한국음향학회지 제27권 제7호 pp. 372~378 (2008)

신호 준공간 모델에 기반한 통계적 음성 검출기

Statistical Voice Activity Detector Based on Signal Subspace Model

류 광 춘*, 김 동 국*
(Kwang-chun Ryu*, Dong-kook Kim*)

*전남대학교 전자컴퓨터공학과
(접수일자: 2008년 7월 30일; 수정일자: 2008년 9월 1일; 채택일자: 2008년 9월 29일)

음성 검출기 (VAD, Voice Activity Detector)는 이동 통신이나 음성신호처리 등에 매우 중요한 기법으로 사용된다. 일반적인 음성 검출방식은 이산 푸리에 변환 (DFT, Discrete Fourier Transform)영역에서 통계적인 모델을 기반으로 하여 우도비 검정 (LRT, Likelihood Ratio Test)을 하게 된다. 그리고 이 값을 임계값과 비교하여 음성인지 아닌지 판단하게 된다. 본 논문에서는 신호 준공간 (Signal Subspace)에 기반한 새로운 통계적 음성 검출 기법을 제안하다. 확률적인 주성분 분석 (PPCA, Probabilistic Principal Component Analysis)은 신호 준공간 방법에서 잡음신호에 대한 확률적인 모델을 얻기 위해 사용된다. 제안된 기법은 신호 준공간 영역에서 우도비검정에 기반을 두는 결정규칙을 적용하였다. 음성 검출 실험 결과는 신호 준공간 모델에 근거한 음성 검출기 기법이 주파수 영역에 기반한 가우시안 (Gaussian) 음성 검출기 보다 향상된 검출 결과를 보여준다.

핵심용어: 음성 검출기, 가우시안 분포, 신호 준공간, Likelihood Ratio Test
투고분야: 음성처리 분야 (2,3)

Voice activity detectors (VAD) are important in wireless communication and speech signal processing. In the conventional VAD methods, an expression for the likelihood ratio test (LRT) based on statistical models is derived in discrete Fourier transform (DFT) domain. Then, speech or noise is decided by comparing the value of the expression with a threshold. This paper presents a new statistical VAD method based on a signal subspace approach. The probabilistic principal component analysis (PPCA) is employed to obtain a signal subspace model that incorporates probabilistic model of noisy signal to the signal subspace method. The proposed approach provides a novel decision rule based on LRT in the signal subspace domain. Experimental results show that the proposed signal subspace model based VAD mathod outperforms those based on the widely used Gaussian distribution in DFT domain.

Keywords: Voice Activity Detection, Gaussian Distribution, Likelihood Ratio Test
ASK subject classification: Speech Signal Processing (2,3)

I. 서 론

음성 검출기(VAD, Voice Activity Detector)는 현재 이동 통신 및 음성신호처리 기술 등에 있어서 매우 중요한 기술로써 사용되고 있다. 음성 검출기는 잡음이 섞여 있는 음성신호에서 음성이 존재하는 부분과 잡음만 존재하는 부분을 판별하는 기술로, 현재까지 활발히 연구가 이루어지고 있다 [1]-[6][15][16].

최근에 연구되고 있는 여러 가지 통계적 기반 음성 검출 기술들의 특징을 살펴보면 다음과 같다. 우선, 입력된 신호들의 전력스펙트럼 (Power Spectrum) 분석을 한 후, 이를 가우시안 (Gaussian) 분포나 라플라시안 (Laplacian) 분포, 혹은 감마 (Gamma) 분포 형태의 확률밀도함수 (PDF, Probability Density Function)를 갖는다고 가정을 한다 [9]. 이러한 분포는 이산 푸리에 변환 (DFT, discrete Fourier transform) 영역에서의 통계모델로 이용하게 된다. 음성의 이산 푸리에 변환 영역에서 통계적 분포를 분석하여 잡음 혹은, 잡음에 오염된 음성에 대한 우도비 검정 (LRT, Likelihood Ratio Test)식을 세우며, 이를 통하여 최종 결정 규칙을 도출하여 음성을 검출하게 된다 [1][2][4][6][13].

신호 준공간 (Signal Subspace)기법은 음성향상에 성공적으로 적용된 기법이다 [7]-[9]. 이것의 기본적인 개념은 잡음신호를 두 개의 준공간으로 분해하여, 잡음만

책임저자: 김 동 국 (dkim@chonnam.ac.kr)
광주 북구 용봉동 300번지 전남대학교 전자컴퓨터 공학과
(전화: 062-530-0263; 팩스: 062-530-1759)

존재하는 준공간과 음성신호가 함께 존재하는 준공간으로 분해하게 되다. 음성향상은 잡음공간을 버리고 신호 준공간 에서의 깨끗한 신호를 추정하게 된다. 신호 분해 방법에는 고유값 분해 (EVD, Eigenvalue Decomposition) 또는 특이값 분해 (SVD, Singular Value Decomposition) 방법을 사용할 수 있다 [7]. 신호 준공간에 기반한 방법은 잡음 환경에 있어 음성 향상과 음성 인식에 매우 유용한 방법으로 알려져 있다 [10].

본 논문에서는 통계적인 음성 검출 방식에 있어 신호 준공간 방법을 적용하였다. 확률적인 주성분분석 (PPCA, Probabilistic Principal Component Analysis)을 사용하여 신호 준공간 영역에서 새로운 통계적 음성 검출 알고리즘을 제안한다 [11]. 확률적인 주성분분석은 신호 준공간 방법에 대한 확률적인 모델을 포함하는 모델을 얻기 위해 사용된다. 이러한 방법은 신호 준공간 영역에서 우도비 검정에 기반하는 새로운 결정 규칙을 제공한다. 실험 결과 제안된 신호 준공간 기법에 기반한 알고리즘이 이산 푸리에 변환에 기반한 가우시안 음성 검출기 보다 음성 검출 능력이 향상됨을 보여주었다.

본 논문의 II장에서는 확률적 주성분분석에 기반하는 신호 준공간 모델 (SSM, Signal Subspace Model)에 대해 소개하고, III장에서는 준공간 모델에 기반한 음성 검출기 방식에 대해 논하였다. IV장에서는 새로운 음성 검출의 실험 및 결과를 보여주며, V장에서 결론을 맺어 본 논문을 마친다.

II. 확률적인 주성분 분석에 기반하는 신호 준공간 모델

이 장에서는 음성 검출을 위한 확률적 주성분분석에 기반한 신호 준공간 모델을 소개한다 [8][11]. N개의 잡음에 오염된 음성벡터는 $\{\mathbf{y}_1,\mathbf{y}_2,\cdots,\mathbf{y}_N\}$이고, 여기서 $\mathbf{y}_t = [y_{t,1},\cdots,y_{t,D}]^T$는 잡음에 오염된 음성신호의 $D\times 1$ 벡터이다. 잡음에 오염된 음성의 벡터는 파라미터 $\phi = \{\mathbf{W},\sigma^2\}$를 갖는 확률적인 주성분분석 모델을 이용하여 다음과 같은 신호로 가정할 수 있다.

$$\mathbf{y} = \mathbf{W}\mathbf{s} + \mathbf{n} \tag{1}$$

여기서 $\mathbf{W} = [\mathbf{w}_1,\cdots,\mathbf{w}_P]$은 잡음에 오염된 음성신호의 준공간을 나타내는 P개의 열벡터로 구성된 행렬이며, s는 $s = [s_1,\cdots,s_P]^T$이며, P<D인 차수를 갖는다. 그리고 n은 s와 독립적인 임의의 가우시안 분포 $p(\mathbf{s}) \sim N(0,\mathbf{I}_P)$을 가진다고 가정한다. 그리고 잡음 n은 백색잡음 가우시안 모델$p(\mathbf{n}) \sim N(0,\sigma_n^2\mathbf{I}_D)$로 가정한다. 여기서 $\mathbf{I}_P$와 $\mathbf{I}_D$는 각각 $P\times P$, $D\times D$의 항등행렬이다.

깨끗한 음성 벡터 $\mathbf{x} = [x_1,x_2,\cdots,x_D]^T$는 다음과 같이 표현 되며,

$$\mathbf{x} = \sum_{k=1}^{P}\mathbf{w}_k s_k = \mathbf{W}\mathbf{s} \tag{2}$$

식 (1)은 백색잡음에 오염된 음성신호로 다음과 같이 표현된다.

$$\mathbf{y} = \mathbf{W}\mathbf{s} + \mathbf{n} = \mathbf{x} + \mathbf{n} \tag{3}$$

위의 가정에 기반하여, y의 확률밀도 함수는 다음과 같다.

$$p(\mathbf{y}|\phi) = \frac{1}{(2\pi)^{D/2}|\mathbf{R}_y|^{1/2}}\exp\left\{-\frac{1}{2}\mathbf{y}^T\mathbf{R}_y\mathbf{y}\right\} \tag{4}$$

잡음에 오염된 음성신호의 공분산 행렬은 다음과 같이 나타낸다.

$$\mathbf{R}_y \equiv E[\mathbf{y}\mathbf{y}^T] = \mathbf{W}\mathbf{W}^T + \sigma_n^2\mathbf{I}_D \tag{5}$$

먼저 관찰 벡터 $\mathbf{Y} = \{\mathbf{y}_1,\cdots,\mathbf{y}_N\}$가 주어지는 경우 확률적 주성분 분석을 이용하여 $\mathbf{S} = \{\mathbf{s}_1,\cdots,\mathbf{s}_N\}$ 값과 모델 파라메타 값을 추정하는 것이 요구된다. 이는 최대 우도비 (ML, Maximum Likelihood)를 이용하여 최적화된 값을 찾을 수 있다. 그러나 $\{\mathbf{s}_i\}$는 은닉 (hidden) 되어있기 때문에 파라미터 값을 반복적으로 업데이트하는 Expectation Maximization (EM) 알고리즘을 사용하게 된다 [12][17]. 이때 우도비의 전역 최대값에 대해 다음과 같은 관계식이 주어진다 [11].

$$\hat{\mathbf{W}} = \mathbf{U}_P(\mathbf{\Lambda}_P - \sigma_n^2\mathbf{I}_P)^{1/2}\mathbf{R} \tag{6}$$

여기서 P 열벡터 $D\times P$ 행렬 $\mathbf{U}_P = [\mathbf{u}_1,\cdots,\mathbf{u}_P]$의 주요 고유벡터들 $(\mathbf{U}_P^T\mathbf{U}_P = \mathbf{I}_P)$의 공분산 행렬은 다음과 같고,

$$\mathbf{C}=\frac{1}{N}\sum_{t=1}^{N}\mathbf{y}_t\mathbf{y}_t^T \tag{7}$$

$P\times P$ 대각행렬 $\mathbf{\Lambda}_{y,P}=diag\{\lambda_{y,1},\cdots,\lambda_{y,p}\}$에 대응하는 $\lambda_{y,1}>\cdots>\lambda_{y,p}$의 고유값들을 포함하고, R은 임의의 $P\times P$ 직교회전 행렬로 $\mathbf{R}^T\mathbf{R}=\mathbf{I}_P$이다. 그리고 $\hat{\mathbf{W}}$에 의해 다음 관계식이 주어진다.

$$\hat{\sigma_n^2}=\frac{1}{D-P}\sum_{k=P+1}^{D}\lambda_{y,k} \tag{8}$$

여기서 $\lambda_{y,P+1}>\cdots>\lambda_{y,D}$는 주요 고유벡터들의 공분산행렬 C의 아주 작은 고유값이다. 이것은 생략되어진 고유값의 평균값을 나타낸다. 즉 EM과정을 통해 $\hat{\mathbf{W}}$값을 구하면 식(6)을 통해 EVD에 의한 고유값과 고유벡터를 추정할 수 있게 된다. 그리고 [17]에서 이론적으로 $\sigma_n^2\rightarrow 0$인 경우 EVD와 EM과정이 동일함을 증명하였다.

식(3)은 준공간 기법에 대한 확률적인 모델을 포함하는 신호 준공간 모델을 정의한다. 신호 준공간의 모델은 다양한 확률분포를 가지며, 깨끗한 신호와 잡음 신호 그리고 잡음에 오염된 신호의 벡터가 포함 되어있고, 이는 신호 준공간에 기반한 음성 검출기와 음성 향상에 매우 유용하다.

잡음에 오염된 음성신호의 공분산행렬은 다음과 같이 표현되다.

$$\mathbf{R}_y=\mathbf{R}_x+\mathbf{R}_n=\mathbf{W}\mathbf{W}^T+\sigma_n^2\mathbf{I}_D \tag{9}$$

$\mathbf{R}_x$와 $\mathbf{R}_n$는 각각 깨끗한 음성신호 x와 잡음 신호 n의 공분산 행렬이다.

$$\mathbf{R}_x=E\{\mathbf{x}\mathbf{x}^T\}=\mathbf{W}\mathbf{W}^T \tag{10}$$

$$\mathbf{R}_n=E\{\mathbf{n}\mathbf{n}^T\}=\sigma_n^2\mathbf{I}_D\ . \tag{11}$$

W를 대신하여 공식 (10)에 (6)을 대입하여 다음과 같이 표현할 수 있다.

$$\mathbf{R}_x=\mathbf{U}_P(\mathbf{\Lambda}_{y,P}-\sigma_n^2\mathbf{I}_P)\mathbf{U}_P^T \tag{12}$$

그러면 식 (9)의 공분산 행렬 $\mathbf{R}_y$를 다음과 같이 다시 표현할 수 있다.

$$\mathbf{R}_y=[\mathbf{U}_P\mathbf{U}_{D-P}]\begin{bmatrix}\mathbf{\Lambda}_P & 0\\ 0 & \sigma_n^2\mathbf{I}_{D-P}\end{bmatrix}[\mathbf{U}_P\mathbf{U}_{D-P}]^T \tag{13}$$

$\mathbf{U}_P$는 주요 고유벡터와 일치하는 잡음에 오염된 신호의 고유값 $\lambda_{y,k},k=1,\cdots,P$의 $D\times P$행렬이고, $\mathbf{U}_{D-P}$는 작은 고유벡터와 일치하는 잡음 신호의 고유값σ_n^2의 $D\times(D-P)$ 행렬이다. 따라서 $\mathbf{U}_P$는 잡음에 오염된 음성신호의 준공간을 말하고, $\mathbf{U}_{D-P}$는 잡음신호의 준공간을 나타낸다. 깨끗한 음성은 P-차원의 준공간 상에 있다고 가정하면, $\mathbf{R}_x$은 P개의 영이 아닌 고유값을 갖는다.

$$\lambda_{x,k}=\begin{cases}\lambda_{y,k}-\sigma_n^2>0 & \text{for } k=1,\cdots,P\\ 0 & \text{for } k=P+1,\cdots,D\end{cases} \tag{14}$$

III. 준공간 모델에 기반한 음성 검출기

3.1. 준공간 모델에 기반한 유사도 비율 평가 (LRT)

음성신호를 분석하기 위해 잡음 음성신호는 식 (3)과 같은 준공간 모델에 의해 발생한다고 가정한다. 이로부터 잡음만 존재하는 준공간과 잡음에 오염된 음성신호가 함께 존재하는 두 가지 가설을 설정할 수 있다.

$$H_0:\text{ speech absent},\quad :\quad \mathbf{y}=\mathbf{n} \tag{15}$$

$$H_1:\text{ speech present}\quad :\quad \mathbf{y}=\mathbf{W}\mathbf{s}+\mathbf{n} \tag{16}$$

여기서 y, n 그리고 $\mathbf{x}(=\mathbf{W}_S)$는 각각 잡음에 오염된 음성신호, 잡음신호, 원래의 음성신호의 벡터를 나타낸다. 두 가설의 확률 분포를 나타내면,

$$p(\mathbf{y}|H_0)=\frac{1}{(2\pi)^{D/2}|\sigma_n^2\mathbf{I}_D|^{1/2}}\exp\left\{-\frac{1}{2}\mathbf{y}^T(\sigma_n^2\mathbf{I}_D)^{-1}\mathbf{y}\right\} \tag{17}$$

$$p(\mathbf{y}|H_1)=\frac{1}{(2\pi)^{D/2}|\mathbf{R}_y|^{1/2}}\exp\left\{-\frac{1}{2}\mathbf{y}^T\mathbf{R}_y^{-1}\mathbf{y}\right\} \tag{18}$$

$\mathbf{R}_y$은 잡음에 오염된 음성신호 (9)의 공분산 행렬이다. (17)과 (18)에 기반하여 우도비식을 얻을 수 있는데, 그 식은 다음과 같다.

$$\Psi\quad \frac{p(\mathbf{y}|H_1)}{p(\mathbf{y}|H_0)}=\frac{|\mathbf{R}_y|^{-1/2}\exp\left\{-\frac{1}{2}\mathbf{y}^T\mathbf{R}_y^{-1}\mathbf{y}\right\}}{|\sigma_n^2\mathbf{I}_D|^{-1/2}\exp\left\{-\frac{1}{2}\mathbf{y}^T(\sigma_n^2\mathbf{I}_D)^{-1}\mathbf{y}\right\}} \tag{19}$$

위의 식 (19)를 정리해 보면

$$\Psi = \left(\prod_{k=1}^{P}\lambda_{y,k}\right)^{-1/2}(\sigma_n^2)^{P/2}\exp\left\{\frac{1}{2}\mathbf{y}^T(\mathbf{R}_y^{-1}-\sigma_n^{-2}\mathbf{I}_D)\mathbf{y}\right\} \quad (20)$$

이고, 식 (19)의 $|\mathbf{R}_y|^{-1/2}$을 정리하면 다음과 같다.

$$|\mathbf{R}_y|^{-1/2} = \left(\prod_{k=1}^{P}\lambda_{y,k}\right)^{-1/2}(\sigma_n^2)^{-(D-P)/2} \quad (21)$$

여기서 역행렬 $\mathbf{R}_y^{-1}$ 계산하고, 식 (21)의 역행렬 부분을 다음과 같이 정리 할 수 있다.

$$\mathbf{R}_y^{-1}-\sigma_n^{-2}\mathbf{I}_D = \left(\sigma_n^2\mathbf{W}\mathbf{W}^T+\mathbf{I}\right)^{-1}-\sigma_n^{-2}\mathbf{I}_D \quad (22)$$

$$= -\frac{1}{\sigma_n^2}\mathbf{W}\left(\mathbf{W}^T\mathbf{W}+\sigma_n^2\right)^{-1}\mathbf{W}^T \quad (23)$$

W를 대신하여 식 (6)을 식 (23)에 대입하면,

$$\frac{1}{\sigma_n^2}\mathbf{W}\left(\mathbf{W}^T\mathbf{W}+\sigma_n^2\right)^{-1}\mathbf{W}^T = \mathbf{U}_P\left(\mathbf{\Lambda}_{y,P}-\sigma_n^2\right)\mathbf{\Lambda}_{y,P}^{-1}\frac{1}{\sigma_n^2}\mathbf{U}^T \quad (24)$$

$$= \mathbf{U}_P\mathbf{G}_P\mathbf{U}_P^T \quad (25)$$

이고, 여기서 $\mathbf{G}_P$는 다음과 같은 $(P\times P)$ 대각행렬이다.

$$\mathbf{G}_P = \left(\mathbf{\Lambda}_{y,P}-\sigma_n^2\right)\mathbf{\Lambda}_{y,P}^{-1}\frac{1}{\sigma_n^2} \quad (26)$$

$$= diag\left\{\frac{\lambda_{x,1}}{\left(\lambda_{x,1}+\sigma_n^2\right)\sigma_n^2},\cdots,\frac{\lambda_{x,P}}{\left(\lambda_{x,P}+\sigma_n^2\right)\sigma_n^2}\right\} \quad (27)$$

(6)과 (9)식 그리고 위의 식을 토대로 최종 우도비식은 다음과 같이 얻어진다.

$$\Psi = \prod_{k=1}^{P}\frac{1}{\sqrt{\lambda_{x,k}/\sigma_n^2+1}}\exp\left\{\frac{\lambda_{x,k}}{2(\lambda_{x,k}+\sigma_n^2)\sigma_n^2}\hat{y}_k^2\right\} = \prod_{k=1}^{P}\frac{1}{\sqrt{\xi_k+1}}\exp\left\{\frac{\xi_k}{2(\xi_k+1)}\gamma_k\right\} \quad (28)$$

$\xi_k=\lambda_{x,k}/\sigma_n^2$과 $\gamma_k=\hat{y}_k^2/\sigma_n^2$는 준공간 영역에서의 a priori SNR 과 a posteriori SNR이며 [3], $\hat{y}_k=\mathbf{u}_k^T\mathbf{y}$는 잡음에 오염된 음성신호의 준공간 영역에서의 k번째 신호의 값을 나타내고, 일반적으로 다음식과 같이 주어진다.

$$\hat{\mathbf{y}} = \left[\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_P\right]^T = \mathbf{U}_P^T\mathbf{y}. \quad (29)$$

우도비의 가하평균을 구하고, 다음과 같은 결정 식을 얻을 수 있다.

$$\log\Psi = \frac{1}{P}\sum_{k=1}^{P}\log\Psi_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \quad (30)$$

여기서 Ψ_k는

$$\Psi_k = \frac{1}{\sqrt{\xi_k+1}}\exp\left\{\frac{\xi_k}{2(\xi_k+1)}\gamma_k\right\} \quad (31)$$

이고, 여기서 η는 음성의 존재 유무를 판별하는 임계값이다.

3.2. 기존 알고리즘과 비교

제안된 준공간 기반 음성 검출 방법은 이산 푸리에변환 영역 기반의 가우시안 방법과 비슷하다. 가우시안 기반의 음성 검출의 우도비는 다음과 같다 [3].

$$\tilde{\Psi}_k = \frac{1}{\tilde{\xi}_k+1}\exp\left\{\frac{\tilde{\xi}_k}{\tilde{\xi}_k+1}\tilde{\gamma}_k\right\},\ k=1,\cdots,D \quad (32)$$

k는 k번째 주파수 bin을 나타내고, $\tilde{\xi}_k$와 $\tilde{\gamma}_k$는 이산 푸리에변환 영역에서의 a priori SNR과 a posteriori SNR이고, $\tilde{\xi}_k$는 Decision-Directed (DD) 방법을 사용하여 계산된다 [3]. 위 식에서 제안한 준공간 기반 우도비와 가우시안 기반의 우도비 사이에 유사성이 존재함을 알 수 있다. 두 음성 검출기의 차이점은 가우시안 기반의 음성 검출기는 이산 푸리에변환 영역에서 사용되며, 준공간 기반의 검출기는 카루넨-루베 변환 (KLT, Karhunen-loève transform)을 사용한 신호 준공간에서 적용된다는 점이다. 일반적으로 이산 푸리에변환은 고정적이고 입력신호에 무관한 변환인 반면, KLT변환은 입력신호에 따라 변환행렬을 구하게 되므로 입력신호에 맞는 최적의 변환을 구할 수 있게 된다 [7]. 그러나 KLT변환은 이산 푸리에 변환 보다 많은 계산량을 요구하는 단점을 갖는다.

IV. 실험 및 결과고찰

4.1. 실험

신호 준공간은 PPCA와 공분산행렬 $\mathbf{R}_y$에 EVD을 적용함으로 얻어질 수 있다. 그러나 PPCA는 신호 준공간을 얻기 위해 반복적인 EM과정이 필요하여 계산량이 많아지게 된다. 따라서 본 연구에서 계산량을 줄이기 위해 EVD 방법을 사용하였다. EVD을 사용하기 위해 잡음에 오염된 음성과 잡음의 공분산 행렬의 정확한 추정이 필요하다. 본 논문에서는 잡음에 오염된 음성의 자시상관 계수의 추정을 통해 테플리츠 (topeliz) 공분산 행렬을 얻을 수 있다. 음성신호의 프레임을 50%로 overlap하였고, 신호 준공간과 잡음 변화를 추정하기 위해 L=160, D=20을 선택하였다. 백색 잡음을 보존하기 위하여 위상 분석을 위해 직사각창을 사용하였다.

신호 준공간 모델의 한 가지 문제는 신호 준공간 모델이 백색잡음 가정에 기반을 둔다는 것이다. 유색잡음을 다루기 위해 본 논문에서는 prewhirening행렬 $\mathbf{R}_n^{-1/2}$을 사용하였다. 그것은 유색 잡음 $\mathbf{R}_n$의 공분산 행렬의 제곱의 루트로 표현된다 [8]. prewhitening행렬은 EVD를 사용하는 잡음의 공분산 행렬의 Cholesky 분해를 사용하여 얻어진다. prewhitening행렬은 음성부재 프레임동안 생성된 잡음 벡터를 사용하여 계산하였다. 또한 신호 준공간에서는 적절한 차원 P값을 추정하는 것이 필요하게 되는데, 본 연구에서는 잡음 분산값을 초과하는 $\mathbf{R}_y$의 고유값의 수로 계산하였다 [7][8]. 즉 P는 다음과 같이 얻어진다.

$$\hat{P} = \underset{1 \le k \le D}{\operatorname{argmax}} \left\{ \lambda_{y,k} - \hat{\sigma}_n^2 > 0 \right\} \tag{33}$$

여기서 $\hat{\sigma}_n^2$는 문장시작 부분의 비음성 구간에서의 잡음의 분산 추정 값이다.

4.2. 결과

제안된 알고리즘의 실험결과를 평가하기 위해 본 논문에서는 각각의 음성 검출 알고리즘에 대해 detection과 false-alarm (FA)확률 P_D와 P_{FA}를 조사 하였다. P_D는 실제로 정확하게 음성이라고 판단할 확률을 뜻하고, P_{FA}는 비음성을 음성이라 잘못 판단한 확률을 뜻한다. P_D와 P_{FA}의 값을 계산하기 위하여 우리는 456초의 음성을 10 ms 단위로 수동 레이블링하여 기준으로 삼았다. 수동 표시된 실제 음성의 비율은 58.2%이고, 이중에 44.85는 유성음 (voiced sounds)이며 13.4%는 무성음 (unvoiced sound)이었다. 잡음에 오염된 음성신호를 만들기 위해, white, babble, factory 그리고 pink 잡음을 NOISEX-92 잡음으로부터 SNR을 변화하면서 원래의 음성신호에 첨가 하였다. 음성 검출 테스트는 10 ms의 프레임에 대하여 수행하였다 [14].

성능을 알아보기 위해서 ROC(Receiver Operating Characteristic)곡선을 그려보았다. 그림 1-4는 각각 white, babble, factory 그리고 pink 잡음의 경우의 ROC 곡선이며, 신호 준공간 기반의 음성검출기 알고리즘은 가우시안 음성 검출기 보다 뛰어난 성능을 보여준다. 그리고 SNR이 더 높을수록 탐지능력이 더 우수한 성능을 갖음을 확인할 수

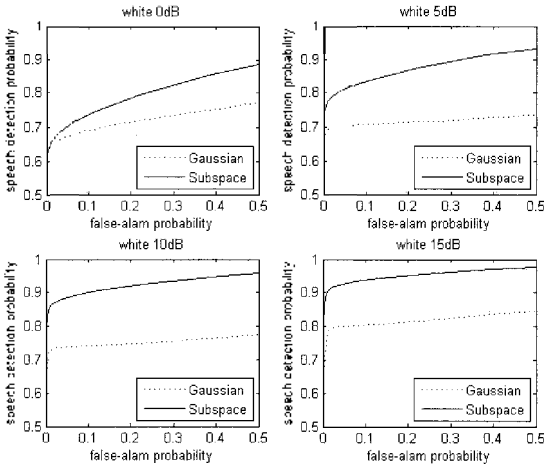

그림 1. white 잡음이 더해진신호의 ROC 곡선, SNR은 위에서부터 0 dB, 5 dB, 10 dB, 15 dB의 경우

Fig. 1. ROC curve of white noisy signal. SNR is 0 dB, 5 dB, 10 dB, 15 dB from top to bottom.

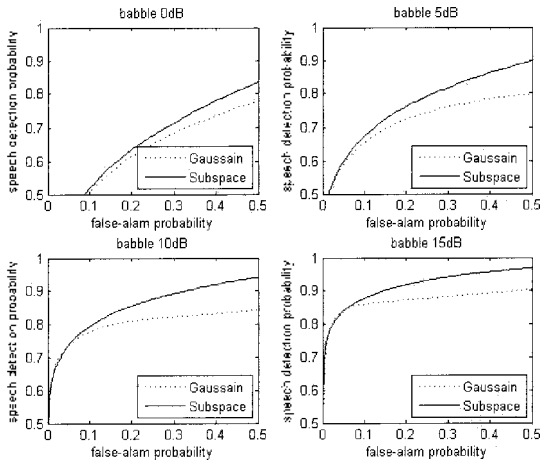

그림 2. babble 잡음이 더해진신호의 ROC 곡선, SNR은 위에서부터 0 dB, 5 dB, 10 dB, 15 dB의 경우

Fig. 2. ROC curve of babble noisy signal, SNR is 0 dB, 5 dB, 10 dB, 15 dB from top to bottom.

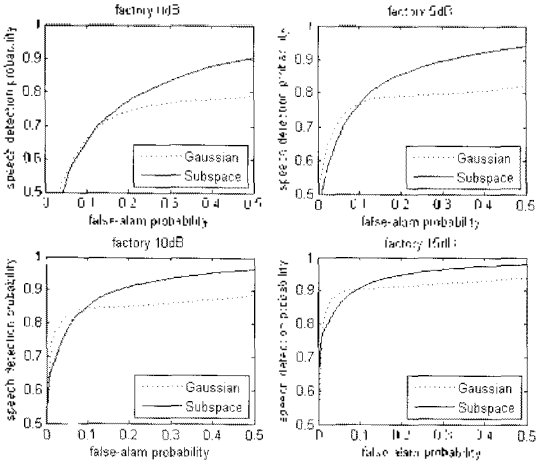

그림 3. factory 잡음이 더해진신호의 ROC 곡선. SNR은 위에서부터 0 dB, 5 dB, 10 dB, 15 dB의 경우

Fig. 3. ROC curve of factory noisy signal, SNR is 0 dB, 5 dB, 10 dB, 15 dB from top to bottom.

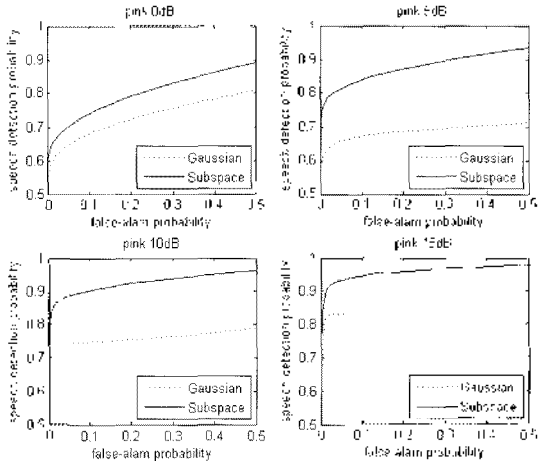

그림 4. pink 잡음이 더해진신호의 ROC 곡선. SNR은 위에서부터 0 dB, 5 dB, 10 dB, 15 dB의 경우

Fig. 4. ROC curve of pink noisy signal, SNR is 0 dB, 5 dB, 10 dB, 15 dB from top to bottom.

표 1. 제안된 음성 검출기와 가우시안 음성 검출기의 P_D와 P_{FA}의 다양한 조건 환경에서 비교

Table 1. P_D and P_{FA} of the proposed and Gaussian VAD's for various environmental conditions.

Environment		ED VAD				Gaussian VAD			
		P_D (%)		P_{FA} (%)		P_D(%)		P_{FA}(%)	
Noise	SNR (dB)	Voiced	UV	Speech	Noise	Voiced	UV	Speech	Noise
Babble	0	68.41	64.19	67.34	24.74	68.00	56.86	65.44	24.94
	5	80.67	76.54	78.43	25.57	74.32	64.78	72.16	25.94
	10	88.98	85.60	88.23	25.59	83.02	68.76	79.72	26.63
	15	93.90	92.15	93.58	26.94	89.70	76.52	86.67	26.42
White	0	86.33	35.81	74.58	11.68	82.61	14.72	66.99	9.26
	5	92.77	55.81	84.19	12.00	87.74	9.37	67.71	9.60
	10	96.16	71.90	90.66	12.41	91.04	13.24	73.14	10.85
	15	97.92	82.14	94.39	13.24	94.34	28.42	79.18	11.25
Factory	0	80.07	77.88	79.23	22.73	70.41	45.96	64.78	22.56
	5	87.37	86.73	87.10	23.14	80.69	44.05	72.26	23.31
	10	92.10	93.17	92.27	23.62	85.13	54.36	78.05	23.74
	15	95.34	96.43	95.66	24.28	90.03	67.60	84.87	23.94
Pink	0	79.36	48.18	72.17	7.34	79.51	28.99	67.88	9.72
	5	88.50	63.87	82.83	7.73	82.16	17.34	67.17	8.70
	10	93.71	76.78	89.81	8.29	86.14	36.81	74.79	10.95
	15	96.60	87.10	94.40	9.06	90.97	59.09	83.64	11.47

있었다. 기존 방식에 비해 제안된 방식이 높은 성능을 나타내는 이유는 제안된 방식이 입력신호에 따라 최적의 KLT 변환을 구하고 변환영역에서 a priori SNR 과 a postriori SNR을 보다 정확히 추정 가능하기 때문이다.

또한 본 논문에서 제안한 방법을 부분별로 가우시안 방법과 비교하였다. 그 결과 위의 표와 같이 요약 된다. 위의 표를 보면 본 논문에서 제안한 방식의 음성 검출기가 기존의 음성 검출기에 비해 비교적 우수한 성능을 보여준다.

실험결과로부터 제안된 신호 준공간 모델에 기반한 방법이 다양한 잡음 조건에서의 가우시안 기반 알고리즘 보다 더 높은 성능을 나타내는 것을 확인 할 수 있었다.

V. 결 론

본 논문에서는 잡음과 음성을 판별하기 위해 신호 준공간 모델에 기반한 통계적 음성 검출 알고리즘을 제안하였다. 신호 준공간 영역에서 우도비 검정에 기반한 새로운 결정 규칙을 유도하여 제시하였다. 그리고 신호 준공간 영역에서의 결정 규칙이 이산 푸리에변환 영역의 가우시안 방법과 유사성과 차이점을 나타내었다. 다양한 잡음환경에서의 실험 결과 제안한 신호 준공간 기반한 음성 검출 알고리즘이 이산 푸리에 변환 영역에서 사용되는 가우시안 음성 검출기 보다 우수한 성능을 나타내었다.

참 고 문 헌

1. A. Dvis, S. Nordholm and R. Togneri, "Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold," IEEE Trans. Audio, Speech, and Language Processing, 14(2), 412-424, March 2006.
2. J. S. Sohn, N. S. Kim and W. Y. Sung, "A stislical Model -Based Voice Activity Detection," IEEE Signal pocessing Lett., 6(1), 1-3, Jan. 1999.
3. N. S. Kim, and J. -H. Chang, "Spectral Enhancement Based on Global Soft Decision," IEEE Signal Process. Lett,, 7(5), 108-110, 2000.
4. J. -H. Chang, J. W. Shin and N. S. Kim "Voice Activity Detector Employing Generalized Gaussian Distribution," IEEE Electronics Lett. 40(24), 1561-1563, Nov. 2004.
5. J. -H. Chang, N. S. Kim and S. K. Mitra, "Voice Activity Detection Based on Multiple Statistical Models," IEEE Trans. Signal Proc., 54(6), 1965-1976, June 2006.
6. S. Gazor and W. Zhang, "A Soft Voice Activity Detector Based on a Laplacian-Gaussian Model," IEEE Trans. Speech and Audio Proc., 11(5), 498-505, Sept. 2003.
7. P. Loizou, Speech Enhancement : Theory and Practice, CRC Press. 2007.
8. Y. Ephraim and H. L. Van Tress, "A Signal Subspace Approach for Speech Enhancement," IEEE Trans. Speech and Audio Proc., 3(4), 251-266, July 1995.
9. F. Jabloun and B. Champagne, "Incorporating the Human Hearing Properties in the Signal Subspace Approach for Speech Enhancement," IEEE Trans. Speech and Audio Proc., 11(6), 700-708. Nov. 2003.
10. K. Hermus, P. Wambacq, and H. V. Hamme, "A Review of Signal Subspace Speech Enhancement and Its Application to Noise Robust Speech Recognition," EURASIP Journal on Advances in Signal Procesing, 2007. Article D 45821, 15 pages. 2007.
11. M. Tipping and C. Bishop, "Mixtures of probabilistic principal component analyzers," Neural Computation, 11, 435-474, 1999.
12. A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, 39, 1-38, 1977.
13. Y. Ephraim and D. Malah, "Speech Enhancement Using A Minimum Mean-square Error Short-time Spectral Amplitude Estimator," IEEE Trans. Acoust., Speech, Signal Proc., ASSP -32, 1109-1121, Dec. 1984.
14. 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, 12(3), 247-251, Jul.1993.
15. 강상익, 조규행, 박승섭, 장준혁, "통계적 모델 기반의 음성 검출기를 위한 변별적 가중치 학습," 한국음향학회지, 26(5), 194-198, 2007년 7월.
16. 장근원, 장준혁, 김동국, "UMP 테스트에 근거한 새로운 통계적 음성검출기," 한국음향학회지, 26(1), 16-24, 2007년 1월.
17. S. Roweis, "EM Algorithms for PCA and SPCA," Neural Inform. Process. System, 10, 626-632, 1997.

저자 약력

•류 광 춘 (Kwang-Chun Ryu)

2007년 2월 : 광주대학교 컴퓨터 전자공학부 학사
2007년 3월~현재 : 전남대학교 전자컴퓨터공학과 석사과정

•김 동 국 (Dong-Kook Kim)

1989년 2월 : 전남대학교 전자공학과 학사
1991년 2월 : 포항공과대학 전자 전기 공학과 석사
2003년 2월 : 서울대학교 전기컴퓨터공학박사
1991년 2월~1993년 3월 : 삼성 전자 정보통신 연구원
1993년 3월~1999년 2월 : 삼성종합기술원 전문 연구원
2000년 2월~2002년 12월 : (주)넷더스 기술이사
2003년 4월~2004년 2월 : 한국전자통신연구원 선임 연구원
2004년 2월~현재 : 전남대학교 전자컴퓨터 정보통신 공학부 조교수