네트워크 환경에서 서버용 음성 인식을 위한 MFCC 기반 음성 부호화기 설계

이길호, 윤재삼, 오유리, 김홍국(광주과학기술원)

<차 례>

<Abstract>

A MFCC-based CELP Speech Coder for Server-based Speech Recognition in Network Environments

Gil Ho Lee, Jae Sam Yoon, Yoo Rhee Oh, and Hong Kook Kim

Existing standard speech coders can provide speech communication of high quality while they degrade the performance of speech recognition systems that use the reconstructed speech by the coders. The main cause of the degradation is that the spectral envelope parameters in speech coding are optimized to speech quality rather than to the performance of speech recognition. For example, mel-frequency cepstral coefficient(MFCC) is generally known to provide better speech recognition performance than linear prediction coefficient(LPC) that is a typical parameter set in speech coding. In this paper, we propose a speech coder using MFCC instead of LPC to improve the performance of a server-based speech recognition system in network environments. However, the main drawback of using MFCC is to develop the efficient MFCC quantization with a low-bit rate. First, we explore the interframe correlation of MFCCs, which results in the predictive quantization of MFCC. Second, a safety-net scheme is proposed to make the MFCC-based speech coder robust to channel error. As a result, we propose a 8.7 kbps MFCC-based CELP coder. It is shown from a PESQ test that the proposed speech coder has a comparable speech quality to 8 kbps G.729 while it is shown that the performance of speech recognition using the proposed speech coder is better than that using G.729.

* Keywords: CELP speech coder, MFCC, Predictive VQ, Safety-net VQ, Speech recognition.

1. 서 론

이동 통신 환경 하에서 설계된 서버기반 음성 인식 시스템에 입력되는 음성은 음성 부호화기에 의해 그 음질이 왜곡되어 진다. 이런 주요한 이유 때문에 서버기반 음성 인식 시스템의 성능은 기존의 전화망이나 PC 환경에서의 음성 인식 시스템의 성능에 비해 열화 되는 것이 일반적이다. 이러한 문제를 극복하는 방법으로 음성 인식에 앞서 음성을 복원하는 대신 음성부호화에 사용되는 음성 파라미터를 이용하는 방법이 제안되어 왔다[1]. 즉, 일반적으로 표준 음성 부호화기에서 음성의 spectral envelope을 나타낼 때 사용하는 Linear Prediction Coefficient (LPC)를 이용하여 음성 인식을 수행하는 것이다. 하지만 이 경우의 음성 인식 시스템은 또 다른 음성 특징 파라미터인 Mel-frequency Cepstral Coefficient (MFCC)를 이용한 음성 인식 시스템에 비하여 상대적으로 나쁜 음성 인식률을 보이게 된다[2].

MFCC는 음성 인식 시스템에서 보편적으로 사용되고 있다. 보통 13차 MFCC가 1초당 100회의 비율로 계산되어 음성 인식에 사용된다. 통신 환경에서 음성 인식을 수행하기 위해서는 이 13차 MFCC는 양자화되어 전송된다. 이 때 양자화에 따른 열화를 최소화하는 것이 주된 연구 목표가 되고 있다. MFCC 양자화에 대한 많은 연구가 진행되고 있으며[3]-[5], 이들 연구 중 ETSI에서는 분산 음성 인식 시스템의 표준안을 제정했다. ETSI 표준 양자화기는 13차 MFCC를 분할 벡터 양자화 (Split vector quantization)를 사용하여 44 비트로 양자화한다.

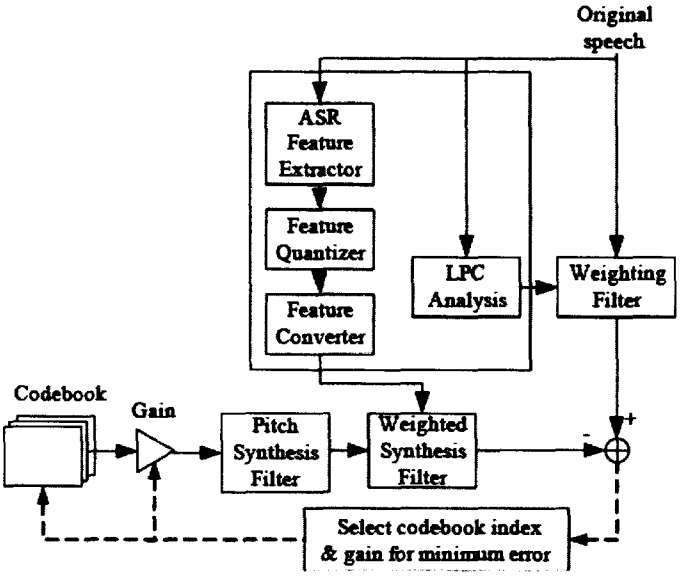
본 논문은 네트워크 환경에서 서버 기반 음성 인식 시스템에 사용되는 고품질 음성 부호화기 설계에 목표를 두고 있다. 즉, 사용자에게 좋은 음질의 음성 통신을 제공하면서 동시에 음성 인식 시스템에 대해서 높은 음성 인식률을 제공하는 음성 부호화기를 설계하는 것이다. 이를 위하여 본 논문은 음성 부호 파라미터로 LPC 대신 MFCC를 이용하는 음성 부호화기를 제안한다. 즉, 클라이언트에서는 입력된 음성 신호에서 MFCC를 추출하고 이 MFCC를 양자화하여 전송을 하며 서버에서는 전송된 MFCC를 이용하여 LPC로 변환 후 음성 재생을 수행한다. 또한 효율적인 MFCC 양자화를 통해 제안된 음성 부호화기의 비트 전송률을 최소화하는데 중점을 두고 있다. 이는 실제 통신 환경에서는 낮은 비트 전송률을 갖는 효율적인 음성 부호화기를 필요로 하기 때문이다. 제안된 음성 부호화기의 성능 평가를 위해 음성 인식은 ETSI DSR 표준 부호화기와, 음성 재생은 8 kbps G.729와 비교하였다.

본 논문은 2장에서 제안된 음성 부호화기의 구조를 설명하고, 3장에서 양자화기 설계의 기준이 되는 분할 벡터 양자화 기법과 양자화에 필요한 비트 수를 줄이기 위해 제안된 safety-net 예측 벡터 양자화 기법을 설명한다. 다음으로 4장에서는 제안된 양자화기를 이용한 음성 부호화기를 이용하여 비잡음 채널과 잡음 채널 상태에서 음질 측정 (PESQ)한 실험 결과와 음성 인식률 평가에 대한 실험 결

과를 제시하고 5장에서는 결론을 맺는다.

2. 제안된 MFCC 기반 음성 부호화기

본 논문은 단순한 음성 부호화기가 아닌 음성 인식 시스템에서도 충분한 성능을 보이는 음성 부호화기를 제안하고 있다. 기존 CELP 음성 부호화기는 음성의 spectral envelope을 LPC를 이용하여 표현하기 때문에 음성으로부터 LPC를 추출한 후 이 LPC를 양자화하여 통신 채널로 전송한다. 반면에 제안된 음성 부호화기는 기존 음성 부호화기에 비해 음성 인식률을 높이기 위하여 음성으로부터 MFCC를 추출한 후 MFCC를 양자화, 전송한다. 제안된 음성 부호화기의 구성은 <그림 1>에 나타나있다.

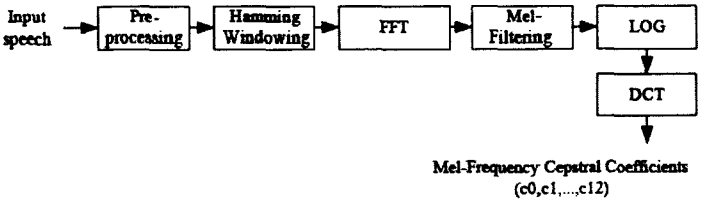


<그림 1> 제안된 음성 부호화기의 구성도

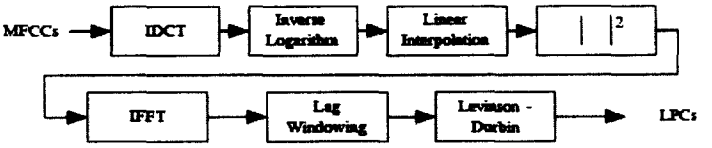
제안된 음성 부호화기는 ITU-T Recommendation G.729를 기본으로 하여 구현하였다[6]. 즉, 10ms 마다 프레임을 구성하며 각 프레임은 long-term prediction과 excitation 모델링을 위해 두 개의 부 프레임으로 나뉘게 된다. 제안된 음성 부호화기가 G.729와 다른 점은 MFCC 추출 과정, MFCC 양자화 과정, 그리고 음성 재생을 위해 LPC를 MFCC로부터 변환하는 과정이다.

<그림 2>는 MFCC 추출 과정을 보여주고 있다. 음성 신호는 140 Hz의 cut-off

주파수를 갖는 고대역 통과 필터를 거쳐 2로 나누어지는 전처리 과정을 하게 된다. 다음으로 G.729에서 사용하는 비대칭 윈도우를 통과하게 된다. 그 후 256개의 샘플로 zero padding을 하고 256 point의 FFT를 하여 magnitude spectrum이 계산된다. 이 magnitude spectrum이 23개의 triangular mel-filterbank와 로그 스케일, 이산 코사인 변환(DCT)의 단계를 거쳐 MFCC가 된다. 이 23개의 MFCC 중 13개가 음성 인식에 사용된다 ($c_0, c_1, \ldots, c_{12}$).



<그림 2> MFCC 추출 과정



<그림 3> MFCC를 LPC로 변환하는 과정

<표 1> 제안된 음성 부호화기의 비트 할당 정보
(β는 MFCC 전송에 소요되는 비트 수이다.)

Parameter	Subframe		Frame
	1	2	
MFCC	-		β
Adaptive codebook index	8	5	13
Pitch parity	1	-	1
Fixed codebook index	13	13	26
Fixed codebook sign	4	4	8
Codebook gain	7	7	14
Total			62+ β

MFCC를 LPC로 변환하는 과정은 <그림 3>에 나타나있다. 먼저 13차 MFCC는 zero padding을 거쳐 23개의 MFCC로 만들어지고 이 값들은 역 DCT와 역 로그 스케일을 거치게 된다. 다음으로 23개의 값들은 256개의 값으로 선형 보간을 하고 각 값들을 제곱하여 역 FFT 과정을 거치면 음성 신호의 autocorrelation을 구할 수 있다. 다음으로 Levinson-Durbin recursion을 이용하면 LPC를 구할 수 있다.

<표 1>은 제안된 음성 부호화기의 비트 할당 정보를 보여주고 있다. 여기서 β는 MFCC 양자화에 필요한 비트 수를 나타내며 다음 장에서 설명될 벡터 양자화 방법에 따라 그 값이 정해지게 된다. 따라서 β에 대한 내용은 다음 장에서 설명한다. 이 밖에 adaptive codebook과 fixed codebook index, codebook gain등 나머지 정보는 G.729의 그것과 동일하다.

3. MFCC 벡터 양자화

서버 기반 방식에서의 음성 인식은 복호화기 측에서 양자화된 MFCC를 이용하여 이루어진다. 따라서 MFCC 양자화기의 성능은 음성 인식률에 밀접한 관련이 있다. 이번 장에서는 두 가지 방법의 벡터 양자화 방법과 이를 사용하여 구현한 MFCC 벡터 양자화기를 설명한다. 먼저 음성 인식 시스템에서 보편적으로 많이 사용되고 있는 분할 벡터 양자화 (SVQ) 방법을 구현, 이를 기준으로 하여 보다 낮은 비트 전송률과 채널 잡음에 대한 고려를 위해 safety-net 예측 벡터 양자화 (Safety-net predictive vector quantization) 방법을 고안, 구현하였다.

3.1. 분할 벡터 양자화 (Split Vector Quantization)

<표 2> MFCC의 분할 벡터 양자화 비트 할당

Sub-vector elements	Codebook size
(c_0)	256
(c_1, c_2)	64
(c_3, c_4)	64
(c_5, c_6)	64
(c_7, c_8)	64
(c_9, c_{10})	64
(c_{11}, c_{12})	64

우리는 먼저 ETSI에서 표준안으로 규정한 분할 벡터 양자화 방법[3]을 이용하

여 MFCC 양자화기를 구현하였다. <표 2>는 13개 MFCC ($c_0, c_1, \ldots, c_{12}$)의 분할된 부벡터와 각 부벡터에 할당된 비트 수를 보여준다. MFCC 양자화기는 LBG 알고리즘 [7]을 이용하여 training 되었으며 이 때 사용된 음성 데이터는 NTT-AT 음성 데이터베이스 중 American, English, Korean 음성 데이터이다[8]. 이처럼 분할 벡터 양자화의 방법으로 양자화기를 구현했을 때 MFCC 양자화에 필요한 총 비트 수는 44 비트이며 이를 <표 1>에 적용하면 제안된 음성 부호화기는 10.6 kbps의 비트 전송률을 갖게 된다.

3.2. Safety-net 예측 벡터 양자화 (Safety-net Predictive Vector Quantization)

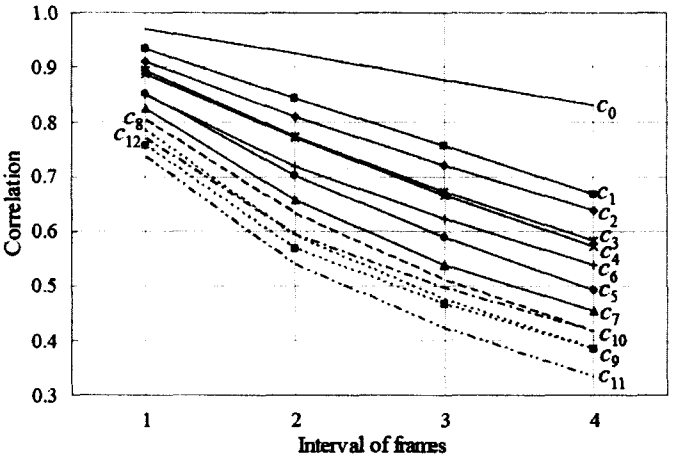
MFCC 양자화에 3.1절에서 설명한 분할 벡터 양자화의 방법을 적용하면 프레임 당 44 비트가 필요하며 결과적으로 이를 적용한 음성 부호화기의 비트 전송률은 10.6 kbps가 된다. 이는 표준 음성 부호화기인 G.729의 비트 전송률이 8 kbps인 것에 비해 2.6 kbps가 높다. 우리는 제안된 음성 부호화기의 비트 전송률을 줄이기 위해 MFCC의 프레임 간 상관관계(correlation)를 이용한 예측 벡터 양자화 방법(PVQ)을 사용하였다[9]. 추가로 MFCC 전송 시 발생할 수 있는 오류의 전파를 최소화하고 음질 및 음성 인식률에 영향을 미치는 프레임 삭제(erasure)의 영향을 완화시킬 수 있는 safety-net 예측 벡터 양자화를 사용하였다[10].

먼저 우리는 PVQ의 사용이 합당한지 판단하기 위해 프레임 간 상관관계를 측정했다. 프레임 간 상관관계는 다음과 같이 정의된다.

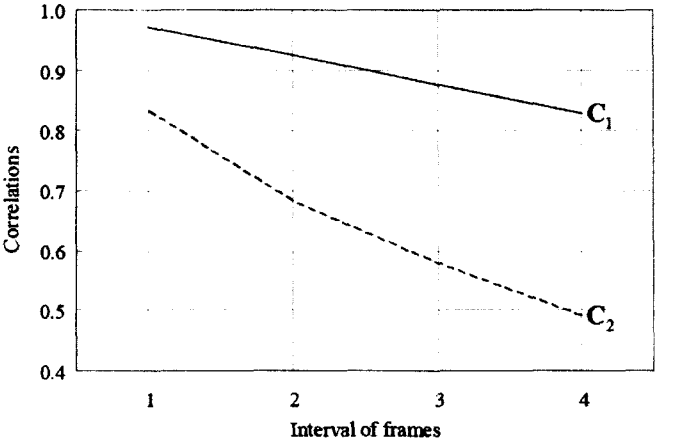
$$corr(i,k) = \frac{\sum_{n=0}^{N-1-k} c_{i,n} c_{i,n+k}}{\sqrt{\sum_{n=0}^{N-1-k} c_{i,n}^2}\sqrt{\sum_{n=0}^{N-1-k} c_{i,n+k}^2}} \quad (1)$$

여기서 i는 quefrency 인덱스이고 k는 프레임간의 간격, N은 전체 프레임 수, $c_{i,n}$은 n번째 프레임의 i번째 MFCC를 나타낸다. <그림 4>는 프레임 간격에 따른 각 MFCC의 프레임 간 상관관계를 보여준다. 측정에 사용된 음성 신호는 남자 2명과 여자 2명이 발성한 자료이며 총 3,200 프레임이 사용되었다. 그림에서 나타나듯이 각 MFCC는 프레임 간격이 1일 때 높은 상관관계를 갖고 있다. 특히 c_0는 MFCC 중 가장 큰 상관관계를 갖고 있으며 그 값은 0.95 이상이다. 또한 나머지 성분에 대하여 벡터를 구성한 뒤 그 벡터의 상관관계를 측정하였다. 즉 13차 MFCC를 1차원 벡터인 C_1과 12차원 벡터인 C_2로 나누었으며 각각의 구성은 $[c_0]$, $[c_1\, c_2 \ldots c_{12}]^T$이다. <그림 5>는 프레임 간격에 따른 각 부벡터의 프레임

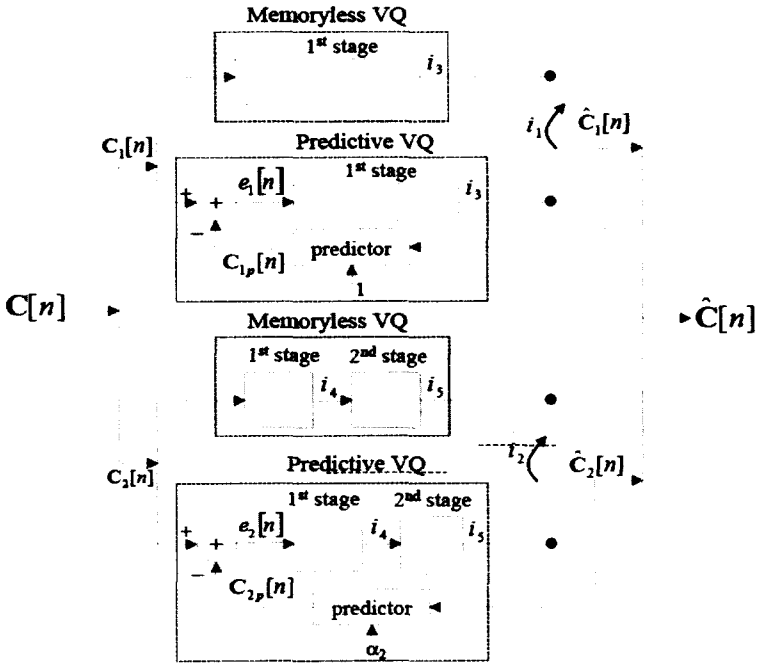
간 상관관계를 보여준다. 1차원 벡터 C_1의 경우 <그림 4>의 c_0과 같은 결과를 보이며 12차원 벡터 C_2의 경우 프레임 간격이 1일 때 0.8 이상의 상관관계를 나타내고 있다. 따라서 우리는 양자화를 하기 전에 MFCC를 두 개의 부벡터로 나누어 양자화를 진행하였다.



<그림 4> 프레임 간격에 따른 각 MFCC의 상관관계



<그림 5> 프레임 간격에 따른 각 부벡터의 상관관계, 여기서
$C_1 = [c_0]$, $C_2 = [c_1\ c_2 \ldots c_{12}]^T$ 임



<그림 6> 예측 VQ와 비기억 VQ가 결합된 제안된 MFCC 양자화기의 구조

다음으로 적용된 예측 VQ와 비기억 (Memoryless) VQ를 결합한 safety-net 예측 벡터 양자화를 소개한다. 여기서 비기억 VQ는 벡터의 각 성분이 이전 벡터의 성분에 독립적으로 양자화되는 방법이며 이는 이전 벡터의 성분에 의존적인 예측 VQ에 대하여 상대적인 성격을 갖고 있다. 이 safety-net 예측 벡터 양자화는 채널 오류의 전파를 줄이는데 효과가 있다[10]. 입력된 음성 신호로부터 추출된 MFCC 벡터는 safety-net 예측 벡터 양자화에서 둘 중 어떤 방식의 VQ (PVQ 또는 비기억 VQ)로 양자화 될 지 결정되어야 한다. 그러므로 우리는 Euclidean 거리 측정을 통해 두 가지 양자화 방법 중 어느 것을 선택할 지 결정하도록 했다. 다시 말해, PVQ 방법을 통해 양자화 된 MFCC 벡터의 측정된 거리가 비기억 VQ의 방법을 통해 양자화 된 MFCC 벡터의 측정된 거리보다 작을 경우 PVQ를 선택하게 되며 반대의 경우 비기억 VQ가 선택된다.

<그림 6>은 제안된 MFCC 양자화기의 구조를 보여준다. n번째 입력 MFCC 벡터는 다음과 같이 두 개의 부벡터로 나눠진다.

$$\mathbf{C}[n] = \begin{bmatrix} C_1[n] \\ C_2[n] \end{bmatrix} = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_{12} \end{bmatrix} \tag{2}$$

여기서 $C_1[n]$과 $C_2[n]$은 앞에서 설명한 대로 각각 1차원 부벡터와 12차원 부벡터이다. 나눠진 각각의 부벡터는 safety-net 예측 벡터 양자화 방식에 의해 양자화된다. 이때 두 가지 양자화 방식 중 어느 방식을 사용할 지는 Euclidean 거리 측정을 통해서 결정하게 된다. PVQ 방식에서 MFCC 벡터의 예측은 이전 양자화된 MFCC 벡터를 통해 이루어진다.

$$C_{ip}[n] = \alpha_i \widehat{C}_i[n-1] \tag{3}$$

여기서 α_i는 과거 첫 번째 프레임의 i번째 부벡터의 예측 계수이다. 특별히 C_2에 대해서는 다중 단계 VQ (multi-stage VQ) 방식을 사용했다. 이는 다중 단계 VQ 방식이 일반적으로 고차원 벡터의 탐색과 훈련에 효과적인 것으로 알려져 있기 때문이다[11].

마지막으로 양자화에 필요한 비트 할당량이 정해져야 한다. <표 3>은 제안된 양자화기의 5개 양자화 인덱스에 대한 비트 할당량을 보여주고 있다. 양자화기의 비트 할당 결정과 training에 사용한 데이터는 NTT-AT 음성 데이터 중 American, English, Korean 데이터를 사용하였다. i_3와 i_4, i_5의 비트 할당을 위하여 우리는 음성 데이터를 두 부분으로 나누었다. 첫 번째 부분은 172,800 프레임으로 구성되었으며 제안된 양자화기의 training에 사용되었고 두 번째 부분은 48,400 프레임으로 구성되었으며 양자화기 성능 평가에 사용되었다. 실제로 MFCC 벡터의 예측 계수 α_i는 PVQ를 위한 비트 할당량과 밀접한 관계가 있다. 그러므로 먼저 최적의 α_i를 찾아야 하며 다음으로 i_3와 i_4, i_5의 적절한 비트 할당량을 찾아야 한다.

<표 3> 제안된 MFCC 벡터 양자화기의 비트 할당 정보

Index	Number of Bits	Function
i_1	1	Prediction selector for C_1
i_2	1	Prediction selector for C_2
i_3	5	VQ index for C_1
i_4	11	First stage VQ index for C_2
i_5	7	Second stage VQ index for C_2
Total	25	

최적의 α_i를 결정하고 적절한 비트 수를 할당하기 위한 측정 방법으로 다음과 같

은 Euclidean 거리 측정을 사용하였다. 측정시 기존 MFCC 양자화의 방법인 SVQ와 Euclidean 거리의 측면에서 비교 후 SVQ에 비해 성능의 저하가 없는 적절한 비트 할당량을 찾는 방법을 사용하였다.

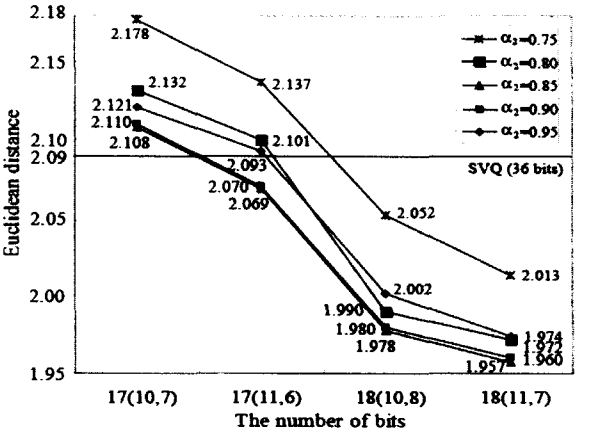
$$D(C, \hat{C}) = \frac{1}{N} \sum_{n=0}^{K-1} \sqrt{\sum_{i=0}^{N-1} (c_{i,n} - \hat{c}_{i,n})^2} \quad (4)$$

여기서 K는 부벡터의 구성 성분 수를 나타내며, C_1에서는 1, C_2에서는 12가 된다. 그리고 N은 전체 프레임 수를 나타내고 $c_{i,n}$과 $\hat{c}_{i,n}$은 각각 n번째 프레임의 양자화 되기 전과 양자화된 후의 부벡터의 i번째 구성 성분을 나타낸다.

<표 4> C_1의 예측 계수 α_1과 할당된 비트 수에 따른 성능 비교

α_1	Safety-net VQ			SVQ
	4 bits	5 bits	6 bits	8 bits
1	0.72	0.36	0.18	0.41
0.95	1.18	0.71	0.42	
0.90	1.54	0.88	0.51	

<표 4>와 <그림 7>은 각각 C_1과 C_2의 비트 할당량과 예측 계수에 따라 식 (4)를 이용해 측정된 Euclidean 거리 비교를 보여준다. 여기서 SVQ는 3.1 장에서 설명한 분할 벡터 양자화 방식이다. SVQ와 비교해 볼 때 부벡터 C_1에서의 예측 계수 α_1은 1, 비트 할당량은 5 이상을 적절한 값으로 고려할 수 있다. 만약 α_1이 1보다 작게 선택된다면 C_1을 위한 비트 할당량은 α_1이 1일 때 보다 더 필요하다. 그 결과 <표 1>에서의 i_3는 α_1을 1로 설정한 상태에서 5비트를 할당하는 것이 적절한 선택으로 판단되었다. 이와 같은 방법으로 C_2에 대해서도 α_2를 0.75와 0.95 사이의 값을 갖게 하면서 18비트를 할당하면 적절한 선택이 될 것이며 실험 결과 α_2를 0.85로 할 때 가장 좋은 결과를 얻을 수 있었다. 게다가 C_2에 할당한 18비트에 대해서 첫 번째 단계(i_4)에서 11비트, 두 번째 단계(i_5)에서 7비트를 할당한 다중 단계 VQ를 사용했을 때 가장 좋은 결과가 나타났다. 그러므로 13차 MFCC를 양자화 하는데 25비트가 소요되며 이는 SVQ 방식과 비교했을 때 19비트의 절감 효과를 보이고 있다. 끝으로 <표 5>에 이번 장에서 설명된 양자화기를 이용한 본 논문의 8.7 kbps 음성 부호화기의 비트 할당 정보를 요약했다.



<그림 7> C_2의 예측 계수 α_2와 할당된 비트 수에 따른 성능 비교. x축의 (a,b)는 각각 다중 VQ의 첫 번째와 두 번째 단계의 비트 수를 나타낸다.

<표 5> 제안된 음성 부호화기의 비트 할당 정보

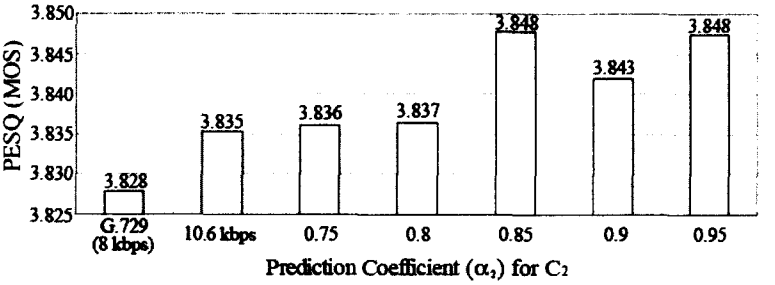
Parameter	Subframe		Frame
	1	2	
MFCC	-		25
Adaptive codebook index	8	5	13
Pitch parity	1	-	1
Fixed codebook index	13	13	26
Fixed codebook sign	4	4	8
Codebook gain	7	7	14
Total			87

4. 성능 평가

4.1. 음질 평가

제안된 음성 부호화기는 perceptual evaluation of speech quality (PESQ) 방법으로 음질이 평가되었다 [12]. 실험에 사용한 음성 데이터는 64개의 한국어 문장 낭독 음성 데이터로 화자는 남자 4명, 여자 4명으로 구성되었다. 각 음성 데이터는 9000 프레임으로 구성되어 있으며 8 kHz로 표본화되었다.

4.1.1. 비잡음 채널



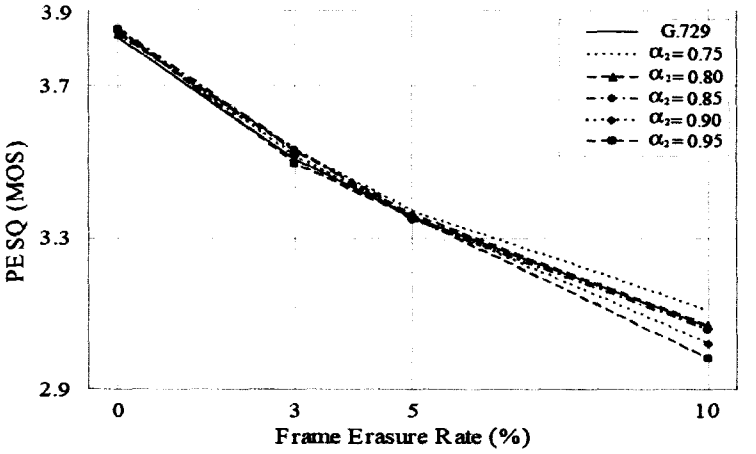
<그림 8> 비잡음에서 C_2의 예측 계수(α_2)에 따른 PESQ 측정

<그림 8>은 비잡음 채널에서 G.729와 분할 벡터 양자화 기법을 이용한 10.6 kbps 음성 부호화기, safety-net 예측 벡터 양자화 기법을 이용한 8.7 kbps 음성 부호화기의 mean opinion score (MOS) 점수를 보여주고 있다. 또한 safety-net 예측 벡터 양자화 기법을 적용한 음성 부호화기에 대해서는 C_2의 예측 계수 α_2의 값에 따른 MOS 변화도 보여주고 있다. 여기서 C_1의 예측 계수 α_1은 3.2장에서 언급한 대로 1로 고정되었다. 그림으로부터 10.6 kbps 음성 부화기는 G.729에 비해 MOS 점수가 약간 높다는 것을 알 수 있다. 그러나 비트 전송률 측면에서 봤을 때 10.6 kbps 음성 부호화기는 8 kbps G.729에 비하여 2.6 kbps 높기 때문에 효율면에서 G.729에 비해 뒤쳐진다고 볼 수 있다. 본 논문의 목표는 기존 표준 음성 부호화기와 비슷한 비트 전송률을 갖는 음성 부호화기를 개발하는 것에 있다. 실험 결과로부터 safety-net 예측 벡터 양자화 기법이 적용된 음성 부호화기가 SVQ 기법이 적용된 음성 부호화기에 비해 비트 전송률이 1.9 kbps 낮지만 MOS 점수에 있어서는 보다 높은 성능을 갖고 있다는 것을 알 수 있다. 또한 α_2가 0.85와 0.95 사이의 값일 때 8 kbps G.729보다 약 0.02 높은 점수를 보이고 있다. 즉, 제안된 음성 부호화기는 α_2의 적절한 설정으로 음질 면에서 G.729에 비해 보다 나은 성능을 갖을 수 있는 것이다. 결론적으로 <그림 6>과 <그림 7>로부터 비잡음 채널 상태에서는 α_2를 0.85로 설정하는 것이 적절하다고 볼 수 있다.

4.1.2. 잡음 채널

실제 통신상에서 음성 부호화기는 채널 잡음을 적절하게 처리하는 것이 주요 사항이다. 8.7 kbps 음성 부호화기를 잡음 채널 상태에서 평가하기 위하여 ITU-T 권고안 G.191[13]을 사용하여 오류 패턴을 만들고 이를 부호화된 비트 스트림에

반영하였다. 오류가 발생하여 프레임이 삭제되었을 때 제안된 음성 부호화기는 정상적으로 받은 가장 최근의 프레임으로부터 파라미터를 보외(extrapolation)하여 음성을 재생하게 된다. 이는 표준 음성 부호화기 G.729가 사용하는 방식과 비슷하다. <그림 9>는 frame erasure rate (FER)와 C_2의 예측 계수 α_2에 따른 MOS 점수를 보여준다. 본 실험에서는 FER을 0%부터 10%까지 변화시켜 측정을 했다. 그림으로부터 α_2의 값이 감소할수록 음성 부호화기가 채널 잡음에 강해짐을 알 수 있다. 따라서 비잡음 채널 상태에의 실험 결과와 본 실험 결과를 종합하여 고려할 때 8.7 kbps 음성 부호화기에서 가장 적절한 C_2의 예측 계수 α_2값으로 0.85를 결정할 수 있다.



<그림 9> 잡음 채널에서 C_2의 예측 계수(α_2)에 따른 PESQ 측정

4.2. 음성 인식률 평가

음성 인식률은 AURORA 4 대용량 데이터베이스[14]를 사용하여 단어 오류율(word error rates)을 측정하는 것으로 평가하였다. 데이터베이스는 training 데이터과 테스트 데이터로 이루어져 있으며 training 데이터는 다시 clean-condition training과 multi-condition training을 위한 데이터로 구분된다. 본 실험에서는 8 kHz로 표본화된 multi-condition training 데이터를 이용하여 음성 인식 모델을 training을 했다. 이때 사용된 training 데이터는 Sennheiser close talking 마이크로 녹음되었으며 6가지 잡음 (car, babble, restaurant, street, airport, train station)이 더해져 있다. AURORA 4의 테스트 데이터는 녹음에 사용된 마이크와 잡음 환경에 따라 14개의 집합으로 구분된다. 본 논문에서는 14개의 집합 중 7개의 집합을 사용했으며 이들 각각은 무잡음과 training 환경과 동일한 6가지 잡음 환경에서 Sennheiser close talking 마이

크를 사용하여 녹음된 330개의 발성 문장들로 구성되어있다.

<표 6> Multi-condition training에서의 AURORA 4 데이터베이스를 이용한 ASR 설정에 따른 단어 오류율 (%) 측정

ASR Configuration / Test Set	Client-based	Client/Server-based	Server-based		
	Baseline ASR	ETSI Quant ASR	MFCC ASR (MFCC-based speech coder)		G.729 ASR
			10.6 kbps	8.7 kbps	8 kbps
Clean (Set1)	18.21	18.92	19.56	19.39	18.87
Car (Set2)	20.34	20.81	22.64	22.98	22.70
Babble (Set3)	29.63	30.79	29.39	30.97	36.52
Restaurant (Set4)	31.70	33.22	32.30	33.03	36.82
Street (Set5)	32.51	32.71	33.10	34.19	36.41
Airport (Set6)	28.21	28.73	30.34	29.93	32.36
Train station (Set7)	32.84	33.79	33.89	35.36	37.18
Average	27.63	28.42	28.75	29.41	31.55

<표 6>은 세 가지 ASR 설정에 따른 단어 오류율을 보여준다. 첫 번째와 두 번째 열은 각각 인식률 측정에 기준이 되는 클라이언트 기반에서의 ASR 시스템(Baseline ASR)의 단어 오류율과 클라이언트/서버 기반인 ETSI Quant ASR 시스템의 단어 오류율을 나타낸다. 또한 서버 기반 음성 인식 시스템은 3장에서 설명한 10.6 kbps와 8.7 kbps의 음성 부호화기, 그리고 G.729를 사용한 ASR 시스템으로 구분, 성능 평가를 진행하였다.

표로부터 ETSI Quant ASR 시스템의 평균 단어 오류율은 기준 ASR 시스템에 비해 약 2.1% 증가하는 것으로 나타났다. 이 단어 오류율의 증가는 ETSI 양자화기의 압축 손실에 의한 것이다. 반면에 MFCC ASR 시스템의 평균 단어 오류율은 기준 ASR 시스템에 비해 10.6 kbps 음성 부호화기 사용시 약 4.1%, 8.7 kbps 음성 부호화기 사용시 약 6.4% 증가하고 있다. 그러나 G.729 ASR 시스템의 평균 단어 오류율은 기준 ASR 시스템에 비해 약 14.2% 증가하며, 이는 10.6 kbps 음성 부호화기에 비해 약 8.9%, 8.7 kbps 음성 부호화기에 비해 약 6.8% 증가한 값이다. 또한 MFCC ASR 시스템은 Set2부터 Set7까지의 잡음에 대해 G.729 ASR 시스템에 비하여 낮은 단어 오류율을 보이고 있다. 특별히 Babble 잡음에서는 10.6 kbps 음성 부호화기에서 약 20%, 8.7 kbps 음성 부호화기에서 약 15% 낮은 단어 오류율을 보이고 있다. 따라서 제안된 음성 부호화기가 기존의 음성 부호화기보다 높은 음성 인식률을 보이고 있다.

5. 결 론

본 논문에서는 MFCC를 사용하여 네트워크 환경에서 서버 기반 음성 인식을 위한 CELP 음성 부호화기를 제안하였다. 즉, 음성의 재생과 인식을 동시에 하면서 그 성능에 있어서도 기존 개별 부호화기에 비하여 성능 저하를 보이지 않는 음성 부호화기를 설계하는 것이다. 기존 음성 인식 시스템에 비하여 인식률 저하를 방지하기 위해 제안된 음성 부호화기는 음성의 spectral envelope을 기존 음성 부호화기와는 다르게 LPC가 아닌 MFCC를 이용하여 나타내었다. 또한 낮은 비트 전송률을 갖으면서도 성능이 기존 방식과 뒤떨어지지 않는 MFCC 양자화기를 설계하고 제안된 음성 부호화기가 채널 잡음에 강인하도록 하기 위해 예측 벡터 양자화와 비기억 벡터 양자화를 결합시킨 safety-net 예측 벡터 양자화를 제안하였다. 그 결과 MFCC 양자화에 프레임 당 25 bits가 소요되며, 제안된 MFCC 양자화기를 이용한 본 논문의 음성 부호화기는 8.7 kbps의 비트 전송률을 갖게 되었다. PESQ 실험으로부터 제안된 음성 부호화기는 비잡음 채널과 잡음 채널 상태에서 8 kbps G.729와 비교했을 때 다소 좋은 음질 성능을 보여주었다. 또한 음성 인식 실험에서도 제안된 음성 부호화기는 음성 인식 파라미터인 MFCC를 추출, 전송하여 서버 측에서 이를 직접 사용하므로 기존 음성 인식 시스템과 비슷한 성능을 보이며 G.729를 이용한 음성 인식 시스템보다 좋은 결과가 나왔다. 본 연구에서 지향하는 최종 목표는 G.729와 동등한 전송률을 갖으면서 음성 인식률을 높이는 것이다. 따라서 향후 현재의 8.7 kbps의 전송률을 현재 나타나고 있는 음성 재생, 음성 인식의 저하가 없이 0.7 kbps를 줄여 8 kbps의 음성 부호화기를 연구할 예정이다.

감사의 글

이 논문은 2004년도 한국학술진흥재단의 지원에 의하여 연구되었음 (KRF-2004-003-D00271).

참 고 문 헌

[1] S. H. Chol, H. K. Kim and H. S. Lee, "Speech recognition using quantized LSP parameters and their transformations in digital communication", *Speech Communication*, vol. 30, pp. 223-233, 2000.

[2] J. He, L. Liu and G. Palm, "On the use of residual cepstrum in speech recognition", *Proc. of ICASSP*, pp. 5-8, 1996.

[3] ETSI Standard ES 201 108 v1.1.3, *Speech processing, transmission and quality aspects; Distributed speech recognition; Front-end feature extraction algorithm; Compression*

algorithm, 2003.

[4] V. Digalakis, L. Neumeyer and M. Perakakis, "Quantization of cepstral parameters for speech recognition over the world wide web", *in Proc. of ICASSP*, pp. 989-992, 1998.

[5] Q. Zhu and A. Alwan, "An efficient and scalable 2D DCT-based feature coding scheme for remote speech recognition", *in Proc. of ICASSP*, pp. 7-11, 2001.

[6] ITU-T Recommendation G.729, *Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP)*, 1996.

[7] Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design", *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84-95, 1980.

[8] NTT-AT, *Multi-lingual speech database for telephonemetry*, 1994.

[9] G. N. Ramaswamy and P. S. Gopalakrishnan, "Compression of acoustic features for speech recognition in network environments", *in Proc. of ICASSP*, pp. 977-980, 1998.

[10] T. Eriksson, J. Linden and J. Skoglund, "Interframe LSF quantization for noisy channels", *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 495-509, 1999.

[11] B. H. Juang and A. H. Gray, "Multiple stage vector quantization for speech coding," *in Proc. of ICASSP*, pp. 597-600, 1982.

[12] ITU-T Recommendation P.862, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, 2001.

[13] ITU-T Recommendation G.191, *Software tools for speech and audio coding standardization*, 2000.

[14] G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task", ETSI STQ Aurora DSR Working Group, 2002.

접수일자 : 2005년 5월 14일
게재결정 : 2005년 6월 14일

▶ 이길호 (Gil Ho Lee)
주소: 500-712 광주광역시 북구 오룡동 1번지 광주과학기술원
소속: 광주과학기술원 정보통신공학과
전화: 062) 970-3121
E-mail: ghlee@gist.ac.kr

▶ 윤재삼 (Jae Sam Yoon)
주소: 500-712 광주광역시 북구 오룡동 1번지 광주과학기술원
소속: 광주과학기술원 정보통신공학과
전화: 062) 970-3121
E-mail: jsyoon@gist.ac.kr

▶ 오유리 (Yoo Rhee Oh)
주소: 500-712 광주광역시 북구 오룡동 1번지 광주과학기술원
소속: 광주과학기술원 정보통신공학과
전화: 062) 970-3121
E-mail: yroh@gist.ac.kr

▶ 김홍국 (Hong Kook Kim)
주소: 500-712 광주광역시 북구 오룡동 1번지 광주과학기술원
소속: 광주과학기술원 정보통신공학과
전화: 062) 970-2228
E-mail: hongkook@gist.ac.kr