분산음성인식 환경에서 서버에서의 스케일러블 고품질 음성복원

*윤재삼, *김홍국, **강병옥
*광주과학기술원 정보통신공학과, **한국전자통신연구원
e-mail : *jsyoon@gist.ac.kr, *hongkook@gist.ac.kr, **bokang@etri.re.kr

Scalable High-quality Speech Reconstruction in Distributed Speech Recognition Environments

*Jae Sam Yoon, *Hong Kook Kim, **Byung-Ok Kang
*Department of Information and Communications
Gwangju Institute of Science and Technology
** Electronics and Telecommunications Research Institute

Abstract

In this paper, we propose a scalable high-quality speech reconstruction method for distributed speech recognition (DSR). It is difficult to reconstruct speech of high quality with MFCCs at the DSR server. Depending on the bit-rate available by the DSR system, we can send additional information associated with speech coding to the DSR server, where the bit-rate is variable from 4.8 kbit/s to 11.4 kbit/s. The experimental results show that the speech quality reproduced by the proposed method when the bit-rate is 11.4 kbit/s is comparable with that of ITU-T G.729 under both ideal channel and frame error channel conditions while the performance of DSR is maintained to that of wireline speech recognition.

I. 서론

분산음성인식 시스템에서는 서버의 풍부한 연산능력을 이용하여 소용량 및 대용량에 관계없이 음성인식을 수행한다. 또한 서버에서의 음성복원은, 음성인식을 통한 은행 또는 중개 업무에서 있어서의 발화자 검증, 사람과 기계 인식이 혼합된 응용 분야 등 그 적용과 응용 범위가 넓어 그 기술 개발이 요구된다.

분산음성인식 환경에서 음성을 복원하는 대표적인 방식 중의 하나로 ETSI (European Telecommunications Standards Institute) DSR 표준이 있다[1]. ETSI DSR 표준에서는 피치와 음성분류 정보 등과 함께 음성인식 파라메터인 MFCC (Mel-Frequency Cepstral Coefficient)를 통해 음성을 복원하기 때문에 음질이 떨어지는 단점을 보인다. 이러한 단점을 보완하기 위해, 본 논문에서는 전송 대역폭에 따라 음성 부호화기의 음성복원 정보를 추가적으로 전송함으로써, 고품질의 음성복원을 가능케 한다.

II. 스케일러블 음성복원 구조

스케일러블 구조를 통해 고품질의 음성을 복원하기 위해 ETSI DSR 표준 방식과 CELP 부호화기[2]의 기본 구조를 이용한다. 즉, 네트워크 환경에 따라 ETSI DSR 표준 음성복원 방식에 추가적인 부호화 데이터를 이용하여 음성을 복원한다. 그림 1에서는 전송 대역폭에 따라 세 가지 모드를 지원하는 스케일러블 음성복원 시스템의 구조를 보여준다. 첫 번째 모드는 ASR front-end에서 추출되어 양자화된 MFCC만을 전송하여 서버에서 음성인식만을 수행하며, 두 번째 모드는 ETSI DSR 표준의 음성복원 방식 지원하는 모드로 MFCC와 함께 ASR front-end에서 추출된 pitch, VAD(Voice Activate Detection), voicing class 등을 전송하여 음성을 복원한다. 마지막으로 세 번째 모드는 CELP 음성부호화기의 residual 정보를 추가적으로 전송함으로써 고품질의 음성을 복원한다. 특히, 세 번째 모드에서 CELP음성부호화기를 이용함에 있어 음성 스펙트럴 정보

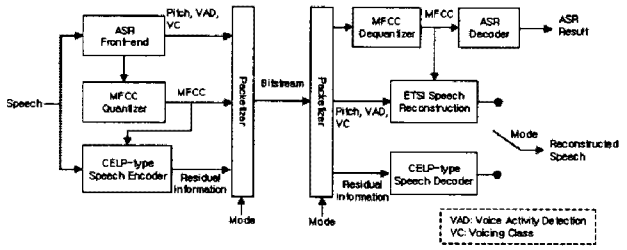

그림 1. 스케일러블 음성복원 시스템 구조

를 표현하는 LPC (Linear Prediction Coefficeint)를 MFCC로부터 변환한다[3]. 이는 MFCC를 전송하여 음성인식 성능은 그대로 유지하면서도, 음성부호화기에서 필요로 하는 LPC는 전송하지 않음으로써 전송 데이터량을 줄일 수 있는 장점이 있다.

III. 패킷 구조

제안된 스케일러블 음성복원 시스템에서 생성되는 패킷 구조는 그림 2와 같다. Mode 0는 Header 및 MFCC를 전송하여 음성인식을 제공하며, Mode 1은 추가 음성복원 정보를 통해 음성복원 기능을 제공한다. 마지막으로 Mode 2는 음성부호화기의 resudual 정보를 추가하여 고품질의 음성복원을 지원한다. 각각의 비트전송률은 4.8 kbp/s, 5.6 kbp/s 그리고 11.4 kbp/s이다.

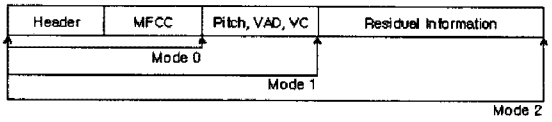

그림 2. 스케일러블 음성복원 전송 패킷 구조

IV. 음질 성능 평가

제안된 방식의 음질을 평가하기 위해 PESQ 평가를 수행하였다. 원어민 남, 녀 1명씩이 각각 발성한 한국어와 영어 문장에 대한 (총 4문장, 각 문장은 8초 발성 음성) 결과를 보면, 그림 3과 같이 프레임 오류가 없는 이상 채널 환경에서 ETSI DSR 표준 방법과 동일한 Mode 1은 MOS로 2.933, 추가적인 음성부호화기 정보

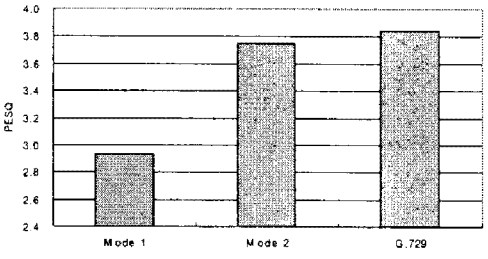

그림 3. 이상 채널 환경에서의 복원 음질 비교

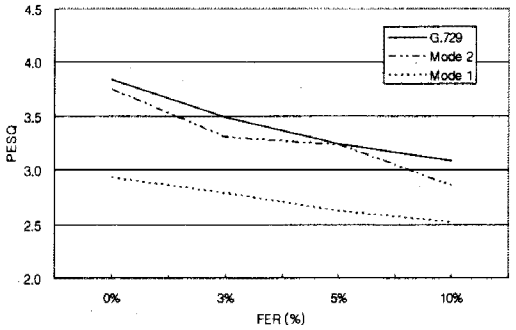

그림 4. 오류 채널 환경에서의 복원 음질 비교

를 이용하는 Mode 2는 MOS로 3.748를 보였다. 여기서 Mode 2는 표준 음성부호화기인 G.729의 성능인 MOS 3.840 과 유사한 결과를 보임을 알 수 있다. 한편, 그림 4와 같이 Mode 2는 3%,5%, 10%의 FER (Frame Error Rate) 환경에서도 G.729와 유사한 성능을 보였으며 Mode 1에 비해 우수한 성능을 보임을 알 수 있다.

V. 결론

본 논문에서는 분산음성인식 환경에서 채널 환경에 따라 스케일러블 구조를 가지면서도 음성 부호화기 수준의 고품질 음성 복원이 가능한 방법을 제안하였다. 제안한 방식에서는 ETSI DSR 표준 음성복원 정보이외에 추가로 음성부호화기 구조로부터 얻은 부가 정보를 스케일러블 패킷 형태로 구성하여, 네트워크 환경에 따라 4.8 kbp/s에서 11.4 kbp/s의 비트전송률을 제공한다. 복원된 음질 성능 평가에서는 오류가 없는 환경과 프레임 오류가 있는 환경에서 모두 제안한 방식이 G.729와 유사한 음질을 보여 우수한 성능을 보였다.

감사의 글

본 연구는 정보통신부 및 정보통신연구진흥원의 IT신성장동력핵심기술개발사업 [신성장동력산업용 대용량/대화형 분산/내장처리 음성인터페이스 기술 개발]의 일환으로 수행하였음.

참고문헌

[1] ETSI ES 202 211, v.1.1.1, *Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Extended frond-end feature extraction algorithm; Compression algorithms*, Nov. 2003.

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

[3] J. S. Yoon, G. H. Lee, H. K. Kim, "A MFCC-based CELP speech coder for server-based speech recognition in network environments," *IEICE Trans. on Electronics, Communications and Computer Sciences*, vol. E90-A, no. 3, pp. 626-632, 2007.