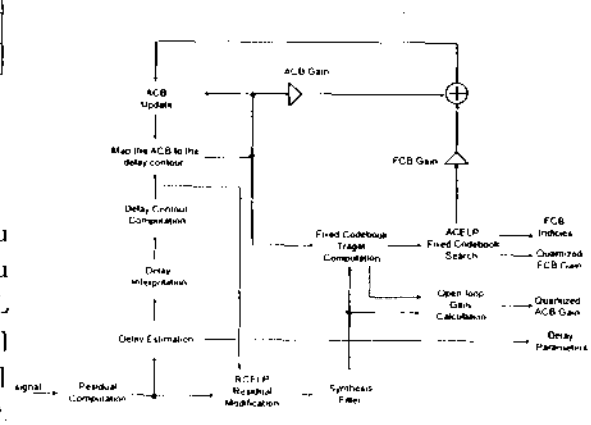

그림 5. EVRC 부호화(encoding) 상세도.

EVRC는 IS-95 Multiplex Option 1에 의해 4가지의 기본적인 전송 패킷(traffic packet) 유형 중에서 3가지를 사용하는데 Rate 1 (171 bits/packet), Rate 1/2 (80 bits/packet) 그리고 Rate 1/8 (16 bits/packet)이다. 외부 rate 명령어가 있으면 부호화기는 blank packet이나 Rate 1외의 다른 rate의 packet(Rate 1/2 maximum)을 만든다. 외부 rate 명령어가 없으면 패킷 유형에 대한 결정은 부호화기 내부의 RDA (Rate Determination Algorithm)에 의해 이루어진

다. 각 패킷 유형에 따른 비트 할당은 표 3과 같다.

표 3. EVRC 비트 할당.

FIELD	Packet Type			
	Rate 1	Rate 1/2	Rate 1/8	Blank
Spectral Transition Indicator	1			
LSP	28	22	8	
Pitch Delay	7	7		
Delta Delay	5			
ACB Gain	9	9		
FCB Shape	105	30		
FCB Gain	15	12		
Frame Energy			8	
(reserved)	1			
TOTAL	171	80	16	0
Bit-Rate (bps)	8550	4000	800	0

3. 현재 진행 중인 표준안 (ITU-T 4 kbps standard)

ITU-T에서는 이동 통신의 수요를 확장하기 위한 목표로 4 kbps 전송율에서 기존의 표준화 시스템과 상응하는 성능을 지니는 표준화 작업을 진행하고 있다. 1995년에 처음 시작된 이 작업은 처음에는 기존의 8 kbps 표준화 작업에서와 같이 32 kbps ADPCM(G.726)을 비교 부호화기로 하여 잡음 환경을 포함한 여러 환경하에서의 요구 사항을 설정하였다[7]. 그러나, 1 년여에 걸친 실험 결과 4 kbps에서 이러한 조건을 만족하는 부호화기를 개발하기는 어렵다는 판단하에 기준 부호화기로서 G.729를 추가하여 주로 잡음 환경하에서의 성능 비교에 사용하기로 하였다. 표 4와 5는 변경된 4 kbps 음성 부호화기의 요구 사항을 정리한 것이다[14].

표 4. 4 kbps 음성 부호화기 요구사항 (experiment 1).

Test conditions(experiment 1)	Reference coder
Clean speech (-26 dBov)	G.726 (ADPCM)
Level dependency (-16 dBov)	G.726 (ADPCM)
Level dependency (-36 dBov)	G.726 (ADPCM)
Tandem capability (x2)	G.726 (ADPCM) -4 tandem
Random frame erasure (3 %)	G.729 (CS-ACELP)
Random bit error rate (0.1 %)	G.729 (CS-ACELP)

표 5. 4 kbps 음성 부호화기 요구사항 (experiment 2).

Test conditions(experiment 2)	Reference coder
Babble noise(30 dB)	G.729 (CS-ACELP)
Car noise(15 dB)	G.729 (CS-ACELP)
Interference talker(20 dB)	G.729 (CS-ACELP)
Interference talker(tandem,20dB)	G.729 (CS-ACELP)

앞에서도 설명했듯이 지금까지 제안된 낮은 전송율 음성 부호화기는 크게 CELP 형의 분석-합성 방법과 하모닉 부호화 방법으로 나눌 수 있다. 보통 4.8 kbps 이상에서는 분석-합성 방법이 주로 적용되어 왔으며 2.4 kbps 이하에서는 하모닉 부호화 방법이 주를 이루어 왔다. 그러나, 4 kbps 에서는 두 방법 중 어떤 방법이 더 나은 지는 아직 결론을 내리지 못하고 있는 실정이다. 이는 1998년 1월 ITU 에 제출된 7 개의 음성 부호화기 중 4 개는 하모닉 방법을 그리고 3 개는 분석-합성 방법을 사용하는 방법이었다는 데에서도 쉽게 찾아볼 수 있다. 본 절에서는 각 방법의 특징을 정리한다.

3.1 분석-합성(ABS) 부호화 방법

분석-합성 방법을 사용하는 4 kbps 부호화기의 기본 알고리듬은 기존의 CELP 알고리듬과 유사한 구조를 가지며, 가장 큰 차이점은 통계적 또는 고정 코드북(stochastic or fixed codebook)을 구성하는 방법에 있다. NTT에서는 PSI(Pitch Synchronous Innovation)-CELP[15]라는 이름으로 그리고, Mitsubishi에서는 PPS(Pitch Position Synchronized)-CELP[16]라는 이름으로 명명하고 있지만, 궁극적인 목표는 고정 코드북을 적응 코드북(adaptive codebook)의 정보를 이용하여 피치 단위로 일치시킨다는 점에서 같다고 할 수 있다. 그림 6은 PSI-CELP의 여기 신호 구성 방법을 보여준다.

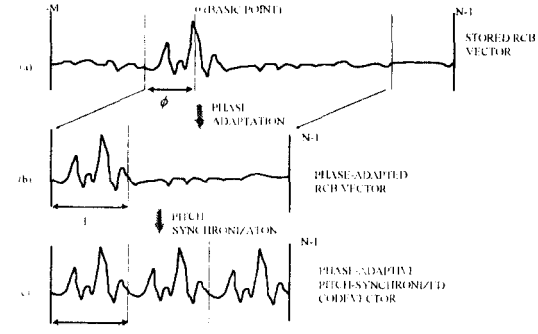

그림 6. PSI-CELP의 여기 신호 구성.

Matsushita에서는 G.729와 유사한 방법을 제안하고 MDP-CELP(Multi-Dispersed-Pulse CELP)[17]라고 명명하였다. 이는 G.729에서 사용하는 대수 코드북(algebraic codebook) 대신 한 개의 위치 펄스를 dispersion 벡터와 컨볼루션을 한 신호를 사용한다는 데 차이가 있다. 그림 7은 dispersed 펄스를 만드는 방법이다.

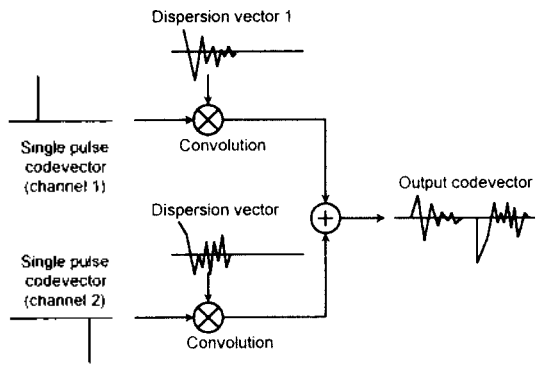

그림 7. dispersed 펄스 여기신호.

3.2 하모닉 부호화 방법

Samsung[18], Voxware, Comsat 등에 의해 제안되고 있는 하모닉 부호화기들은 대부분 STC나 MBE와 유사한 기본 구조를 가지며 각 방법의 차이점은 LPC 잔차 신호의 스펙트럼의 크기와 위상을 부호화하고 합성하는 방법에 있다. 또한 이 방법들은 앞장에서 설명한 MPEG-4 HVXC과도 같은 계열 부호화기라고 할 수 있다.

1998년 1월에 제출된 각 부호화기의 실험을 살펴보면, 잡음 없는 음성(clean speech) 규정을 통과하는 부호화기는 존재하지만, 잡음 환경 하에서는 대부분의 부호화기가 규정을 통과하지 못하고 있는 실정이다. CELP 방법의 경우에는 코드북 훈련(training)의 문제점이 주요 이유가 되며, 하모닉 부호화 방법의 경우에는 피치 탐색 알고리듬과 위상 모델링의 어려움이 그 주요한 이유가 될 수 있다.

4. 새로운 연구 동향

4.1 인터넷 전화(internet phone)

파형 부호화기(예를 들면 LPC)를 기본으로 하는 음성 부호화 알고리듬은 디지털 셀룰라 시스템이 활발히 상용화됨에 따라 연구 분야가 확장되어왔다. 이러한 응용 분야에서 음성 부호화기의 역할은 단지 고정된 몇 개의 비트율에서 고음질의 합성음을 얻는 것이었다. 따라서, 음성 부호화 방법도 이를 해결하기 위한 연구에 집중되어 있었다. 그러나, 최근에 인터넷이 활발히 대중화되고 네트웍을 통해 음성 정보를 전달 혹은 저장할 필요가 발생에 따라 네트웍 시스템을 이해하고 이를 기반으로 하는 음성 부호화기를 효과적으로 응용하기 위한 연구가 활발해 지고 있는 실정이다. 그 중의 한 예가 인터넷 전화기이다.

1995년 2월 13일 이스라엘에서 최초의 인터넷 전화기가 발표되었다. 물론 음질을 매우 좋지 않았고, 지연 시간도 수초에 이르렀으며, 일방 통신 모드로만 동작하였지만 이는 전 세계의 이목을 집중시키기에 충분하였다. 왜냐하면 인터넷 전화기는 국제 전화를 무료로 걸 수 있다는 하나의 시작점이었기 때문이다. 1996년, IMTC(International Multimedia Teleconferencing Consortium)에서는 인터넷을 통한 원격회의를 위한 표준화 작업에 착수하였고, 오늘날 H.323 이라고 하는 인터넷 전화기의 표준안이 만들어지게 되었다. 이것의 표준 음성 부호화기는 ITU G.723.1로서 PSTN(Public Switched Telephone Network)을 이용한 멀티미디어 원격회의 (H.324)의 기본 부호화기이기도 하다.

물론 세계 각국에서 인터넷 전화기를 상용하고 있기는 하지만 기존의 전화기와 같은 실시간 통신에 응용하기 위해서는 아직도 해결해야할 문제점이 산재해 있다. 가장 근본적인 문제는 네트웍 자체에 있으므로 이러한 문제점은 음성 부호화 분야의 문제라고 인식하기보다는 네트웍 분야에서 해결해야 한다는 것이 옳은 판단일 것이다. 그러나, 음성 부호화 분야에서도 이러한 문제를 해결 혹은 축소할 수 있는 방법을 찾을 수 있다. 예를 들면, 패킷 지연 혹은 패킷 손실로 발생하는 음질 저하 문제는 효과적인 가변 전송율 부호화기 혹은 에러 은닉(error concealment) 기법을 통해 축소할 수 있을 것이다.

4.2 아주 낮은 전송율을 갖는 음성 부호화기

음성 부호화 방법을 연구하는 사람들의 궁극적인 목표는 낮은 전송율에서 고음질의 합성음을 만들어내는 것이다. 기존의 LPC 부호화 방법은 2 kbps 내외에서는 우수한 성능을 유지하지만 그 이하의 전송율에서는 음질이 급격히 저하된다는 단점이 있다. 음성 인식 분야와 음성 합성 분야의 연구 결과가 확장되고 개선됨에 따라 이 두 분야의 연구를 종합하여 하나의 음성 부호화기를 만들려고 하는 시도가 이루어지고 있다[18]. 즉, 부호화단에서는 음성 인식기를 사용하여, 인식된 어휘 혹은 음소를 전송하고, 복호화단에서는 음성 합성 방법을 사용함으로써 하나의 음성 부호화 시스템을 구성하는 것이다. 그러나, 100 % 정확한 음성 인식 시스템을 구현하는 것은 불가능하며, 음소 정보만을 보낼 경우에는 음색을 구분할 수 없으므로 음성 합성 시스템도 효과적이지 못하므로 이러한 방법의 음성 부호화 시스템을 구현하기 어렵다는 의견이 나오고 있다. 그렇지만, 현재의 음성 인식 기술에서 화자 종속 시스템의 성능에는 인식율이 90 % 이상이며, 음소와 더불어 그 사람의 음색을 나타낼 수 있는 피치와 에너지 정보를 보내나면 실제 시스템을 개발하기는 그리 어려운 일이 아닐 것이다. 또한 이러한 연구는 음성 합성 분야에도 많은 도움을 줄 수 있다는 장점이 있다. 왜냐하

면, 아직도 자연성(naturalness)라는 면에서 부족한 점이 많은 현재의 음성 합성 분야에 실제 음성 신호에서 추출한 음색 정보를 추가함으로서 이러한 문제를 해결하는데 많은 도움을 줄 수 있기 때문이다. 그림 8은 이러한 시스템의 기본 블록도로서 인식 및 합성 부분에 사용되는 알고리듬에 따라 성능 및 계산량에 차이가 발생한다.

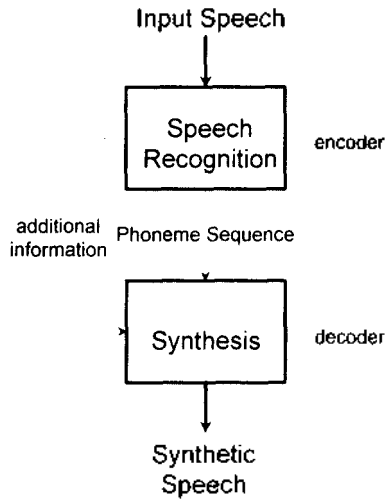

그림 8. 인식-합성 부호화기.

사람의 음성은 평균적으로 1초에 12-15개의 음소로 이루어졌고, 각 음소는 약 6 비트로 부호화 할 수 있다. 그러므로 완벽한 음성 인식-합성(recognition-synthesis)을 가정하면 약 100 bps 전송율은 갖게 된다. 또 운율 정보, 피치, 에너지 등의 부가 정보를 전송하면 수백 bps의 음성부호화기를 생각할 수 있다. 아주 낮은 전송율 음성 부호화기는 기존의 음성 부호화기에서 스펙트럼 정보의 양자화 방법을 개선하거나, 음성 인식-합성 방법[19]을 사용한다.

기존의 음성 부호화기는 고정된 길이를 갖는 프레임을 가지고, 음성신호의 스펙트럼 포락선 성분을 추출하고 합성한다. 포락선 성분은 발성되는 음소에 따라 느린 변화를 갖는다. 그러므로 이의 변화를 관찰하여, 가변길이를 갖는 구간에 대하여 포락선 성분을 양자화 하면 보다 낮은 전송율로 음성을 전송할 수 있다. 효과적인 구간화(segmentation)를 위하여 연속되는 프레임 사이의 포락선 변화량을 관찰하거나[20], 선형 대수를 이용하여 음성을 사건 벡터(event vector)와 사건 함수(event function)로 분리하는 시간 분리법(temporal decomposition)[21]등을 사용한다.

음성 인식-합성 방법은 부호화 과정에서는 음성 인식 기법을, 복호화 과정에서는 음성 합성 기법을 사용한다. 그림 8에서 부호화기는 주로 HMM을 이용하여 음성을 인식하여 부호화 한다.

부호화기에 사용되는 음성 인식 시스템은 화자 독립, 다양한 주변 환경 등에 대하여 일관된 성능이 요구되며, 음성 합성에 적합한 음성 특징 벡터(feature vector)를 사용한다. 복호화기는 부호화기로부터 전달된 음소열(phoneme sequence)을 합성하여 재생음을 만든다. 합성 방법으로는 직접 음소들을 결합(concatenation)시키거나, SOLA(Pitch Synchronous OverLap Add) 기법[22]이나, MLSA(Mel Log Spectrum Approximation)[19] 필터 등을 사용한다. 음성합성 방법으로 부호화기의 인식 오차에 민감하지 않는 합성 방법이 연구되고 있다. 복호화기로 전달되는 부가정보에는 화자의 음성 특징을 나타내는 운율정보, 음성 피치, 에너지, 구간 길 이등이 전송되어 재생음의 음질을 높인다.

음성 인식-합성 방법을 이용한 아주 낮은 전송율 음성 부호화기는 아직 초기 연구 단계로 아직 미진한 부분이 많다. 계산량, 음성 합성 방법, 다국어 환경에서의 화자 적응, 전송 지연 등의 문제를 개선하면 수백 bps에서 양질(intelligibility and naturalness)의 음을 전송할 수 있을 것으로 예상된다.

5. 결 론

본 논문에서는 LPC 보코더 계열의 US DoD 2.4 kbps MELP와 하모닉 계열의 MPEG-4 HVXC, 그리고 CELP 계열의 CDMA용 IS-127 EVRC 음성 부호화기 들을 비교 설명함으로써 최근의 표준안 동향을 살펴 보았고, 현재 진행 중인 ITU-T 4 kbps 표준안으로 제안된 부호화 방법들의 경향을 살펴보았다. 또한 새로운 연구 분야인 인터넷 전화기와 인식-합성 기법을 이용한 부호화에 대한 연구 동향을 소개하였다.

참고 문헌

[1] M. R. Schroeder, B. S. Atal, "Code-Excited Linear Prediction (CELP) High-Quality Speech at Very Low Bit Rates," *IEEE Proc. Int. Conf. Acoust. Speech and Signal Proc.*, pp937-940, 1985.

[2] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. on Acoust. Speech and Signal Proc.*, Vol. ASSP-34, No. 4, pp.744-754, Aug. 1986.

[3] IMBE Vcoder description, DVSI, Jul. 1993.

[4] Technical Description of Sony IPC's Proposals for MPEG-4 Audio and Speech Coding, November 1995.

[5] A. V. McCree, T. P. Barnwell III, " A Mixed Excitation Vocoder LPC Model for Low Bit Rate Speech Cosing," *IEEE Trans. on Acoust. Speech and Signal Proc.*, Vol. 3, No. 4, p

p.242-250. 1995.

[6] W. B. Kleijn, K. K. Paliwal, "Waveform Interpolation for Coding and Synthesis," *Speech Coding and Syntheis*, Elsevier, 1995.

[7] 윤대희, 최용수, "국내외 저전송율 음성 부호화 연구 동향," 제 13회 음성통신 및 신호처리 워크샵, pp.63-72, 1996년 8월.

[8] QUALCOMM Inc., Proposed TIA/EIA/PN-3292 Standard - Enhanced Variable -Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems, Official Ballot Version, April 19, 1996.

[9] Draft of MPEG-4 Requirements, March, 1996.

[10]한국전자통신연구소, "주간 기술동향", TIS-96-21, 749, Jun 5, 1996.

[11] QUALCOMM Inc., Proposed EIA/TIA Interim Standard-Wideband Spread Spectrum Digital Cellular System Dual-Mode Mobile Station - Base Station Compatibility Standard, Submitted to the TIA TR45.5 Subcommittee, 20 Apr. 1992.

[12] W. B. Kleijn, P. kroon, and D. Nahumi, "The RCELP Speech-Coding Algorithm", European Transactions on Telecommunications, Vol 5, Number 5, pp. 573-582, Sept/Oct 1994.

[13] W. B. Kleijn, P. Kroon, L. Cellario, D. Sereno, "A 5.85 kb/s CELP Algorithm for Cellular Applications", *Proc. Int. Conf. Acoust. Speech Sign. Process.*, pp II596-II599, Minnneapolis 1993.

[14] ITU-T Standardization Study Group 16, Subjective Qualification Test Plan for the ITU-T 4-kbit/s Speech Coding Algorithm, Revision 2.1, 30 Sep. 1997.

[15] NTT, High Level Description and the Qualification Test Results of NTT 4-kbit/s Speech Coder, Feb. 1998.

[16] Mitubishi Electric Co., High Level Description of Mitubishi 4-kbit/s Speech Coder and the Qualification Test Results, Feb. 1998.

[17] Matsushita Electric Industirial Co. Ltd., High Level Description of Proposed 4-kbit/s Speech Coder and Qualification Test Results, Feb. 1998.

[18] Samsung Electronics Co. (SEC)/SAIT, High Level Description of Samsung's Spectrally Mixed Excitation (SMX) Coder for ITU-T 4-kbit/s Speech Coding Standard, Feb. 1998.

[19] Keiichi Tokuda, Takashi Masuko, Jun Hiro, Takao Kobayashi and Tadashi Kitamura, "A Very Low Bit Rate Speech Coder Using HMM-Based Speech Recognition/Synthesis Techniques", *IEEE Proc. Int. Conf. Acoust. Speech and Signal Proc.*, pp609-612. 1998.

[20] S. Roucous, R. Schwarts and J Makhoul, "Segment Quantization for Very-Low-Rate Speech Coder", *IEEE Proc. Int. Conf. Acoust. Speech and Signal Proc.*, pp1565-1568, 1982.

[21] Yan Ming Cheg and Douglas O'Shaughnessy, "Shrot-Term Temporal Decomposition and its Properties for Speech Comopression", *IEEE Trans. Signal Processing*, Vol39, No6, pp.1282-1291, June 1991.

[22] W. B. Kleijn, K. K. Paliwal, "Time-Domain and Frequency-Domain Techniques for Prosodic Modification of Speech," *Speech Coding and Syntheis*, Elsevier, 1995.