JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 19, No. 5 : 1047~1054 May. 2015

미전사 음성 데이터베이스를 이용한 가우시안 혼합 모델 적응 기반의 음성 인식용 음향 모델 변환 기법

김우일*

Acoustic Model Transformation Method for Speech Recognition Employing Gaussian Mixture Model Adaptation Using Untranscribed Speech Database

Wooil Kim*

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

요 약

본 논문에서는 음성 인식 성능 향상을 위해 미전사된 음성 데이터베이스를 이용한 효과적인 음향 모델 변환 기법을 기술한다. 본 논문에서 기술하는 모델 변환 기법에서는 기존의 적응 기법을 이용하여 환경에 적응된 GMM을 얻는다. HMM의 가우시안 요소와 유사한 요소를 선택하여 선택된 가우시안 요소의 변환 벡터를 구하고 이를 평균 파라미터 변환에 이용한다. GMM 적응 기반의 모델 변환 기법을 기존의 MAP, MLLR 적응 기법과 결합하여 적용한 결과, 자동차 잡음과 음성 Babble 잡음 환경에서 기존의 MAP, MLLR을 단독으로 사용할 경우보다 높은 음성 인식 성능을 나타낸다. 온라인 음향 모델 적응 실험에서도 MLLR과 결합할 경우 기존의 MLLR을 단독으로 사용할 때보다 효과적인 모델 적응 성능을 나타낸다. 이와 같은 결과는 본 논문에서 소개한 GMM 적응 기반의 모델 변환 기법을 채용함으로써 미전사된 음성 데이터베이스를 음향 모델 적응 기법에 효과적으로 활용할 수 있음을 입증한다.

ABSTRACT

This paper presents an acoustic model transform method using untranscribed speech database for improved speech recognition. In the presented model transform method, an adapted GMM is obtained by employing the conventional adaptation method, and the most similar Gaussian component is selected from the adapted GMM. The bias vector between the mean vectors of the clean GMM and the adapted GMM is used for updating the mean vector of HMM. The presented GAMT combined with MAP or MLLR brings improved speech recognition performance in car noise and speech babble conditions, compared to singly-used MAP or MLLR respectively. The experimental results show that the presented model transform method effectively utilizes untranscribed speech database for acoustic model adaptation in order to increase speech recognition accuracy.

키워드 : 음성 인식, 잡음 환경, 모델 적응, 음향 모델, 가우시안 혼합 모델

Key word : Speech recognition, Noisy environment, Model adaptation, Acoustic model, Gaussian mixture model

Received 22 January 2015, **Revised** 16 February 2015, **Accepted** 02 March 2015

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

Open Access http://dx.doi.org/10.6109/jkiice.2015.19.5.1047 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

지난 1980년대 이후로 음성 인식에 관한 연구가 활발히 진행되어 왔으며, 최근 Google의 Voice Search, Apple의 Siri 등의 성공적인 출시와 더불어 일반 사용자의 관심이 증대되고 있다. 스마트폰 사용을 위한 소프트웨어, 자동차의 네비게이션과 내부 조작을 위한 음성 명령 장치, 콜 센터에서 자동 응답 장치, 게임 및 오락을 위한 장치, 로봇 인터페이스 등에 음성 인식을 이용한 다양한 종류의 애플리케이션들이 상용화되어 등장하고 있다. 특히 최근에는 Google 글래스, 스마트 와치와 같은 웨어러블 컴퓨터의 상용화에 따라 보다 정확한 음성 인식 기능이 급격히 요구되고 있다. 하지만, 아직은 그 인식 성능이 일반 사용자의 기대에 미치지 못하는 실정이다.

인식 성능이 저하되는 가장 큰 원인 중 하나는 음성 인식 시스템에 장착되어지는 음향 모델을 훈련하는 환경과 실제 시스템을 적용하는 환경이 음향학적 측면에서 불일치(Mismatch) 한다는 점이다. 즉, 일반적으로 음성 인식 시스템을 위한 음향 모델을 훈련하기 위해서는 다양한 화자(Speaker)에서 발생하는 다양한 종류의 음향적 특성을 보편적으로 나타내기 위한 음성 데이터베이스를 사용하게 된다. 이와 같은 대용량 음성 데이터베이스는 다양한 발화 환경에서 수집하기가 용이하지 않으므로, 일반적으로 잡음이 없는 깨끗한 환경에 수집을 하게 된다. 따라서 훈련된 음향 모델은 깨끗한 발화 환경만을 표현하게 되어 실제 잡음 환경에서는 그 차이로 인한 오류가 불가피해지고 이는 음성 인식 성능 하락의 주요한 원인이 된다.

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.

본 논문에서는 잡음 환경에 강인한 음성 인식을 위한 효과적인 음향 모델 적응 기법을 소개하고 이에 관한 실험 결과를 기술한다. 음향 모델 적응 기술을 적용하기 위해서는 잡음 환경과 일치하는 음향 환경을 가지는 음성 데이터가 필요하며, 모델 적응 기법의 성능을 높이기 위해서는 적응에 사용되는 음성 데이터에 대한 정확한 전사 (Transcription) 정보가 요구된다. 특히 최근 음성 인식 기술의 애플리케이션 분야로 주목 받고 있는 자동차, 스마트폰, 로봇 인터페이스 등의 동작 환경에서는 다양한 종류의 잡음이 존재하고 시간 및 장소에 따라 잡음의 종류가 변하므로 이를 효과적으로 처리할 수 있는 음향 모델 적응 기법이 절실하게 요구된다.

본 논문에서는 미전사 (Untranscribed) 음성 데이터베이스를 이용한 가우시안 혼합 모델(Gaussian Mixture Model, GMM) 적응 기반의 음향 모델 변환 기법을 소개한다[9]. 기존의 전통적인 음향 모델 적응 기법은 음성 인식기의 은닉 마르코프 모델(Hidden Markov Model, HMM)을 새로운 음향 환경에 적응하기 위해 적응 음성 데이터베이스에 대한 정확한 전사 정보를 요구한다. 하지만 다양한 적용 환경을 요구하는 음성 인식 시스템의 경우 적응 환경에 대해 전사된 음성 데이터베이스를 취득하는 것은 많은 수고와 비용을 요구한다. 본 논문의 선행 연구로서 GMM 적응 기반의 HMM 파라미터 변환 기법이 제안되었고[9], 본 논문에서는 이 기법의 원리와 구현 과정에 대해 상세히 소개하며 다양한 조건에서 그 성능을 평가하고 기존의 음향 적응 기법과 비교하고자 한다.

본 논문은 다음과 같이 이루진다. II장에서는 기존의 전통적 음향 모델 적응 기법에 대해 기술한다. III장에서는 미전사 음성 데이터베이스를 효과적으로 이용할 수 있는 GMM 적응을 이용한 음향 모델 변환 기법에 대해 상세히 설명한다. IV장에서는 음성 인식 시스템을 이용한 실험과 결과를 기술하고, V장에서 논문의 결론을 맺는다.

Ⅱ. 음향 모델 적응 기법

음향 모델 적응 기법에서는 음성 인식 시스템의 음향 모델이 훈련된 환경과 실제 인식 시스템이 적용되는 테스트 환경의 배경 잡음 등의 음향적 조건이 동일할 때 최고의 성능을 가지는 것을 가정한다. 따라서 실제 환경과 동일한 음향적 특성을 가지는 음성 데이터를 취득하여 이를 적응된 모델 파라미터를 예측하는데 사용한다. 모델 적응에 사용하는 데이터는 실제 환경과 유사한 환경에서 모델 적응을 위해 별도로 수집을 하거나 실제 입력 음성을 사용한다. 모델 적응 기술은 소량의 음성 데이터로부터 신뢰성이 있는 모델 파라미터를 효과적으로 예측하는 것을 목표로 하며 대표적인 적응 기술로 MAP 기반 적응 기법, MLLR 기반 적응 기법 등이 있다.

2.1. MAP 기반 적응 기법

MAP 예측 기반의 적응 기법에서는 사후 확률 (A Posteriori), 즉 모델 적응에 사용되는 음성 데이터가 주어졌을 때 해당 모델의 확률 값을 최대로 하는 모델 파라미터를 예측하며 다음과 같이 나타낼 수 있다.

$$\begin{aligned}\lambda_{MAP} &= \arg\max_{\lambda} f(\lambda|x) \\ &= \arg\max_{\lambda} f(\lambda) f(x|\lambda)\end{aligned} \tag{1}$$

MAP 기반 적응 기법에서는 모델에 대한 사전 확률(Prior Probability)과 반복적인 EM (Expectation Maximization) 알고리즘을 적용함으로써 “불완전한 (Incomplete)” 데이터로부터 모델 파라미터를 예측할 수 있다. 각 모델이 가우시안 혼합 모델로 표현된다고 가정하면, 각 가우시안 요소의 평균 (Mean) 파라미터는 MAP 기반 적응 기법에 따라 다음의 식으로 갱신된다.

$$\hat{\mu_k} = \frac{\tau_k \mu_k + \sum_{t=1}^{T} \zeta_t(k) x(t)}{\tau_k + \sum_{t=1}^{T} \zeta_t(k)} \tag{2}$$

식 (2)에서 μ_k, τ_k, $\zeta_k(t)$는 각각 적응 이전의 초기 모델의 k번째 가우시안 요소의 평균 벡터, 적응 계수, 입력 데이터 $x(t)$가 주어졌을 때 k번째 가우시안 요소가 발생할 확률을 나타낸다. 따라서 적응 계수 τ_k가 크면 사전 모델에 의존적인 적응 파라미터가 얻어지고, 반대로 적응 계수가 작은 값이면 관찰 값 즉 적응 데이터에 의존적인 파라미터 값을 얻을 수 있다. 적응 계수가 0과 같은 값이면 적응 데이터만을 이용한 최대 우도 (Maximum Likelihood, ML) 기반의 예측 기법과 같아진다.

2.2. MLLR 기반 적응 기법

또 다른 대표적인 적응 기법인 MLLR 기반의 예측 기법에서는 모델 파라미터의 “변환 (transformation)”에 의해 새로운 파라미터를 얻는다. MLLR 기법은 MAP 기법에 비해 보다 소량의 데이터로부터 효과적인 모델 적응 성능을 얻을 수 있는 것으로 알려져 있다. 적응 데이터로부터 선형 Regression 변환을 위한 행렬을 추정하여 취득하고 이를 이용하여 다음과 같은 식으로 평균 벡터를 갱신할 수 있다.

$$\hat{\mu_k} = W\mu_k \tag{3}$$

위 식에서 μ_k는 바이어스 요소를 포함한 확장된 평균 벡터 (n+1차원의 벡터)이고, W는 n×(n+1) 크기를 갖는 행렬이다. 파리미터 변환 행렬 W는 적응 데이터의 우도 (Likelihood)를 최대화하는 EM 알고리즘을 통해서 얻을 수 있다.

Ⅲ. 가우시안 혼합 모델 적응을 이용한 모델 변환 기법

본 논문의 선행 연구로서 전사되지 않은 (Untranscribed) 데이터를 이용한 효과적인 모델 적응 기법[9]에 관한 연구를 소개한다. 선행 연구에서는 전사되지 않은 데이터를 모델 적응에 효과적으로 이용하기 위해 가우시안 혼합 모델(GMM)을 이용한다. 우선 깨끗한 훈련용 음성 데이터를 이용하여 다음과 같이 가우시안 혼합 모델을 얻는다.

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \omega_k N(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x},k}, \boldsymbol{\Sigma}_{\boldsymbol{x},k}) \tag{4}$$

위 식에서 $\boldsymbol{x}$는 깨끗한 음성의 켑스트럼(Cepstrum) 특징 벡터를 나타낸다. 식 (4)에서 취득한 깨끗한 음성의 GMM을 기반으로 기존의 MAP 또는 MLLR 모델 적응 기법을 이용하여 적응용 훈련 데이터로부터 적응된 GMM을 얻을 수 있다. 모델 적응에 HMM 대신 GMM을 음향 모델로 사용하므로 사전 작업에 의해 적응용 데이터에 대한 전사 정보(Transcription)를 얻거나 음성 인식 과정을 통해 발음 정보 취득할 필요가 없다. 모델 적응 후 얻어지는 오염 음성 $\boldsymbol{y}$에 대한 확률 밀도 함수 GMM은 다음과 같이 나타낼 수 있다.

$$p(\boldsymbol{y}) = \sum_{k=1}^{K} \omega_k N(\boldsymbol{y}; \boldsymbol{\mu}_{\boldsymbol{y},k}, \boldsymbol{\Sigma}_{\boldsymbol{y},k}) \quad (5)$$

제안된 기법에서는 위 식과 같이 각 가우시안 요소의 대한 사전 확률 ω_k은 갱신하지 않는 것을 가정한다. 깨끗한 음성을 표현하는 초기 모델과 오염 환경에 적응된 음성 모델의 각 가우시안 요소는 다음과 같이 대응된다.

$$\{\boldsymbol{\mu}_{\boldsymbol{x},k}, \boldsymbol{\Sigma}_{\boldsymbol{x},k}\} \leftrightarrow \{\boldsymbol{\mu}_{\boldsymbol{y},k}, \boldsymbol{\Sigma}_{\boldsymbol{y},k}\} \quad (6)$$

제안된 모델 변환 기법에서는 깨끗한 음성의 켑스트럼 특징 벡터 $\boldsymbol{x}$와 오염된 음성의 켑스트럼 특징 벡터 $\boldsymbol{y}$ 사이에 다음과 같은 관계가 있음을 가정한다.

$$\boldsymbol{y} = \boldsymbol{x} + g(\boldsymbol{x}, \boldsymbol{n}) \quad (7)$$

위 식에서 $\boldsymbol{n}$는 배경 잡음 요소를 나타내고, 함수 $g(\bullet)$는 일반적으로 비선형 함수로 가정하는 깨끗한 음성 및 배경 잡음과의 관계를 나타낸다. 함수 $g(\bullet)$를 상수 벡터로 가정할 경우, 깨끗한 음성 모델과 오염된 음성 모델의 평균 파라미터 사이에는 다음과 같은 관계를 가정할 수 있다.

$$\boldsymbol{\mu}_{\boldsymbol{y},k} = \boldsymbol{\mu}_{\boldsymbol{x},k} + \boldsymbol{r}_k \quad (8)$$

식 (8)과 같이 켑스트럼 도메인에서 가정되는 평균 파라미터의 바이어스 이동 변환은 이전의 다양한 모델 기반의 음성 처리 연구에서 사용된 가정이다[4, 5].

제안된 GMM 적응 기반의 모델 변환 기법에서는 깨끗한 음성 HMM의 각 가우시안 요소와 가장 유사한 가우시안 요소를 식 (4)의 깨끗한 음성 GMM에서 선택한다. 선택된 가우시안 요소와 대응되는 적응된 모델의 가우시안 요소의 평균 파라미터 사이의 바이어스 이동 벡터 $\boldsymbol{r}_k$를 계산하여 이를 HMM 파라미터 갱신에 이용한다. 유사한 가우시안 요소는 식 (9)와 같이 다양한 모델 유사도 측정 기법 $sim(\bullet)$에 의해 선택할 수 있으며, 선행 연구에서는 Kullbeck-Leibler (KL) 거리를 이용하였다[9].

$$k_{s.i}^{sim} = \underset{k}{\arg\max} \{sim(p(\boldsymbol{x}|k), p(\boldsymbol{x}|s,i))\} \quad (9)$$

위 식에서 $p(\boldsymbol{x}|k)$는 $\{\boldsymbol{\mu}_{\boldsymbol{x},k}, \boldsymbol{\Sigma}_{\boldsymbol{x},k}\}$로 구성된 깨끗한 GMM의 k번째 가우시안 요소를 나타내며, $p(\boldsymbol{x}|s,i)$는 깨끗한 HMM의 상태 s의 i번째 가우시안 요소를 나타낸다. 식 (9)에 의해 선택된 가우시안 요소의 평균 벡터를 이용하여 다음과 같이 이동 벡터를 계산한다.

$$\boldsymbol{r}_{k_{s,i}^{sim}} = \boldsymbol{\mu}_{\boldsymbol{y},k_{s,i}^{sim}} - \boldsymbol{\mu}_{\boldsymbol{x},k_{s,i}^{sim}} \quad (10)$$

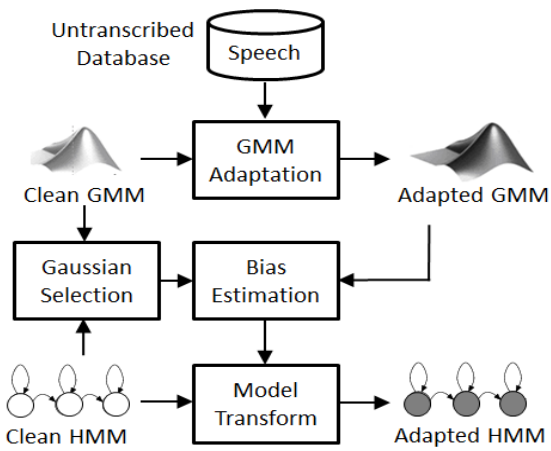

그림 1. 미전사 음성 데이터베이스를 이용한 GMM 적응 기반의 음향 모델 변환 기법의 블록 다이어그램
Fig. 1 Block diagram of the model transformation method employing GMM adaptation using untranscribed speech database

최종적으로 적응된 HMM 모델의 평균 벡터는 식 (11)과 같은 변환에 의해 얻을 수 있다. 분산 행렬은 선택된 가우시안 요소의 분산을 그대로 가져다 쓴다.

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{y},s,i} &= \boldsymbol{\mu}_{\boldsymbol{x},s,i} + \boldsymbol{r}_{k_{s,i}^{sim}} \\ \boldsymbol{\Sigma}_{\boldsymbol{y},s,i} &= \boldsymbol{\Sigma}_{\boldsymbol{y},k_{s,i}^{sim}} \end{aligned} \quad (11)$$

제안된 모델 변환 기법은 그림 1과 같은 블록 다이어그램으로 나타낼 수 있다.

Ⅳ. 실험 및 결과

4.1. 실험 환경 및 음성 인식 시스템

본 논문에서는 잡음 환경에서의 음성 인식 성능 평가를 위해 TIMIT[10] 음성 데이터베이스를 이용하여 잡음 오염 음성 데이터를 생성하였다. TIMIT 데이터베이스는 630명의 서로 다른 화자가 발음한 영어 낭독체 문장이 녹음된 6,300개의 음성 샘플로 구성되어 있으며 이는 약 5.6시간 길이의 녹음 분량에 해당된다. 본 논문에서는 이 중 4,620개의 발음을 음성 인식기 음향 모델의 기본 훈련에 사용하고, 1,000개의 발음은 인식 성능 테스트에, 나머지 680개는 모델 적응을 위한 데이터로 사용했다. 원래의 TIMIT 데이터는 16kHz로 녹음되어 있으나 본 연구에서는 8kHz로 다운 샘플링을 실시하였다. 잡음 환경을 반영하기 위해 잡음 데이터를 인공적으로 첨가함으로써 잡음에 오염된 음성 데이터를 생성하였다. NOISEX-92[11] 데이터베이스에 수록된 자동차 잡음과 음성 Babble 잡음을 10dB의 신호대 잡음비(SNR)로 부가적으로 오염시켰다. 또한 시간에 따라 변하는 배경 음악 잡음 환경을 모의하기 위해 배경 음악 샘플을 10dB SNR로 부가적으로 오염시켜 오염 음성 데이터베이스를 생성하였다. 배경 음악은 비트와 빠르기가 다양한 유명 한국 가요 10곡의 전주 부분에서 샘플링 하였다.

본 연구에서는 음성 인식 성능 평가를 위해 SPHINX3 [12]를 이용하여 음성 인식 시스템을 구축하였다. SPHINX3는 카네기 멜론 대학교에서 제공하는 HMM 기반의 음성 인식 Toolkit이다. TIMIT 데이터베이스를 이용하여 기본 음성 인식기를 구성하였으며, 깨끗한 환경의 훈련 데이터 4,620개를 훈련하여 HMM을 생성하였다. 각 HMM은 Tri-phone 모델을 나타내며 각 음소 모델은 3개의 상태(State)로 구성되고 각 상태는 8개의 가우시안 요소로 구성된 GMM을 출력 확률 함수로 갖는다. 효과적인 음향 모델을 위해 최종적으로 1,138개의 엮인 (Tied) 구조의 HMM을 생성하였다. 기본 인식기는 총 6,233개의 어휘를 가지며 BN (Broadcast News) 언어 모델을 기초로 TIMIT 데이터베이스에 적응하여 생성한 Trigram 형식의 언어 모델을 채용하였다. ETSI 표준 방식의 Mel-Frequency Feature Extraction (MFCC) 특징 추출 기법[13]을 채용하여 13차 static 특징(c0~c12)과 미분 계수를 포함한 총 39차원의 특징 벡터를 추출하며, 특징 추출 단계에 CMN 기법을 적용하였다. 기본 인식 시스템은 깨끗한 환경의 음성 데이터에 대해 8.05%의 오인식률 (Word Error Rate, WER)을 갖는다.

4.2. 전사 정보를 이용한 기존 모델 적응 기법 성능

표 1은 기본 인식 시스템의 성능과 전사 정보를 이용한 기존의 전통적 음향 모델의 훈련 및 적응 기법을 채용한 시스템의 인식 성능 평가 결과를 나타낸다. 성능의 수치는 음성 인식 평가에서 일반적으로 사용되는 WER을 사용하였다. 본 실험에서 기본적으로 채용한 CMN 기법 외에 아무 처리를 하지 않은 경우 세 환경에 대해 평균 37.95%의 WER을 나타낸다. 표의 결과에서 Matched Train은 훈련 데이터 (4,620 샘플)를 각 잡음 환경과 동일한 잡음 데이터를 이용하여 인공적으로 오염시킨 후 음성 인식 시스템의 음향 모델을 훈련하여 인식한 결과를 나타낸다. Matched Train의 경우 세 환경에 대해 평균 20.17% WER의 성능을 나타내었다.

적응 데이터 세트 (680 샘플)와 이에 대한 전사정보가 제공되었음을 가정하여 기존의 적응 기법인 MAP 적응 기법과 MLLR 적응 기법을 적용한 결과도 관찰하였다. MAP 적응 기법을 적용 했을 때는 Matched 훈련 성능과 유사한 성능을 나타냈다. MLLR 적응 기법의 경우 자동차 잡음 환경에서는 31.47%로 MAP에 비해 인식 성능 향상이 낮았지만, 배경 음악 환경에서는 13.58%의 WER로 MAP 적응 및 Matched 훈련 조건 보다 월등한 음성 인식 성능을 나타냈다. 본 실험에서 사용한 배경 음악 잡음이 시간에 따라 그 음향 특성이 매우 크고 다양하게 변하기 때문에 MLLR 모델 적응 기법을 이용한 파라미터 공간의 전체적인 변환이 보다 효과적임을 추정할 수 있다.

표 1의 결과는 적응용 음성 데이터에 대한 정확한 전사정보가 제공될 경우 효과적으로 음향 모델을 배경 잡음 환경에 적응함으로써 음성 인식 성능을 높일 수 있음을 입증한다.

표 1. 전사된 음성 데이터베이스를 이용한 기존 모델 적응 기법 성능 평가 결과 (WER, %)
Table. 1 Speech recognition performance with conventional model adaptation methods using transcribed database (WER, %)

	Car	Babble	Music	Avg.
No processing	50.53	40.89	22.42	37.95
Matched Train	21.61	20.06	18.84	20.17
MAP + 전사정보	21.71	21.05	18.99	20.58
MLLR + 전사정보	31.47	22.10	13.58	22.38

4.3. 미전사 적응 데이터베이스를 이용한 음성 인식 성능 평가

표 2와 3은 미전사 (Untranscribed) 적응 데이터베이스를 이용한 음성 인식 성능 평가 결과를 나타낸다. 표 2의 실험에서는 깨끗한 환경에서 훈련한 기본 음성 인식기를 이용하여 적응 데이터에 대한 전사 정보를 취득하고, 이를 기반으로 하여 MAP 적응 기법과 MLLR 적응 기법을 각각 적용하였다. 성능 평가 결과 MAP 기법은 평균 30.00%, MLLR 기법은 평균 24.01% WER로서 MLLR 기법이 미전사 모델 적응 조건에서는 MAP에 비해 상당히 우수한 인식 성능을 나타내었다. 깨끗한 환경에서 훈련한 기본 인식기로 적응 데이터에 대한 전사 정보를 얻게 되므로 각 환경에 대한 전사 정보는 표 1에 나타낸 No Processing에서와 유사한 오인식률을 가질 것으로 예상된다.

따라서 오인식된 전사정보를 기반으로 모델 적응을 할 경우 정확한 전사 정보를 이용하여 적응을 한 경우 (표 1)의 MAP 및 MLLR 결과보다 성능이 뒤떨어지는 것은 쉽게 예상할 수 있다. MAP 적응 기법의 경우 모든 음향 모델 파라미터를 적응 데이터에 맞춰 각각 갱신하므로 모델 파라미터 공간 전체를 이동시키는 MLLR에 비해 보다 정확한 전사정보를 필요로 한다. 이와 같은 이유로 미전사 적응 데이터 조건에서 MLLR 기법이 MAP 기법에 비해 우수한 모델 적응 성능을 나타낸 것으로 판단된다.

표 2. 미전사된 음성 데이터베이스를 이용한 기존 모델 적응 기법 성능 평가 결과 (WER, %)
Table. 2 Speech recognition performance with conventional model adaptation methods using untranscribed database (WER, %)

	Car	Babble	Music	Avg.
MAP	37.21	30.35	22.44	30.00
MLLR	34.31	24.00	13.71	24.01

표 3은 미전사 적응 데이터베이서 조건에서 본 논문에서 설명한 GMM 적응 정보를 이용한 모델 변환 기법을 채용한 음성 인식 성능 결과를 나타내며, 편의상 GAMT (GMM Adaptation-based Model Transform)라고 칭한다. GAMT+MAP과 GAMT+MLLR은 각각 MAP와 MLLR 기법과 결합하여 적용한 결과를 나타낸다. GAMT 기법을 단독으로 적용한 경우 MAP 기법과 비교해서 자동차 잡음 환경에서는 하락한 성능을 나타내지만 음성 Babble 및 배경 음악 환경에서는 비교적 향상된 성능을 나타낸다. GAMT 기법을 MAP 기법 또는 MLLR 기법과 결합한 경우에는, MAP, MLLR 기법을 각각 단독으로 사용했을 때와 비교하여 배경 음악 환경을 제외한 두 잡음 환경 향상된 성능을 나타낸다. 따라서 본 논문에서 소개한 GAMT 기법이 미전사 데이터베이스 조건에서 기존의 MAP, MLLR 적응 기법과 결합할 경우 천천히 변하는 배경 잡음 환경에 대해 향상된 음성 인식 성능을 가져오는 것을 확인할 수 있다.

표 3. 미전사된 음성 데이터베이스를 이용한 GMM 적응 기반 모델 변환 기법 (GAMT) 성능 평가 결과 (WER, %)
Table. 3 Speech recognition performance with GMM adaptation-based model transformation method (GAMT) using untranscribed database (WER, %)

	Car	Babble	Music	Avg.
GAMT	42.92	28.49	20.38	30.60
GAMT+MAP	35.83	26.96	22.28	28.36
GAMT+MLLR	33.14	22.51	17.06	24.24

4.4. 온라인 음향 모델 적응 성능 평가

4.3 섹션의 결과는 본 논문에서 소개하는 GAMT 기법을 온라인 음향 모델 적응 기법에 적용할 수 있는 가능성을 시사한다. 온라인 음향 모델 적응은 입력 음성

데이터에 대해 아무런 정보가 주어지지 않는 환경을 가정하고, 입력 음성에 대해 음성 인식기의 음향 모델을 적응을 실시하여 향상된 음성 인식 성능을 얻고자 하는 방법이다. 즉, 입력 음성에 대해 음성 인식기를 이용하여 초기 음성 인식 결과를 얻고 인식된 발음 정보를 바탕으로 음향 모델을 적응하는 것을 말한다. 입력 음성 데이터만을 음향 모델 적응에 이용하므로 소량의 데이터에 우수한 모델 적응 성능을 보이는 MLLR 기법을 이용한다.

표 4의 첫 행의 결과는 입력 음성의 전사정보를 이용하여 온라인 모델 적응을 실시한 결과이다. 이 경우와 같이 전사 정보를 알고 있다는 경우는 입력 음성에 대해 정확한 인식 결과가 주어지는 비현실적인 경우를 가정하지만 온라인 모델 적응 기법의 최대 성능 한계를 보여주는 결과이다. 두 번째 행은 MLLR 기법을 이용한 기본적인 온라인 모델 적응 기법의 결과를 나타내는데, 표 1의 첫 행과 동일한 결과, 즉 아무런 성능 향상이 없음을 나타낸다. 이러한 결과는 본 실험에 사용한 잡음 환경이 기존의 모델 적응 기법인 MLLR을 단독으로 채용하여 온라인 음향 모델 적응을 적용하기에 매우 어려운 조건임을 말해준다. 세 번째 행은 MLLR 기법을 이용한 온라인 모델 적응에 GAMT 기법을 결합한 결과이다.

이 실험에서는 깨끗한 GMM을 MLLR 기법을 이용하여 입력 음성에 적응하여 적응된 GMM을 생성하였다. 이 실험의 결과에서도 배경 음악 조건을 제외하고 자동차 잡음 환경, 음성 Babble 환경에서 상당한 성능 향상이 있음을 확인할 수 있다. 따라서 본 논문에서 소개하는 GAMT 기법이 온라인 모델 적응과 같은 미전사 데이터베이스 조건에서 기존의 적응 기법과의 결합을 통해 음성 인식 성능 향상을 가져올 수 있음을 확인할 수 있다.

표 4. 온라인 모델 적응 성능 평가 결과 (WER, %)
Table. 4 Speech recognition performance of online model adaptation techniques (WER, %)

	Car	Babble	Music	Avg.
MLLR + Transcript	6.39	3.55	1.18	3.71
MLLR	50.53	40.89	22.42	37.95
GMM+MLLR	47.72	36.99	22.57	35.76

Ⅴ. 결 론

본 논문에서는 음성 인식 성능 향상을 위해 미전사된 음성 데이터베이스를 이용한 효과적인 음향 모델 적응 기법을 소개하였다. 소개된 모델 적응 기법에서는 기존의 적응 기법을 이용하여 잡음 환경에 적응된 GMM을 얻고, HMM의 가우시안 요소와 유사한 요소를 선택한다. 선택된 가우시안 요소의 변환 벡터를 구하여 평균 파라미터 변환에 이용한다. TIMIT 데이터베이스와 SPHINX3 음성 인식 시스템을 이용하여 실험한 결과, 소개한 모델 변환 기법을 기존의 MAP, MLLR 적응 기법과 결합하여 적용한 경우, 자동차 잡음과 음성 Babble 잡음 환경과 같이 천천히 변하는 잡음 환경에서 기존의 MAP, MLLR을 단독으로 사용할 경우보다 높은 인식 성능을 나타냈다. 또한 온라인 음향 모델 적응 실험에서도 MLLR과 결합할 경우 기존의 MLLR을 단독으로 사용할 때보다 효과적인 모델 적응 성능을 나타내었다. 이와 같은 결과는 본 논문에서 소개한 GMM 적응 기반의 모델 변환 기법을 채용함으로써 미전사된 음성 데이터베이스를 음향 모델 적응 기법에 효과적으로 활용할 수 있음을 입증한다.

감사의 글

이 논문은 인천대학교 2013년도 자체연구비 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] W. Kim and J.H.L. Hansen, “Gaussian Map based Acoustic Model Adaptation Using Untranscribed Data for Speech Recognition in Severely Adverse Environments,” *Interspeech-2012*, pp.1764-1767, Sept. 2012.

[10] https://catalog.ldc.upenn.edu/LDC93S1

[11] http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html

[12] http://cmusphinx.sourceforge.net

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

김우일(Wooil Kim)

2003년 고려대학교 전자공학과 공학박사
2004년 ~ 2005년 미국 카네기 멜론 대학교 박사 후 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