일반논문 (Regular Paper)
방송공학회논문지 제23권 제4호, 2018년 7월 (JBE Vol. 23, No. 4, July 2018)
https://doi.org/10.5909/JBE.2018.23.4.511
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

음성 명료도 향상을 위한 분류 모델의 잡음 환경 적응

정 준 영[a)], 김 기 백[a)‡]

Adaptation of Classification Model for Improving Speech Intelligibility in Noise

Junyoung Jung[a)] and Gibak Kim[a)‡]

요 약

본 논문에서는 잡음 환경의 음성 신호를 시간-주파수 영역으로 분해한 후 0 또는 1로 표현되는 이진 마스크를 적용하여 음성의 명료도를 높이는 방법에 대해 다룬다. 시간-주파수 영역으로 분해된 신호에 대해 상대적으로 잡음이 많이 섞인 경우는 마스크 “0”을 할당하여 제거하고, 그렇지 않은 경우는 마스크 “1”을 할당하여 보존하는 방식을 채택한다. 이러한 이진 마스크의 추정은 가우시안 혼합 모델로 학습된 베이지안 분류기를 사용한다. 가우시안 혼합 모델 학습에 포함된 잡음 환경에 대해서는 학습된 모델을 이용하여 추정된 이진 마스크의 적용을 통해 잡음 환경에서 음성 명료도를 높일 수 있으나 학습에 포함되지 않은 잡음 환경에 대해서는 음성 명료도를 향상시키지 못하는 문제가 있다. 본 논문에서는 이러한 문제를 해결하기 위해 학습 모델을 잡음 환경에 적응시키고자 한다. 새로운 잡음 환경에 대처하고자 음성 인식에서 사용되는 대표적인 화자 적응 방법을 적용하였으며 실험을 통해 새로운 잡음 환경에 적응함을 확인하였다.

Abstract

This paper deals with improving speech intelligibility by applying binary mask to time-frequency units of speech in noise. The binary mask is set to “0” or “1” according to whether speech is dominant or noise is dominant by comparing signal-to-noise ratio with pre-defined threshold. Bayesian classifier trained with Gaussian mixture model is used to estimate the binary mask of each time-frequency signal. The binary mask based noise suppressor improves speech intelligibility only in noise condition which is included in the training data. In this paper, speaker adaptation techniques for speech recognition are applied to adapt the Gaussian mixture model to a new noise environment. Experiments with noise-corrupted speech are conducted to demonstrate the improvement of speech intelligibility by employing adaption techniques in a new noise environment.

Keyword : speech intelligibility, noise suppression, binary mask, Gaussian mixture model, adaptation

a) 숭실대학교 전기공학부(School of Electrical Engineering, Soongsil University)
‡ Corresponding Author : 김기백(Gibak Kim)
E-mail: imkgb27@ssu.ac.kr
Tel: +82-2-828-7266
ORCID:https://orcid.org/0000-0001-5114-4117

※ 이 논문은 한국 산업통상자원부의 로봇산업융합핵심기술사업 프로그램 No.10048474, ‘고령화 세대에게 개인별 특화된 복지 서비스를 제공하기 위한 빅데이터 기반의 서비스 로봇개발’의 지원으로 수행되었음.
※ This material is based upon work supported by the Ministry of Trade, Industry & Energy(MOTIE, Korea) under Industrial Technology Innovation Program. No.10048474, 'Development of a Service Robot based on Big Data for Providing Aging Generation with Personalized Welfare Services'
· Manuscript received March 16, 2018; Revised May 8, 2018; Accepted May 8, 2018.

Ⅰ. 서 론

잡음 환경에서 취득된 음성 신호의 잡음 제거는 오래 전부터 연구되어 온 음성 신호 처리의 대표적인 관심 분야이다. 두 개 이상의 마이크를 이용하여 빔포밍 등 어레이 신호 처리 기법을 사용하여 보다 의미 있는 잡음 제거 성능을 얻을 수 있으나 여전히 하나의 마이크를 이용하는 시스템이 음성 취득 방법의 대부분을 차지하고 있어 하나의 마이크를 이용한 잡음 제거 구현이 절실히 요구되고 있다. 잡음 제거의 목적은 크게 음질 향상과 음성 명료도 향상으로 나눌 수 있다. 두 가지 목적이 서로 깊은 관련이 있으나 기본적으로 다음과 같이 구분될 수 있다. 음질 향상은 잡음을 제거하여 청취자가 좀 더 편안하게 들을 수 있게 하는데 그 목적이 있다. 반면에 명료도 향상은 청취자가 음성에 담겨있는 내용을 얼마나 정확하게 의미를 파악할 수 있는가에 초점을 맞추고 있다. 음성이 왜곡되어 청취자가 듣기 불편하나 의미는 정확하게 파악되는 상황이라면 음질은 좋지 않으나 명료도는 우수하다고 판단할 수 있다.

하나의 마이크를 이용한 잡음 제거 알고리즘의 성능은 오랜 연구에도 불구하고 여전히 많은 개선의 여지를 남겨두고 있다. 잡음의 종류에 따라 차이는 있으나 정상 청력을 가진 사람들은 5 dB SNR (Signal-to-Noise Ratio) 보다 약한 정도의 잡음 환경에서는 100%에 가까운 인식률을 보이고 있다. 따라서 음질 향상을 위한 잡음 제거 연구는 주로 5~15 dB SNR 수준의 잡음 환경을 대상으로 하고 있는 반면 명료도 향상은 이보다 낮은 수준 (‐5~0 dB SNR)의 심한 잡음 환경에서의 청취자에 의한 음성 인식률을 높이는 것을 목표로 한다. 연구 결과에 따르면 기존의 대표적인 음질 향상을 위한 잡음 제거 알고리즘들은 잡음 환경에서 유의미한 음성 명료도 향상을 기대하기 어려움을 알 수 있다[1-3].

2000년대 이진 마스크를 이용하여 음성 명료도 향상을 도모하는 연구들이 이어졌는데, 이진 마스크는 CASA (Computational Auditory Scene Analysis) 연구에서 처음 제안되었다[4-6]. 잡음이 섞인 음성 신호를 시간-주파수 영역으로 변환한 다음 각각의 시간-주파수 영역에 대해 잡음이 음성 신호보다 상대적으로 큰 영역은 마스크 “0”을 적용하여 제거하고, 나머지 영역은 마스크 “1”을 적용하여 그대로 보존한다. 마스크 “0”을 적용할지 “1”을 적용할지에 대해서는 영역의 신호 대 잡음 비에 의존하는데 문턱값은 주파수에 따라 다르게 적용될 수도 있으며 보통 0dB 이하로 설정된다. 이후 마스킹된 시간-주파수 영역의 신호를 다시 시간 영역으로 신호를 합성하여 잡음을 제거한다. 이진 마스크의 추정을 두 개의 클래스에 대한 분류 문제로 정의하고, 학습 데이터를 이용하여 분류기를 학습시킨 후 입력 데이터의 시간-주파수 각 영역에 대해 마스크를 추정하여, 낮은 신호 대 잡음 비를 갖는 음성 신호의 음성 명료도를 의미있게 향상시킬 수 있음이 확인되었다[7-11].

학습 데이터로부터 학습된 분류기 모델을 이용하여 이진 마스크를 추정하는 경우는 학습에 포함된 잡음 환경에 대해서는 이진 마스크의 적용을 통해 잡음 환경에서 음성 명료도를 높일 수 있으나 학습에 포함되지 않은 잡음 환경에 대해서는 음성 명료도를 향상시키지 못하는 문제가 있다. 이진 마스크 추정을 위한 분류 모델을 학습하기 위해서는 수백 문장 수준의 잡음 환경 음성 데이터가 필요하므로 새로운 잡음 환경에 대해 빠르게 적용되기 어렵다. 이러한 문제를 해결하기 위해 새로운 잡음 데이터를 점진적으로 모아서 학습시키는 점증 학습 (incremental training) 방법이 제안되었다[12]. 가우시안 혼합 모델 (Gaussian Mixutre Model: GMM)의 파라미터들을 재귀적 베이즈 추정 방식의 유사 베이즈 (quasi-Bayes) 방법으로 추정하기 위해 여러 개의 하이퍼 파라미터를 설정하고 순차적으로 계산한 후, 이들 파라미터들로부터 각 가우시안 분포의 가중치, 평균벡터, 분산행렬 등을 구한다. 이러한 점증 학습 방법으로 새로운 잡음 환경에 대해 이진 마스크 추정 성능이 점진적으로 향상됨을 확인하였다. 그러나 이러한 방법은 점진적으로 성능이 향상되기는 하지만 적은 양의 데이터를 이용한 초기 학습 과정에서는 충분한 성능을 발휘하지 못하는 단점이 있다.

이러한 점증학습과는 달리 음성 인식 모델의 화자 적응 방법을 적용하여 새로운 잡음 환경에 적응시키는 방법을 고려할 수 있다. 화자 적응에서 널리 사용되는 eigenvoice 방법을 적용하여 새로운 잡음 환경에서 이진 마스크 분류 모델을 적응시킬 수 있다[13]. Eigenvoice 방법은 여러 잡음 환경에 대해 학습된 가우시안 혼합 모델로부터 가우시안 분포들의 평균벡터를 연결하여 생성한 슈퍼벡터를 사용한

다. 이렇게 생성된 슈퍼벡터들을 PCA (Principal Component Analysis) 분석을 통해 eigenvoice를 추출한다. 이렇게 추출된 eigenvoice의 가중합과 여러 잡음 환경 데이터를 이용하여 만든 모델을 결합하여 새로운 환경의 가우시안 평균벡터를 추정한다. 이와 같은 eigenvoice 적응 방법은 고유 벡터의 가중치만 구하면 되므로 추정해야 할 파라미터의 수가 적고 그에 따라 음성 인식의 화자 적응에서는 적은 양의 데이터에 대해서도 높은 성능을 나타내는 것으로 알려져 있다[14]. Eigenvoice의 기본 아이디어는 신호의 특성을 몇 개의 주된 모델의 가중합으로 나타낼 수 있다는 사실을 전제로 하여 사전에 구한 고유 벡터에 대한 의존도가 높다. 이와 같은 아이디어는 음성 신호에는 효과적으로 적용되었으나, 잡음의 형태는 음성 신호보다 훨씬 다양하여 몇 개의 고유 벡터의 가중합으로 표현하는데 한계가 있다. 해당 논문의 결과에서 보듯이 적은 양의 적응 데이터로 모델 적응이 가능하나 데이터 양이 늘어나더라도 성능 향상이 제한적임을 알 수 있다. 본 논문에서는 신호를 몇 개의 고유벡터 가중합으로 표현하여 적응시키는 eigenvoice와 달리 가우시안 모델을 직접 적응시키는 MAP (Maximum a posterior)과 가우시안 모델의 평균을 변환시키는 MLLR (Maximum Likelihood Linear Regression) 등의 보다 전통적인 화자 적응 방법을 이용하고자 한다[15,16]. 실험을 통하여 비대각 공분산행렬 (full covariance matrix) 가우시안 혼합 모델의 적응을 위한 반복적인 MLLR 적응방법이 MAP보다 우수하고, MLLR 적용 후 순차적으로 MAP을 적용하는 것이 가장 우수한 성능 나타냄을 확인하였다.

본 논문의 나머지 구성은 다음과 같다. 2장에서는 분류 모델 기반의 이진 마스크 추정에 대해 설명하고, 3장에서는 MAP, MLLR 적응 방식에 대해 설명한다. 이진 마스크 분류 성능에 대한 실험결과는 4장에서 제시한다.

Ⅱ. 이진 마스크 추정을 통한 잡음 제거

본 장에서는 잡음 환경에서 취득된 음성 신호에 이진 마스크를 적용하여 잡음을 제거하는 방법에 대해 설명한다. 이진 마스크를 적용하기 위해서는 우선 음성 신호를 시간-주파수 영역으로 분해하고 각 시간-주파수 영역에서 특징

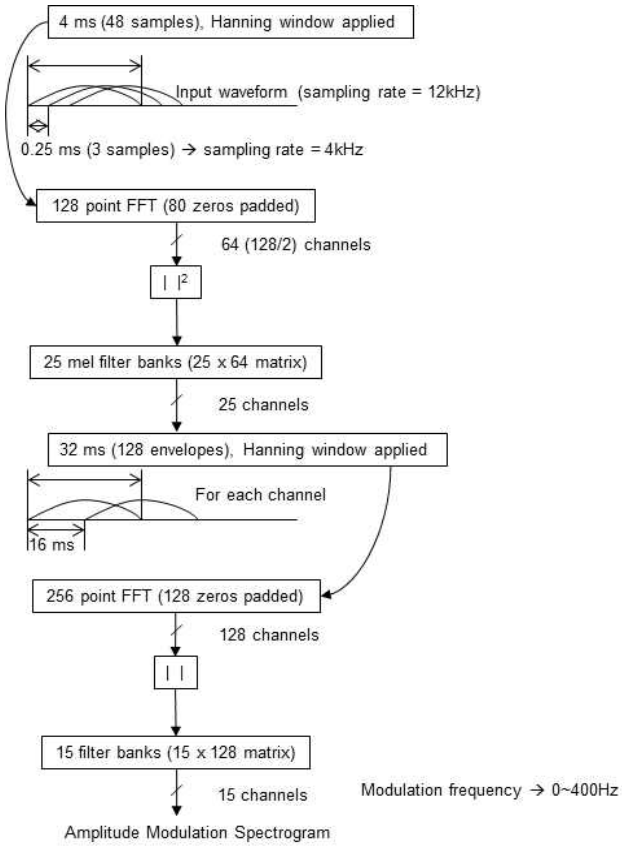

그림 1. 진폭 변조 스펙트로그램의 추출 과정
Fig. 1. Extraction of amplitude modulation spectrogram

벡터를 추출해야 한다. 본 논문에서 사용하는 특징 벡터는 Tchorz와 Kollmeier에 의해 제안된 방법을 변형한 진폭 변조 스펙트로그램 (AMS: Amplitude Modulation Spectrogram)으로서[17,18] 구체적인 추출 과정은 그림 1로 정리할 수 있으며 각 과정에 대한 설명은 다음과 같다.

우선, 마이크로 입력되는 신호에 대해 4ms 크기의 Hann (또는 Hamming) 윈도우를 적용한다.

4ms는 48샘플에 해당하며(12kHz 샘플링 주파수인 경우), 80개의 제로를 추가하여 128 샘플을 만든 후, 128 포인트 FFT (Fast Fourier Transform)를 적용한다. 그 다음 FFT 계수의 절대값에 제곱을 취하여 전력 스펙트럼 (power spectrum)을 얻는다. 0.25ms가 진행될 때마다 Hann윈도우를 적용하여 FFT를 수행하게 되며, 이런 과정을 통해 전력 스펙트럼의 크기에 대한 표본 주파수는 4kHz (1/0.25ms)가 된다. 이 후, mel 스케일 주파수를 기반으로 한 25개의 필터

로 구성된 필터뱅크를 전력 스펙트럼에 적용한다. 이렇게 해서 25개의 mel 스케일 주파수 대역에 대해 4kHz로 표본화된 전력 스펙트럼을 얻게 된다. 각 주파수 대역에 대해 32ms (128샘플) 동안의 전력 스펙트럼들을 모아서 다시 Hann 윈도우를 적용하고, 이러한 과정을 16ms마다 진행하여 수행하게 된다. 이제 Hann 윈도우를 적용한 128개의 데이터에 128개의 제로를 추가하여 256 샘플을 만든 후, 256 포인트 FFT를 적용한다. 이 후, 절대값을 취하여 진폭 변조 스펙트럼을 얻는다. 특징벡터의 크기를 줄이기 위해 15개의 삼각필터로 이루어진 필터뱅크를 적용하여 15차원의 진폭 변조 스펙트럼을 얻는다. 15차원의 특징벡터를 주파수와 시간에 대해 각각 차분 성분을 추가하여 25개의 주파수 대역은 각각 총 45차원의 특징벡터를 갖게 된다.

추출된 특징 벡터를 입력으로 하여 마스크 “0” 또는 “1”로 분류하기 위해서는 베이지안 (Bayesian) 분류기를 사용한다. 학습 데이터로부터 가우시안 혼합 모델을 사용하여 마스크 “0”과 “1”에 해당하는 모델(λ_0, λ_1)을 생성하고, 이진 마스크를 분류할 시간-주파수 영역의 특징 벡터(AMS)가 주어지면 “0”과 “1”로 분류될 사후 확률 (posterior probability)을 각각 구한 후, 두 확률값을 비교하여 마스크를 결정한다. 시간 t, 주파수 k 영역에 대한 이진 마스크 $I(t,k)$는 다음 식에 의해 추정된다.

$$I(t,k) = \begin{cases} 0, \text{if } P(\lambda_0|\boldsymbol{o}(t,k)) > P(\lambda_1|\boldsymbol{o}(t,k)) \\ 1, \text{otherwise} \end{cases} \quad (1)$$

여기서 $o(t,k)$는 입력 신호의 시간 t, 주파수 k영역에 해당하는 특징 벡터를 나타내고 사후 확률 $P(\lambda_0|\boldsymbol{o}(t,k))$는 베이즈 정리를 이용하여 다음 식과 같이 계산된다.

$$P(\lambda_0|\boldsymbol{o}(t,k)) = \frac{P(\lambda_0, \boldsymbol{o}(t,k))}{P(\boldsymbol{o}(t,k))} = \frac{P(\lambda_0)P(\boldsymbol{o}(t,k)|\lambda_0)}{P(\boldsymbol{o}(t,k))} \quad (2)$$

$P(\boldsymbol{o}(t,k)|\lambda_0)$는 GMM으로 학습된 모델에 특징 벡터를 대입하여 계산할 수 있고 $P(\lambda_0)$는 학습데이터로부터 구할 수 있는 사전 확률이다. $P(\lambda_1|\boldsymbol{o}(t,k))$도 유사한 방법으로 구할 수 있다.

Ⅲ. 이진 마스크 추정 모델의 환경 적응

1. MAP (Maximum a posteriori) 적응 방법

MAP은 광범위한 분야에서 통계 모델의 파라미터를 추정하는 방법으로 널리 적용되고 있다. 학습 데이터 양이 증가할수록 ML (Maximum Likelihood) 방법의 결과에 접근하는 것으로 알려져 있다. ML 방법에서는 관찰 데이터를 랜덤 벡터로 가정하고 각 모델에 대한 likelihood가 최대가 되도록 하는 모델 파라미터를 찾는다. 이와 달리 MAP 적응 방법은 모델 파라미터를 랜덤 벡터로 가정하고, 데이터의 변화에 따른 확률 분포 함수의 변화에 따라 사후 확률 분포 (posterior distribution)가 최대가 되는 모델 파라미터를 찾는다. 관찰 데이터가 적을 때는 사전 확률 분포가 최대가 되는 파라미터를 선택하고, 데이터가 늘어날수록 사전 확률 분포의 영향은 줄어들면서 ML의 결과에 수렴하게 된다. 가우시안 혼합 모델의 가중치, 평균 벡터, 공분산 행렬 등을 업데이트하므로 업데이트해야하는 파라미터의 수가 많다. 성능 확보를 위해서는 비교적 많은 양의 데이터를 필요로 한다. 가우시안 혼합 모델의 경우 m번째 가우시안 분포의 평균 벡터 업데이트 식은 다음과 같다[16].

$$\boldsymbol{\mu}_{m,MAP} = \frac{r\boldsymbol{\mu}_m + \sum_{t=1}^{T} P(m|\boldsymbol{o}(t,k),\lambda)\boldsymbol{o}(t,k)}{r + \sum_{t=1}^{T} P(m|\boldsymbol{o}(t,k),\lambda)} \quad (3)$$

위 식에서 $\boldsymbol{\mu}_m$은 사전 모델의 m번째 가우시안 분포 평균 벡터이고, $\boldsymbol{o}(t,k)$는 길이 T의 적응 데이터 중 시간 t, k번째 주파수 대역의 특징 벡터이다. r은 적응 데이터에 의한 ML 평균 벡터 추정값과 사전 모델의 평균 벡터 중 어느 쪽에 더 가중치를 많이 둘 지를 결정하는 하이퍼 파라미터이다. 즉 r이 크면 사전 모델에 가중치를 많이 두는 것이고 r이 작은 경우 적응을 위해 입력된 데이터로 ML 추정한 값에 가중치를 많이 두게 된다. 여기서 $P(m|\boldsymbol{o}(\tau,k),\lambda)$는 데이터와 모델($\lambda$)이 주어졌을 때 m번째 가우시안 분포에 해당할 확률이며 아래와 같이 계산할 수 있다.

$$P(m|\boldsymbol{o}(t,k),\lambda)=\frac{c_m b_m(\boldsymbol{o}(t,k))}{\sum_{l=1}^{M} c_l b_l(\boldsymbol{o}(t,k))} \quad (4)$$

여기서 c_m은 가우시안 혼합 모델에서 m번째 가우시안 분포에 대한 가중치이며 $b_m(\cdot)$은 관찰데이터에 대한 m번째 가우시안 분포의 확률값이다.

2. MLLR (Maximum Likelihood Linear Regression) 적응 방법

MLLR 적응 방법에서는 MAP방법과는 달리 모델 파라미터를 직접 업데이트하는 것이 아니라 파라미터를 다음과 같은 선형 회귀 (linear regression) 식으로 업데이트한다[15].

$$\bar{\boldsymbol{\mu}}_m = \mathbb{W}\tilde{\boldsymbol{\mu}}_m, \quad \tilde{\boldsymbol{\mu}}_m = [\boldsymbol{\mu}_m \quad 1]^T \quad (5)$$

$\mathbb{W}$는 EM (Expectation Maximization)을 이용한 ML방법으로 추정되며 하나의 $\mathbb{W}$ 행렬을 이용하여 모든 가우시안 파라미터를 업데이트한다.

이진 마스크 모델링을 위한 가우시안 혼합 모델에서는 대각 공분산 행렬보다 비대각 공분산 행렬이 더 좋은 성능을 보임이 확인되었다[7]. 본 논문에서는 비대각 공분산 행렬을 가진 가우시안 혼합 모델의 MLLR 적응을 위해 Povey와 Saon에 의해 제안된 반복적인 (iterative) 방법에 의한 $\mathbb{W}$ 업데이트 방법을 이용한다[15]. 이 방법에서는 먼저, 식(6)과 같은 보조 함수를 설정한다.

여기서 $E_m(\cdot)$은 m번째 가우시안 분포에 대한 기댓값을 의미하고 $\boldsymbol{\Sigma}_{\mathrm{m}}$은 m번째 가우시안 분포의 공분산 행렬 (covariance matrix)을 의미한다.

보조함수의 $\mathbb{W}$에 대한 그레디언트 (gradient)는 다음 식으로 나타나며

$$\boldsymbol{L}=\nabla_{\mathbb{W}} f=-\sum_{m=1}^{M} c_m \boldsymbol{\Sigma}_{\mathrm{m}}^{-1}\left(\mathbb{W}\tilde{\boldsymbol{\mu}}_m - E_m(\boldsymbol{o}(t,k))\right)\tilde{\boldsymbol{\mu}}_m^T \quad (7)$$

다음 식으로 n번째 업데이트 과정을 나타낼 수 있다.

$$\mathbf{w}_{\mathrm{i}}^{(\mathrm{n})} = \mathbf{w}_{\mathrm{i}}^{(\mathrm{n}-1)} + \alpha \mathbf{G}_{\mathrm{i}}^{-1}\mathbf{l}_{\mathrm{i}} \quad (8)$$

여기서 $\mathbf{w}_{\mathrm{i}}^{(\mathrm{n})}$은 n번째 반복과정에서 업데이트된 행렬 $\mathbb{W}$의 i번째 행벡터이고, $\mathbf{l}_{\mathrm{i}}$는 $\mathbb{L}$의 i번째 행벡터이다. α는 반복과정을 거칠 때마다 1/2로 줄어드는 값이고 $\mathbf{G}_{\mathrm{i}}$는 다음 식으로 정의된다.

$$\mathbf{G}_{\mathrm{i}} = \sum_{\mathrm{m}=1}^{\mathrm{M}} \frac{c_{\mathrm{m}} \tilde{\boldsymbol{\mu}}_{\mathrm{m}} \tilde{\boldsymbol{\mu}}_{\mathrm{m}}^{\mathrm{T}}}{\sigma_{\mathrm{m,i}}} \quad (9)$$

여기서 $\sigma_{m,i}$는 m번째 가우시안 분포 공분산 행렬의 i번째 대각원소이다.

Ⅳ. 실험 결과

1. 실험 환경

이진 마스크 분류를 위한 모델의 환경 적응 실험에 사용한 IEEE 문장은 영어로 구성되어 있고[19], 원어민에 의해 발성된 것을 25kHz 샘플링 주파수로 녹취한 후, 12kHz로 다운 샘플링하였다. 음질 향상을 다루는 실험에서는 16kHz 샘플링이 널리 이용되나 음성의 내용을 파악하는 명료도를 다루는 문제에서는 그보다 낮은 12kHz 샘플링을 사용하는 것이 바람직하다. 64개의 가우시안 분포를 이용하여 가우시안 혼합 모델을 구성하였다. 사전 모델을 생성을 위하여 34가지 잡음을 이용하였고, 테스트를 위해서는 사전 모델 학습에 포함되지 않은 babble, factory, speech-shaped 잡음을 사용하였다. Babble은 남녀 각각 10명이 동시에 서로 다른 문장을 읽는 것을 혼합하여 생성하였고, factory 잡음은 NOISEX 데이터베이스에서 발췌하였다[20]. NOISEX 데이터베이스는 20kHz 샘플링 주파수로 수집되었으며 본 실험에서는 12kHz로 다운 샘플링되었다. Speech-shaped 잡음은 IEEE 데이터베이스 음성 데이터들의 평균 스펙트럼을 갖도록 백색 잡음을 변조한 stationary 잡음이다. IEEE 데이

$$f=-0.5\sum_{m=1}^{M} c_m\left(\mathbb{W}\tilde{\boldsymbol{\mu}}_m - E_m(\boldsymbol{o}(t,k))\right)^T \boldsymbol{\Sigma}_{\mathrm{m}}^{-1}\left(\mathbb{W}\tilde{\boldsymbol{\mu}}_m - E_m(\boldsymbol{o}(t,k))\right) \quad (6)$$

터베이스 각 문장에는 10개 내외의 단어들이 포함되어 있고 2~3초의 길이로 녹음되어 있다. 사전 모델 생성을 위한 학습 데이터로 200 개의 문장을 사용하였고, 신호 대 잡음비 -5, 0, 5 dB 등으로 잡음을 섞어 학습 데이터를 생성하였다. 테스트에 사용한 데이터는 모델 적응에 사용한 데이터를 포함하지 않는다. 잡음 환경 적응 테스트를 위해서 -5dB 10개 문장부터 10개씩 증가시키며 50개 문장까지 사용하였다.

2. 실험 결과

1절에서 설명한 실험 환경에 대해 MAP, MLLR 적응 알고리즘을 적용하였다. III장에서 설명한 MAP과 MLLR 방법을 이용하여 잡음 환경 적응 성능을 평가하였다. MAP적응 후 MLLR을 적용하거나 MLLR 후 MAP을 적용하는 실험도 추가하였다. 성능 평가를 위해서는 알고리즘 적용 전후의 음성 명료도를 평가해야 하는데, 가장 바람직한 방법은 기존의 논문에서와 같이 정상청력을 가진 영어 원어민들을 대상으로 하는 평가이다. 피험자들에게 음성 파일을 들려주고 들은 내용을 받아 적도록 하여 얼마나 많은 단어들이 제대로 인식되었는지 평가하는 것이다. 그러나 영어 원어민 피험자들을 모집하여 청취 실험하는 것은 시간과 비용 측면에서 부담이 될 뿐만 아니라 개인 간의 오차도 발생하므로 큰 성능 향상이 있는 경우를 제외하고는 의미 있는 실험 결과를 도출하기 어려운 단점이 있다. 따라서 본 실험에서는 청취 실험 대신 추정된 이진 마스크의 정검출율(Hit)과 오검출율(False Alarm:FA)을 계산하여 그 차이(Hit-FA)로서 성능을 평가한다. Hit는 실제 마스크가 “1”인 영역이 “1”로 추정되는 비율이며, FA는 실제 마스크가 “0”인 영역이 “1”로 잘못 추정되는 비율이다. Hit-FA는 음성 명료도와 높은 상관관계가 있는 것으로 확인되어 Hit-FA를 측정하여 이진 마스크를 이용한 잡음 제거의 음성 명료도를 가늠할 수 있다[7].

모델 적응에 사용된 문장 수에 대해 각 적응 방법의 성능을 그림 2에 나타내었다. MAP, MLLR과 함께 두 가지 적응 과정을 순차적으로 적용한 MAP+MLLR, MLLR+MAP 결과도 포함하였다. 비교를 위해 eigenvoice(K=5)의 결과도 함께 나타내었다. 잡음 환경 적응 전 모델에 대해서는

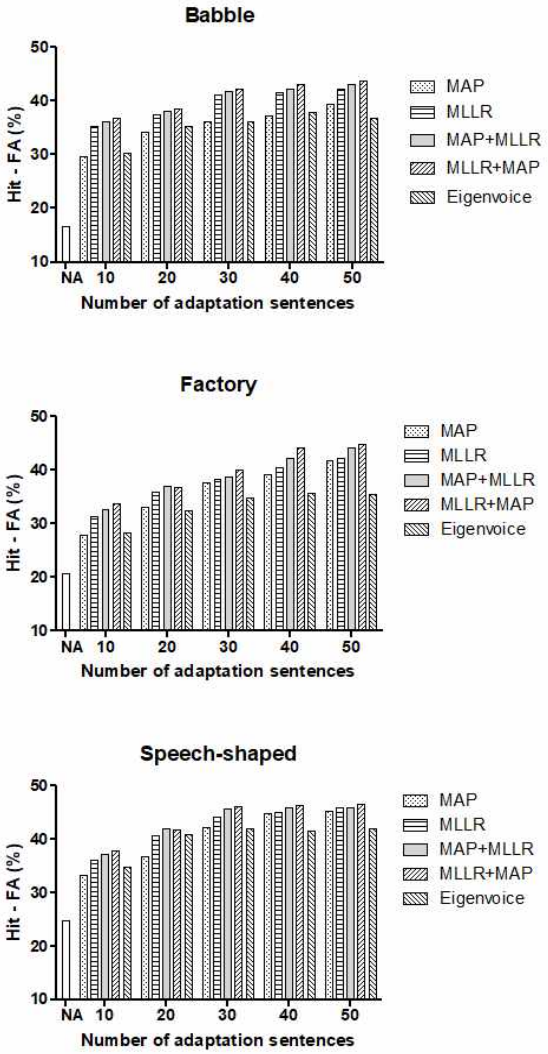

그림 2. 모델 환경 적응 결과 (정검출율 - 오검출율: Hit - FA, NA: No Adaptation)
Fig. 2. Performance of model adaptation in new noises

20% 내외의 낮은 Hit-FA 성능을 나타내었다. 테스트 잡음 환경이 초기 모델에 포함되어 있는 경우의 성능[7]에는 미치지 못하나, 문장의 개수를 늘려가며 환경 적응시킨 결과 40% 이상으로 성능이 확연히 향상된 것을 알 수 있다. 모든 적응 방법에 대해 적응 데이터 양의 증가에 따라 성능이 향상되는 것을 확인할 수 있으며, MLLR이 MAP보다 우수한 성능을 보이며, MLLR 후에 추가로 MAP 적응을 적용하여 추가 성능 향상을 기대할 수 있음을 확인하였다.

V. 결 론

본 논문에서는 잡음 환경 음성에 대해 이진 마스크 적용을 통한 음성 명료도를 향상하는 방법 중, 사전 모델 생성을 위한 학습에 포함되지 않은 새로운 잡음 환경에 적응시키기 위해 음성 인식에서 사용하는 대표적인 화자 적응 방법을 이용하였다. 음성 신호에 비해 다양한 잡음 신호의 특성상, 고유값분해를 통한 고유벡터를 이용한 eigenvoice 적응 방법보다는 적응 데이터에 의존하는 MAP, MLLR 방법이 더 효과적임을 확인할 수 있었다. 또한 두 가지 적응 방법을 순차적으로 적용하여 추가 성능 향상을 기대할 수 있음을 확인하였다.

참 고 문 헌 (References)

[1] Y. Hu and P. Loizou, “Subjective comparison and evaluation of speech enhancement algorithms,” Speech communication, vol. 49, no. 7, pp. 588–601, Jul. 2007. https://doi.org/10.1016/j.specom.2006.12.006

[2] Y. Hu and P. Loizou, “Evaluation of objective quality measures for speech enhancement,” IEEE Transactions on Speech and Audio Processing, vol. 16, no. 1, pp. 229–238, 2008. https://doi.org/10.1109/tasl.2007.911054

[3] Y. Hu and P. Loizou, “A comparative intelligibility study of single-microphone noise reduction algorithms.” The Journal of the Acoustical Society of America, vol. 122, no. 3, p. 1777, Sep. 2007. https://doi.org/10.1121/1.2766778

[4] G. Brown and M. Cooke, “Computational auditory scene analysis,” Computer speech and language, vol. 8, pp. 297–336, 1994. https://doi.org/10.1006/csla.1994.1016

[5] D. Wang and G. Brown, Computational Auditory Scene Analysis : Principles, Algorithms, and Applications, Wiley, Hoboken, NJ, 2006. https://doi.org/10.1109/tnn.2007.913988

[6] D. Wang, “On ideal binary mask as the computational goal of auditory scene analysis,” In Divenyi P. (ed.), Speech Separation by Humans and Machines, pp. 181-197, Kluwer Academic, Norwell MA, 2005. https://doi.org/10.1007/0-387-22794-6_12

[7] G. Kim. Y. Lu, Y. Hu and P. Loizou, “An algorithm that improves speech intelligibility in noise for normal-hearing listeners,” The Journal of the Acoustical Society of America, vol. 126, no. 3, pp 1486-1494, 2009. https://doi.org/10.1121/1.3184603

[8] Y. Hu, P. Loizou, “Environment-specific noise suppression for improved speech intelligibility by cochlear implant users”, The Journal of the Acoustical Society of America, vol. 127, no. 6, pp 3689-3695, 2010. https://doi.org/10.1121/1.3365256

[9] K. Han, D. Wang, “A classification based approach to speech segregation”, The Journal of the Acoustical Society of America, vol. 132, no. 5, pp 3475-3483, 2012. https://doi.org/10.1121/1.4754541

[10] Y. Wang, K. Han, D. Wang, “Exploring monaural features for classification-based speech segregation”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 2, pp 270-279, 2013. https://doi.org/10.1109/tasl.2012.2221459

[11] G. Kim, “A Post-processing for Binary Mask Estimation Toward Improving Speech Intelligibility in Noise,” Journal of Broadcast Engineering, Vol. 18, No.2, pp.311-318, March, 2013. https://doi.org/10.5909/jbe.2013.18.2.311

[12] G. Kim, P. Loizou, “Improving speech intelligibility in noise using environment-optimized algorithms”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 8, pp 2080-2090, 2010. https://doi.org/10.1109/tasl.2010.2041116

[13] G. Kim, “Eigenvoice Adaptation of Classification Model for Binary Mask Estimation,” Journal of Broadcast Engineering, Vol.20, No.1, pp.164-170, 2015. https://doi.org/10.5909/jbe.2015.20.1.164

[14] R. Kuhn, J. Junqua, P. Nguyen, N. Niedzielski, "Rapid Speaker Adaptation in Eigenvoice Space," IEEE Transactions on Speech and Audio Proceeding, vol. 8, no. 6, pp. 695-707, November 2000. https://doi.org/10.1109/89.876308

[15] D. Povey and G. Saon, “Feature and model space speaker adaptation with full covariance Gaussians,” Proceedings of Interspeech, USA, september 2006.

[16] K. Shinoda, “Speaker Adaptation Techniques for Speech Recognition Using Probabilistic Models,” Electronics and Communications in Japan, Part 3, Vol. 88, No. 12, pp.25-42, 2005. https://doi.org/10.1002/ecjc.20207

[17] J. Tchorz and B. Kollmeier, “Estimation of the signal-to-noise ratio with amplitude modulation spectrograms,” Speech Communication, vol. 38, no. 1–2, pp. 1–17, Sep. 2002. https://doi.org/10.1016/s0167-6393(01)00040-1

[18] J. Tchorz and B. Kollmeier, “SNR estimation based on amplitude modulation analysis with applications to noise suppression,” IEEE Transactions on Speech and Audio Processing, vol. 11, no. 3, pp. 184–192, May 2003. https://doi.org/10.1109/tsa.2003.811542

[19] IEEE, “IEEE recommended practice for speech quality measurements",” IEEE Trans. Audio Electroacoust., vol. 17, pp. 225-246, 1969. https://doi.org/10.1109/tau.1969.1162058

[20] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Communication, vol. 12, pp. 247-251, 1993. https://doi.org/10.1016/0167-6393(93)90095-3

저 자 소 개

정 준 영

- 2017년 : 숭실대학교 전기공학부 학사
- 2017년 ~ 현재 : 숭실대학교 전기공학부 석사과정
- ORCID : https://orcid.org/0000-0002-8327-7181
- 주관심분야 : 머신러닝 및 딥러닝, 음성신호처리, 전력신호처리

김 기 백

- 1994년 : 서울대학교 전자공학과 학사
- 1996년 : 서울대학교 전자공학과 석사
- 2007년 : 서울대학교 전기컴퓨터공학부 박사
- 1996년 ~ 2000년 : LG전자기술원 연구원
- 2000년 ~ 2003년 : (주)보이스웨어 선임연구원
- 2008년 ~ 2010년 : Univ. of Texas at Dallas, Research Associate
- 2011년 ~ 현재 : 숭실대학교 전기공학부 교수
- ORCID : https://orcid.org/0000-0001-5114-4117
- 주관심분야 : 머신러닝 및 딥러닝, 음성신호처리, 전력신호처리