잡음환경에서의 음성인식 성능 향상을 위한 이중채널 음성의 CASA 기반 전처리 방법

*박지훈, 윤재삼, 김홍국
광주과학기술원 정보통신공학과
e-mail : [jh_park, jsyoon, hongkook]@gist.ac.kr

CASA-based Front-end Using Two-channel Speech for the Performance Improvement of Speech Recognition in Noisy Environments

*Ji Hun Park, Jae Sam Yoon, and Hong Kook Kim
Department of Information and Communication
Gwangju Institute of Science and Technology

Abstract

In order to improve the performance of a speech recognition system in the presence of noise, we propose a noise robust front-end using two-channel speech signals by separating speech from noise based on the computational auditory scene analysis (CASA). The main cues for the separation are interaural time difference (ITD) and interaural level difference (ILD) between two-channel signal. As a result, we can extract 39 cepstral coefficients are extracted from separated speech components. It is shown from speech recognition experiments that proposed front-end has outperforms the ETSI front-end with single-channel speech..

I. 서론

홈 오토메이션, 텔레메틱스, 지능형 로봇 등의 발전과 더불어 이를 효율적으로 조정할 수 있는 인터페이스로 음성인식의 중요성이 증대되고 있다. 그러나 핵심 인터페이스 수단으로 사용되기 위해서는 음성인식시스템이 잡음환경에서의 안정적인 인식성능을 가져야 한다.

본 논문에서는 잡음환경에서의 음성인식 성능을 향상시키기 위해 사람의 청각특성을 반영한 computational auditory scene anlysis (CASA) 기반의 전처리 (front-end) 방법을 제안한다. 특히 제안된 전처리 방법은 이중마이크 환경에서 녹음된 이중채널의 음성신호를 입력으로 하며, 각 채널 별 음성신호의 시간 및 크기차이를 이용하여 특정방향에 위치한 음성신호를 CASA 시스템을 통해 잡음으로부터 분리해 낸다.

II. CASA 기반 전처리 방법

제안된 전처리 구성도는 그림 1과 같다. 우선 이중채널의 음성신호는 각 채널별로 인간의 청각특성을 반영한 gammatone 필터뱅크 [1]를 통해 시간-주파수 영역별 신호로 변환되고, 변환된 신호로부터 각 채널 별로 주파수 포락선 정보를 추출한다. 또한 채널 간의 방향정보(ITD, ILD)를 구하여 잡음으로부터 음성신호의 분리에 사용되는 마스크 패턴을 추정한다[2]. 그리고 나서 추정된 마스크 패턴으로 부터 잡음신호의 영향이 최소화된 음성신호의 새로운 포락선을 구하고, 이 포락선에 로그변환을 가

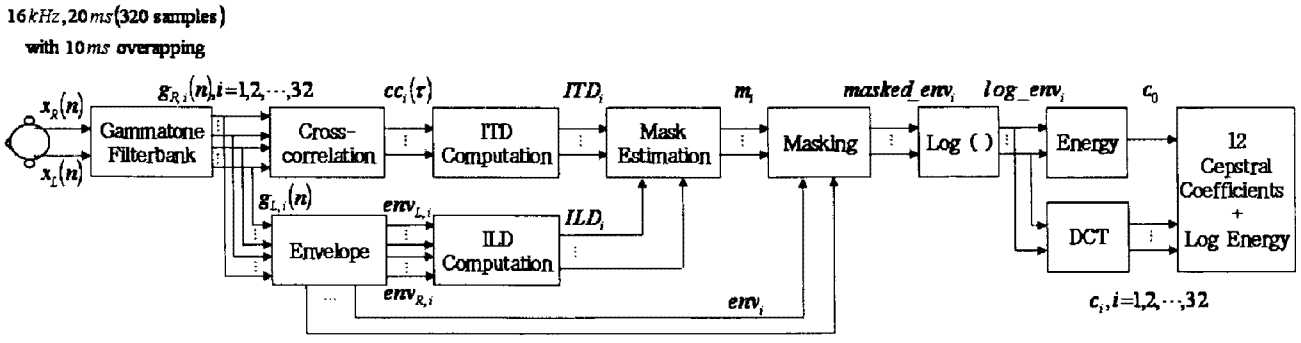

그림 1. CASA 기반 전처리 구성도

한다. 마지막으로 discrete cosine transform (DCT) 변환을 통해 음성인식에 사용될 12차 켑스트럼 계수와 에너지 파라메터를 구하고, 이 들 13개 파라메터에 대해 차분 및 차분의 차분 계수들을 더하여 총 39차의 특징벡터를 추출한다.

Ⅲ. 음성인식 실험 및 결과

제안된 전처리 방법의 성능을 검증하기 위해 음성인식 실험을 실시하였다. 이중채널 환경에서의 인식성능을 평가하기 위해 ETRI 한국어 헤드셋 인식용 단어DB [3]를 사용하여 인위적으로 이중채널용 음성DB를 구축하였다. 음성인식 시스템의 학습에는 18,240개의 단어음성을, 인식 테스트에는 570개의 단어음성을 각각 사용하였으며, 각 음성신호에 대해 0°에 위치하게 하는 머리 전달함수를 적용하였다 [4]. 또한 여성의 낭독체 음성과 음악을 잡음신호로 사용하여, 10°, 20° 및 40°에 위치하도록 하는 머리 전달함수를 적용하여 음성신호와 방향이 전환된 잡음신호를 더해 이중채널용 테스트 잡음음성 DB를 제작하였다. 이때 잡음은 0 dB, 10 dB 및 20 dB의 신호 대 잡음비(SNR)를 갖도록 가공하였다.

음성인식 시스템은 트라이폰 (triphone) 단위의 hidden Markov model (HMM)을 기반으로 하며, 각 트라이폰은 3개의 상태 (state)를 갖는 left-to-right로 표현되었다. 이때 각 상태는 4개의 Gaussian mixture를 가지며, 결정트리(decision tree)를 통해 트라이폰들의 상태를 결합하여 총 2,296개의 상태를 갖는 음향모델을 구성하였다. 인식 시스템에서 사용된 어휘수는 2,250 단어이며 유니그램을 사용하였다.

그림 2는 제안된 전처리 방법과 ETSI 표준 전처리 방법 [5]의 인식성능을 비교한다. 그림 2(a)는 여성의 낭독음을 잡음으로 하였을 경우, 각 SNR에서 잡음의 위치에 따른 인식성능을 비교하며, 그림 2(b)는 여성의 낭독음 이외에 음악을 잡음으로 할 경우의 인식성능을 비교한다. 특히, 그림 2(b)의 결과는 잡음이 원음에 비해 40°일 때를 비교한다. 따라서 그림 2(b)의 왼쪽 결과는 그림 2(a)의 각 40°에서의 결과와 동일하다. 그림 2(a)에서 보는 바와 같이 여성의 낭독체 음성을 잡음으로 하는 환경에서, SNR이 10 dB이고 음성신호와 위치차이가 40°일 때 72.3%의 인식률을 보였으며, 이는 ETSI 표준 전처리 방법에 비해 50.6%의 오인식률 개선에 상당한 결과였다. 또한, 그림 2(a)와 2(b)에서 볼 수 있듯이 각기 다른 음성신호와 잡음 간의 위치차이, SNR, 잡음종류 등의 요소들의 결과를 평균한 결과, 제안된 전처리 방법이 ETSI 표준 전처리 방법에 비해 평균적으로 13.4% 오인식률의 개선을 보였다.

Ⅳ. 결론

인간의 청각 시스템은 잡음환경에서도 강인한 특성을 지닌다. 본 논문에서는 잡음환경에서의 음성인식 성능 향상을 위해 청각특성을 반영한 전처리 방법을 제안하였다. 제안된 전처리 방법은 이중채널 음성의 채널간의 시간 및 크기 차이를 이용하여, 특정방향에 위치한 잡음음성을

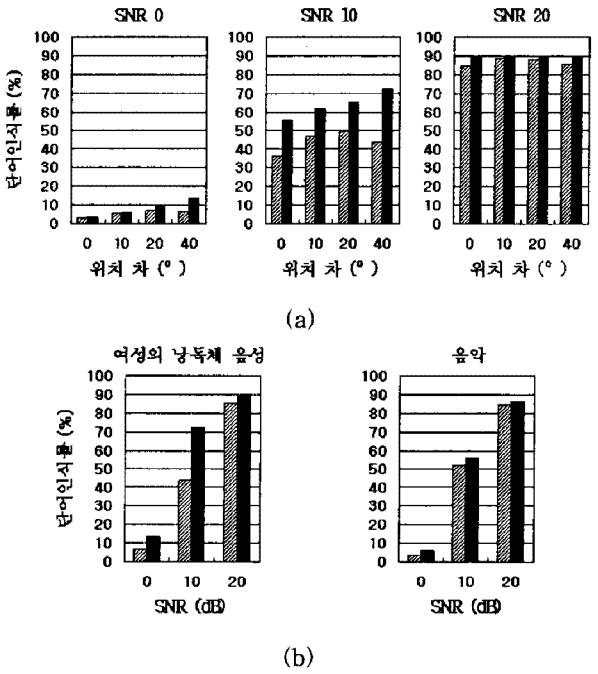

그림 2. 인식성능: (a) 여성의 낭독체 음성 잡음 환경에서 음성과 잡음의 위치 차에 따른 인식성능 (b) 잡음종류에 따른 인식성능 (범례 : ▩ ETSI 표준 전처리 방법, ■ CASA 기반 전처리 방법)

노이즈 마스크 패턴을 이용해 잡음으로부터 분리해냈다. 제안된 방법의 성능검증을 위해 음성인식 실험을 실시하였으며, 그 결과 기존 ETSI 표준 전처리 방법의 성능 대비 13.4%의 오인식률 개선을 보였다.

감사의 글

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

참고문헌

[1] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," *Appendix B of SVOS Final report: The Auditory Filterbank*. APU Report 2341, 1987.

[2] K. Palomaki, G. Brown, and D. Wang, "A binaural processor for missing data speech recognition in the presence of noise ans small-room reverberation," *Speech Communication*, vol. 43, no. 8, pp. 361-378, Mar. 2004.

[3] 김상훈, 오승신, 정호영, 전형배, 김정세, "공통음성 DB 구축," *한국음향학회 하계학술대회논문집*, 제21권, 제1(s)호, pp. 21-24, July 2002.

[4] W. Gardner and K. Martin, "HRTF measurements of a KEMAR," *J. Acoust. Soc. Am.*, vol. 97, no. 6, pp. 3907-3908, June 1995.

[5] ETSI ES 201 108, v.1.1.3, *Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms*, Sept. 2003.