http://dx.doi.org/10.7840/kics.2016.41.8.851

확률적 비음수 행렬 인수분해를 사용한 통계적 음성검출기법

김 동 국◆, 신 종 원°, 권 기 수*, 김 남 수*

Statistical Voice Activity Detection Using Probabilistic Non-Negative Matrix Factorization

Dong Kook Kim◆, Jong Won Shin°, Kisoo Kwon*, Nam Soo Kim*

요 약

본 논문은 비음수 행렬 인수분해(NMF)의 확률적 해석에 근거한 새로운 통계적 음성검출기법을 제안한다. NMF의 기저와 부호화 행렬들이 주어졌을 때, 데이터 행렬의 분포를 Poisson 분포로 가정한 로그 우도는 Kullback-Leibler 발산을 이용한 NMF의 목적 함수와 일치한다. 이러한 NMF의 확률모델에 근거하여 음성검출을 위해 DFT영역에서 잡음과 음성의 크기 스펙트럼을 Poisson 분포로 모델링하여 새로운 우도비 검출 규칙을 유도한다. 실험 결과를 통해 제안된 기법이 0-15dB 신호 대 잡음비의 시뮬레이션 환경에서 기존 Gaussian과 NMF을 사용한 기법보다 향상된 음성검출 결과를 보여준다.

Key Words : voice activity detection, NMF, Poisson distribution, likelihood ratio test.

ABSTRACT

This paper presents a new statistical voice activity detection (VAD) based on the probabilistic interpretation of nonnegative matrix factorization (NMF). The objective function of the NMF using Kullback-Leibler divergence coincides with the negative log likelihood function of the data if the distribution of the data given the basis and encoding matrices is modeled as Poisson distributions. Based on this probabilistic NMF, the VAD is constructed using the likelihood ratio test assuming that speech and noise follow Poisson distributions. Experimental results show that the proposed approach outperformed the conventional Gaussian model-based and NMF-based methods at 0-15 dB signal-to-noise ratio simulation conditions.

Ⅰ. 서 론

음성검출기법(voice activity detection, VAD)은 잡음이 섞여있는 음성신호에서 음성이 존재하는 부분과 잡음만 존재하는 부분을 검출하는 기술이다. 현재 음성검출은 음성신호처리와 관련된 음성코딩, 음성향상, 음성인식 등에서 널리 사용되고 있으며, 특히 이동통신 시스템에서 대역폭을 효율적으로 사용하기 위해 필수적으로 요구되고 있다.

※ 이 연구는 방위사업청 및 국방과학연구소의 재원에 의해 설립된 신호정보 특화연구센터 사업의 지원을 받아 수행되었음.
◆ First Author : Chonnam National University, School of Electronic and Computer Engineering, dkim@jnu.ac.kr, 정회원
° Corresponding Author : Gwangju Institute of Science and Technology, School of Electrical Engineering and Computer Science, jwshin@gist.ac.kr, 정회원
* Seoul National University, Department of Electrical and Computer Engineering and the Institute of New Media and Communications, kskwon@hi.snu.ac.kr, 학생회원, nkim@snu.ac.kr, 종신회원
논문번호 : KICS2016-05-092, Received May 12, 2016; Revised July 7, 2016; Accepted July 12, 2016

음성검출 기술은 크게 통계적 모델기반과 기계학습 기반 기법으로 분류된다[1-4]. 통계적 모델기반 기법은 DFT(discrete Fourier transform)와 같은 변환영역에서 동작하며, 우도비 검증(likelihood ratio test, LRT)을 결정규칙으로 사용한다[1,2]. 이 방법에 의해 음성검출을 잘 수행하기 위해서는 각 영역에서 음성과 잡음에 대한 통계적 특성을 정확히 모델링할 수 있어야 한다. 이 기법의 특징은 통계적 모델과 우도비 검증에 근거해 테스트 잡음음성만을 사용하여 검출과정을 수행하기 때문에 음성과 비음성에 대한 프레임 레벨의 정보를 사용하는 학습과정이 필요 없는 장점을 갖는다. 그러나 신호 대 잡음비(signal-to-noise ratio, SNR)이 높은 환경 하에서는 비교적 높은 음성검출 성능을 보이지만, 신호 대 잡음비가 낮은 열악한 잡음하에서는 성능이 저하되는 단점을 갖고 있다.

한편 기계학습 기반 기법은 SVM(support vector machine)이나 DNN(deep neural network)와 같은 기계학습 분야에서 사용되고 있는 분류기(classifier)를 사용하여 음성과 잡음을 분류하는 기법이다[3,4]. 이러한 기법을 사용하기 위해 음성과 잡음의 변별력을 갖는 특징벡터들을 추출하는 과정이 중요하다. 일반적으로 DFT계수, MFCC, 피치 등 음성으로부터 추출된 다양한 특징벡터을 사용할 뿐아니라, 통계적 음성검출 모델로부터 사전 신호 대 잡음비(a priori SNR), 사후 신호 대 잡음비(a posteriori SNR) 그리고 우도비 값 등이 사용된다[3,4]. 이 기법은 다양한 환경하에서의 잡음음성과 해당되는 음성 또는 잡음을 나타내는 목적(target) 값을 사용하여 학습과정을 통해 검출기를 구축하기 때문에 통계적 모델에 비해 신호 대 잡음비가 낮은 환경에서 더 뛰어난 성능을 나타낸다. 그러나 이러한 기계학습 기반 음성검출기법은 학습에 사용되는 음성파일에 대해 프레임 단위로 음성과 비음성에 대한 레이블 정보가 필요하다. 따라서 학습에 필요한 데이터 베이스를 구축하는 데 많은 시간과 노력이 요구되는 단점이 있다.

최근 비음수 행렬 인수분해(NMF)는 영상처리 뿐아니라 음원분리와 잡음제거 등 음성처리 분야에서 널리 쓰이고 있는 기법이다[5-11]. NMF는 비음수 값을 갖는 데이터 행렬을 비음수 성분을 갖는 기저벡터들의 선형결합으로 근사적으로 분해하는 기법이다[5]. 이러한 NMF을 이용한 몇 가지 음성검출기법들이 최근에 제안되었다[6,7]. 확률적(probabilistic) NMF (PNMF)는 관측된 비음수 데이터에 대해 Poisson 확률분포를 가정하고, 최대유사도(maximum likelihood, ML) 기법과 EM(expectation maximization)알고리즘을 통해 원래의 NMF의 목적함수와 곱 갱신 규칙과 같은 학습 알고리즘을 갖는다[8,9]. 즉 NMF의 기저과 부호화행렬이 주어졌을 때의 데이터 행렬의 분포를 Poisson 분포로 가정하면, 이때의 로그 우도는 Kullback-Leibler (KL) 발산(divergence)을 이용한 원래의 NMF의 목적함수와 일치한다.

본 논문에서는 NMF의 확률적 해석에 근거한 PNMF을 이용한 통계적 음성검출 기법을 제안한다. 원래 음성과 잡음에 대해 Poisson 확률모델을 가정하고 이로부터 잡음음성 모델을 유도한다. 이러한 확률모델에 근거하여 우도비 검증을 통해 통계적 음성기법을 제시한다. 이 기법은 기계학습기반 음성검출과 같은 프레임 레벨의 정보가 필요하지 않고, 음성과 잡음으로 분류된 데이터만을 사용하여 학습하는 특징을 갖는다. 실험결과 기존의 같은 DFT 영역에서 Gaussian 분포를 사용한 기법과 NMF을 사용한 기법에 비해 특히 낮은 신호 대 잡음비에서 성능이 우수함을 나타낸다.

본 논문의 본론 II장에서는 NMF와 PNMF을 간단히 소개하고, PNMF에 근거한 음성모델링과 이에 근거한 통계적 음성검출기법을 제시한다. III장에서는 실험과 결과에 대해 나타내고, IV에서는 결론을 맺는다.

Ⅱ. 본 론

2.1 NMF

두 단원에서 원래의 NMF[5]와 PNMF[8]에 대해 간단히 살펴본다. 원래의 NMF에서는 비음수 성분으로 구성된 데이터 행렬 $V \in R^{M \times N}$을 두 개의 비음수 행렬 $W \in R^{M \times R}$와 $H \in R^{R \times N}$의 곱($V \approx WH$)으로 분해한다. 여기서 W는 기저행렬을, H는 부호화 행렬을 나타낸다[5].

데이터 행렬 V가 주어진 경우, W와 H행렬을 추정하기 위해 V와 WH사이의 차를 나타내는 목적함수 $D(V\|WH)$가 필요하다. 음원분리와 같은 응용분야에서는 목적함수로서 KL 발산을 가장 많이 사용한다. KL 발산 목적함수는 다음과 같다.

$$D(V\|WH) = \sum_{i,j}\left(V_{ij}\log\frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij} \right) \quad (1)$$

위의 목적함수를 최소화하는 학습 방식인 곱갱신(multiplicative update) 규칙은 각 단계에서 W와 H은 다음과 같이 반복적으로 갱신된다.

$$H_{jk} \leftarrow H_{jk} \frac{\sum_i W_{ij} V_{ik} / (WH)_{ik}}{\sum_m W_{mj}} \quad (2)$$

$$W_{ij} \leftarrow W_{ij} \frac{\sum_k H_{jk} V_{ik} / (WH)_{ik}}{\sum_v H_{jv}} \quad (3)$$

여기서 H_{jk}와 W_{ij}는 각각 행렬 H와 W의 jk번째와 ij번째의 원소를 나타낸다.

2.2 PNMF

PNMF는 데이터에 대한 확률적인 모델과 최대유사도에 근거한 목적함수을 사용함으로 원래의 NMF와 같은 기능을 수행한다[8]. 확률모델을 사용하기 위해 먼저 비음수 은닉변수 Z_{ikj}은 평균이 $W_{ik}H_{kj}$인 Poisson 분포를 아래와 같이 갖는다고 가정한다.

$$Z_{ikj} \sim PO(Z_{ikj}; W_{ik}H_{kj}) \quad (4)$$

여기서 $PO(x;\lambda)$는 파라메타 λ를 갖는 다음과 같은 Poisson 분포를 나타낸다.

$$PO(X;\lambda) = \frac{e^{-\lambda}\lambda^X}{X!} \quad (5)$$

그리고 $X!$는 X의 계승(factorial)을 나타낸다. Poisson 분포는 정수 값을 갖는 확률변수에 대해서 정의되므로 실수 값을 갖는 관측 값에 대해서는 적당한 스케일링을 통해 가장 가까운 정수 값으로 근사화 시켜야 한다. 관측행렬 ij번째 성분 V_{ij}는 R개의 독립적인 비음수 은닉변수 Z_{ikj}의 합, $V_{ij} = \sum_k Z_{kit}$으로 표현된다고 가정한다. 독립적인 Poisson분포를 갖는 확률변수의 합은 또한 Poisson 확률변수이며, 이때의 파라메타는 각 파라메타의 합과 같다. 따라서 관측행렬 V_{ij}의 확률분포는 다음과 같다.

$$p(V_{ij};\ W_{i*}, H_{*j}) = PO(V_{ij}; \sum_{k=1}^{R} W_{ik}H_{kj}) \quad (6)$$

여기서 W_{i*}는 행렬 W의 i번째 행을 나타내고, H_{*j}는 행렬 H의 j번째의 열을 나타낸다. W와 H가 주어진 조건하에서 모든 관측행렬 성분들이 서로 독립이라 가정한다면 전체 유사도함수는 다음과 같다.

$$p(V;\ W,H) = \prod_{i,j} \frac{e^{-(WH)_{ij}} (WH)_{ij}^{V_{ij}}}{V_{ij}!} \quad (7)$$

여기서 $(WH)_{ij} = \sum_k W_{ik}H_{kj}$는 행렬 WH의 ij번째 성분을 나타낸다. 파라메타 W와 H의 최대유사도 추정치는 EM알고리즘을 통해 구할 수 있고 그 결과는 잘 알려진 (1)식의 KL 발산 목적함수와 (2)-(3)식과 같은 파라메터 곱갱신 규칙과 같다[8].

2.3 PNMF기반 음성검출기법

2.3.1 PNMF기반 음성모델

시간영역에서 원래의 신호 $s(t)$에 잡음신호 $n(t)$가 더해져 잡음음성 $y(t)$가 주어진다. 이들을 DFT를 통해 주파수 영역으로 변환한다. PNMF을 사용하기 위해 시간 프레임 t에서 S_t, N_t 그리고 Y_t은 원래음성, 잡음 그리고 잡음음성의 각각 DFT 계수의 크기(magnitude) 벡터라 하자. 또한 k번째 주파수 채널의 원래음성과 잡음의 DFT 크기 스펙트럼 $S_{k,t}$와 $N_{k,t}$은 다음과 같이 각각 Poisson 분포를 따른다고 가정한다.

$$p(S_{k,t};\ W_s, H_{s,t}) = PO(S_{k,t}; (W_s H_{s,t})_k) \quad (8)$$

$$p(N_{k,t};\ W_n, H_{n,t}) = PO(N_{k,t}; (W_n H_{n,t})_k) \quad (9)$$

여기서 W_s와 W_n는 각각 원래음성과 잡음을 위한 기저 행렬이고, $H_{s,t}$와 $H_{n,t}$는 시간 프레임 t에서 부호화 벡터이다. 그리고 $(W_s H_{s,t})_k$와 $(W_n H_{n,t})_k$는 k번째 주파수 채널의 각각 $S_{k,t}$와 $N_{k,t}$의 평균을 나타낸다. 각각 원래음성과 잡음의 기저 행렬은 각각의 데이터를 사용하여 훈련을 통해 얻게 된다. NMF 신호 모델 하에서는 잡음음성의 DFT 계수의 크기가 원래음성과 잡음의 DFT 계수의 크기의 합에 의해 근사화 된다고 가정한다[9],[10].

$$Y_{k,t} \approx S_{k,t} + N_{k,t} \quad (10)$$

그러면 Poisson 분포의 성질에 의해 잡음음성 $Y_{k,t}$도 다음과 같은 Poisson 분포를 갖는다.

$$p(Y_{k,t};\, W_s, H_{s,t}, W_n, H_{n,t}) = PO(Y_{k,t};(W_sH_{s,t}+W_nH_{n,t})_k) = PO(Y_{k,t};(WH_t)_k) \quad (11)$$

여기서 $W=[W_s,\ W_n]$와 $H_t=[H_{s,t};H_{n,t}]$이다. 즉 잡음음성을 위한 기저벡터 W는 W_s와 W_n을 결합함으로 얻어진다. 잡음음성의 크기 벡터 Y_t가 주어진 경우, 고정된 W에 대해 부호화 행렬 H_t을 추정화하기 (2)식을 반복적으로 수행한다.

2.3.2 PNMF기반 음성검출기

이 장에서는 위에서 주어진 PNMF을 사용하여 새로운 통계적 음성검출기법을 제안한다. 주어진 가설 H_0, H_1이 각각 음성 부재와 존재를 나타낼 때, 시간 프레임 t에서 각 주파수 채널별로 DFT 계수의 크기는 PNMF 가정 하에서 다음과 같이 표현된다.

$$H_0 : \text{음성부재} \ :\ Y_t = N_t \quad (12)$$

$$H_1 : \text{음성존재} \ :\ Y_t = S_t + N_t \quad (13)$$

여기서 $Y_t=[Y_{1,t},\cdots,Y_{K,t}]^T$, $S_t=[S_{1,t},\cdots,S_{K,t}]^T$ 그리고 $N_t=[N_{1,t},\cdots,N_{K,t}]^T$는 각각 잡음에 오염된 음성신호, 원래의 음성신호 및 잡음의 DFT 계수의 크기 벡터를 나타낸다. 그리고 K은 전체 주파수대역의 개수이고 T는 전치행렬을 나타낸다. 음성과 잡음신호의 DFT 계수의 크기가 각 주파수 채널별로 Poisson 분포를 따른다는 가정으로부터 각각의 가설 H_0와 H_1을 조건으로 한 분포함수는 다음과 같다.

$$p(Y_{t,k}|H_0) = \frac{e^{-(W_nH_{n,t})_k}(W_nH_{n,t})_k^{Y_{k,t}}}{Y_{k,t}!} \quad (14)$$

$$p(Y_{k,t}|H_1) = \frac{e^{-(W_sH_{s,t}+W_nH_{n,t})_k}(W_sH_{s,t}+W_nH_{n,t})_k^{Y_{k,t}}}{Y_{k,t}!} \quad (15)$$

위와 같은 가설하에서 시간 프레임 t에서 k번째 주파수 대역별 우도비는 다음과 같다.

$$\Psi_{k,t} = \frac{p(Y_{k,t}|H_1)}{p(Y_{k,t}|H_0)} = e^{-(W_sH_{s,t})_k}\left(\frac{(W_sH_{s,t}+W_nH_{n,t})_k}{(W_nH_{n,t})_k}\right)^{Y_{k,t}} \quad (16)$$

(16)식의 유사도를 계산하기 위해 잡음과 원래 음성의 부호화 벡터 $H_{n,t}$와 $H_{s,t}$을 잘 추정해야 한다. 이를 위해 먼저 잡음음성의 크기 벡터 Y_t와 기저벡터 W가 주어진 경우, (11)식의 유사도를 최대화 하는 부호화 벡터의 추정치, $\hat{H}_t=[\hat{H}_{s,t};\hat{H}_{n,t}]$는 (2)식을 반복적으로 갱신하므로 얻어진다. 따라서 이러한 추정치를 바탕으로 시간 프레임 t에서 잡음과 원래음성의 평균 추정치는 각각 다음과 같이 구해진다.

$$\hat{N}_{k,t} = (W_n\hat{H}_{n,t})_k \quad (17)$$

$$\hat{S}_{k,t} = (W_s\hat{H}_{s,t})_k \quad (18)$$

실제로 잡음과 원래음성이 시불변(stationary)하다는 가정 하에 부호화 벡터를 다음과 같이 평활화를 수행한다[11].

$$\hat{N}_{k,t} = \alpha_n\hat{N}_{k,t-1} + (1-\alpha_n)(W_n\hat{H}_{n,t})_k \quad (19)$$

$$\hat{S}_{k,t} = \alpha_s\hat{S}_{k,t-1} + (1-\alpha_s)(W_s\hat{H}_{s,t})_k \quad (20)$$

여기서 α_n와 α_s는 각각 잡음과 음성을 위한 평활화 상수이다. 최종적으로 우도비의 기하평균을 임계값 η와 비교함으로서 다음과 같은 결정 식을 얻을 수 있다.

$$\ln\Psi_t = \frac{1}{K}\sum_{k=1}^{K}\ln\Psi_{k,t} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \quad (21)$$

여기서 $\ln\Psi_{k,t}$는 (16), (19) 그리고 (20)식으로부터 다음과 같이 주어진다.

$$\ln\Psi_{k,t} = -\hat{N}_{k,t} + Y_{k,t}\ln\left(\frac{\hat{S}_{k,t}}{\hat{N}_{k,t}}+1\right) \quad (22)$$

2.3.3 기존 NMF 기반 음성검출기법과 비교

기존에 NMF을 이용한 음성검출 기법들은 기저벡터들을 이용하는 기법과 부호화 값을 이용하는 기법으로 분류된다[6,7]. 기저벡터를 이용하는 기법[6]은 먼저 시작부분의 프레임들을 음성부재 구간으로 가정한 후 배경 잡음에 대한 기저벡터들을 추출한 후 평균 기저벡터를 구한다. 그리고 각 입력신호의 기저벡터를 구해 잡음 기저벡터와의 거리를 계산하여 음성 활성

화 여부를 판단한다. 이 기법의 특징은 미리 깨끗한 음성과 잡음에 대한 기저벡터들을 구하지 않고 검출하는 과정에서 잡음에 대한 기저만을 추출하여 음성검출에 이용한다는 것이다. 반면 부호화 값을 이용하는 기법[7]은 깨끗한 음성 또는 잡음에 대한 기저벡터들을 미리 학습을 통해 구한다. 그리고 이들을 이용하여 음성검출을 위해 입력신호의 각 프레임에 대해 음성기저에 해당되는 부호값들을 아래와 같이 합하여 이를 임계값과 비교하여 음성 존재 여부를 판단한다.

$$\overline{H}_{s,t} = \sum_{k=1}^{R_s} (\hat{H}_{s,t})_k \tag{23}$$

여기서 R_s는 음성기저벡터의 수를 나타낸다. 위의 두 가지 기존 NMF 기반 음성검출기법들은 경험적으로 NMF의 기저벡터와 부호값들을 이용하여 음성을 검출한다. 반면에 제안된 PNMF 기반 기법은 음성과 잡음에 대한 Poisson 확률 모델을 사용하여 통계적 모델기반으로 음성을 검출한다는 부분이 가장 큰 차이점이라 할 수 있다.

Ⅲ. 실험 및 결과

제안된 음성검출 알고리즘을 평가하기 위해 여러 가지 잡음 환경하에서 실험을 수행하였다. 먼저 PNMF을 학습하여 음성과 잡음에 대한 각각의 기저행렬을 구하기 위해 TIMIT과 NoiseX-92 데이터베이스를 사용하였다. 8 kHz로 샘플링된 음성과 잡음에 대해 한 프레임으로 20 msec(160샘플)을 사용 하였고 50% 중첩(overlap) 하였다. 160-포인트 DFT을 수행한 후 각 주파수 대역별로 크기를 구하였다. 각각의 잡음 형태에 대한 기저 행렬을 구하기 위해 테스트에 포함되지 않은 60초 길이의 잡음 파형이 사용되었고, 음성에 대해서는 56(남자 :28, 여자:28)명의 화자에 의해 발성된 130초 길이의 음성파형이 학습에 사용되었다. 잡음과 음성에 대한 기저벡터의 수는 각각 128개로 설정하였다. 음성검출을 위한 테스트 음성으로는 별도로 녹음되어진 10 ms단위로 수동으로 얻어진 음성/비음성 레이블을 갖는 456초의 음성을 사용하였다. 전체 신호 중에서 음성의 비율은 58.2%이고, 이 중 44.85%는 유성음(voiced sounds)이며 13.4%는 무성음(unvoiced sound)이었다. 잡음에 오염된 음성신호를 만들기 위해 babble, factory 그리고 pink잡음을 NOISEX-92 잡음으로부터 신호 대 잡음비를 0, 5, 10, 15dB로 변화하면서 원래의 음성신호에 첨가하였다. 잡음과 원래 음성을 추정하기 위한 평활화 상수 값, $\alpha_n = 0.98$와 $\alpha_s = 0.1$을 사용하였다.

음성검출의 성능의 지표로 검출(detection)과 오경보(false-alarm, FA) 확률 P_D와 P_{FA}를 사용하여 평가 하였다. P_D는 실제로 정확하게 음성이라고 판단할 확률을 뜻하고, P_{FA}는 비음성을 음성이라 잘못 판단할 확률을 뜻한다. 제안된 PNMF 기반 음성검출기의 성능을 평가하기 위해 기존의 Gaussian 분포를 사용한 DFT 기반 기법[1]와 (23)식을 이용하는 기존 NMF 기반 기법[7]을 사용하여 비교하였다. 또한 ETSI

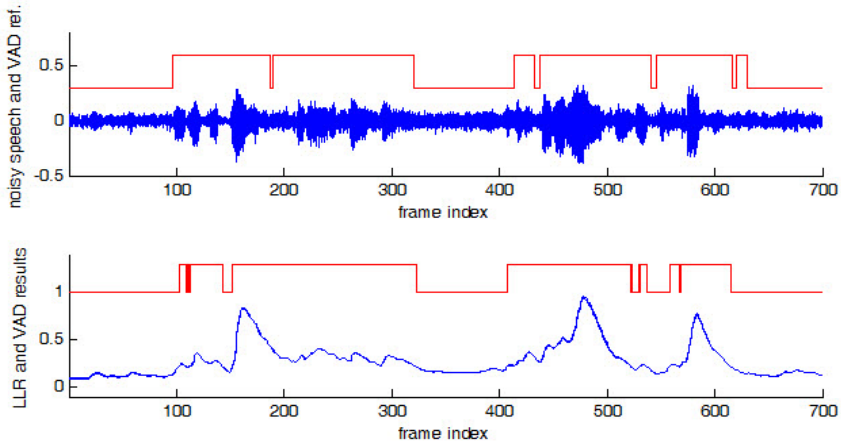


그림 1. Babble 잡음이 더해진 5dB 신호 대 잡음비 환경 하에서 잡음음성, 음성검출 기준값, 그리고 제안된 알고리즘의 로그우도비와 음성검출 결과값.
Fig. 1. Noisy speech, VAD reference, log-likelihood ratio (LLR) and VAD results of the proposed method for the babble noise at 5dB SNR.

AMR VAD option 2[12]와 같은 표준 음성검출기를 비교를 위해 포함하였다.

그림 1은 제안된 PNMF 기반 음성검출기의 실제 수행 과정을 나타낸 것이다. 그림의 위 부분은 5dB 신호 대 잡음비 환경 하에서 babble 잡음이 더해진 잡음 음성과 수동으로 표시된 음성검출 기준값을 나타낸다. 그림의 아래 부분은 제안된 알고리즘을 이용한 경우, (22)식에 해당되는 로그 우도비와 이를 임계값과 비교하여 얻어진 음성검출 결과를 나타낸다.

그림 2-4는 babble, factory 및 pink 잡음에서 AMR, DFT 영역에서 Gaussian 기반, NMF 기반, 그리고 제안된 PNMF 기반 음성검출 기법에 대한 ROC(receiver operating characteristic) 곡선을 나타내다. 그림으로부터 제안된 PNMF 기반 음성검출기는 표준 AMR이나 기존의 Gaussian 분포, 그리고 기존 NMF을 사용한 기법보다 뛰어난 성능을 보여준다. 특히 신호 대 잡음비가 더 낮을수록 PNMF을 사용한 음성검출 기법이 신호 대 잡음비가 더 높은 환경에 비해 성능 향상이 더 크게 나타남을 확인할 수 있다. PNMF 기법이 다른 기법에 비해 성능이 향상된 이유로는 학습을 통해 잡음과 음성에 대한 기저벡터를 생성하고 이를 기반으로 하여 각 프레임별 음성신호와 잡음의 통계적인 값들을 잘 추정하여 음성검출을 위한 우도비 검증을 이용하기 때문이다.

표 1은 앞에서 언급된 Gaussian 기법과 제안된 알고리즘에 대한 계산량을 나타낸다. 표에 나타난 계산시간은 456초의 전체 음성에 대해 각각의 알고리즘을

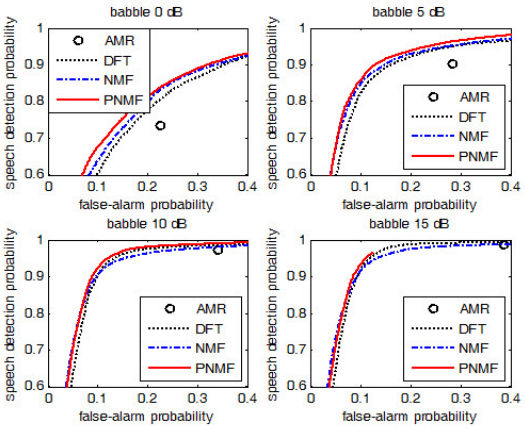


그림 2. Babble 잡음이 더해진 0, 5, 10 및 15dB 에서 AMR, Gaussian 기반, NMF 기반 그리고 제안된 PNMF 기반의 음성검출기를 사용하는 경우의 ROC 곡선.
Fig. 2. ROC curves of AMR, Gaussian, NMF and PNMF methods for the babble noise at 0, 5, 10, and 15dB SNR.

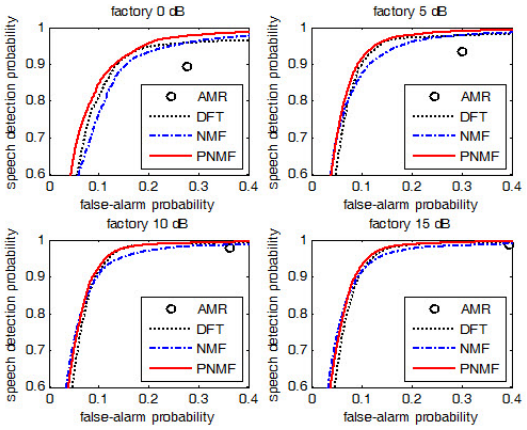


그림 3. Factory 잡음이 더해진 0, 5, 10 및 15dB에서 AMR, Gaussian 기반, NMF 기반 그리고 제안된 PNMF 기반의 음성검출기를 사용하는 경우의 ROC 곡선.
Fig. 3. ROC curves of AMR, Gaussian, NMF and PNMF methods for the factory noise at 0, 5, 10, and 15dB SNR, respectively.

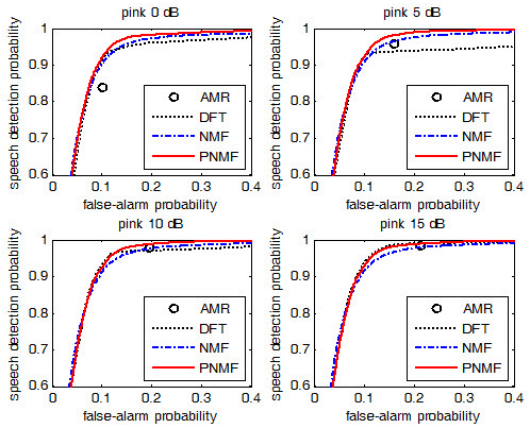


그림 4. Pink 잡음이 더해진 0, 5, 10 및 15dB에서 AMR, Gaussian 기반, NMF 기반 그리고 제안된 PNMF 기반의 음성검출기를 사용하는 경우의 ROC 곡선.
Fig. 4. ROC curves of AMR, Gaussian, NMF and PNMF methods for the pink noise at 0, 5, 10, and 15dB SNR, respectively.

표 1. 456초의 음성에 대해 Gaussian 기법과 제안된 PNMF 기법을 사용한 경우 음성검출 계산시간
Table 1. Computation time of the Gaussian and PNMF methods for the 456 sec. of speech.

methods	Gaussian	PNMF
computation time(sec)	35.052700	34.674740

사용하여 음성검출을 수행하는 경우 소요되는 시간을 나타낸다. 계산을 위해 MATLAB R2014와 인텔 Xeon 3GHz CPU을 사용하였다. 측정결과 보면 두 알

고리즘의 계산시간은 비슷함을 알 수 있다.

Ⅳ. 결 론

본 논문에서는 잡음과 음성의 잘 판별하기 위해 PNMF을 신호 모델링에 사용하였고 이를 음성검출에 적용하였다. PNMF 모델에 의해 잡음과 음성의 DFT 계수 크기에 대한 확률적 분포로 Poisson 분포를 사용하였고, 이를 기반으로 하여 음성검출을 위한 새로운 우도비 검증 식을 제안하였다. 다양한 잡음환경에서의 실험 결과 제안한 PNMF 기반 음성검출 알고리즘이 기존의 알고리즘 보다 우수한 성능을 나타내었다.

제안된 기법은 NMF 기반의 음성향상 기법과 결합하여 사용할 수 있으며, 기계학습 기반의 음성검출을 위한 특징벡터를 추출하기 위한 전처리 단계로 사용할 수 있다.

References

[1] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1-3, Jan. 1999.

[2] J. -H. Chang, N. S. Kim, and S. K. Mitra, “Voice activity detection based on multiple statistical models,” *IEEE Trans. Sign. Process.*, vol. 54, no. 6, pp. 1965-1976, Jun. 2006.

[3] Q. -H. Jo, J. -H. Chang, J. Shin, and N. S. Kim, “Statistical model-based voice activity detection using support vector machine,” *IET Sign. Process.*, vol. 3, no. 3, pp. 205-210, May 2009.

[4] L. Zhang and J. Wu, “Deep belief networks based voice activity detection,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 3371-3408, Apr. 2013.

[5] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, pp. 788-791, Oct. 1999.

[6] S. -I. Kang and J. -H. Chang, “Voice activity detection based on non-negative matrix factorization,” *J. KICS*, vol. 35, no. 8, pp. 661-666, 2010.

[7] F. G. Germain, D. L. Sun, and G. J. Mysore, “Speaker and noise independent voice activity detection,” *Interspeech*, pp. 732-736, Aug. 2013.

[8] A. T. Cemgil, “Bayesian inference for nonnegative matrix factorisation models,” *Computational Intelligence and Neuroscience*, vol. 2009, no. 785152, p. 17, 2009.

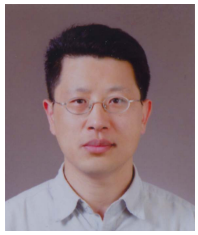
[9] T. Virtanen, A. T. Cemgil, and S. J. Godsill. “Bayesian extensions to non-negative matrix factorisation for audio signal modelling,” in *Proc. IEEE Int. Conf. Acoust. Speech and Sign. Process. 2008,* pp. 1825-1828, Las Vegas, Apr. 2008.

[10] N. Mohammadiha, T. Gerkmann, and A. Leijon, “A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization,” *IEEE WASPAA*, pp. 45-48, 2011.

[11] K. Kwon, Y. G. Jin, S. H. Bae, and N. S. Kim, “A NMF-based speech enhancement method using a prior time varying information and gain function,” *J. KICS*, vol. 38C, no. 6, pp. 503-511, 2013.

[12] ETSI EN 301708-1999: Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels, v7.1.1 (European Telecommunications Standards Institute, France, 1999).

김 동 국 (Dong Kook Kim)



1989년 2월 : 전남대학교 전자공학과 학사

1991년 2월 : 포항공과대학 전자전기공학과 석사

2003년 2월 : 서울대학교 전기컴퓨터공학부 박사

2004년 2월~현재 : 전남대학교 전자컴퓨터공학부 교수

<관심분야> 음성처리, 음성인식, 기계학습

신 종 원 (Jong Won Shin)

2002년 2월 : 서울대학교 전기공학부 학사
2008년 8월 : 서울대학교 전기컴퓨터공학부 박사
2008년 12월~2012년 8월 : Qualcomm Inc., Senior Engineer
2012년 9월~현재 : 광주과학기술원 전기전자컴퓨터공학부 조교수
<관심분야> 음성/음향/오디오처리

권 기 수 (Kisoo Kwon)

2011년 2월 : 서울대학교 전기공학부 졸업
2011년 3월~현재 : 서울대학교 전기·정보공학부 석박통합과정
<관심분야> 음성 신호처리, 음원 분리, 음질 향상

김 남 수 (Nam Soo Kim)

1988년 2월 : 서울대학교 전자공학과 졸업
1990년 2월 : 한국과학기술원 전기공학과 석사
1994년 8월 : 한국과학기술원 전기공학과 박사
1998년 3월~현재 : 서울대학교 교수
<관심분야> 음성 신호처리, 음성 인식, 통계적 신호처리, 패턴 인식, 휴먼 인터페이스