정규논문 (Regular Paper)
방송공학회논문지 제20권 제1호, 2015년 1월 (JBE Vol. 20, No. 1, January 2015)
http://dx.doi.org/10.5909/JBE.2015.20.1.164
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

Eigenvoice를 이용한 이진 마스크 분류 모델 적응 방법

김 기 백[a)‡]

Eigenvoice Adaptation of Classification Model for Binary Mask Estimation

Gibak Kim[a)‡]

요 약

본 논문에서는 잡음 환경에서 취득된 음성 신호에서 잡음을 제거하기 위한 방법으로 사용되는 이진 마스크 분류 모델의 적응과정에 대해 다루고자 한다. 기존 연구결과에 의하면, 잡음 환경 데이터에 이진 마스크 기법을 적용하면 음성 명료도를 향상시킬 수 있다고 알려져 있다. 하지만 이진 마스크 분류 모델 학습 시 테스트 환경 데이터가 포함되어야 한다는 단점을 안고 있다. 본 논문에서는 새로운 잡음 환경에서 이진 마스크 분류 모델을 적응하기 위해, 음성 인식에서 널리 사용되는 화자 적응 기법인 eigenvoice 방법을 적용하고자 한다. 실험결과에서는 모델 적응에 사용되는 데이터량에 따른 성능을 정검출율과 오검출율 관점에서 평가하였고, 그 결과 새로운 잡음 환경에서 데이터량을 증가시켜 모델을 적응함으로써 향상된 성능을 나타냄을 확인할 수 있었다.

Abstract

This paper deals with the adaptation of classification model in the binary mask approach to suppress noise in the noisy environment. The binary mask estimation approach is known to improve speech intelligibility of noisy speech. However, the same type of noisy data for the test data should be included in the training data for building the classification model of binary mask estimation. The eigenvoice adaptation is applied to the noise-independent classification model and the adapted model is used as noise-dependent model. The results are reported in Hit rates and False alarm rates. The experimental results confirmed that the accuracy of classification is improved as the number of adaptation sentences increases.

Keyword : Noise reduction, Binary mask estimation, Environment 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: http://orcid.org/0000-0001-5114-4117
· Manuscript received October 17, 2014 Revised November 26, 2014 Accepted December 2, 2014

Ⅰ. 서 론

잡음 환경에서 취득된 음성의 음질을 개선하기 위한 많은 연구들이 진행되어 왔다. 하지만 잡음을 제거하여 음질을 개선하는 많은 연구들이 명료도는 의미 있게 개선하지 못함을 알 수 있다[1-3]. 최근에는 이진 마스크 기법을 잡음

이 섞인 음성에 적용하여 명료도를 개선하는 연구가 제안되었다[4-6]. 제안된 방법에서는 입력 신호를 시간-주파수 영역으로 분해하여 각 시간-주파수 영역에서 목적 신호 (음성)의 에너지가 간섭 잡음 신호의 에너지보다 큰 경우는 그대로 두고, 그렇지 않은 경우는 제거하는 방식을 적용한다. 이러한 이진 마스크를 추정하기 위해서는 분류기를 설계하고 특정한 잡음 환경에서 학습 데이터를 수집하여 분류 모델을 학습시키는 방법을 이용할 수 있다. 그러나 이러한 방법은 학습된 잡음 환경 하에서만 동작한다는 단점을 지니고 있다.

본 논문에서는 이러한 단점을 보완하기 위해 학습데이터에 포함되지 않은 새로운 잡음 환경에 노출되었을 때, 분류 모델을 적응시키는 방법을 다루고자 한다. 즉, 여러 잡음 환경에서 취득된 데이터로 일차적인 분류 모델을 생성한 후, 새로운 잡음 환경 데이터를 이용하여 기존 분류 모델을 적응시키는 방법을 시도하고자 한다. 이진 마스크 추정을 위한 분류 모델은 가우시안 혼합 모델을 사용하고, 이를 새로운 환경에 적응시키기 위해서 음성 인식 화자 적응에 널리 사용되는 eigenvoice 기법을 사용한다.

Eigenvoice 기법은 얼굴 인식 방법으로 제안된 eigenface 라는 알고리즘을 음성 신호 처리에 적용한 것으로서 화자들 간의 변이를 대표하는 기저벡터를 설정하고 적응하고자 하는 화자에 대해 기저벡터 성분의 가중치를 추정하여 화자에 적응된 음향 모델을 추정하는 방법이다[7]. 본 논문에서는 eigenvoice를 이용한 화자 적응 알고리즘을 새로운 잡음 환경에서의 이진 마스크 추정을 위한 분류 모델 적응에 적용해보고자 한다. 이진 마스크는 “0”과 “1”로 구분되므로 두 개의 가우시안 혼합 모델이 필요하다. 먼저, 여러 잡음 환경에서 취득한 데이터로부터 잡음 환경 독립 가우시안 혼합 모델을 생성한다. 이렇게 생성된 잡음 환경 독립 가우시안 혼합 모델을 eigenvoice 알고리즘을 적용하여 새로운 잡음 환경에 대해 적응시킨다.

제안하는 방법의 성능 검증을 위해 새로운 잡음 환경에 대해 10개의 문장에서부터 50개까지의 문장을 사용하여 모델을 업데이트하고, 이진 마스크 분류기를 적용한 후, 정검출율 (Hit)과 오검출율 (False Alarm)을 계산한다. 실험 결과, 제안하는 방법을 적용하여 새로운 잡음 환경에서 모델 적응을 위한 데이터량을 증가시킬수록 향상된 이진 마스크 분류 성능을 나타냄을 알 수 있다.

본 논문의 나머지 구성은 다음과 같다. 2장에서는 분류 모델 기반의 이진 마스크 추정에 대해 설명하고, 3장에서는 eigenvoice 알고리즘을 이용한 모델 적응에 대해 설명한다. 이진 마스크 분류 성능에 대한 실험결과는 4장에서 제시한다.

II. 분류 모델 기반의 이진 마스크 추정

잡음 환경에서 녹음된 음성 신호를 시간-주파수 영역으로 분해하여 각 유닛의 이진 마스크를 추정하는 과정은 크게 특징 벡터 추출과 학습과 판별 과정으로 나뉘게 된다. 본 논문에서 사용하는 특징 벡터는 Tchorz와 Kollmeier가 사용했던 진폭 변조 스펙트로그램 (AMS)[8-10]을 기초로 하고 있으며 구체적인 특징 벡터 추출과정은 그림 1과 같이 정리할 수 있다[6,11]. 이진 마스크의 분류를 위해서는 베이지안 (Bayesian) 분류기를 사용하였다. 즉, 마스크 “0”에 해당하는 확률 모델과 마스크 “1”에 해당하는 확률 모델을 생성하고, 테스트 음성의 특징 벡터가 주어졌을 때, 마스크 “0”에 해당할 사후 (a posteriori) 확률과 마스크 “1”에 해당할 사후 확률을 각각 구하여 서로 비교하여 이진 마스크를 추정한다. 각 이진 마스크의 확률 모델로는 가우시안 혼합 모델 (Gaussian Mixture Model: GMM)을 사용하였다. 학습단계에서 이진 마스크에 대한 확률 모델이 λ_0와 λ_1이 학습되고 나면, 각 시간 프레임 m, 주파수 대역 n에 대한 이진 마스크 $I(m,n)$은 다음 식에 의해 추정된다.

$$I(m,n)=\begin{cases}0, & \text{if } P(\lambda_0|\vec{o}(m,n)) > P(\lambda_1|\vec{o}(m,n)) \\ 1, & \text{otherwise}\end{cases}. \quad (1)$$

여기서 $\vec{o}(m,n)$은 잡음 환경 신호의 특징벡터를 나타내고 사후확률 $P(\lambda_0|\vec{o}(m,n))$은 베이즈 정리를 이용하여 다음 식으로 계산된다.

$$P(\lambda_0|\vec{o}(m,n)) = \frac{P(\lambda_0,\vec{o}(m,n))}{P(\vec{o}(m,n))} = \frac{P(\lambda_0)P(\vec{o}(m,n)|\lambda_0)}{P(\vec{o}(m,n))}. \quad (2)$$

$P(\vec{o}(m,n)|\lambda_0)$는 GMM으로 학습된 모델에 특징벡터를 대입하여 계산할 수 있고 $P(\lambda_0)$는 학습데이터로부터 구할 수 있는 사전확률이다. $P(\vec{o}(m,n)|\lambda_1)$도 유사한 방법으로 구할 수 있다. 보다 구체적인 내용에 대해서는 기존에 발표되었던 논문[6]을 참고하기 바란다.

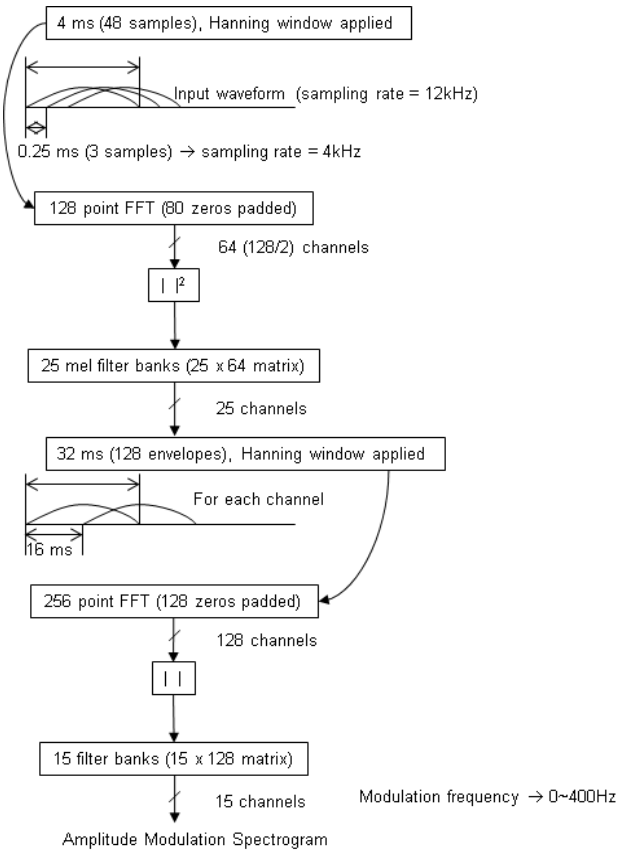

그림 1. 진폭 변조 스펙트로그램의 추출 과정
Fig. 1. Extraction of amplitude modulation spectrogram

III. Eigenvoice를 이용한 모델 적응

1. 화자 적응을 위한 eigenvoice

그림 2에 화자 적응을 위한 eigenvoice 알고리즘을 요약하였다[12,13]. 먼저, R개의 화자 종속 모델과 하나의 화자 독립 모델을 구성한다. 각 화자 종속 모델로부터 모델 업데이트에 필요한 모든 파라미터를 포함하는 슈퍼벡터를 추출한다. 추출된 R개의 슈퍼벡터는 PCA (Principal Component Analysis)를 적용하여 R개의 eigenvector를 얻고 그 중에서 eigenvalue가 큰 K개만을 선택한다. R개의 슈퍼벡터 평균과 PCA를 적용해서 얻은 K개의 eigenvector들을 모은 것을 eigenvoice라 부른다. 새로운 화자에 대해 모델을 적응하기 위해서는 먼저 화자에 대한 슈퍼벡터는 eigenvoice들의 선형 가중 조합으로 구할 수 있다고 가정한다. 그 다음 가중치들을 추정하여 슈퍼벡터로부터 화자 적응 모델의 파라미터를 구한다. 슈퍼벡터에서 표현되지 않는 파라미터는 화자 독립 모델의 파라미터를 사용한다.

보다 구체적인 내용은 다음과 같다. 슈퍼벡터는 다음 식과 같이 K개 eigenvoice들의 선형 가중 조합으로 표현한다.

$$\boldsymbol{p} = \boldsymbol{e}(0) + \sum_{k=1}^{K} w(k)\boldsymbol{e}(k) \tag{3}$$

여기서 $\boldsymbol{e}(0)$는 R개의 화자 종속 모델로부터 추출된 슈퍼벡터들의 평균이고, $\boldsymbol{e}(k)$는 k번째 eigenvoice이다. M개의 가우시안 모델 평균값을 업데이트하는 경우는, 가중치를 추정하기 위해 MLED (Maximum Likelihood Eigen-Decomposition) 방법이 사용된다. 가우시안 혼합 모델의 공분산 행렬과 혼합 가중치는 화자 독립 모델의 값들을 그대로 사용한다. 각 주파수 대역에서, 가우시안 혼합 모델 (평균) 적응을 위해 T개의 시계열 데이터 $(\boldsymbol{o}_1 \cdots \boldsymbol{o}_T)$ 가 주어졌을 때, 가중치 $\boldsymbol{W}(=[w_1 \cdots w_K]^T)$를 추정하여 식 (3)과 같은 새로운 슈퍼벡터를 업데이트한다. ML (Maximum Likelihood) 방법으로 가중치 $\boldsymbol{W}_s$를 추정하기 위해서는 다음과 같은 보조함수 Q를 최대화한다.

$$Q(\lambda|\bar{\lambda}) = -\sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_m(t)(\boldsymbol{o}_t - \hat{\boldsymbol{\mu}}_m)^T \boldsymbol{C}_m^{-1}(\boldsymbol{o}_t - \hat{\boldsymbol{\mu}}_m) \tag{4}$$

여기서

$$\hat{\boldsymbol{\mu}}_m = \boldsymbol{\mu}(0) + \sum_{k=1}^{K} w(k)\boldsymbol{\mu}_m(k) \tag{5}$$

이고, $\gamma_m(t)$는 o_t가 m번째 혼합 가우시안에서 나왔을 확률이며, $\boldsymbol{\mu}_m, \boldsymbol{C}_m^{-1}$ 은 m번째 가우시안의 평균과 공분산 행렬을 나타낸다. $\overline{\lambda}, \lambda$는 각각 초기 모델과 재추정된 모델을 나타낸다. 식 (4)에 주어진 Q함수를 최대화하는 가중치 $\boldsymbol{W}$를 찾는 방식이 바로 MLED방법이다.

2. 이진 마스크 분류 모델 적응

Eigenvoice[1] 방법을 이용하여 새로운 잡음 환경에서 이진 마스크 분류 모델을 적응하는 과정은 다음과 같다. 이진 마스크는 0 또는 1의 값을 가지므로 마스크 "0"과 마스크 "1"에 해당하는 모델이 필요하다. 먼저, R개의 잡음 환경에 대해 마스크 "0"과 마스크 "1"에 해당하는 잡음 환경 종속 가우시안 혼합 모델을 생성한다. 이렇게 생성된 R개의 모델을 이용하여 R개의 슈퍼벡터를 만든다. 잡음 환경 종속 모델은 M개의 가우시안을 갖는 가우시안 혼합 모델을 사용하며, 가우시안들의 평균들을 모두 연결하여 슈퍼벡터를 만든다. 본 연구에서는 38차원의 AMS 특징벡터를 사용하므로 슈퍼벡터의 차원은 38×M이 된다. R개의 잡음 환경에 대해 각각의 잡음 환경 종속 모델을 생성하고, 추가로 R개 잡음 환경 데이터를 모두 학습에 참여시켜 하나의 잡음 환경 독립 모델을 만든다. 그 과정을 정리하면 다음과 같다.

1) R개의 잡음 환경 종속 모델로부터 eigenvoice를 추출하기 위해 PCA (Principal Component Analysis)를 적용하여 K개의 eigenvoice를 채택한다.
2) 새로운 잡음 환경에서 데이터가 수집되면, MLED (Maximum Likelihood Eigen-Decomposition) 방법으로 eigenvoice들의 가중치들을 추정한다.
3) 이렇게 추정된 가중치들과 기존에 추정된 eigenvoice를 이용하여 잡음 환경 종속 모델의 가우시안 평균들을 업데이트한다.
4) 업데이트된 가우시안 평균들과 잡음 환경 독립 모델로부터 가져온 공분산 행렬과 가우시안 혼합 가중치들

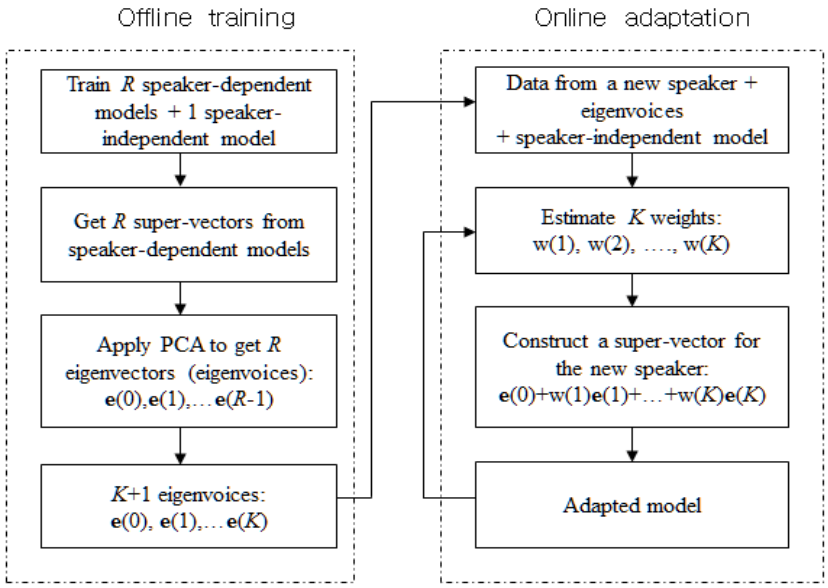

그림 2. Eigenvoice를 이용한 모델 적응 과정
Fig. 2. Block diagram for eigenvoice model adaptation

1) Eigenvoice는 화자 변이 특성을 대표하는 슈퍼벡터들을 지칭하는 용어이다. 본 논문에서 다루는 내용인 새로운 잡음 환경에서의 이진 마스크 분류 모델 적응에 적용할 때는, eigenvoice는 화자 변이가 아니라 잡음 환경 변이 특성을 대표하는 슈퍼벡터들을 지칭하고 있다.

이 새로운 잡음 환경 적응 모델로 사용된다.

IV. 실험 결과

1. 실험 환경

성능 검증을 위한 실험에 사용한 문장은 IEEE문장 데이터베이스[14]로서 영어로 구성되어 있고, 영어 원어민에 의해 발성된 것을 25kHz 샘플링 주파수로 녹음하여 12kHz로 다운 샘플링하였다. 가우시안 혼합 모델에 사용한 가우시안 개수(M)는 64이다. 여러 잡음 환경 모델 생성을 위해 34가지 잡음[2]을 사용하였고, 테스트를 위해서는 모델 생성에 참여하지 않은 Babble 잡음, factory 잡음, speech-shaped 잡음을 사용하였다. Babble 잡음은 남녀 각각 10명이 동시에 서로 다른 문장을 읽는 것을 녹음하여 생성한 잡음을 사용하였고, factory 잡음은 NOISEX 데이터베이스[15]에서 발췌하였으며, speech-shaped 잡음은 IEEE문장 데이터베이스의 음성들의 평균 스펙트럼을 갖는 stationary 잡음이다. 한 문장에는 10 개 내외의 단어들이 포함되어 있고 (2~3초 정도의 길이), 학습 데이터로는 200개의 문장을 사용하였으며, 신호 대 잡음 비 (SNR) –5dB, 0dB, 5dB 등으로 잡음을 섞어 사용하였다. 테스트를 위해서는 –5dB 10개의 문장을 사용하였다. 테스트를 위해 사용한 데이터는 모델 적응을 위해 사용된 데이터와는 겹치지 않는다. 그림 1에서 보듯이 신호를 32ms 윈도우를 이용하여 16ms씩 이동하며 시간영역으로 분해하였고, 주파수영역으로는 mel스케일 25개의 필터뱅크를 이용하여 분해하였다.

2. 실험 결과

제안하는 방법을 평가하기 위해, 잡음 환경 적응에 사용된 문장 수와 각 경우에 대해 eigenvoice의 개수(K)를 변화시키며 이진 마스크의 정검출율 (Hit) 오검출율 (FA)을 계산하였다. 정검출율은 높을수록 좋은 성능을 나타내고, 오검출율은 낮을수록 좋은 성능을 나타내는 것을 의미하므로 두 측정값의 차이 (Hit-FA)가 클수록 분류기의 성능이 좋은 것이라고 할 수 있다. 또한 Hit-FA는 음성 명료도와 높은 상관관계를 보이는 것으로 알려져 있어 이진 마스크를 이용한 명료도 향상 알고리즘의 성능을 Hit-FA를 측정하여 가늠할 수 있다.

모델 적응에 사용된 문장 수와 eigenvoice의 개수에 따른 성능을 그림 3에 나타내었다. 세 가지 잡음 환경 모두, 새로

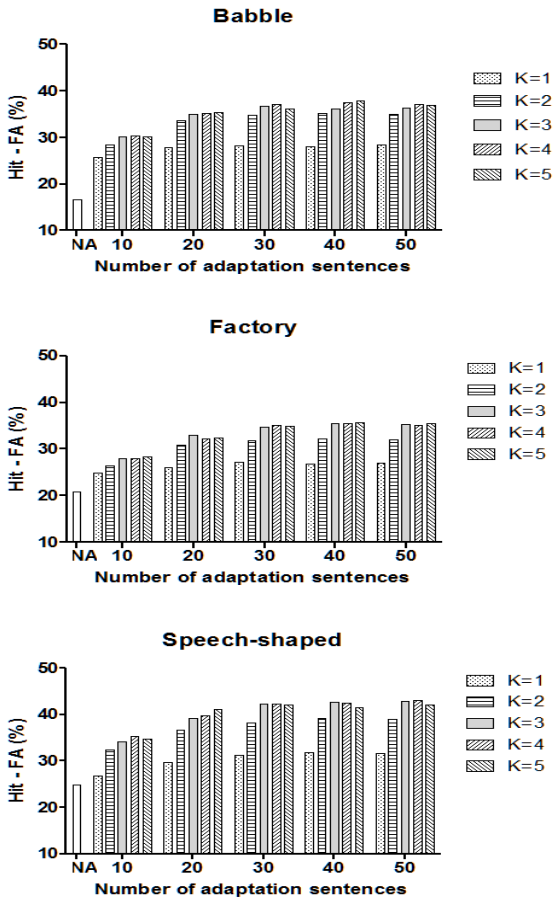

그림 3. Eigenvoice를 이용한 모델 적응 결과 (정검출율 - 오검출율: Hit - FA, NA: No Adaptation)
Fig. 3. Performance of eigenvoice model adaptation

2) 34가지 잡음은 다음과 같다. Airplane cabin in flight, Alarm (burglar alarm), Applause (in theater), Boat engine room, Car ventilation fan, Car wash, Cooker hood, Crowd at airport, Crowd at convention, Crowd outdoor, Dragster idle, Drum dryer, Electricity transformer, Extractor fan, Fan heater, Food mixer, Hair cutter, Hair dryer, Hand dryer, Helicopter idle, Helicopter hovering, Lawn mower, Motorcycle 1, Motorcycle 2, Nose trimmer, Outdoor 1, Outdoor 2, Rain, Restaurant, Shaver, Slot machine, Traffic, horn, Vacuum, Washing machine

운 잡음 환경 적응을 하기 전에는 성능이 20% 내외로서 상당히 낮은 결과가 나왔다. 적응에 사용된 문장이 증가함에 따라 성능이 향상되나 20개를 넘어서면서부터는 성능향상이 크지 않음을 알 수 있다. 또한 eigenvoice의 개수(K)가 1일 때와 2일 때의 성능차이는 크게 나타나지만 3~5범위에서는 유의미한 차이를 발견할 수 없었다.

V. 결 론

기존 연구에 따르면, 잡음 환경에서 취득된 신호에 이진 마스크 기법을 적용하여 음성 명료도를 향상시킬 수 있다. 이진 마스크를 적용하기 위해서는 잡음 환경 데이터를 사용하여 분류 모델을 학습이 필요한데, 본 논문에서는 새로운 잡음 환경에 대한 분류 모델 적응을 위해서 eigenvoice 화자 적응 방법을 적용하였다. 여러 잡음 환경 데이터를 학습시켜 생성한 잡음 환경 독립 모델을 이용하여 이진 마스크를 추정했을 때에 비해 eigenvoice 화자 적응 방법을 적용하여 잡음 환경 독립 분류 모델을 적응시킴으로써 이진 마스크 분류 성능을 향상시킬 수 있었다. 본 연구에 대한 후속으로 다양한 화자 적응 방법을 적용하여 최적의 성능을 발휘하는 모델 적응에 대한 연구를 진행할 예정이다.

참 고 문 헌 (References)

[1] Y. Hu and P. C. Loizou, “Subjective comparison and evaluation of speech enhancement algorithms,” Speech communication, vol. 49, no. 7, pp. 588–601, Jul. 2007.

[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.

[3] Y. Hu and P. C. 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.

[4] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, “Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation,” The Journal of the Acoustical Society of America, vol. 120, no. 6, p. 4007, 2006.

[5] N. Li and P. C. Loizou, “Factors influencing intelligibility of ideal binary-masked speech: implications for noise reduction.,” The Journal of the Acoustical Society of America, vol. 123, no. 3, pp. 1673–82, Mar. 2008.

[6] G. Kim, Y. Lu, Y. Hu, P. C. Loizou, “An algorithm that improves speech intelligibility in noise,” Journal of Acoustical Society of America, September 2009.

[7] R. Kuhn, J. Junqua, P. Nguyen, N. Niedzielski, "Rapid Speaker Adaptation in Eigenvoice Space," IEEE Trans. Speech and Audio Proc., vol. 8, no. 6, pp. 695-707, November 2000.

[8] 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.

[9] J. Tchorz and B. Kollmeier, “SNR estimation based on amplitude modulation analysis with applications to noise suppression,” IEEE Trans. on Speech and Audio Processing, vol. 11, no. 3, pp. 184–192, May 2003.

[10] M. Kleinschmidt and V. Hohmann, “Sub-band SNR estimation using auditory feature processing,” Speech Communication, vol. 39, no. 1–2, pp. 47–63, Jan. 2003.

[11] Gibak Kim, “A Post-processing for Binary Mask Estimation Toward Improving Speech Intelligibility in Noise,” JBE, Vol. 18, No.2, pp.311-318, March, 2013.

[12] N. Iwahashi, A. Kawasaki, “Speaker Adaptation in noisy environments based on parameter estimation using uncertain data,” In Proc. Intl. Conf. on Spoken Language Processing, Vol. 4, pp. 528-531, October 2000.

[13] M. Kirby and L. Sirovich, “Application of the Karhunen–Loève procedure for the characterization of human faces,” IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 12, No. 1, pp. 103–108, January 1990..

[14] IEEE, “IEEE recommended practice for speech quality measurements",” IEEE Trans. Audio Electroacoust., vol. 17, pp. 225-246, 1969.

[15] 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 Commun., vol. 12, pp. 247-251, 1993.

저 자 소 개

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
- 2010년 ~ 2011년 : 대구대학교 전자공학부 전임강사
- 2011년 ~ 현재 : 숭실대학교 전기공학부 조교수
- ORCID : http://orcid.org/0000-0001-5114-4117
- 주관심분야 : 음성신호처리, 영상신호처리, 멀티모달신호처리, 어레이신호처리