피치예측과 점진적 복원 기법을 이용한 EVRC 음질개선

EVRC Speech Quality Enhancement Using Pitch Prediction and Gradual Increase of the Decoded Speech

민 병 준*, 김 재 원*
(Byung Jun Min*, Jae Weon Kim*)

요 약

SK Telecom에서 현재 서비스중인 EVRC 보코더는 유선전화 수준의 음질을 제공하는 우수한 음성 부호화기이나, 약전계에서 급격한 음질 저하를 보인다.

본 논문에서는 실제 서비스 상황에서 발생하는 EVRC 보코더의 음질 저하 현상 및 그 원인을 분석하였고, 해결책으로 피치 예측과 점진적 복원 기법을 제안하였다. 다양한 전파환경에 대한 음질 평가방법으로 선호도 실험을 수행하였고, 제안한 방법이 효과적임을 확인하였다.

ABSTRACT

The EVRC vocoder is a toll quality coder, but it shows significant degradation of the quality in weak RF environment. In this paper, the speech quality degradation phenomenon of the EVRC is analyzed, and two methods are proposed as the solution - the pitch prediction and the gradual increase. The preference tests for various RF environment are performed for speech quality assessments and both the methods show better performance.

I. 서 론

디지털 이동통신 시스템에서 통화 품질에 가장 큰 영향을 미치는 부분은 음성 신호의 부호화 기능을 담당하는 보코더 부분이다. 보코더 기술은 음성 신호처리기술 및 디지털 이동통신기술의 발달에 따라 비약적으로 발전해 왔으며, 앞으로도 더욱 뛰어난 성능의 보코더가 개발되어 더 좋은 서비스가 가능하게 될 것이다.

현재, 국내에서 셀룰러 사업자들은 8Kbps의 최대 전송률을 사용하는 QCELP(Qualcomm Code Excited Linear Prediction)와 EVRC(Enhanced Variable Rate Codec)의 두 가지 보코더를 사용하고 있고, PCS 사업자들은 13Kbps QCELP를 사용하고 있다. 이 중 EVRC는 1998년 초부터 서비스 되고 있으며[1], 최대 8Kbps의 전송률을 사용하면서도, 13Kbps QCELP와 같이 toll-quality 수준의 우수한 음질을 보인다.

EVRC는 Lucent, Nokia, Morotola, Qualcomm등 여러 회사들이 모여 제안한 보코딩 알고리즘으로, 8, 4,1Kbps의 전송률을 사용한다. 기본적으로 Lucent 에서 개발한 RCELP(Relaxed Code Excited Linear Prediction) 방식을 따르며[2][3], 고정 코드북 부분은 Nokia, University of Sherbrooke에서 개발한 ACELP(Algebraic Code Excited Linear Prediction) 방식으로 되어있다[4]. 전송률 결정 알고리즘은 13Kbps QCELP에서 쓰이는 것과 동일하며, 잡음 감쇄 기능은 Motorola에서 제안한 방식이다. 그림 1에 EVRC 상위구조의 인코딩 과정을 보였고, 그림 2에 인코딩 블록도를 보였다.

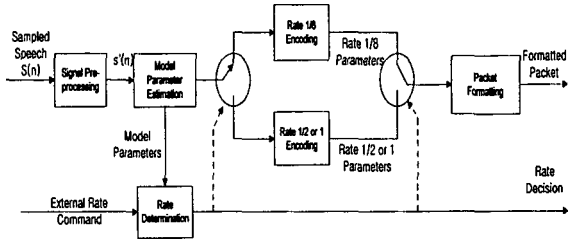


그림 1. 상위 구조의 인코딩 과정
Fig. 1. Speech encoder top-level diagram.

대부분의 디지털 이동통신용 음성 부호화기와 마찬가지로 EVRC도 3%이내의 FER(Frame Error Rate)을 기준으로 개발되었다[5]. 이처럼 정상적인 상황에서는 EVRC는 우수한 음질을 보이지만, 전파 환경이 나쁜 상황에서는 FER이 10% 이상으로 증가하게 되고, 이 때 음질열화가

* SK 텔레콤 중앙연구원
접수일자 : 1999년 3월 5일

심각하게 제기되었다. 본 연구에서는 이처럼 나쁜 전파환경(약전계)에서의 EVRC 음질 열화를 막기위한 두가지 방법이 제안되었고, 실험을 통해 음질 열화 정도를 완화시킴을 확인하였다.

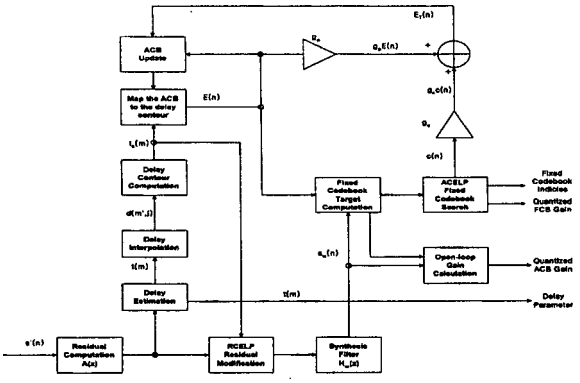


그림 2. EVRC 인코딩 블록도
Fig. 2. EVRC encoding block diagram.

본 논문의 구성은 다음과 같다. II장에서는 EVRC 보코더의 디코딩 알고리즘을 살펴보고, III장에서는 약전계에서의 EVRC 음질 저하 현상 및 그 원인에 대한 분석을 한다. IV장에서 피치 예측과 점진적 복원 기법을 이용한 음질 개선 방법에 대하여 기술한다. V장에서 음질 평가 실험 및 결과에 대해 논하고, VI장에서 결론 및 향후 연구 과제에 대해 논한다.

II. EVRC 디코더 알고리즘

본 장에서는 IS-127 표준안에 규정되어 있는 EVRC 디코더 알고리즘 및 오류 프레임 처리 방식에 대해 기술한다. EVRC 디코더는 프레임 단위로 디코딩을 수행한다. 보코더는 CPU로부터 패킷을 전송받고, 이를 오류 프레임인지 정상 프레임인지를 구분하여, 각각을 다른 방식으로 디코딩한다.

2.1 프레임 오류 검사

보코더가 수신하는 패킷은 전송률과 패킷데이터로 구성된다. 수신한 패킷의 전송률이 1, 1/2, 1/8(각각 8, 4, 1Kbps) 중 하나가 아니면, 이프레임은 오류 프레임으로 판정한다. 또, 전송률이 1/8이면서 모든 패킷데이터의 값이 1이면, 이 프레임은 오류프레임이 된다. 전송률이 1또는 1/2이며, 패킷에 포함된 DELAY 인자의 값이 100을 초과하면, 이 프레임 또한 오류 프레임으로 결정한다. 보코더가 패킷을 수신할 때, 비트 에러가 있는 전송률이 1인 경우와, 삭제 프레임의 경우도 오류 프레임으로 판단한다. 이러한 정보들은 패킷의 전송률 부분에 들어있다.

2.2 정상 프레임 디코딩

수신된 패킷의 전송률에 따라 다른 방식으로 디코딩한다. EVRC또한 CELP 계열의 보코더 이므로, 대체적으로 일반적인 CELP 계열의 디코딩 방식을 따른다. Fixed codebook과 adaptive codebook은 부 프레임 단위로 재생된다.

가. 전송률 1/2 또는 1을 갖는 프레임

대부분의 음성 구간은 전송률이 1로 인코딩되고, 전송률 1/2는 전이 구간에서 주로 사용된다. 디코딩은 다음의 단계에 따라 이루어 진다.

1) LSP(Line Spectral Pair) 디코딩
2) LSP interpolation
3) LSP를 LPC(Linear Prediction Coefficient)로 변환
4) 현 프레임의 delay값 추출
5) 피치 interpolation하여 피치 contour 구성
6) 이전 프레임이 오류 프레임이면, 이 전 프레임의 pitch contribution 재생
7) 피치 contour로 부터 adaptive codebook 합성
8) Fixed codebook 합성
9) Total 여기신호 = 7) + 8)
10) LPC filtering하여 음성 합성
11) 후처리 필터로 음질 개선

나. 전송률 1/8을 갖는 프레임

전송률 1/8은 묵음 구간일 때 주로 사용된다. 따라서, 피치 정보가 없고, 고정 코드북 여기 신호도 random noise로 구성된다. 디코딩 과정은 다음과 같다.

1) LSP 디코딩
2) LSP interpolation
3) LSP to LPC 변환
4) Gaussian random noise 발생
5) 4)의 신호를 LPC filtering
6) 후처리 필터

2.3 오류 프레임 디코딩

오류 프레임 이전 프레임의 전송률 및 종류에 따라, 현재 프레임의 해당 전송률을 결정한다. 이전 프레임이 1/8이었으면, 전송률 1/8 디코딩 방식에 따라 현재 오류 프레임을 디코딩하고, 그 외의 경우는 전송률 1로 가정하고 디코딩한다.

두 경우 모두 LSP는 이 전 프레임의 값을 그대로 사용한다. 전송률 1로 결정된 경우의 디코딩은 정상적인 프레임의 디코딩 방식을 따르며, 몇 가지 인자들을 바꾸어서 사용한다. 사용되는 인자들은 다음과 같다.

1) Delay 값은 이전 프레임의 값을 사용한다.
2) 두 프레임이 연속해서 오류 프레임이면,평균 adaptive codebook gain을 0.75배로 감소시킨다.
3) 평균 adaptive codebook gain이 0.3보다 적어지면 delay contour를 pseudo random 값을 사용한다.
4) LPC 합성을 위한 합성 코드북 여기신호는 adaptive codebook만을 사용한다.

즉, fixed codebook은 사용하지 않는다.

$$E_T(n) = g_p E(n) \quad (1)$$

5) 평균 adaptive codebook gain이 0.4보다 적어지면, random fixed codebook 여기 신호를 사용한다. 즉,

$$E_T(n) = E_T(n) + 0.1 g_{avg} ran\_g\{seed\} \quad (2)$$

6) 그 외 과정은 정상적인 프레임의 경우와 같다.

III. 약전계에서의 EVRC 음질 저하 현상 및 원인

3.1 약전계 상황

우리나라는 지리적·사회적인 요인으로 인해 전파환경이 나쁜 곳이 많다. 통칭하여 약전계라고 볼 수 있는데, 이런 곳의 통화 품질은 매우 나쁘다. FER이 30%이상 되는 경우도 빈번히 발생한다. 그림 3에 단말기에서 수신한 패킷의 예를 보였다. 전송률(8, 4, 1Kbps가 각각 4, 3, 1로 표시됨)과 오류 프레임 여부(0으로 표시됨)를 볼 수 있으며, 연속적으로 오류 프레임이 발생하는 경우를 볼 수 있다. 그림 3에 나타난 패킷의 경우 FER은 65.2%에 달했다.

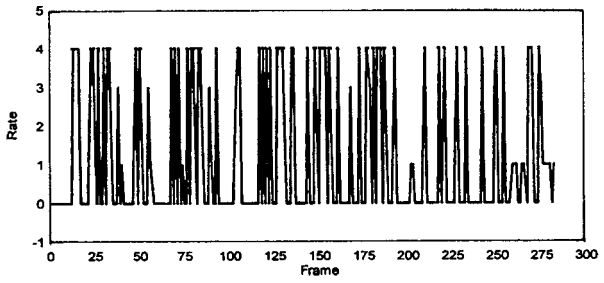


그림 3. 약전계에서 수신한 패킷 상태
Fig. 3. Status of packet received in the weak RF environment.

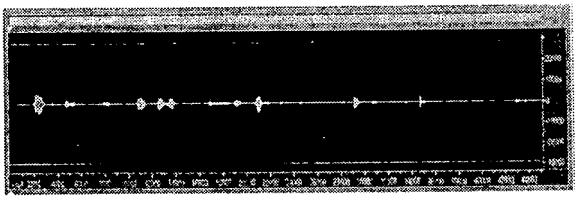


그림 4. 수신된 패킷으로부터 합성한 음성
Fig. 4. Decoded speech waveform from the received packet.

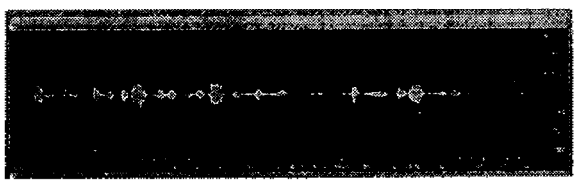


그림 5. 원 음성 파형
Fig. 5. Original speech waveform.

3.2 EVRC 음질 저하 현상 및 원인

그림 3에 보여진 패킷을 디코딩한 음성 파형은 그림 4와 같고, 오류가 없는 원 음성 파형은 그림 5에 보였다. 대부분의 음성이 제대로 복구되지 못하고, 잡음성분처럼 존재함을 볼 수 있다.

오류 프레임인 경우는 II장에서 설명한 방식대로 디코딩된다. 그런데, 이 방식은 FER이 낮은 경우에는 문제없이 좋은 음질을 유지하지만, FER이 높아지면 음질이 급격히 나빠진다. 오류 프레임 부분에서는 음성의 내용을 거의 알아들을 수 없을 뿐 아니라, 재생음이 귀에 거슬리는 느낌이 든다.

이렇게 약전계에서 급격히 음질이 떨어지는 원인은 EVRC 보코더의 구조 및 오류 프레임 처리 방식에 기인한다. II장에서 설명한 것처럼, 오류 프레임이 발생하면 피치 delay값을 이 전 프레임의 값을 사용한다. EVRC는 기본적으로 RCELP 방식을 따르므로, 한 프레임에 하나의 피치 delay값을 사용한다. 이를 이전 프레임의 피치 delay값과 interpolation하여, 각 샘플별로 피치 contour를 구성하여 adaptive codebook을 구성한다. 이를 위해 전송률 1인 경우는 이전 프레임의 delay값과 현재 프레임의 delay값의 차이인 delta delay를 전송하고, 한 프레임이 오류인 경우는 다음 정상 프레임의 delay값과 delta delay 값을 이용해 adaptive codebook을 제대로 재생해내어 좋은 음질을 유지할 수 있다. 그러나, 연속적으로 오류프레임이 발생하면 adaptive codebook을 제대로 구성할 수 없고, 계속해서 같은 delay값을 사용하게 되어, 피치가 너무 강조되는 현상이 생긴다. 이로 인해 연속적인 오류 프레임의 경우 음이 삑삑 거리는 듯 하는 느낌을 주게 된다. 이 문제는 IV장에서 설명할 피치예측 기법을 이용하여 해결하였다.

또, EVRC는 오류 프레임 이후에 정상 프레임이 수신되면, 그 프레임은 정상적인 방법으로 디코딩한다. 즉, 2~3프레임정도 정상 프레임이 수신되면, 합성 음성은 정상적인 경우의 에너지를 거의 갖게 된다. 이는 좋은 환경에서 한 두 프레임의 오류 프레임이 발생할 때, 음질에 거의 영향을 받지않게 되지만, 약전계에서는 오히려 나쁜 영향을 끼치게 된다.

그림 3에서 볼 수 있듯이, 오류 프레임 사이에 몇 프레임의 정상 프레임이 수신되고, 또 오류 프레임이 수신되는 형태일 때, 연속되는 오류 프레임으로 인해, 합성 음성의 에너지는 매우 적어지게 된다. 이 때 2~3프레임 정도 큰 에너지를 갖는 음성이 합성되고, 또 오류 프레임이 몇 개 연속해서 수신되면 정상 프레임에 의해 디코딩된 소리만 갑자기 존재하게 된다. 이는 매우 짧은 시간이므로 사람은 제대로 듣지 못하고, 전자음 같은 잡음으로 인식하게 되어 귀에 거슬리게 된다. 이 문제는 점진적 복원 기법을 이용하여 해결하였고, IV장에서 설명한다.

IV. 피치예측 및 점진적 복원을 이용한 음질개선

4.1 피치예측 기법

음성은 크게 유성음과 무성음으로 나눌 수 있다. 유성

음의 경우는 피치성분이 존재하고, 무성음은 존재하지 않는다. 무성음은 음의 지속시간이 짧으며, 에너지도 적어, 대부분의 음성 구간은 유성음 구간이 된다. 따라서, 피치는 음성 신호의 특징을 결정짓는 매우 중요한 요소이다.

III장에서 기술한 바와 같이, 오류 프레임이 발생하면, EVRC 디코더는 이 전 프레임의 피치 값을 현재 프레임에서 사용한다. 이렇게 하면, 유성음의 피치는 급작스럽게 변하지 않으므로, 한 두 프레임의 오류인 경우에는 별 문제 없으나, 약전계 상황처럼 연속적으로 오류 프레임이 발생할 때는 단일 톤 비슷한 효과를 내게 되어 음질에 나쁜 영향을 주게 된다.

그림 6에 EVRC로 인코딩된 음성 신호의 피치값을 보였다. 전송률 1/8은 피치가 없으므로 10으로 표시했다. EVRC에서는 20~120까지의 피치값을 사용한다. 음성 구간에서 피치는 대개 일정하지만, 완전히 동일한 값은 아니며 서서히 변화함을 알 수 있다. 뒷부분의 급격히 변화하는 부분은 무성음 구간으로 피치가 존재 하지 않는 부분이다.

오류 프레임시 피치의 변화를 예측하기 위해 간단한 1차 선형 예측 기법을 사용했다. 즉, 현재 프레임 번호를 m, 피치를 $\tau(m)$이라 하면,

$$\tau(m) = \begin{cases} DELAY(m)+20; & FER(m) = FALSE \\ 2\tau(m-1)-\tau(m-2); & FER(m) = TRUE \end{cases} \quad (3)$$

이 된다. 여기서, DELAY(m)은 m번째 프레임의 delay값이고, FER은 TRUE이면 오류 프레임, FALSE이면 정상 프레임을 나타내는 지시자이다.

그림 6의 음성 패킷을 약전계에서의 오류 프레임 발생 패턴에 따라 오류 프레임을 만들고, 이를 기존 방법과 피치 예측 기법을 이용한 방법에 따른 피치 값을 그림 7, 8에 각각 보였다. 125~150프레임의 경우처럼, 연속적으로 오류 발생시 기존 방법으로는 일정한 피치를 유지하나, 피치 예측 기법을 사용한 것은 자연스럽게 변화하여 원음성에 더 가까움을 볼 수 있다.

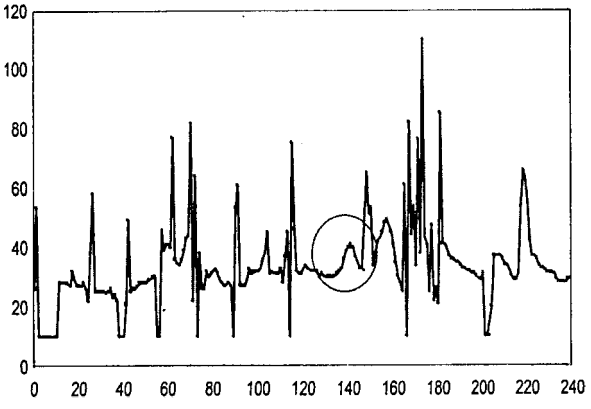


그림 6. EVRC로 인코딩된 음성의 피치
Fig. 6. Pitch of the original speech encoded with EVRC.

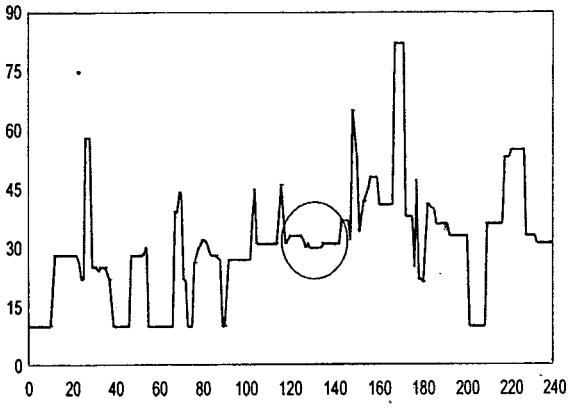


그림 7. 오류 프레임 발생시 피치
Fig. 7. Pitch of the synthesized speech from erasure frames.

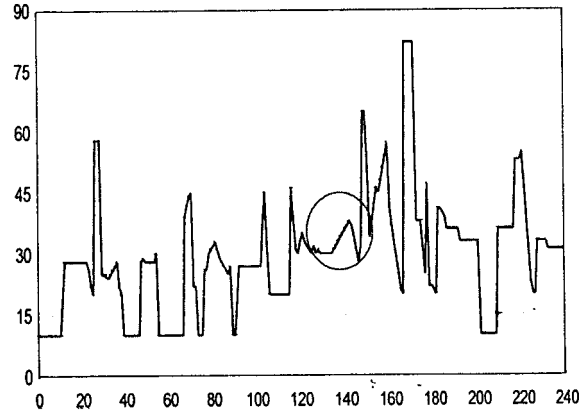


그림 8. 피치 예측을 사용한 피치
Fig. 8. Predicted pitch of the erasure frames.

4.2 점진적 복원 기법

연속 적으로 오류 프레임이 발생하고, 그 사이 사이에 정상 프레임이 수신 될 때, 정상 프레임으로 인해 오히려 음질이 저하되는 현상을 막기위해 점진적 복원 기법을 제안하였다.

연속적으로 ERR_THLD 프레임 이상의 오류 프레임이 수신 된 후, 정상 프레임이 수신되면, 그 때부터 N 프레임 동안의 최종 출력 gain을 0부터 점진적으로 1까지 증가시킨다. 한 프레임의 음성 샘플 수는 160(8 KHz * 20 ms)이므로, 합성 음성 신호를 s(n)이라 하면, 연속된 오류 프레임 수신 후, 첫 정상 프레임부터 N프레임까지의 출력 음성 신호는 다음 식 (4)와 같이 된다.

$$s(n) = s(n) \times \frac{n}{160 \times N} \quad (4)$$

여기서, n은 $0 \le n < 160 \times N$의 범위를 가지는 sample index이다.

이러한 방식으로 연속되는 오류 프레임으로 인해 출력되는 음성의 에너지가 거의 0이었다가 갑자기 보통 음성

크기로 증가해서 생기는 잡음 성분을 줄일 수 있다. N 프레임이 지난 후에도 계속해서 정상적인 프레임이 수신되면, 그 때부터는 전파환경이 좋은 경우이므로 정상적인 출력을 내보낸다. 점진적 복원 기법의 flow chart는 그림 9와 같다.

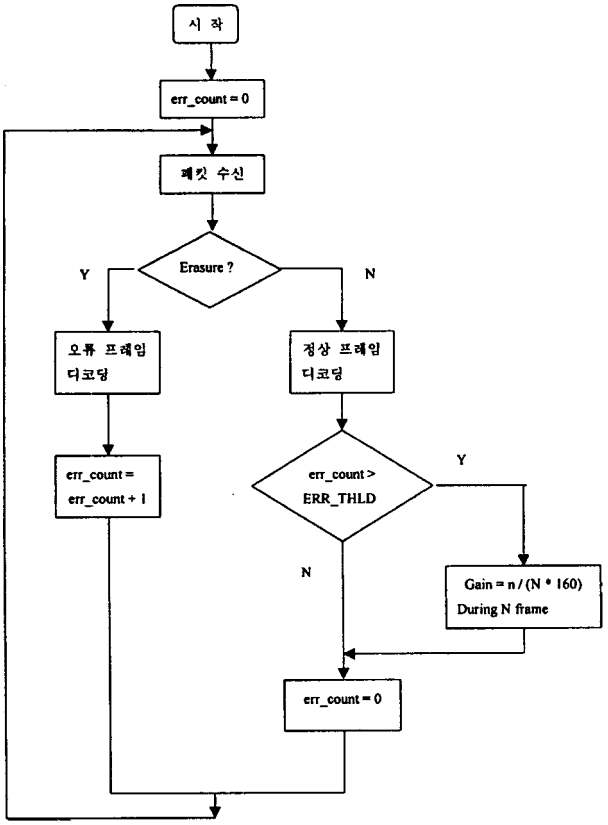


그림 9. 점진적 복원 기법의 flow chart
Fig. 9. Flow chart of gradual increase.

FER이 64.6%인 또 다른 패킷을 기존 방법과 점진적 복원을 이용한 방법으로 각각 디코딩한 음성 파형을 그림 10, 11에 보였다. 점진적 복원을 사용한 파형에서는 오류 프레임 사이의 잡음 성분(정상 프레임으로 인한)이 줄어듦을 볼 수 있다.

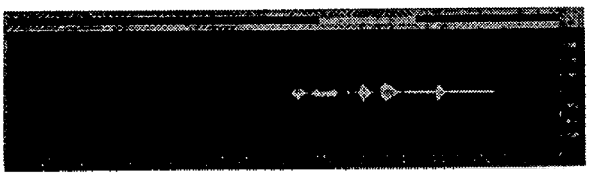


그림 10. 기존 방법으로 디코딩한 음성 파형
Fig. 10. Speech waveform decoded by EVRC.

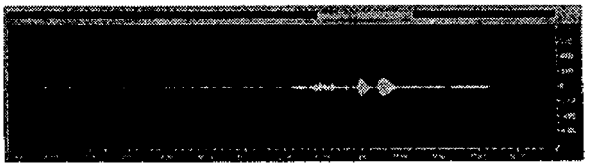


그림 11. 점진적 복원을 이용한 디코딩한 음성 파형
Fig. 11. Speech waveform decoded by gradual increase method.

IV. 실 험

현장의 다양한 환경에서 녹음한 패킷을 제안한 방법 및 기존 방법으로 디코딩하여 음질 선호도를 평가하였다. FER이 2, 5, 10, 20, 30, 50%인 각기 다른 패킷을 기존 방법(EVRC, IS-718), 피치 예측을 사용한 방법, 점진적 복원을 사용한 방법, 그리고 피치 예측과 점진적 복원을 동시에 사용한 방법으로 각각 디코딩 하고, 이를 21명(남자 11명, 여자 10명)이 동일한 패킷에 대해 선호도를 평가하였다. 가장 선호하는 방법을 4, 가장 나쁜 것을 1로 하여, 순서를 정하는 방법으로 평가하였다.

점진적 복원은 5프레임 이상의 오류 프레임이 발생하면, 그 후 첫번째 정상 프레임부터 5프레임 동안의 출력 음성을 점진적으로 복원시켰다.

평가에 사용한 음성 패킷은 남자 아나운서가 발음한 3문장과, 여자 아나운서가 발음한 3문장을 다양한 FER로 섞어 사용하였다. 그림 12에 각 경우의 선호도 실험 결과를 보였다.

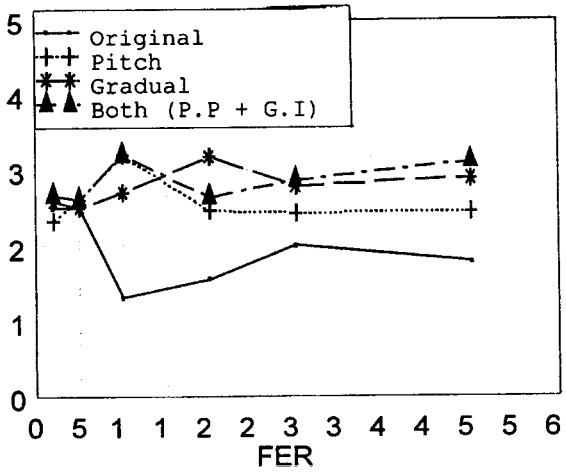


그림 12. 선호도 실험 결과
Fig. 12. Result of the preference test.

좋은 전파환경에서는 어떤 디코딩 방법을 사용하던지 차이가 없으나, FER이 높아지면 많은 차이가 난다. 피치 예측 기법과 점진적 복원 기법을 모두 사용한 경우 가장

선호도가 높았으며, FER이 10%정도인 경우는 피치 예측이, 20%이상에서는 점진적 복원 기법이 선호도가 높았다.

이는 10%정도의 FER은 매우 높은 값으로, 전파환경이 나쁘지만, 어느 정도의 통화는 가능한 상황이므로, 피치 예측을 하는 경우가 좀 더 자연스럽게 들리고, 그 이상인 경우는 정상적인 통화가 어렵기 때문에, 간간이 발생하는 잡음 성분을 줄이는 것이 더 체감 통화 품질을 높이기 때문이다. 극단적인 상황인 FER 50%인 경우에서 보다 분명히 확인할 수 있다.

V. 결 론

본 논문에서는 EVRC 상용화 후, 약전계에서 나타난 EVRC 음질 저하 문제에 대한 원인을 분석하였고, 이에 대한 해결책으로 피치 예측 기법과 점진적 복원 기법을 제시하였다. 피치 예측을 이용하여 오류 프레임의 합성음을 더욱 부드럽게 했으며, FER이 매우 높은 상황에서 연속적으로 오류 프레임이 발생할 때 정상 프레임으로 인한 잡음 성분을 점진적 복원 기법을 이용하여 줄였다.

음질 평가 결과 두 가지 방법 모두를 사용한 경우가 가장 선호도가 높았으며, 알고리즘의 단순성에 의해 전체 계산량도 조금밖에 증가하지 않았다.

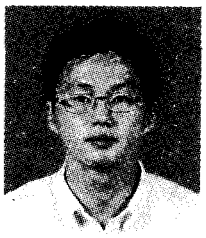
본 논문에서는 피치 예측 기법으로 1차 선형 예측 만을 사용했으나, 2차 이상 또는 비선형 예측 기법등을 사용할 수도 있고, 점진적 복원의 인자들도 더욱 다양하게 변화시켜가며 실험해 보아야 한다. 이상의 내용들은 향후 지속적으로 연구될 예정이다.

또한 개선된 EVRC를 이용한 현장 시험 및 세부 조정 작업 또한 상용화를 위해 필수적이라 하겠다.

참 고 문 헌

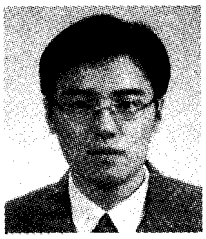
1. 김재원, 권우성, 이인홍, 이주식, 이진익, "EVRC Impelementation 및 시험", SK Telecom Technical Journal, pp. 69-79, 1997년 12월.
2. TIA/EIA/IS-127, Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems
3. W. B. Kleijn, P. Kroon,, and D. Nahumi, "The RCELP Speech-Coding Algorithm", European Transactions on Telecommunications, Vol 5, Number 5. Sept/Oct 1994, pp. 573-582
4. R. Salami, C. Laflamme, J.-P. Adoul, and D. Massaloux, "A Toll Quality 8 kbps Speech coder for Personal Communication System(PCS)", IEEE Trans. On Vehicle Technology, 1994.
5. TIA/EIA/IS-718(PN-3648), Minimum Performance Specification for the Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems

▲민 병 준 (Byung Jun Min) 1962년 2월 20일생



1981년 ~1985 : 서울대학교 공과대학 전자공학과(공학사)
1985년~1987년 : 한국과학기술원 전기 및 전자공학과(공학석사)
1987년~1993년 : 한국과학기술원 전기 및 전자공학과(공학박사)
1987년~1997년 : 디지콤 정보통신연구소 책임연구원
1997년~현재 : SK텔레콤 중앙연구원 수석연구원
※주관심분야 : 음성 데이터 압축, 음성 인식, 음성합성

▲김 재 원(Jae Weon Kim) 1972년 12월 23일생



1990년~1994년 : 한국과학기술원 과학기술대학 전기 및 전자공학과(공학사)
1994년~1996년 : 한국과학기술원 전기 및 전자공학과(공학석사)
1996년~현재 : SK 텔레콤 전임연구원
※주관심분야 : 음성 데이터 압축, 음성 인식