64kbit/s(7 kHz) Codec을 경유한 연속음성의 인식

°정현열*, Richard M. Stern**
*영남대학교 전자공학과,
**The robotics Institute, Carnegie Mellon University

Recognition of Continuous speech via 64kbit /s(7 kHz) Codec

° Hyun-Yeol chung*, Richard M Stern**
*Dept. of Electronic Eng. Yeungnam University,
** The robotics Institute, Carnegie Mellon University

요약

오디오 혹은 비디오회의, 방송 고품질전화 등의 음성신호의 전송을 위해 마련된 CCITT Recommendation G.722에 의거 Codec을 구성하고 이를 통과한 연속음성을 CMU의 불특정 화자 연속음성인식 시스템인 SPHINX에 입력하여 인식률을 조사 한 후 CODING전의 인식결과와 비교하였다. 이때 CODEC은 크게 네 부분(Trans Quarature Mirror Filter, Encoder, Decoder, Receive QMF)으로 구성하고 입력음성 데이타는 150화자에 의한 1018문장을 훈련용으로, 140문장을 테스트용으로 하였을 때의 단어 인식률을 인식률로 하였다. 또 이때 특징벡터로는 12차 Melcepstrum 계수를 사용하였다. 인식결과 코딩전(close talk Mic를 이용하여 직접입력)의 단어 인식률이 86.7%인데 비해 코딩후의 인식률은 85.6%로 나타나 약 1%의 인식률 저하를 가져와 코딩으로 인한 Error에 비해 비교적 양호한 결과를 얻을 수 있었다. 인식률 저하의 원인으로서는 코딩시의 BER(Bit Error Rate)에 의한 것으로 생각된다.

1. 서론

가입자 상호간의 64kbits/s로 연결하는 디지탈 네트워크의 출현과 함께 1970년도초에 제정된 PCM 코딩법이 아직도 유효한가에 대한 의문이 지속적으로 제기되어 왔으며 디지탈 신호처리기술의 발전과 더불어 64kbits/s bit rate 에 대한 코딩에 있어서도 현재의 대역폭과 양자화 잡음 사이의 타협점은 적당한 가에 대한 의문이 꾸준히 제기되어 왔었다.

현재의 CCITT Recommendation G.711[1]의 PCM 코딩법(A- 혹은 μ-law)은 analog/digital 혼합형네트웍으로 1970년대초에 제정되었다. 이를 완전디지탈 네트에 적용했을 때는 양자화잡음을 최소화할 수 있으나 전송대역은 300-3400Hz로 제한된다. 따라서 이를 고품질 전화, 오디오 혹은 비디오회의, 방송등에 적용하였을 경우 음성의 품질이 크게 떨어지게 된다.

이를 고려하여 CCITT Study Group XVIII는 1980년대초부터 64 kbit/s내에서 대역폭 제한에 따른 일그러짐을 거의 제거할 수 있고 고품질 음성전송을 실현할 수 있는 필요한 대역폭에 대한 논의를 시작하여 1986년 3월 G.72X "7 kHz audio coding within 64 kbit/s" 보고서를 제출하게 되었고 이 보고서는 그해 7월 승인되어 CCITT Recommendation G.722로 되었다.

한편 음성인식분야에서도 이방법을 통해 전송되어진 음성을 인식하였을 때 그결과가 코딩전의 음성을 인식하였을 때와 어느정도 차이가 있을 까에 대한 관심이 집중해 왔으나 이에대한 검토는 아직 행해지지 않고있다.

본 논문에서는 이를 확인하기위해 G.722에 따라 CODEC을 구성하고 이를 통과시켰을 경우와 그렇지 않을 경우의 연속음성인식 결과를 비교검토하기로 한다.

2. CODEC의 구성[2]

CODEC은 G.722에 따라 그림1 과같이 크게 4부분으로 구성되며 그 각각의 특성을 간략하면 다음과 같다.

그림1. CODEC의 구성

2.1 Transmit QMF(Quadrature Mirror Filter)

14bit 16 kHz로 샘플링된 PCM값 xin을 24탭 Quadrature Mirror Filter(QMF)를 통해 두개의 Subband(하측 xl:0-4000Hz, 상측xh: 4000-8000Hz)로 분배한다. 이때 출력 xl, xh는 각각 8 kHz로 샘플링된다.

2.2 SB-ADPCM Encoder

그림2.는 SB-ADPCM Encoder의 블럭도로서 다음과 같은 서브블럭으로 구성된다.

2.2.1 하측 ADPCM Encoder

하측 Subband 입력신호 xl과 추정신호 sl의 차신호 eL을 60-level nonlinear quantizer를 이용하

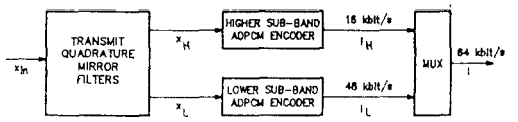


그림2. SB ADPCM Encoder의 블럭도

여 6개의 2진 digit로 변환하여 48 kbit/s신호 il을 만든다. 이때 6개의 bit중 상위 4bit를 adaptation용으로 이용한다.

2.2.2 상측 ADPCM Encoder
하측 subband encoder와 마찬가지로 차신호 eh를 얻은 후 4 level linear quantizer를 이용하여 2개의 2진 digit로 변환하여 16 kbit/s신호 ih를 만든다. 이 2bit는 adaptation용으로도 이용한다.

2.2.3 Multiplexer
하측 및 상측 subband의 출력 il과 ih를 결합하여 64 kbit/s octet형신호 i를 만든다.

2.3 SB ADPCM Decoder
그림3. 은 SB ADPCM Decoder의 블럭도로서 다음과 같은 서브블럭으로 구성된다.

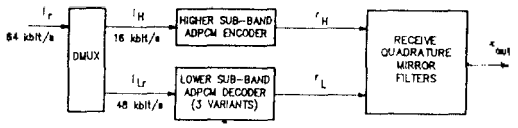


그림3. SB-ADPCM Decoder의 블럭도

2.3.1 Demultiplexer
octet형 신호 i로부터 il ih신호를 재구성한다.

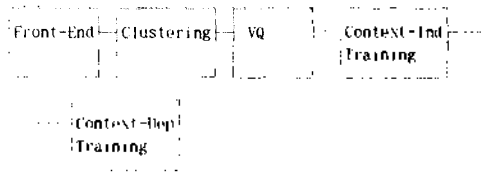
2.3.2 하측 subband ADPCM decoder
2.2.1과 동일한 방법으로 추정신호 sl을 생성하여 양자화된 차신호 dl을 더하여 출력신호 rl을 재구성한다.

2.3.3 상측 subband ADPCM decoder
2.2.2와 동일하게 출력신호 rh를 재구성한다.

2.3.4 Receive QMF
두개의 non-recursive 필터를 이용하여 하상측 subband ADPCM decoder의 출력 rl, rh신호를 16kHz신호로 변환한다.

3. 인식기의 구성
인식기는 CMU의 블록정확자 연속음성인식 시스템인 Sphinx[3]를 이용하였다. Sphinx System은 그림4와 같이 크게 Training 5부분Recognition 3부분으로 구성되며 이하 각 부분의 기능을 개관하기로한다.

Training:



Recognition:

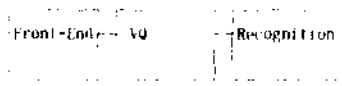


그림4. Sphinx System의 구성

3.1 Training 부
음성신호는 16kHz 로 샘플링되어 FrontEnd부에서 2개의 CODEC을 경유하여 pre emphsis(전달함수: 1 0.97z⁻¹)된후 1 frame 20ms로하여 스펙트럴 분석된다. 이를 LPC분석을 거쳐 12차 LPC cepstrum 계수를 구하여 Melcepstrum계수로 변환한다. Clustring부에서는 LPC cepstrum계수로 부터 1차차분, 2차차분 cepstrum계수 및 power성분을 구하여 이를 특징벡터로 한다. 각각의 특징벡터는 8bit index로 mapping되어 256 vector 코드북이 작성된다. 이때 VQ 알고리즘은 Linde-Buzo Gray[4] 알고리즘을 약간 변형한 것을 이용한다.
Training의 흐름을 그림5 에 보인다. 먼저 레이블링된 TIMIT 문장으로부터 추출된 48개의

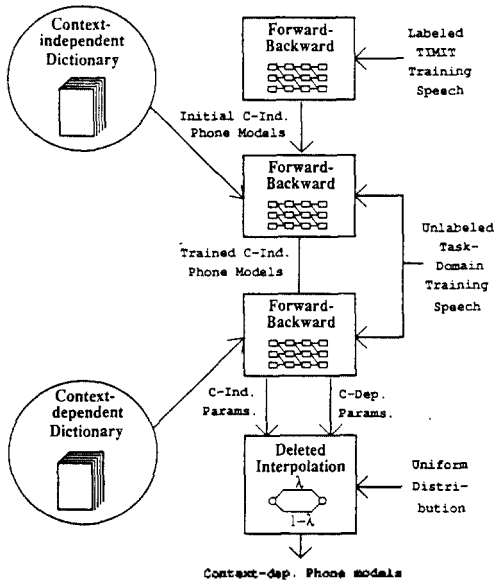


그림5. Training의 흐름도

Context-independent phone 모델을 AN4 데이터 베이스의 1048 문장을 이용하여 Train한후 Context-Dependent phone 모델 (Function-word-dependent model 또는 generalized triphone model등)을 생성한다. 이를 다시 forward-backword 알고리즘을 이용하여 interpolated context-dependent모델을 생성하여 인식용으로 이용한다.

3.2 인식부

Training 에서와 똑같은 처리를 거쳐 VQ된 음성신호는 HMM-Based Viterbi beam search 알고리즘으로 인식한다.

4. 인식 결과

인식에 사용된 음성데이터는 CMU의 AN4데이타 베이스로부터 추출된 150화자에의한 1048문장을 Training용으로, 140문장을 인식용으로 하여 인식한 후 단어인식률을 인식률로 하였다.
이 결과를 표1에 나타낸다. 여기서 Close-talk Mic를 이용하여 직접입력한 경우를 CTLK, CCITT Recommendation G.722에의거 구성한 CODEC를 경유하여 인식하였을 경우를 CODEC으로 표시하기로 한다.

표1 인식결과(%)

	BASELINE		CODEC	
	CLSTK	CRPZM	CLSTK	CRPZM
Correct	88.0	55.1	87.3	51.0
Substitution	11.0	39.9	11.7	43.8
Deletion	0.9	5.0	0.9	5.1
Insertion	1.3	20.7	1.7	21.4
Error	13.3	65.6	14.4	70.3
Word Accuracy	86.7	34.4	85.6	29.7

표1로부터 BASELINE의 단어 인식률이 CLOSE TALK MIC를 사용한 경우 86.7% CODEC을 경유한 한 후의 인식률은 85.6%로 나타나 약 1%의 인식률이 저하했음을 알 수 있다. 그러나 CRPZM MIC를 사용한 경우는 각각 34.4%, 29.7%로 나타나 4%이상의 차가있음을 알수 있다. 이는 CRPZM MIC를 사용한 경우 배경잡음을 많이 포함하고있어 CODEC을 경유함에의해 더욱 잡음이 부가됐음을 의미하며 이로 인해 인식률 저하를 가져왔음을 알 수 있다. 그러나 CLSTK의 경우 코딩으로 인한 인식률 Error가 1% 정도에 그쳐 비교적 양호한 결과를 얻을 수 있다고 할 수 있다. 인식률 저하의 원인으로서는 코딩시의 BER(Bit Error Rate)에 의한 것으로 생각된다.

5. 결론

오디오 혹은 비디오회의, 방송 고품질전화 등의 음성신호의 전송을 위해 마련된 CCITT Recommendation G.722에 의거 Codec을 구성하고 이를 통과한 연속음성을 CMU의 불특정 화자 연속음성인식 시스템인 SPHINX에 입력하여 인식률을 조사 한 후 CODING전의 인식결과와 비교하였다.
인식결과 CLOSE-TALK 마이크를 사용한 경우 CODEC을 경유하기전의 단어 인식률이 86.7%인데 비해 경유후의 인식률은 85.6%로 나타나 약 1%정도의 인식률 저하를 가져와 코딩으로 인한 Error에 비해 비교적 양호한 결과를 얻을 수 있었다. 인식률 저하의 원인으로서는 코딩시의 BER(Bit Error Rate)에 의한 것으로 생각된다. 현재 인식률 제고를 위해 에러의 원인에 대해 검토하고 있다.

6. 참고문헌

[1]CCITT Fascicle III.3 Red Book, Recommendation G 711" Pulse code modulation of voice frequencies," 1984
[2]Xavier Maitre,"7 kHz audio coding within 64 kbis/s," IEEE Journal on selected area in communications, vol.6, No.2, Feb.1988, 283-298
[3]Kai-Fu Lee,"Automatic Speech Recognition: The Development of the Sphinx System," Kluwer Academic Publisher 1989
[4]Y.Linde, A.Buzo and R.M.Gray,"An algorithm for vector quantization,"IEEE Tran. on Communications, vol. COM-28, No.1 Jan. 1980 84-95