음성 압축기를 사용한 통신 시스템에서의 음성 인식 성능 분석

한상욱, 정희석*, 박호종

광운대학교 전자공학과, 광운대학교 전자통신공학과*

Performance Analysis of Speech Recognition in Communication Systems using Speech Coder

Sang-Wook Han, Heui Suck Jung*, Hochong Park

Dept. of Electronics Engineering, Dept. of Telecommunication Engineering*

Kwangwoon University

E-mail : hcpark@mail.gwu.ac.kr

요 약

본 논문에서는 음성 압축기를 사용하는 디지털 이동 통신 환경에서 한글 음성 인식기의 성능을 분석하기 위하여 다양한 표준 음성 압축기를 이용하여 음성 압축기의 구조, 전송률, 전송 채널의 에러율에 대한 성능을 측정하여 비교하였다. 동일한 구조의 음성 압축기에 대하여 전송률의 증가에 따라 음성 인식률이 증가하지만, 음성 압축기의 구조에 따라 동일 전송률에서도 많은 성능 차이가 발생하는 것을 확인하였다. 특히 IS-127 EVRC의 인식 성능이 매우 떨어지는 것을 알 수 있고, EVRC의 잡음 제거기와 가변 전송률에 의하여 음성 인식 성능이 저하되는 것을 확인하였다. 이를 통하여 청취 음질과 음성 인식 성능 사이의 상관 관계가 높지 않는 것을 알 수 있다. 모든 음성 압축기에 대하여 채널 에러율과 음성 인식기의 성능은 매우 밀접한 관계가 있음을 확인하였고, 평균적으로 채널 에러율 1.0%에서 인식률이 0.6% 감소하고, 에러 5.0%에서 인식률이 1.8% 감소한다.

1. 서 론

최근 음성 인식 기술의 발달에 따라 많은 분야에서 음성 인식을 사용하여 새로운 제품 및 서비스를 개발하고 있다. 지금까지의 음성 인식은 주로 음성 신호를 직접 음성 인식기에 입력하는 형태를 가지고 있다. 예로, 컴퓨터의 명령어 인식과 자동 문서 입력 장치, 휴대폰의 전화 번호 인식, 음성 인식 장난감 등이 이와 같은 형태에 해당한다. 그러나, 최근에 통신망을 통하여 음성 명령을 입력하는 경우가 많이 발생하고 있고, 이 경우는 통신망을 통과한 음성을 이용하여 서버에 위치한 음성 인식기에서 인식을 하게 된다. 특히, 최근 디지털 이동 통신의 확산에 따라 이동 통신의 통화 빈도가 급격하게 증가하고 있어 앞으로 이동 통신 단말기를 이용하여 음성 명령을 입력하는 경우가 크게 증가할 것으로 예상된다.

이동 통신망을 통한 음성 인식은 기존의 직접 입력 시스템에 비하여 매우 다른 구조를 가지고 있다. 이동 통신 시스템에서 효율적인 통신을 위하여 입력 음성 신호를 압축하여 전달하므로 원 음성 신호에 비하여 음질이 저하되고 음성의 특성이 왜곡된다. 또한, 무선 채널을 통하여 데이터를 전달하는 과정에 전달 오류가 발생하므로 이로 인하여 추가적인 음질 저하 및 왜곡이 발생한다. 또한, 현재 이동 통신 시스템에서 여러 종류의 음성 압축기가 사용되고 있고, IMT-2000 등의 새로운 통신 서비스가 제공되면 새로운 음성 압축기가 사용되므로, 특정 음성 압축기의 특성의 적용하여 최적의 음성 인식기를 설계하는 것도 문제가 있다.

더욱이, 현재까지 개발된 모든 음성 압축기들은 사람이 청취하는 음질을 향상시키는 것을 목표로 설계되었으며 음성 인식 성능에 대하여서는 전혀 고려하지 않고 설계되었다. 따라서, 음질이 우수한 음성 압축기가 반드시 높은 음성 인식률을 제공하지는 않게 된다. 또한, 현재 개발된 음성 인식 알고리즘은 음성 압축기에 의한 특성 왜곡 및 오류를 보상하는 기술이 적용되지 않아 이동 통신 환경에서의 음성 인식기 성능은 매우 저하된다.

따라서, 본 논문에서는 이동 통신 환경에서 다양한 종류의 표준 음성 압축기를 통한 음성 인식의 성능을 측정하여 음성 압축기의 구조, 전송률, 채널 에러율에 대하여 음성 인식 성능을 비교 분석하고, 앞으로 음성 인식 성능이 우수한 새로운 음성 압축기를 개발하고,

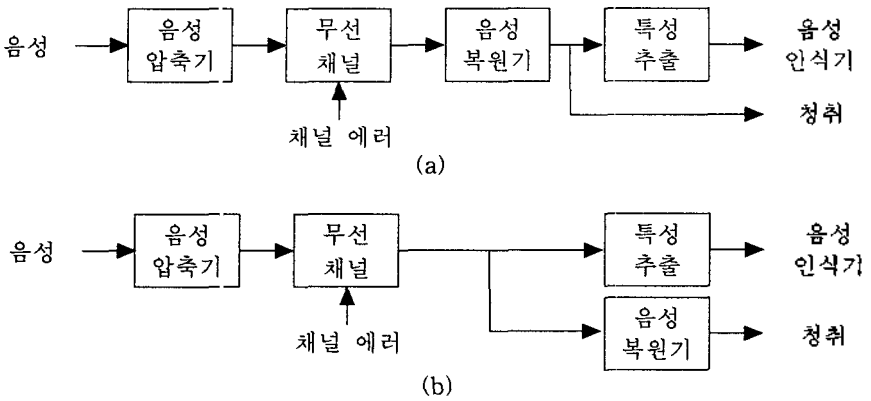

그림 1. 이동 통신 환경에서의 음성 인식 시스템의 구조. (a) 복원된 신호로부터 특성을 추출하여 인식하는 구조. (b) 음성 비트 스트림으로부터 직접 특성을 추출하여 인식하는 구조

이동 통신망에서의 음성 인식 시스템을 개발하기 위한 기초 자료를 제시하고자 한다.

2.이동 통신에서의 음성 인식 시스템 구조

이동 통신 환경에서의 음성 인식 시스템은 그림 1과 같은 구조를 가진다[1]. 음성 전달은 반드시 음성 압축기와 무선 채널을 통과하고, 음성 인식기의 입력은 복원된 음성 신호 또는 전달된 음성 비트 스트림이 된다. [1]의 연구에 의하면 복원된 음성 신호를 입력하여 이로부터 다시 인식을 위한 특성 파라미터를 추출하는 것보다 비트 스트림으로부터 음성 인식에 필요한 특성 파라미터를 직접 추출하는 것이 더 우수한 성능을 가지고, 특히 채널 에러가 많을 경우 성능 차이가 더욱 커진다.

그러나, 비트 스트림에서의 특성 추출을 하기 위하여 현재 이동 통신 교환기 내부에서 음성을 복원하는 모듈에 음성 인식기를 장착하여야 하며, 이는 일반적인 음성 인식 서비스 업체가 쉽게 접근 할 수 없는 영역이므로 구현하기가 어렵다. 따라서, 일반적으로 복원된 음성 신호를 입력하여 인식하는 형태를 가지게 되고, 이에 따라 본 논문에서는 복원된 음성 신호로부터 음성을 인식하는 구조에 대하여 음성 인식 성능을 측정한다.

3. 음성 압축기

본 논문에서 사용하는 음성 압축기는 현재 국내 디지털 이동 통신에서 사용하고 있는 음성 압축기와 IMT-2000의 표준 음성 압축기, 그리고 현재 디지털 이동 통신에서는 사용하고 있지 않지만 ITU의 표준 음성 압축기로서 VoIP 등에서 많이 사용하는 음성 압축기 등이다. 각 음성 압축기에 대한 간단한 설명은 다음과 같다.

(a) IS-96 QCELP와 IS-733 QCELP[2]

현재 국내 이동 통신에서 사용되고 있는 음성 압축기로서 최대 8.55kbps(IS-95)와 13.3kbps(IS-733) 전송률을 가지는 가변 전송률(Variable Rate)을 음성 압축기이다. 20msec 프레임 길이와 CELP (Code-Excited Linear Prediction) 구조를 가지고 있고, 순환 구조의 코드북(Circular Codebook)을 사용하며, 매우 오래 전에 개발되어 최근에 개발된 동일 전송률의 다른 음성 압축기에 비하여 청취 음질은 떨어진다.

(b) IS-127 EVRC[3]

현재 국내 이동 통신에서 사용되고 있는 음성 압축기이며, 최대 8.55kbps 전송률을 가지는 가변 전송률 압축기로서 IS-96의 성능을 향상시키기 위하여 IS-96과 유사한 구조로 개발되었다. Algebraic 코드북을 사용하는 ACELP 구조이고, LSP 양자화 및 이득의 양자화에서 많은 성능의 향상을 가져온다. EVRC의 가장 큰 특징은 입력단에 잡음 제거기를 가지고, 측정된 피치 구조에 입력 신호가 일치하도록 입력 신호를 시간적으로 Warping을 시켜 피치 잔여 신호를 최소로 하는 것이며, 이를 통하여 코드북이 표현하여야 하는 신호의 오차를 최소화시켜 품질의 향상을 가져온다.

(c) G.729[2]

ITU의 8kbps 표준 음성 압축기로서, 10msec 프레임 길이와 ACELP 구조를 가지며, 프레임 길이가 짧아 전달 지연 시간이 짧은 장점이 있고 VoIP 등에서 많이 사용되고 있다.

(d) AMR[4]

비동기 IMT-2000 이동 통신의 표준 음성 압축기이며, 최대 12.2kbps와 최소 4.75kbps 사이에 총 8개의 전송률(모드)을 가진다. 이는 채널의 전송 상태를 고려하여 소스 코딩(음성 신호 압축)과 채널 코딩에 적절하게 비트를 할당하는 구조를 가지며, 20msec 프레임 길이를 가지고 ACELP 구조에 기반을 두고, 각 모드는 거의 유사한 구조로 동작하여 모드 사이의 변환을 용이하게 하였다.

(e) G.723.1[2]

ITU의 표준 음성 압축기로서 6.3kbps와 5.3kbps의 두

가지 전송률을 가지는 음성 압축기로 구성되며, 각각 ACELP 구조의 ML 구조를 가진다. 30msec 프레임 길이를 가지고 현재 VoIP에서 많이 사용되고 있다.

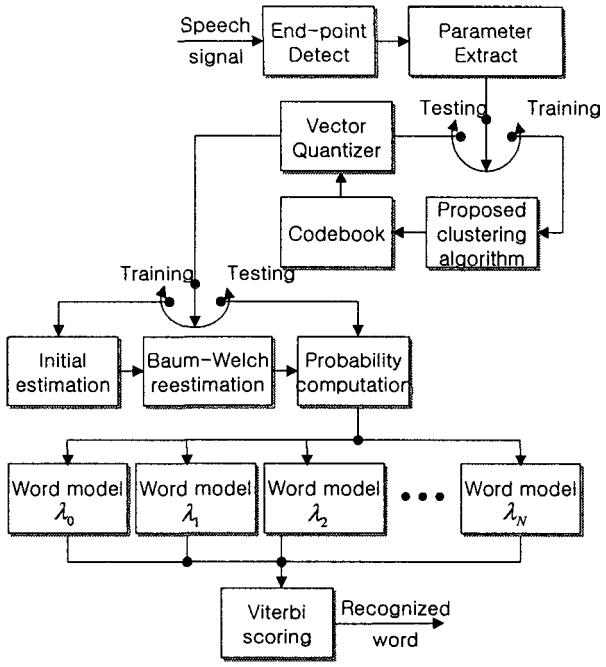

그림 2. 음성 인식기 시스템 구성도

4. 음성 인식기

본 논문에서 사용한 음성 인식기는 VQ/HMM 기반 고립 단어 인식기를 이용하였으며 전화 음성 대역폭을 고려하여 13차의 PLP 켑스트럼과 13차의 델타 켑스트럼을 특징 파라미터로 이용하였다. 전화 채널을 고려하여 음성 특징 파라미터 추출을 위한 분석 구간은 입력 음성의 20msec 간격으로 분석하고 10msec씩 이동하여 분석하였다. 전체적인 구성도는 그림 2에 주어졌다.

5. 음성 인식기의 성능 및 분석

본 논문에서 실시한 실험 환경은 다음과 같다.

- 한글 화자 독립 고립 단어 인식
  - PDA에서 사용하는 명령어

- 훈련 데이터 :
  - 남성 10명, 여성 5명
  - 각 화자 당 40 단어, 총 600 단어

- 인식 실험 :
  - 14가지 음성 압축기, 6가지 채널 에러율 : 총 84 가지의 실험 항목
  - 남성 15명, 여성 10명
  - 각 화자 당 40 단어, 각 실험 항목 당 1,000 단어
  - 총 실험 데이터 : 84,000 단어

채널 에러율	0%	0.1%	0.2%	1.0%	2.0%	5.0%
PCM(128k)	86.6	-	-	-	-	-
IS-733 QCELP(13.3k)	84.2	83.7	84.3	83.4	83.1	83.0
AMR#1(12.2k)	86.3	85.9	85.9	85.9	85.7	85.3
AMR#2(10.2k)	86.8	86.3	86.2	85.9	85.7	85.4
IS-96 QCELP(8.55k)	83.3	82.4	82.9	81.9	80.7	75.0
IS-127 EVRC(8.55k)	77.8	77.9	77.4	78.1	76.5	76.6
ITU G.729(8k)	84.8	84.0	83.9	84.1	83.2	82.3
AMR#3(7.95k)	85.8	85.8	85.5	85.5	84.5	85.1
AMR#4(7.40k)	84.8	85.6	84.9	83.6	84.3	82.5
AMR#5(6.70k)	82.9	83.1	82.8	82.1	82.4	82.0
ITU G.723.1(6.3k)	84.6	85.1	86.6	85.2	85.3	85.1
AMR#6(5.90k)	82.5	82.6	82.1	83.7	81.8	81.0
ITU G.723.1(5.3k)	83.9	83.9	83.4	83.4	83.4	83.6
AMR#7(5.15k)	81.2	80.4	80.2	79.7	78.9	79.0
AMR#8(4.75k)	80.8	80.1	79.4	80.5	79.3	78.6

표 1. 여러 음성 압축기의 채널 에러율에 대한 음성 인식기의 인식률(%)

전체 실험 항목에 대한 인식률은 표 1에 정리되어 있다. 128kbps PCM은 음성 압축기를 통과하지 않은 신호를 나타내고 실험에 대한 기준 성능을 제시한다. 채널 에러율이 0%가 아닌 경우에는 PCM 신호는 의미가 없다. 모든 음성 압축기는 채널 에러율이 증가하면 인식률이 저하되며, 특히 IS-96 QCELP는 채널 에러율 5%에서 매우 심하게 성능이 저하되는 성질을 가진다. 평균적으로 채널 에러율 1.0%에서 인식률이 0.6% 감소하고, 에러 5.0%에서 인식률이 1.8% 감소한다.

그림 3은 채널 에러율 0%, 1%, 5%에 대하여 각 음성 압축기에 대한 성능을 비교한 것이며, x 축은 PCM과 14개의 음성 압축기를 표 1의 순서로 나열한 것이고 검은 막대는 AMR을 나타낸다. AMR의 전송률과 음성 인식률은 높은 상관 관계를 가지지만, G.723.1은 비슷한 전송률의 AMR 보다 성능이 우수하고, QCELP, EVRC, G.729는 비슷한 AMR보다 성능이 떨어진다. G.723.1이 AMR보다 성능이 우수한 것은 G.723.1의 프레임 크기가 길고 AMR과는 달리 주어진 전송률에 최적으로 설계되었기 때문으로 판단된다. 이를 통하여 음성 인식기의 성능은 음성 압축기의 전송률뿐만 아니라 음성 압축기의 구조에 따라 많은 차이를 가지는 것을 알 수 있다.

EVRC는 모든 채널 에러율에서 다른 음성 압축기에 비하여 매우 낮은 인식률을 가지는 특이한 현상을 보여준다. EVRC는 청취 음질(MOS)로 측정한 성능에서 IS-96보다 훨씬 우수하고 G.729와 동등한 성능을 가지

는 것으로 알려져 있고[3], 실험에 사용된 음성 신호를 직접 청취하여보면 IS-96보다 음질이 우수한 것을 확인 할 수 있다. 그러나, 음성 인식의 성능은 IS-96에 비하여 매우 떨어지며, G.723.1, 저 전송률 AMR 등과 같이 청취 성능이 매우 낮은 음성 압축기보다도 오히려 음성 인식 성능이 떨어진다. 이에 대한 원인은 가변 전송률, 잡음 제거기, 시간에 대한 Warping을 실행하므로 이로 인하여 인식에 필요한 특성이 심하기 왜곡되어 음성 인식 성능이 저하되는 것으로 추측된다.

표 2에서는 이 세 가지 예상되는 원인 중 첫 번째와 두 번째의 예상되는 원인을 고려하여 가변 전송률을 최대 전송률로 고정한 것과 잡음 제거기를 제거하고 측정한 결과이다. 또한 IS-96도 가변 전송률을 최대 고정 전송률로 고정하여 측정하였다. 표 2에 측정된 결과로 보면 EVRC는 평균적으로 5.8% 정도 인식률이 증가하였고 IS-96은 2% 정도 인식률 증가를 확인할 수 있고 특히 채널 에러율 5%에서는 7.5%나 증가하였다. 그리고 세 번째로 예상되는 원인에 대한 연구는 진행 중이다.

채널 에러율	0%	0.1%	0.2%	1.0%	2.0%	5.0%
PCM(128k)	86.6	-	-	-	-	-
IS-127 EVRC	77.8	77.9	77.4	78.1	76.5	76.6
IS-127 EVRC (잡음 제거기 제거)	83.1	82.5	82.3	82.8	82.6	81.5
IS-127 EVRC (잡음 제거기 제거 + 최대 전송률)	83.9	83.5	83.5	82.9	83.5	81.8
IS-96 QCELP	83.3	82.4	82.9	81.9	80.7	75.0
IS-96 QCELP (최대 전송률)	85.0	84.8	84.7	83.9	82.8	82.5

표 2. IS-127 EVRC 와 IS-96 QCELP의 알고리즘 변경에 대한 음성 인식기의 인식률(%)

6. 결 론

본 논문에서 다양한 음성 압축기를 사용한 이동 통신 환경에서 음성 인식기의 성능을 음성 압축기의 구조, 전송률, 채널 에러율에 따라 측정하여 각 요소들과 음성 인식 성능과의 관계를 분석하였다. 인식기 성능과 채널 에러율은 밀접한 관계가 있으며 채널의 상태가 음성 인식에 큰 영향을 미치는 것을 확인하였다. 인식률은 음성 압축기의 전송률에 따라 다르며, 특히 음성 압축기의 구조와 높은 상관 관계가 있는 것을 확인하였다. 특히, EVRC의 인식 성능이 다른 음성 압축기에 비하여 매우 저하되는데, 이는 EVRC의 특이한 구조에 의한 것이다.

감사의 글

본 연구는 2001년 정보통신부 대학 IT 연구센터의 지원으로 이루어졌습니다.

참 고 문 헌

[1] H. Kim and R. Cox, "A bitstream-based front-end for wireless speech recognition on IS-136 communications system," *IEEE Trans. on Speech and Audio Processing*, July, 2001.

[2] L. Hanzo, F. Somerville and J. Woodard, *Voice Compression and Communications*, New York : John Wiley and Sons, 2001.

[3] D. Nahumi and W. B. Kleijn, "An improved 8kb/s RCELP coder," *Proc. IEEE Workshop on Speech Coding for Telecommunications*, Sept. 1995.

[4] GSM 06.90, "Digital cellular telecommunications systems(phase2+); adaptive multi-rate(AMR) speech transcoding," 1998.

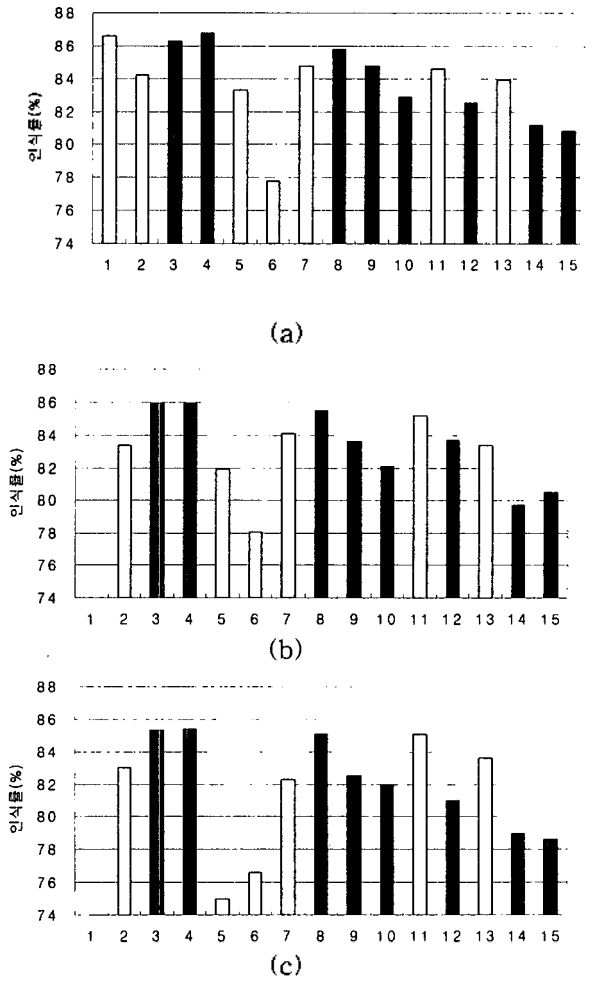

그림 3. 음성 압축기 구조 및 전송률에 대한 음성 압축기의 음성 인식 성능. (a) 채널 에러율 0%. (b) 채널 에러율 1%. (c) 채널 에러율 5%.