JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 18, No. 10 : 2367~2374 Oct. 2014

이기종 음성 인식 시스템에 독립적으로 적용 가능한 특징 보상 기반의 음성 향상 기법

김우일*

Speech Enhancement Based on Feature Compensation for Independently Applying to Different Types of Speech Recognition Systems

Wooil Kim*

School of Computer Science & Engineering, Incheon National University, Incheon 406-772, Korea

요 약

본 논문에서는 이기종 음성 인식 시스템에 독립적으로 적용할 수 있는 음성 향상 기법을 제안한다. 잡음 환경 음성 인식에 효과적인 것으로 알려져 있는 특징 보상 기법이 효과적으로 적용되기 위해서는 특징 추출 기법와 음향 모델이 음성 인식 시스템과 일치해야 한다. 상용화된 음성 인식 시스템에 부가적으로 전처리 기법을 적용하는 상황과 같이, 음성 인식 시스템에 대한 정보가 알려져 있지 않은 상황에서는 기존의 특징 보상 기법을 적용하기가 어렵다. 본 논문에서는 기존의 PCGMM 기반의 특징 보상 기법에서 얻어지는 이득을 이용하는 음성 향상 기술을 제안한다. 실험 결과에서는 본 논문에서 제안하는 기법이 미지의 (Unknown) 음성 인식 시스템 적용 환경에서 기존의 전처리 기법에 비해 다양한 잡음 및 SNR 조건에서 월등한 인식 성능을 나타내는 것을 확인한다.

ABSTRACT

This paper proposes a speech enhancement method which can be independently applied to different types of speech recognition systems . Feature compensation methods are well known to be effective as a front-end algorithm for robust speech recognition in noisy environments. The feature types and speech model employed by the feature compensation methods should be matched with ones of the speech recognition system for their effectiveness. However, they cannot be successfully employed by the speech recognition with “unknown” specification, such as a commercialized speech recognition engine. In this paper, a speech enhancement method is proposed, which is based on the PCGMM-based feature compensation method. The experimental results show that the proposed method significantly outperforms the conventional front-end algorithms for unknown speech recognition over various background noise conditions.

키워드 : 음성 향상, 특징 보상, 음성 인식, 잡음 환경, 미지 시스템

Key word : Speech enhancement, Feature compensation, Speech recognition, Noisy environment, Unknown system

접수일자 : 2014. 07. 11 심사완료일자 : 2014. 07. 16 게재확정일자 : 2014. 07. 31

* **Corresponding Author** Wooil Kim(wikim@incheon.ac.kr, Tel:+82-32-835-8459)
School of Computer Science and Engineering, Incheon National University, Incheon 406-772, Korea

Open Access http://dx.doi.org/10.6109/jkiice.2014.18.10.2367 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

음성 인식 기술은 Google의 Voice Search, Apple의 Siri 등의 성공적인 출시와 더불어 일반 사용자의 관심이 증대되고 있으며, 자동차의 내비게이션과 내부 조작을 위한 음성 명령 장치, 로봇 인터페이스 등에 다양한 종류의 애플리케이션들이 상용화되어 등장하고 있다. 특히 최근 자동차 업계의 가장 큰 이슈가 되고 있는 텔레매틱스 (Telematics) 용 단말 장치의 인터페이스를 위해 음성 인식 기술의 적용을 시도하고 있으나, 아직 만족할만한 수준이 되지 못하고 있다. 그 주요한 이유는 자동차 주행 환경이 엔진 소음, 바람 소리, 지면 마찰에 의한 소음, 차량 내부의 소리 울림 등 입력 음성의 음향적 특성을 왜곡시키는 요소들로 이루어졌기 때문이다.

음성 인식 시스템의 인식 성능이 저하되는 가장 큰 원인 중 하나는 인식 시스템에 장착되어지는 음향 모델을 훈련하는 환경과 실제 시스템을 적용하는 환경이 음향학적 측면에서 불일치 (Mismatch) 한다는 점이다. 이러한 음향학적 불일치를 줄이고 음성 인식 성능 향상을 위해 다양한 연구가 진행되어 왔다[1-8]. 이러한 연구는 두 가지 측면으로 나눌 수 있는데, 하나는 음성 인식 시스템의 전처리 단계에서 음성 신호로부터 잡음을 제거하고 음성을 향상시키거나, 잡음에 강인한 음성 특징 (Feature)을 추출하거나, 또는 특징 영역에서 잡음을 제거하거나 보상 (Compensation)하는 방법이다. 이러한 기법에는 주파수 차감법 (Spectral Subtraction)[1], 켑스트럼 평균 정규화 (Cepstrum Mean Normalization, CMN), 다양한 종류의 특징 보상 (Feature compensation) 기법[4, 5] 등이 포함된다. 두 번째 접근 방법은 이미 훈련되어진 음향 모델을 새로운 잡음 환경과 일치하도록 적응 (Adaptation) 해주는 기법이다. 최대 사후 확률 (Maximum A Posteriori, MAP) 예측법[6], 최대 우도 선형 회귀 (Maximum Likelihood Linear Regression, MLLR) 기법[7], 병렬 모델 결합 기법 (Parallel Model Combination, PMC)[8] 등이 이 접근 방법에 속한다.

일반적으로 특징 보상 기반의 전처리 기술은 음질 향상 기술에 비해 음성 인식 성능 향상에 월등히 효과적인 것으로 알려져 있으나, 특징 기반의 전처리 기술은 음성 인식 시스템과 밀접하게 연동되어 개발 및 작동되어야 효과적으로 동작할 수 있다. 즉, 음성 인식 시스템과 동일한 음성 특징 추출 기법을 사용해야한다. 또한 많은 특징 보상 기술이 가우시안 혼합 모델 (Gaussian Mixture Model, GMM)과 같은 음성 음향 모델을 채용하고 있는데, 음성 모델을 채용한 전처리 기술의 경우에는 음성 인식 시스템의 음성 음향 모델인 은닉 마르코프 모델 (Hidden Markov Model, HMM)과 동일한 음향적 특성을 갖는 음성 모델을 채용해야 한다.

하지만 대부분의 상용 음성 인식 시스템은 이러한 내부 동작 방식, 음향 모델의 특성 등이 공개되어 있지 않기 때문에 독립적으로 개발된 특징 보상 기반의 전처리 기술을 그대로 적용하기가 어렵다. 또한 같은 이유로 인해 특정 음성 인식 시스템과 연동되어 개발된 특징 기반의 전처리 기술은 다른 인식 시스템에 적용했을 때 최적의 성능을 나타내기가 어렵다. 본 논문에서는 이 기종 음성 인식 시스템에 독립적으로 적용할 수 있는 특징 보상을 기반으로 하는 음성 전처리 기술을 제안하고자 한다. 따라서 음성 인식 시스템의 세부 정보가 알려져 있지 않은 미지의 (Unknown) 음성 인식 시스템 적용 조건에서 효과적인 성능을 가지는 음성 향상 기술을 제안한다.

본 논문은 다음과 같이 이루어져있다. II장에서는 미지의 음성 인식 시스템 환경에서 기존의 전처리 기술을 적용하는 한계점을 기술한다, III장에서는 본 논문에서 제안하는 음성 향상 기술의 기반이 되는 병렬 결합된 가우시안 혼합 모델 (Parallel Combined GMM, PCGMM) 기반의 특징 보상 기법[5]에 대해 설명하고, IV장에서는 제안하는 음성 향상 기술에 대해 설명한다. V장에서 실험과 결과를 기술하고, VI장에서 논문의 결론을 맺는다.

Ⅱ. 미지의 (Unknown) 음성 인식 시스템 조건에서 기존 전처리 기술 적용의 한계

본 논문에서는 상용화된 음성 인식 시스템을 채용하는 상황과 같이 시스템의 내부 동작 방식, 음향 모델의 특성 등이 알려져 있지 않은 상황에서 전처리 기술을 추가적으로 적용하는 상황을 미지의 (Unknown) 음성 인식 시스템 적용 조건으로 정의한다. 이러한 Unknown 음성 인식 시스템 조건에서 기존 전처리 기

술을 적용하고자 할 때 발생하는 한계점에 대해 자세히 기술한다.

2.1. 음성 향상 기법

주파수 차감법[1, 9]과 같은 음성 향상 기법은 향상된 음성 신호의 파형을 출력으로 하기 때문에 Unknown 음성 인식 시스템에 적용하는 데 큰 제약이 없다. 단, 기존의 특징 보상 기법에 비해 상대적으로 낮은 성능을 보이는 단점을 가진다.

2.2. 음성 특징 기반의 전처리 기법

특징 기반의 전처리 기법을 적용하기 위해서는 음성 인식 시스템의 구조가 특징 추출 단계와 인식 단계로 모듈화 되어 있어야 한다. 본 논문에서는 Unknown 음성 인식 시스템이 이와 같이 모듈화 되어 있는 것을 가정한다.

인식 시스템이 모듈화 되어 있다고 해도, 특징 기반의 전처리 기법을 적용하는 것은 쉽지 않다. 대표적인 특징 보상 기법의 하나인 Vector Taylor Series(VTS) 알고리즘은 로그 스펙트럼 도메인 (즉, 필터뱅크 출력에 log를 취한 것)에 적용된다[4]. 이를 Unknown 음성 인식 시스템에 적용하기 위해서는 VTS 알고리즘을 적용한 후, 음성 인식 시스템과 동일한 특징 도메인으로 변환해야 한다. 하지만, 음성 인식 시스템의 특징 추출 처리 과정에 관한 정보가 알려져 있지 않은 상황에서는 동일한 특징 도메인으로 변환은 불가능하다. 따라서 음성 인식 시스템이 탑재하고 있는 음향 모델과의 불일치가 발생하여 성능 저하를 가져올 것을 예상할 수 있다. 또한, VTS 알고리즘의 경우 GMM으로 훈련된 음향 모델을 채용하게 되는데, 특징 추출 기법이 알려져 있지 않고 음성 인식 시스템의 음향 모델이 알려져 있지 않은 상태에서는 VTS 알고리즘의 GMM과 음성 인식 시스템의 HMM과의 음향 모델 불일치로 인한 성능 하락이 자명하다. 이와 같은 이유로 VTS를 포함하여 PCGMM 기반 특징 보상 기법[5]과 같은 대다수의 음향 모델 기반의 특징 보상 기법들은 Unknown 음성 인식 시스템의 전처리로서 효과적으로 적용하는 것이 불가능 하다.

이 외의 대표적인 전처리 기법의 하나인 European Telecommunications Standards Institute (ETSI) Advanced Front-End (AFE)[10] 기술의 경우에도 음성 인식 시스템과의 일치(Match)된 조건이 매우 중요하다. 즉, 음성 인식 시스템의 음향 모델 역시 ETSI AFE 전처리 기법을 적용한 특징으로 훈련되었을 때 최적의 성능을 나타낼 수 있다. 따라서 ETSI AFE 기법 또한 Unknown 음성 인식 시스템에 적용이 힘들다.

Ⅲ. PCGMM 기반의 특징 보상 기법

본 논문에서는 Unknown 음성 인식 시스템의 전처리로 기법으로 적용하기 위해 PCGMM 기반의 특징 보상 [5]을 이용한 음성 향상 기법을 제안한다. 본 절에서는 제안하는 기법의 기반이 되는 PCGMM 기반의 특징 보상 기법을 설명한다. PCGMM 기반의 특징 보상 기법은 음성 모델을 이용하는 특징 보상 기법의 하나로서, 깨끗한 음성 $\boldsymbol{x}$의 켑스트럼 특징 벡터의 분포를 다음과 같이 K개의 가우시안 요소로 이루어진 가우시안 혼합 모델 (GMM)을 이용하여 표현된다.

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \omega_k N(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x},k}, \boldsymbol{\Sigma}_{\boldsymbol{x},k}) \tag{1}$$

PCGMM 기반 특징 보상 기법에서는 병렬 모델 결합 (Parallel Model Combination) 기법[8]을 이용하여 깨끗한 음성 모델과 잡음 모델을 수학적으로 결합함으로써 잡음에 오염된 음성 모델을 생성한다.

$$(\boldsymbol{\mu}_{\boldsymbol{y},k}, \boldsymbol{\Sigma}_{\boldsymbol{y},k}) = F[(\boldsymbol{\mu}_{\boldsymbol{x},k}, \boldsymbol{\Sigma}_{\boldsymbol{x},k}), (\boldsymbol{\mu}_{\boldsymbol{n}}, \boldsymbol{\Sigma}_{\boldsymbol{n}})] \tag{2}$$

식 (2)에서 $F[\bullet]$은 모델 결합을 위한 함수를 나타내며, 모델 파라미터 $(\boldsymbol{\mu}_{\boldsymbol{n}}, \boldsymbol{\Sigma}_{\boldsymbol{n}})$는 단일 가우시안 확률 분포 함수로 얻어지는 잡음 모델을 나타낸다. PCGMM 기반 특징 보상 기법에서는 모델 결합을 위해 로그-노말 가정법 (Log-normal approximation)을 채용한다. 또한, 오염 음성과 깨끗한 음성 모델의 평균 파라미터 사이에는 다음과 같은 바이어스 변환 관계를 가정한다.

$$\boldsymbol{\mu}_{\boldsymbol{y},k} = \boldsymbol{\mu}_{\boldsymbol{x},k} + \boldsymbol{r}_k \tag{3}$$

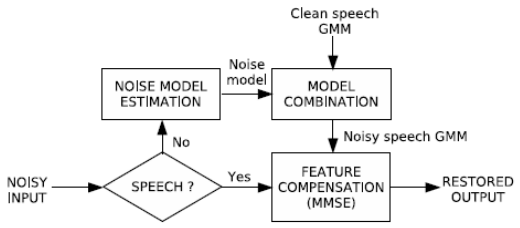

그림 1. PCGMM 기반 특징 보상 기법의 블록 다이어그램[5]
Fig. 1 Block diagram of the PCGMM-based feature compensation scheme[5]

깨끗한 음성 모델과 잡음에 오염된 음성 모델을 이용하여, 최소 평균 제곱 오차 (Minimum Mean Squared Error) 기반의 예측 기법에 의해 다음과 같은 식으로 입력 음성을 깨끗한 음성으로 복구할 수 있다.

$$\tilde{\boldsymbol{x}}_{MMSE} = \int_X \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x} \simeq \boldsymbol{y} - \sum_{k=1}^{K} \boldsymbol{r}_k p(k|\boldsymbol{y}) \quad (4)$$

위 식에서 사후확률인 $p(k|\boldsymbol{y})$는 다음과 같이 계산한다.

$$p(k|\boldsymbol{y}) = \frac{\omega_k p(\boldsymbol{y}|k)}{\sum_{k=1}^{K} \omega_k p(\boldsymbol{y}|k)} \quad (5)$$

그림 1은 PCGMM 기반의 특징 보상 기법의 다이어그램을 나타낸다.

Ⅳ. 제안하는 음성 향상 기법

제안하는 음성 향상 기법에서는 PCGMM 기반의 특징 보상 기법에서 얻어지는 특징 벡터의 이득 (Gain)을 이용한다. PCGMM 기반의 특징 보상 기법은 켑스트럼 도메인에서 이루어지므로, 입력 음성 파형에 적용하기 위해서는 도메인 변환이 이루어져야 한다. 본 연구에서 이득은 다음과 같이 DCT (Discrete Cosine Transform) 역변환을 통해 로그 스펙트럼 도메인에서 얻어진다.

$$\boldsymbol{g} = \boldsymbol{C}^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{x}}_{MMSE}) \quad (6)$$

식 (6)에서 $\boldsymbol{y}$와 $\tilde{\boldsymbol{x}}_{MMSE}$는 PCGMM 특징 보상 기법에서의 입력 음성과 식 (4)에 의해 얻어지는 깨끗한 음성을 말하며, 켑스트럼 특징 벡터이다. 또한 $\boldsymbol{C}$는 DCT 변환 행렬을 나타낸다. 이와 같이 얻어진 이득 벡터 $\boldsymbol{g}$는 입력된 오염 음성 파형의 스펙트럼에 다음과 같이 적용되어 깨끗한 음성을 복구한다.

$$\tilde{X}(k,t) = \frac{Y(k,t)}{\exp(g_i)}, \text{ if } k \in i\text{th Mel-filterbank} \quad (7)$$

위 식에서 $Y(k,t)$와 $\tilde{X}(k,t)$는 각각 입력 오염 음성과 향상된 깨끗한 음성의 시간 t에서의 k번째 주파수 요소를 나타낸다.

이와 같이 제안 하는 음성 향상 기법은 향상된 음성 파형을 출력으로 하기 때문에 특징 추출 및 음향 모델의 불일치에 대한 우려없이 미지의 음성 인식 시스템에 적용이 가능하다. 그림 2는 본 논문에서 제안하는 PCGMM 특징 보상을 기반으로 하는 음성 향상 기법의 블록 다이어그램을 나타낸다.

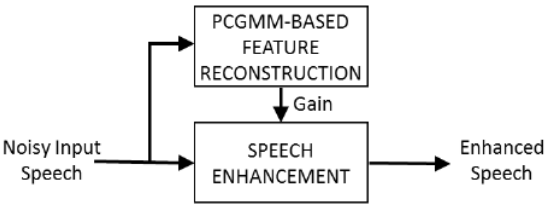

그림 2. 본 논문에서 제안하는 PCGMM 기반 특징 보상 기법을 채용한 음성 향상 기법의 블록 다이어그램
Fig. 2 Block diagram of the proposed speech enhancement scheme employing the PCGMM-based feature compensation method

Ⅴ. 실험 및 결과

5.1. 실험 환경 및 베이스라인 시스템

객관적인 성능 평가를 위해서 Aurora 2.0에서 제공하는 평가 방식을 따랐다. Aurora 2.0의 평가 방식의 주요

특징은 다음과 같다[11].

- 영어 음성, 연속 숫자음 인식, 11단어+묵음 구간(Silence)+짧은 휴지(Short pause)
- ETSI 표준 방식의 Mel-Frequency Feature Extraction (MFCC) 특징 추출[12]
- 13차 static 특징(c1~c12+로그 에너지) 추출 후 인식단에서 미분계수 추출(총 39차) : 본 논문의 실험에서는 PCGMM 구현의 편의를 위하여 로그 에너지 대신 켑스트럼의 0차 계수를 사용하였다.
- 3-mixture, 16-state의 단어 모델, 2종류의 silence 모델

Aurora 2.0에서 제공하는 Clean-condition Training, Multi-condition Testing 방식에 따라 음향 모델은 깨끗한 환경에서 수집된 8,840개의 음성 데이터를 이용하여 훈련하였다. 본 논문에서는 Aurora 2.0의 SetA에 포함되어 있는 지하철, 자동차, 웅성거림 (Speech Babble) 외에, 시간에 따라 변하는 잡음 환경을 반영하기 위해 배경 음악을 잡음 환경으로 사용하였다. 음악 오디오 샘플에 5가지의 신호 대 잡음 비 (SNR, 20, 15, 10, 5, 0dB)에 따라 부가적으로 오염시켜 새로운 테스트 데이터베이스를 제작하였다. 배경 음악은 비트와 빠르기가 다양한 유명 한국 가요 10곡의 전주 부분에서 샘플링하였다.

제안하는 음성 향상 알고리즘과의 성능 비교를 위해 대표적 전처리 알고리즘으로 가장 일반적으로 사용되는 주파수 차감법 (Spectral Subtraction, SS)과 켑스트럼 정규화 (Cepstral Mean Normalization, CMN) 기법을 선택하였다. 주파수 차감법에서는 배경 잡음을 추정하기 위해 250msec의 시간 지연을 갖는 최소 통계 (Minimum statistics) 기법을 적용하였다[9]. 성능 비교를 위해 기존의 대표적인 특징 보상 기법인 Vector Taylor Series (VTS) 기반 알고리즘을 평가하였다[4]. VTS 기법에서는 EM (Expectation Maximization) 기법을 이용하여 적응적으로 잡음 성분을 추정하는 것으로 알려져 있다. 또한 ETSI에서 개발한 Advanced Front-End (AFE) 알고리즘도 최신 기법의 하나로 평가하였다[10]. AFE에서는 반복적인 Wiener 필터와 Blind Equalization 기법을 채용하여 잡음 환경에서 음성 인식 성능을 높이는데 월등한 성능을 가지는 것으로 알려져 있다.

5.2. "Known" 음성 인식 시스템 조건에서의 성능 평가

본 논문에서는 음성 인식 성능의 지표로 단어 오인식율 (Word Error Rate, WER)을 사용하였다. 표 1과 그림 3은 "Known" 자동 음성 인식 (Automatic Speech Recognition, ASR) 시스템에 대한 성능 평가 결과이다. 본 연구에서는 ASR 시스템에 대한 정보가 충분히 알려져 있어서 ASR 시스템과 동일한 특징 추출 기법을 적용할 수 있는 경우를 "Known" ASR 시스템이라 가정하였다. 또한, "Known" 시스템 조건에서는 ASR 시스템의 음향 모델 (즉, HMM) 훈련에 사용한 것과 동일한 음성 데이터베이스를 사용할 수 있다고 가정하여, VTS나 PCGMM과 같이 음향 모델을 사용하는 특징 보상 기법에서 음향 모델 훈련에 동일한 음성 데이터를 사용하는 것이 가능한 것을 가정하였다. 따라서, "Known" ASR 적용 조건은 ASR 시스템을 개발하는 개발자가 직접 전처리 기법을 개발하는 것을 가정하는 상황으로 볼 수 있으며, 많은 전처리 기법 논문들의 실험 조건이 이와 같은 가정을 사용한다.

위에서 언급한 것과 같이 "Known" ASR 시스템 환경에서는 음성 특징과 음향 모델을 동일하게 (Match) 사용할 수 있는 전제 하에 각 전처리 기법 별로 다음과 같이 적용하였다.

- SS+CMN: CMN을 적용하기 위해서는 MFCC 특징 추출이 필요하므로, ASR 시스템과 동일한 ETSI MFCC를 추출하여 CMN을 적용하였다. 이 경우 ASR 시스템의 음향 모델도 CMN을 적용하여 훈련된 것을 가정하였다.
- VTS: 로그 스펙트럼 도메인 (즉, 특징 추출 단계에서 필터뱅크 분석 결과에 log를 취한 것)으로 변환하여 VTS 알고리즘을 적용한 후 다시 켑스트럼 도메인으로 변환하여 ASR 시스템에 입력하였다. 따라서 ASR 시스템과 동일한 ETSI MFCC 특징 추출 알고리즘을 사용하여 VTS 알고리즘을 적용하였다. 또한, VTS 알고리즘에 필요한 음향 모델은 ASR 시스템 훈련에 사용한 동일한 Aurora 2.0의 깨끗한 음성 데이터베이스 8,840개를 사용하여 로그 스펙트럼 도메인에서 훈련하여 얻어진 GMM을 사용하였다.
- ETSI-AFE: ASR 시스템의 음향 모델을 ETSI-AFE 전처리 기법을 적용한 특징을 사용하여 훈련하였다.
- PCGMM: ETSI MFCC 특징 추출 기법을 사용하였고, ASR 시스템과 동일한 Aurora 2.0 훈련 음성 데이

터베이스를 사용하여 얻어진 GMM을 사용하였다.

- Proposed: Gain 예측에 필요한 PCGMM 특징 보상 기법을 적용하기 위해 위 PCGMM 적용과 동일한 조건을 사용하였다 (즉, ETSI MFCC 특징, Aurora 2.0 데이터를 이용한 GMM 훈련)

표 1은 본 논문에 사용한 Aurora 2.0 데이터베이스의 4개의 잡음 환경에 대해 모든 SNR 조건을 평균한 성능이다. 결과로부터 알 수 있듯이 각 전처리 기법을 적용했을 때 상당한 성능 향상이 있는 것을 알 수 있다. 본 실험에서는 ETSI AFE와 PCGMM 기법이 가장 좋은 성능을 나타내었다. 또한 본 논문에서 제안한 음질 향상 기법 (Proposed)도 ETSI AFE 기법에 대응할 만한 성능을 보이는 것을 확인하였다. 그림 3은 각 SNR 조건에 대해 모든 잡음 환경을 평균한 성능 결과이다.

5.3. "Unknown" 음성 인식 시스템 조건에서의 성능 평가

표 2와 그림 4는 "Unknown" ASR 시스템 조건에서의 성능평가 결과이다. 본 연구에서 "Unknown" ASR 시스템 조건은 ASR 시스템에 대한 정보가 알려져 있지 않은 상태로 전처리 기법을 개발하는 조건을 가정하였다. 즉, 상용화된 음성인식 엔진을 그대로 가져다 쓰고, 이에 추가적으로 전처리 기법을 적용하고자 하는 것과 유사한 상황을 가정하였다.

이러한 조건을 시뮬레이션하기 위해 음성 인식 시스템은 앞에서 설명한 것과 동일한 시스템을 사용하였다. 즉, ETSI MFCC 특징 추출을 사용하고, 음향 모델 훈련에는 Aurora 2.0의 깨끗한 음성 데이터를 사용하였다. 음성 인식 시스템에 대한 정보가 알려져 있지 않은 상황을 가정하므로, 전처리 기법은 음성인식기와 상이한 HTK[13]로 구현한 특징 추출 기법을 사용하였다. HTK로 구현한 특징 추출 기법에서 필터뱅크의 개수 (23개)와 켑스트럼의 차수 (13)를 ETSI MFCC와 동일하게 적용하였다.

VTS와 PCGMM 전처리 기법과 제안한 기법에 필요한 음향 모델도 음성인식기 훈련에 사용된 것과 다른 TIMIT 데이터베이스를 사용하여 훈련함으로써 불일치 (Mismatch)되는 상황을 의도적으로 구현하였다. TIMIT 음성 데이터는 462명의 서로 다른 화자로부터 취득된 4,620개의 발음으로 구성되어 있고 이는 총 4.1 시간에 해당되는 분량이다.

표 1. "Known" ASR 조건에서 인식 성능 평가 (WER, %)
Table. 1 Speech recognition performance with the "known" ASR system condition (WER, %)

	Subway	Car	Babble	Music	Avg.
No processing	31.23	35.65	46.98	34.22	37.02
SS+CMN	13.82	15.41	20.18	24.62	18.51
VTS	14.15	14.93	21.10	25.73	18.98
ETSI-AFE	7.70	7.18	16.36	21.14	13.10
PCGMM	8.58	7.97	18.23	15.17	12.49
Proposed	**9.24**	**8.87**	**19.97**	**15.98**	**13.51**

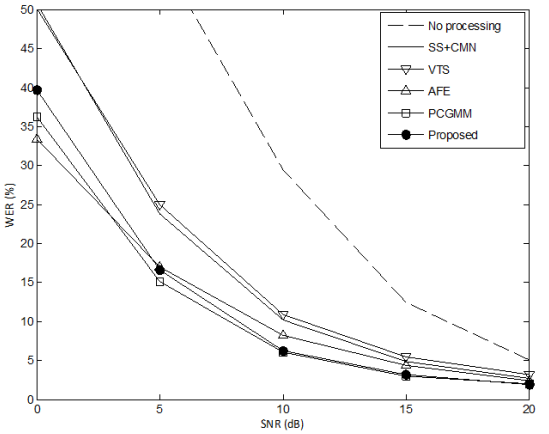

그림 3. "Known" ASR 조건에서 SNR에 따른 인식 성능 평가 (WER, %)
Fig. 3 Speech recognition performance over different SNRs with the "known" ASR system (WER, %)

따라서 각 전처리 기법은 음성 특징 추출 기법과 필요에 따라 사용된 음향 모델이 음성인식기와 Mismatch 되는 성질을 갖는다.

표 2는 "Unknown" ASR 조건에서 각 잡음 환경에 대해 모든 SNR 조건을 평균한 결과이다. 아무런 처리를 하지 않은 경우, "Known ASR" 조건에 비해 소폭 상승한 결과를 보인다. 이는 HTK로 구현한 MFCC 특징 추출 처리 단계에서 기본 세팅값으로 적용된 필터링, 정규화 등의 과정이 잡음 환경에 의한 Mismatch를 감소시킨 것으로 판단된다. 예상과 같이 실험에 사용된 모든 전처리 기법이 "Known" ASR 조건과 비교하여 대폭 하락하는 것을 알 수 있다. "Unknown" ASR 조건에서는 ETSI-AFE 특징에 대한 결과는 제외되었는데, 이는

ETSI-AFE 특징이 음성 인식기의 음향 모델과 상이할 경우에는 잡음 환경에 관계없이 극심한 성능 저하를 나타내기 때문이다.

제안한 음성 향상 기법은 다른 전처리 기법들에 비해 “Unknown” ASR 조건에서 가장 좋은 인식 성능을 보이고, “Known” ASR 조건과 비교하여 상대적으로 낮은 성능 하락을 보인다. 그림 4에서와 같이 SNR 변화에 따른 성능 평가에서도 모든 SNR에 대해 제안한 기법이 다른 전처리 기법들보다 낮은 WER을 나타내는 것을 알 수 있다. 특히, “Known” ASR 조건에서 가장 일반적으로 많이 사용되는 주파수 차감법과 CMN의 결합 (SS+CMN) 보다도 우수한 성능을 보이는 것은 주목할 만한 결과이다. 이와 같은 결과는 본 논문에서 제안하는 특징 보상을 이용한 음질 향상 기법이 사양이 알려져 있지 않은 음성 인식시스템의 전처리 기법으로서 효과적으로 사용될 수 있음을 입증하는 것이다.

Ⅵ. 결 론

본 논문에서는 이기종 음성 인식 시스템에 독립적으로 적용할 수 있는 음성 향상 기법을 제안하였다. 잡음 환경 음성 인식에 효과적인 것으로 알려져 있는 특징 보상 기법이 효과적으로 적용되기 위해서는 특징 추출 기법과 음향 모델이 음성 인식 시스템과 일치해야 한다. 상용화된 음성 인식 시스템에 부가적으로 전처리 기법을 적용하는 상황과 같이, 음성 인식 시스템에 대한 정보가 알려져 있지 않은 상황에서는 기존의 특징 보상 기법을 적용하기가 어렵다. 본 논문에서는 기존의 PCGMM 기반의 특징 보상 기법에서 얻어지는 이득을 이용하는 음성 향상 기술을 제안하였다. 실험 결과에서는 본 논문에서 제안하는 기법이 미지의 (Unknown) 음성 인식 시스템 적용 환경에서 기존의 전처리 기법에 비해 다양한 잡음 및 SNR 조건에서 월등한 인식 성능을 나타내는 것을 확인하였다. 이러한 결과는 제안한 음성 향상 기법이 세부정보가 알려져 있지 않은 이기종 음성 인식 시스템에 독립적으로 적용할 수 있음을 입증하는 것이다.

표 2. “Unknown” ASR 조건에서 인식 성능 평가 (WER, %)
Table. 2 Speech recognition performance with the “unknown” ASR system condition (WER, %)

	Subway	Car	Babble	Music	Avg.
No processing	28.12	31.50	40.33	40.58	35.13
SS+CMN	20.09	22.34	29.27	41.25	28.24
VTS	19.06	18.47	31.67	45.31	28.63
PCGMM	14.27	15.18	26.97	31.11	21.88
Proposed	**11.05**	**11.70**	**27.49**	**16.96**	**16.80**

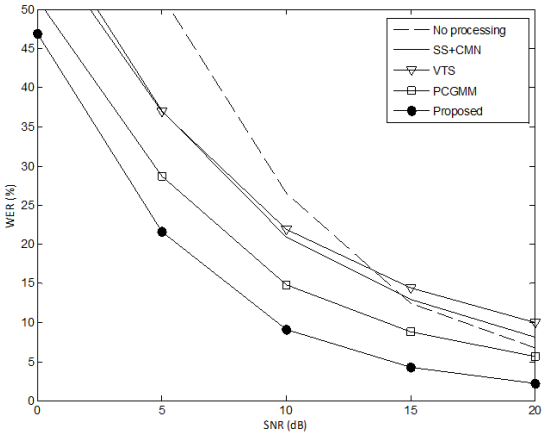

그림 4. “Unknown” ASR 조건에서 SNR에 따른 인식 성능 평가 (WER, %)
Fig. 4 Speech recognition performance over different SNRs with the “unknown” ASR system (WER, %)

감사의 글

이 논문은 산학협동재단 2013년도 신진교수연구비 지원에 의하여 연구되었음.

REFERENCES

[1] S. F. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction,” *IEEE Trans. on Acoustics, Speech and Signal Proc.,* vol.27, pp.113-120, 1979.

[2] Y. Ephraim and D. Malah, “Speech Enhancement Using Minimum Mean Square Error Short Time Spectral Amplitude Estimator,” *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol.32, no.6, pp.1109-1121, 1984.

[3] J. H. L. Hansen and M. Clements, “Constrained Iterative Speech Enhancement with Application to Speech Recognition,” *IEEE Trans. on Signal Proc.*, vol.39, no.4, pp.795-805, 1991.

[4] P. J. Moreno, B. Raj, and R. M. Stern, “Data-driven Environmental Compensation for Speech Recognition: A Unified Approach,” *Speech Communication*, 24(4), pp.267-285, 1998.

[5] W. Kim and J. H. L. Hansen, “Feature Compensation in the Cepstral Domain Employing Model Combination,” *Speech Communication*, 51(2), pp.83-96, 2009.

[6] J. L. Gauvain and C. H. Lee, “Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains,” *IEEE Trans. on Speech and Audio Proc.*, vol.2, no.2, pp.291-298, 1994.

[7] C. J. Leggetter and P. C. Woodland, “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs,” *Computer Speech and Language*, 9, pp.171-185, 1995.

[8] M. J. F. Gales and S. J. Young, “Robust Continuous Speech Recognition Using Parallel Model Combination,” *IEEE Trans. on Speech and Audio Proc.*, vol.4, no.5, pp.352-359, 1996.

[9] R. Martin, “Spectral Subtraction Based on Minimum Statistics,” *EUSIPCO-94*, pp.1182-1185, Sep. 1994.

[10] ETSI Standard Document, ETSI ES 202 050 v1.1.1 (2002-10), 2002.

[11] H. G. Hirsch & D. Pearce, “The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions”, *ISCA ITRW ASR2000*, Sep. 2000.

[12] ETSI standard document, ETSI ES 201 108 v1.1.2 (2000-04), Feb. 2000.

[13] http://htk.eng.cam.ac.uk

김우일(Wooil Kim)

2003년 고려대학교 전자공학과 공학박사
2004년 ~ 2005년 미국 카네기 멜론 대학교 박사후 연구원
2005년 ~ 2012년 미국 텍사스 주립대 (University of Texas at Dallas) 연구원 및 연구교수
2012년 ~ 현재 인천대학교 컴퓨터공학부 조교수
※관심분야 : 신호처리, 패턴인식, 음성인식, 휴먼 컴퓨터 인터페이스