유색 잡음에 오염된 음성의 향상을 위한 백색 변환을 이용한 일반화 부공간 접근

이정욱* · 손경식** · 박장식***· 김현태****

A Generalized Subspace Approach for Enhancing Speech Corrupted by Colored Noise Using Whitening Transformation

Jeong-wook Lee* · Kyung-sik Son** · Jang-sik Park*** · Hyun-tae Kim****

이 논문은 부산대학교 자유과제 학술연구비(2년)에 의하여 연구되었음.

요 약

본 논문에서는 유색잡음에 의해 오염된 음성신호의 음성향상 알고리즘을 제안한다. 유색잡음과 음성신호가 서로 상관이 없을 경우 유색잡음은 백색화 변환을 통해 무색잡음으로 변환된다. 이 변환된 신호를 음성신호 향상을 위한 일반화된 부공간 접근법에 적용한다. 전처리 과정에서의 백색화 변환으로 발생되는 음성 스펙트럼 왜곡은 제안한 알고리즘의 후처리를 통해 역 백색화하여 복구한다. 제안한 알고리즘의 성능을 컴퓨터 시뮬레이션으로 확인하였다. 사용한 유색잡음은 자동차 잡음과 멀티 토커 배블 잡음이다. AURORA 및 TIMIT 데이터 베이스에서 취득한 데이터로 실험했을 때 제안하는 방법이 신호대잡음비 및 스펙트럼 왜곡 측면에서 기존 접근법보다 개선됨을 확인하였다.

ABSTRACT

In this paper, we proposed an algorithm for speech enhancement of speeches corrupted by colored noise. When there is no correlation between colored noise and speech signal, the colored noise turns into white noise through whitening transformation. This transformed signal has been applied to the generalized subspace approach for speech enhancement. The speech spectral distortion, produced by the whitening transformation as pre-processing, has been restored by using the inverse whitening transformation as post-processing of the proposed algorithm. The performance of the proposed algorithm for speech enhancement has been confirmed by computer simulation. The colored noises used in this experiment were car noise and multi-talker babble. It is confirmed that the proposed algorithm shows better performance from SNR and SSD viewpoint over the previous approach with the data from the AURORA and TIMIT data base.

키워드

음성 향상, 일반화 부공간 접근, 백색 변환, 후처리

Key word

Speech Enhancement, Generalized Subspace Approach, Whitening Transformation, Post-processing

* 준회원 : 부산대학교 전자공학과

** 정회원 : 부산대학교 전자공학과

*** 정회원 : 경성대학교 전자공학과

**** 정회원 : 동의대학교 (교신저자, htaekim@deu.ac.kr)

접수일자 : 2011. 03. 10

심사완료일자 : 2011. 07. 13

Ⅰ. 서 론

스펙트럼 차감(spectral subtraction)법은 낮은 복잡성과 높은 효율성으로 인해 오늘날 널리 사용되는 음성 향상(speech enhancement : SE) 알고리즘이다. 스펙트럼 차감법의 주요 결점 중 하나는 음성왜곡의 발생이며, 특히 뮤지컬 잡음(musical noise)의 발생이다. 음성왜곡을 줄이기 위하여 여러 방법들이 제안되었다[1,2]. 그러나 이 방법들은 또 다른 문제인 잔여잡음을 발생시킨다. 이 문제를 해결하기 위해 잔여잡음과 음성왜곡간의 절충점을 찾는 방법들이 제안되었다. 구체적으로는 미리 정해진 임계값 이하로 잔여잡음을 유지하면서 음성왜곡을 최소화하는 부공간(subspace)에 근거한 여러 접근법들이 제안되었다[3,4,5,6]. 부공간 접근법은 잡음이 섞인 신호를 신호 부공간 및 잡음 부공간으로 투영한다. 잡음 부공간은 잡음 프로세서로부터 오는 신호만을 포함하고 있다. 따라서 깨끗한 신호의 추정치는 잡음 부공간의 성분을 제거하고 신호 부공간에 있는 성분만을 취하여 얻을 수 있다. 두 부공간 즉, 잡음 부공간과 신호 부공간으로의 분해는 특이값 분해 (singular value decomposition, SVD)[7,8] 혹은 고유값 분해 (eigen value decomposition, EVD)[3,4,5,6]에 의해 주로 수행된다.

SVD에 근거한 방법은 Dendrinos 등에 의해 제안되었다[7]. 가장 큰 특이값에 상응하는 고유벡터들은 신호 정보를 포함하고 가장 적은 특이값에 상응하는 고유벡터들은 잡음 정보를 포함하고 있다는 개념에 근거한다. 추정되는 신호는 가장 큰 특이값을 포함하는 고유벡터를 사용하여 다시 재구성된다. 이런 방법으로 큰 신호대잡음비(SNR) 이득을 백색잡음에 오염된 음성으로부터 얻을 수 있다. Dendrinos 등에 의해 제안되었던 이 방법을 Jensen 등은 지수 특이값 분해(quotient SVD, QSVD)를 사용하여 유색잡음에 적용할 수 있는 알고리즘으로 확장하였다[8]. QSVD를 사용하는 이 방법은 계산량이 대단히 많고 잔여잡음 제어가 어려운 문제점이 있다.

음성 향상을 위한 부공간에 근거하는 새로운 접근법이 Epharaim 및 Van trees(EV)[3]에 의해 제안되었다. EV는 미리 정해진 임계치 이하로 잔여잡음을 유지하면서 음성왜곡을 최소화하는 최적 추정자를 찾았다. 그들은 유색잡음에 오염된 신호의 상관행렬을 EVD로 분해하기 위하여 Karhumen-Loêve 변환(KLT)을 사용하였다. 신호 부공간을 나타내는 KLT 성분들은 추정자에 의해 결정되는 이득 함수에 의해 수정되는 반면, 잡음 부공간을 나타내는 나머지 KLT 성분들은 영(zero)이 된다. 이렇게 하여 개선된 신호는 수정되어진 성분들의 역 KLT에 의해 얻을 수 있다. EV는 그들의 알고리즘의 구현에서 잡음을 백색이라 가정하였다.

한편 Mittal와 Phamdo[4] 및 Rezayee 와 Gazor[5]는 Epharaim 및 Van trees의 알고리즘을 유색잡음에 적용될 수 있는 알고리즘으로 확장하였다. Mittal 및 Phamdo는 오염된 신호를 음성이 우선시되는 프레임과 잡음이 우선시되는 프레임으로 나누고 추정자를 얻기 위하여 프레임 마다 다른 KLT 행렬을 적용하였다[4]. Rezayee 및 Gazor는 유색잡음을 다루기 위해 KLT 변환된 잡음벡터들의 상관행렬을 대각선 행렬로 근사화하는 방법으로 EV 알고리즘을 확장하였다. 이 방법에 의해 얻어진 추정자는 최적화된 추정자가 아닌 차선에 해당한다.

최근 Yi 및 Loizou(YL)는 유색잡음에 의해 오염된 음성에 대한 음성 향상를 위해 새로운 방법을 제안하였다[9]. 그들은 깨끗한 음성신호의 상관행렬과 유색잡음의 상관행렬을 동시에 대각행렬로 바꿀 수 있는 방법에 근거를 두는 비일원(nonunitary) 변환을 사용하여, 잡음에 오염된 신호를 신호-잡음(signal-plus-noise) 부공간과 잡음 부공간으로 분리하였다. 깨끗한 신호는 잡음부공간 성분을 영으로 하고 신호부공간 성분을 취하여 재구성할 수 있었다. 이 방법은 알고리즘 자체에 유색잡음신호를 백색화시키는 과정이 포함되어 있기 때문에 가장 일반화된 알고리즘이라고 볼 수 있다. 잡음이 백색일 경우 YL 알고리즘은 EV 알고리즘이 된다.

본 연구에서는 유색잡음에 오염된 신호의 음성 향상을 위해, 잡음에 오염된 음성신호의 전처리로 백색화 변환(whitening transformation : WT)을 이용하는 방법을 제안한다. 음성과 유색잡음이 서로 상관이 없다면 음성신호와는 상관없이 유색잡음은 백색화 시킬 수 있다. 이때 음성신호는 왜곡될 수 있지만 WT에 의한 전처리 다음에 수행되는 음성 향상 후 역 WT에 의해 이 음성왜곡은 원상 복구될 수 있다. 또, WT 후에 수행되는 음성 향상은 EV접근법이나 YL접근법 중 어느 것이나 택할 수 있다. 또 제안하는 알고리즘은 VAD(Voice Activity Detector)을 사용하여 구현하였고, VAD를 사용한 경우와 하지 않는 경우를 YL알고리즘과 컴퓨터 실험으로 성능을 비교하였다.

Ⅱ. 시간영역 부공간 접근법

2.1. 원리

잡음이 섞이지 않았을 때 선형 신호 모델 **x**는

$$\mathbf{x} = \Psi \cdot \mathbf{s} \tag{1}$$

로 표현할 수 있다. 여기서 Ψ는 랭크가 M이고 K X M 사이즈의 행렬이다. 또, M < K이고 **s**는 사이즈가 M X 1인 벡터이다. **x** 의 공분산 행렬(covariance matrix) $\mathbf{R_x}$는

$$R_{\mathbb{X}} \triangleq E\{\mathbb{x} \cdot \mathbb{x}^T\} = \Psi \cdot R_{\mathbb{S}} \cdot \Psi^T \tag{2}$$

로 쓸 수 있는, 여기서 $\mathbf{R_s}$는 벡터 **s**의 공분산 행렬이고 각 인자들은 양수 값을 가진다. $\mathbf{R_x}$의 랭크는 M이고, 따라서 $\mathbf{R_x}$는 K-M개의 영인 고유치(eigenvalue)을 갖는다.

깨끗한 신호 **x**와 잡음신호 **n**가 서로 상관관계에 없다고 하면, 오염된 신호 **y**는,

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

가 된다. 여기서 **y**, **x** 및 **n**는 각각 K-차원의 오염된 음성신호벡터, 오염안된 음성신호 벡터 및 잡음신호 벡터이다. $\hat{\mathbf{x}} = H \cdot \mathbf{y}$ 를 깨끗한 음성신호 **x**의 선형 추정자(linear estimator)라하고, H는 K X K 행렬이다. 이 추정에 의해서 얻어지는 오차신호 ε 는

$$\varepsilon = \mathbf{x} - \mathbf{x} = (H - I) \cdot \mathbf{x} + H \cdot \mathbf{n} = \varepsilon_{\mathbf{x}} + \varepsilon_{\mathbf{n}} \tag{4}$$

가 된다. 여기서 $\varepsilon_{\mathbf{x}}$는 음성왜곡을 나타내고 $\varepsilon_{\mathbf{n}}$ 는 잔여잡음을 나타낸다. 신호왜곡 에너지 $\overline{\varepsilon_{\mathbf{x}}^2}$ 는

$$\overline{\varepsilon_{\mathbf{x}}^2} = E[\varepsilon_{\mathbf{x}}^T \varepsilon_{\mathbf{x}}] = tr(E[\varepsilon_{\mathbf{x}} \varepsilon_{\mathbf{x}}^T]) \tag{5}$$

이고, 잔여잡음 에너지 $\overline{\varepsilon_{\mathbf{n}}^2}$ 는

$$\overline{\varepsilon_{\mathbf{n}}^2} = E[\varepsilon_{\mathbf{n}}^T \varepsilon_{\mathbf{n}}] = tr(E[\varepsilon_{\mathbf{n}} \varepsilon_{\mathbf{n}}^T]) \tag{6}$$

이다. 널리 알려진 시간영역 제한 최적화 문제[3, 5]

$$\text{subject to } : \frac{1}{K}\overline{\varepsilon_{\mathbf{x}}^2} \le \sigma^2 \tag{7}$$

의 해를 구하면 최적 선형 추정자 **H**를 얻을 수 있다. 여기서 σ^2 는 양의 실수이다. 식(7)의 해는

$$H_{opt} = R_{\mathbf{x}}(R_{\mathbf{x}} + \mu R_{\mathbf{n}})^{-1} \tag{8}$$

이다[3]. 여기서 $R_{\mathbf{x}}$ 및 $R_{\mathbf{n}}$ 은 각각 깨끗한 음성 및 잡음의 공분산 행렬이다. μ 는 라그랑지 승수(Lagrange multiplier) 이다.

식 (8)은 고유값 분해(EVD) $R_{\mathbf{x}} = U \Delta_{\mathbf{x}} U^T$ 을 이용하면

$$H_{opt} = U \Delta_{\mathbf{x}} (\Delta_{\mathbf{x}} + \mu U^T R_{\mathbf{x}} U)^{-1} U^T \tag{9}$$

로 간략화 된다. 여기서 U 는 일원(unitary) 고유벡터 행렬이고 $\Delta_{\mathbf{x}}$ 는 $R_{\mathbf{x}}$ 의 대각선 고유치 행렬이다. $R_{\mathbf{n}} = \sigma_{\mathbf{n}}^2 I$ 인 백색잡음일 경우, 식 (9)는 Ephraim 및 Van Tree 선형 추정자와 일치한다[3]. 행렬 $U^T R_{\mathbf{n}} U$ 의 대각선 행렬 $\Delta_{\mathbf{n}}$ 는

$$\Delta_{\mathbf{x}} = diag(E(|u_1^T n|), E(|u_2^T n|), \cdots, E(|u_k^T n|)) \tag{10}$$

로 근사시킨다[5]. 여기서 u_k 는 $R_{\mathbf{x}}$ 의 k번째 고유벡터이다. 그리고 음성신호의 음성결여 세그먼트로 추정된 잡음벡터이다. 식 (10)의 근사로 식 (9)는

$$H_{opt} \approx U \Delta_{\mathbf{x}} (\Delta_{\mathbf{x}} + \mu \Delta_{\mathbf{x}})^{-1} U^T \tag{11}$$

가 된다[5]. 근사식 식 (10)을 사용했기 때문에 유도된 추정자는 준 최적값(Suboptimal)이 된다[5]. 유색잡음에 대해서도 최적 추정자를 유도 할 수 있다[9]. 식 (9)에 있는 행렬 $U^T R_{\mathbf{n}} U$ 는 대각선 행렬에 가까우나, 대각선 행렬은 아니다. 행렬 U는 대칭행렬 $R_{\mathbf{x}}$ 의 고유벡터 행렬로서 $R_{\mathbf{x}}$ 을 대각선 행렬로 만드나 $R_{\mathbf{n}}$ 을 대각선 행렬화 하지는 않는다. $U^T R_{\mathbf{n}} U$ 를 식 (10)에 의해 근사시키는 대

신에 R_x 및 R_n 을 동시에 대각선 행렬화 시킬 수 있는 행렬을 구했다[9]. 그러한 행렬 V 는 존재하고,

$$V^T R_x V = \Lambda_x,\ V^T R_n V = I \tag{12}$$

가 된다[10]. 여기서 Λ_x 및 V 는 각각 $\Sigma = R_n^{-1} R_x$ 즉,

$$\Sigma V = V \Lambda_x \tag{13}$$

의 고유치 행렬 및 고유벡터 행렬이다. R_n 이 양수 값을 가지면 Λ_x 는 실수 행렬이 된다[11]. 고유벡터 행렬 V 는 직교행렬이 아니다. R_x 의 랭크가 M이기 때문에 행렬 Σ 도 랭크가 M이다. 식 (8)에 Σ 도 고유값 분해를 적용하고 식 (12)을 이용하면 최적 추정자는

$$\begin{aligned} H_{opt} &= R_n V \Lambda_x (\Lambda_x + \mu I)^{-1} V^T \\ &= V^{-T} \Lambda_x (\Lambda_x + \mu I)^{-1} V^T \end{aligned} \tag{14}$$

로 얻어진다. 여기서, 상수 μ 는,

$$\sigma^2 = \frac{1}{K} tr\{(V^T V)^{-1} \Lambda_x^2 (\Lambda_x + \mu I)^{-2}\} \tag{15}$$

를 만족해야한다.

깨끗한 음성신호의 추정신호 $\hat{\mathbf{x}}$ 는 잡음이 섞인 신호 $\mathbf{y}$ 에 V^T 를 사용하여 변환시키고, 이득함수를 $V^T\mathbf{y}$ 의 성분들에 곱하고, 여기에 수정된 성분들을 위에서 구한 최적 추정자를 이용해 역변환시켜 얻을 수 있다. 그림 1 은 이 과정들을 설명한다.

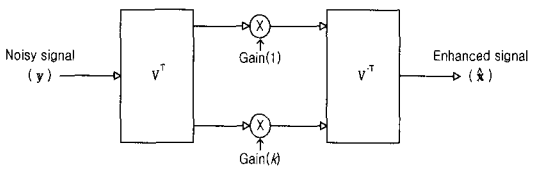

그림 1. 신호 부공간 선형 추정자
Fig. 1. Signal subspace linear estimator

이득 행렬 $G = \Lambda_x (\Lambda_x + \mu I)^{-1}$ 는 대각선 행렬이고, 이것의 k번째 대각선 요소 g_{kk} 는

$$g_{kk} = \begin{cases} \dfrac{\lambda_x^{(k)}}{\lambda_x^{(k)} + \mu} & ,k = 1,2,\cdots,M \\ 0 & ,k = M+1,\cdots,K \end{cases} \tag{16}$$

이다. 여기서 $\lambda_x^{(k)}$ 는 고유치행렬 Λ_x 의 k번째 대각선 요소이고, M은 행렬 Σ 의 랭크이고 음성신호 부공간의 차원이다. 이 경우 $V^T\mathbf{y}$ 는 $\mathbf{y}$ 의 KLT이 아니다. 그러나, 잡음신호가 백색잡음이면 $V^T\mathbf{y}$ 는 $\mathbf{y}$ 의 KLT이 된다.

식 (14)에서 얻어진 추정자와 참고문헌 [3]에서 잡음이 백색일 때 얻어지는 선형 추정자를 비교하면 두 추정자는 같은 형태를 가진다는 것을 알 수 있다. Ephrain 및 Van Tree[3]의 추정자는 식 (4)에 서 주어진 추정자의 특별한 경우가 된다. 백색잡음 ($R_n = \sigma_n^2 I$) 인 경우 Σ 의 고유벡터 행렬 V 는 R_x 의 일원 고유벡터 행렬 U 가 된다. 이것은 $\Sigma = (1/\sigma_n^2) R_x$ 및 대각선 행렬 Λ_x 가 $(1/\sigma_n^2)\Delta_x$ 가 되기 때문이다. Δ_x 는 R_x 의 대각선 고유치 행렬이다. 따라서 백색잡음인 경우 식 (14)는

$$H_{opt} = U \Delta_x (\Delta_x + \mu \sigma_n^2 I)^{-1} U^T \tag{17}$$

로 표시되고 Ephrain 및 Van Tree[3] 추정자이다. 선형추정자 식 (14)는 백색잡음이나 유색잡음에 모두 적용될 수 있는 추정자이고, 참고문헌 [3]에 개발되었던 부공간 접근법의 일반화가 식 (14)에 주어진 추정자이다.

식 (14)의 실제 구현에는 행렬 Σ 를 추정해야 하고, 실제에서는 깨끗한 음성신호의 공분산 행렬을 구할 수 없다. 따라서, 음성신호가 잡음과 상관관계가 없다고 가정하면

$$R_y = R_x + R_n \tag{18}$$

을 얻게 되고 요구되는 Σ 는

$$\Sigma = R_n^{-1} R_x = R_n^{-1}(R_y - R_n)$$
$$= R_n^{-1} R_y - I \tag{19}$$

이다.

2.2. μ 값의 추정

식 (16)에 주어진 이득 함수의 μ 값은 실제 알고리즘의 구현을 위해 추정되어야 하고, 잔여잡음과 음성왜곡 사이에서 상충되기 때문에 추정된 음성신호 Λ_x의 음질을 좌우한다. μ 값이 크면 잔여잡음을 많이 제거되나, 음성왜곡 또한 커진다. 역으로 μ 값을 적게 선택하면 음성왜곡은 최소화 할 수 있으나, 잔여잡음은 커진다. 따라서 적절한 μ 값을 선택해 잔여잡음과 음성왜곡 사이에서 타협점을 찾아야한다.

음성신호는 배경잡음에 대해서 마스킹 효과가 있기 때문에 음성신호가 강조된 프레임에서는 음성왜곡을 최소화해야하고, 반면 배경잡음이 강조된 프레임에서는 잔여잡음을 감소시켜야 한다. μ 값을 단구간 SNR에 의존하도록 만들어야 한다. 따라서, μ 값은

$$\mu = \mu_0 - (SNR_{dB})/s \tag{20}$$

에 의해 추정한다. 여기서, μ_0 및 s는 실험적으로 선택해야할 상수[9]이고, $SNR_{dB} = 10\log_{10} SNR$ 이다. 스펙트럼 차감법에 과차감 요소(over-subtraction factor)를 추정하기 위해 식 (20)이 사용되었다[1]. 또한 프레임 *SNR*에 관계없이 고정된 μ 값을 사용한 경우도 있다[3,6].

식 (12)에서, 고유치 $\lambda_x^{(k)}$ 는 고유벡터 $\mathbf{V}_k$ [즉, $\lambda_x^{(k)} = E(|\mathbf{V}_k^T \mathbf{x}|^2)$]에 상응하는 신호 전력이다. 그러므로, SNR값의 추정은 변환된 영역에서

$$SNR = \frac{\mathrm{tr}(V^T R_x V)}{\mathrm{tr}(V^T R_n V)} = \frac{\sum_{k=1}^{M} \lambda_x^{(k)}}{K} \tag{21}$$

에 의해 구할 수 있다[9].

III. 백색화 변환에 의한 부공간 접근법의 제안

3.1. 백색화 변환

음성신호가 잡음신호와 상관관계가 없다면, 이 두 신호를 합한 신호의 공분산 행렬 식 (18)이 된다. 식 (18)에서 잡음신호의 공분산 행렬 R_n 은

$$R_n = E \Lambda_n E^T \tag{22}$$

로 분해된다. 여기서, E 및 Λ_n 은 각각 행렬 R_n 의 직교 고유벡터 행렬 및 대각선 고유치 행렬이다. 잡음이 백색이면 E는 식 (17)의 고유벡터 행렬 U 가 된다. 식 (22)의 행렬 E 및 Λ_n 을 사용하여 식 (3)의 **y** 에 다음과 같은 선형변환을 적용하면

$$\mathbf{z} = \Lambda_n^{-\frac{1}{2}} \cdot E^T \cdot \mathbf{y} \tag{23}$$

을 얻는다. 행렬 $\Lambda_n^{-\frac{1}{2}}$ 은 대각선 행렬 $\Lambda_n^{\frac{1}{2}}$ 의 역행렬이고 $\Lambda_n^{\frac{1}{2}}$ 의 각 요소들은 고유치의 제곱근이다. 식 (23)의 변환은 백색화 변환으로 알려져 있다. 백색화 변환된 영역에서, 유색잡음은 백색잡음으로 변환된다. 즉

$$R_x' = \Lambda_n^{-\frac{1}{2}} E^T E \Lambda_n E^T E \Lambda_n^{-\frac{1}{2}} = I \tag{24}$$

가 되고, 음성신호 또한 선형변환 되어 그 공분산 행렬은

$$R_x' = \left(\Lambda_n^{-\frac{1}{2}} E^T\right) R_x \left(\Lambda_n^{-\frac{1}{2}} E^T\right)^T$$
$$= \Lambda_n^{-\frac{1}{2}} E^T R_x E \Lambda^{-\frac{1}{2}} \tag{25}$$

가 된다. 백색화 변환을 음성 향상 알고리즘의 전처리기로 사용하면 모든 유색잡음신호는 백색잡음신호로 변환시킬 수 있다. 음성신호도 또한 식 (23)에 따라 변환되지만 그림 2 및 그림 3에서 보는바와 같이 최종단계에서 역 화이트닝 변환하면 원 음선신호로 복구 할 수 있다. 따라서 음성 향상 알고리즘은 EV 알고리즘이나 YL 알고리즘 중 어느 것이나 사용될 수 있다.

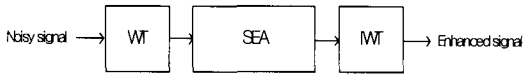

그림 2. 화이트닝 변환을 가지는 음성 인헨스먼트시스템
WT : 백색화 변환
IWT : 역 백색화 변환
SEA : 음성 향상 알고리즘
Fig. 2. Speech enhancement system with whitening transformation
WT : Whitening transformation
IWT : Inverse whitening transformation
SEA : Speech enhancement algorithm

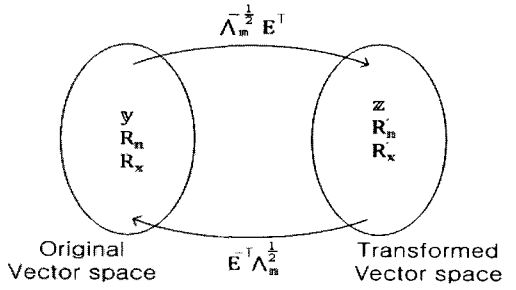

그림 3. 유색잡음에 있는 음성신호의 백색화 변환
Fig. 3. Whitening transformation for signals in colored noise

3.2. 알고리즘 구현

제안하는 알고리즘은 잡음에 오염된 신호 y 의 매 프레임마다 다음 아홉 단계의 과정을 거친다.

단계 1: 매 프레임의 에너지를 구한다. 즉

$$\alpha = \sum_{n=0}^{k} |y(n)|^2 \tag{26}$$

α 값이 임계치 α_{TH}보다 적으면 그 프레임을 잡음 프레임으로 보고 잡음 공분산 행렬 $\mathbf{R_n}$을 구한다. 여기서 α_{TH}는 실험으로 정해진다.

단계 2: $\mathbf{R_n}$을 고유분해(EVD)하여 고유벡터 행렬 E 및 대각선 고유치행렬 $\Lambda_{\mathbf{n}}$을 구하고, 이것을 이용해 식 (23)에서처럼 선형변환 한다.

단계 3: 잡음이 섞인 음성신호의 공분산 행렬 $\mathbf{R_y}$을 구하고 식 (19)을 사용하여 행렬 Σ을 추정한다.

단계 4: Σ을 고유분해(EVD)한다

$$\Sigma \mathbf{V} = \mathbf{V}\Lambda_{\mathbf{x}} \tag{27}$$

단계 5: Σ의 고유치가 큰 순서대로 놓았다면 (즉, $\lambda_{\mathbf{x}}^{(1)}, \lambda_{\mathbf{x}}^{(2)}, \cdots, \lambda_{\mathbf{x}}^{(K)}$), 다음 식 (28)을 이용하여 음성신호 부공간의 차원 M을 구한다.

$$\mathbf{M} = \arg \max_{1 \le k \le K} \{\lambda_{\mathbf{x}}^{(k)} > 0\} \tag{28}$$

단계 6: 다음 식 (29)에 따라 μ 값을 추정한다.

$$\mu = \begin{cases} \mu_0 - (SNR_{dB})/s & , -5 < SNR_{dB} < 20 \\ 1 & , SNR_{dB} \ge 20 \\ 5 & , SNR_{dB} \ge -5 \end{cases} \tag{29}$$

여기서 μ_0 및 s는 실험적으로 선택하며[9], $\mu_0 = 4.2$, $s = 6.25$ 로 두었다. $SNR_{dB} = 10\log_{10} SNR$ 이며 SNR 은 식 (21)에서와 같이 구한다.

단계 7: 최적 선형 추정자를 계산한다.

$$g_{kk} = \begin{cases} \dfrac{\lambda_{\mathbf{x}}^{(k)}}{\lambda_{\mathbf{x}}^{(k)} + \mu} & , k = 1, 2, \cdots, M \\ 0 & , k = M+1, \cdots, K \end{cases}$$

$$\mathbf{G}_1 = \mathrm{diag}\{g_{11}, \cdots, g_{MM}\}$$

$$\begin{aligned} \mathbf{H}_{OPT} &= \mathbf{R_n}\,\mathbf{V} \begin{bmatrix} \mathbf{G}_1 & 0 \\ 0 & 0 \end{bmatrix} \mathbf{V}^{\mathrm{T}} \\ &= \mathbf{V}^{-\mathrm{T}} \begin{bmatrix} \mathbf{G}_1 & 0 \\ 0 & 0 \end{bmatrix} \mathbf{V}^{\mathrm{T}} \end{aligned} \tag{30}$$

단계 8: 잡음이 섞인 신호 y 을 식 (30)에 적용하여, 잡음이 제거된 음성신호를 추정한다.

단계 9: 단계 2에서 구한 E 및 $\Lambda_{\mathbf{n}}$ 행렬들을 이용하여 역 백색화 변환하여 원 음성신호를 복구시킨다.

Ⅳ. 컴퓨터 실험 및 성능 평가

알고리즘의 구현에는 잡음에 오염된 신호의 공분산 행렬 R_y와 잡음신호의 공분산 행렬 R_n을 정확하게 추정해야한다. 본 연구에서는 비편향 자기상관 시퀀스(autocorrelation sequence)을 관측되는 신호로부터 얻었다. 이 시퀀스의 K샘플을 이용하여 Toeplitz 공분산 행렬을 구성하였다. 공분산 행렬을 구성할 때 과거나 미래 프레임을 사용하지 않았다. 음성신호와 잡음신호는 모두 16KHz에서 샘플링 되었다. K=40샘플이었다. 참고문헌[9]에서는 잡음신호의 공분산 행렬 R_n을 추정하기 위해 테스트 문장에 있는 첫 몇 프레임을 사용하여 R_n을 추정하고, VAD를 사용하여 R_n을 갱신하지 않았다. 그러나, 본 논문에서는 제 3장 2절의 단계 1 및 2에서 언급한 바와 같이 프레임마다 그 프레임의 에너지를 식 (26)과 같이 구하여 실험적으로 정해지는 임계값보다 작은 값이면 그 프레임을 잡음 프레임으로 간주하여 R_n을 갱신하였다.

알고리즘의 구현에는 공분산 행렬을 추정하기 위하여 사각 윈도우를 사용하였고, 매 프레임을 50% 중첩시켰다. 잡음에 오염된 음성신호는 최종적으로 해밍(Hamming) 윈도우와 음성 향상된 후 중첩 및 가산(overap-and-add) 접근법을 사용하여 복구하였다.

4.2. 성능평가

4.2.1. 평가척도

성능평가는 음성 스펙트럼 왜곡(speech spectral distortion : SSD)과 음성 향상 시스템 출력에서의 신호 대 잡음비인 SNR을 평가 척도로 이용하였다. 먼저 2차원 SSD 척도는 다음과 같이 정의된다[4]. $\mathbf{a}$ 및 $\mathbf{b}$ 를 N차원 벡터화 한다. 각 벡터는 먼저 에너지가 1(0dB)이 되도록 정규화 한다. 그리고 -30dB의 에너지를 가지는 또 다른 백색잡음 벡터 $\mathbf{c}$ 를 백터 $\mathbf{a}$ 및 $\mathbf{b}$ 에 보탠다. 이 백색 잡음 벡터 $\mathbf{c}$ 을 보태는 이유는 내수적으로 계산될 척도에서 log(0)의 계산을 막기 위함이다. 따라서 벡터 $\mathbf{a}$ 및 $\mathbf{b}$ 는 각각 $\tilde{\mathbf{a}} = \mathbf{a}/\|\mathbf{a}\| + \mathbf{c}$ 및 $\tilde{\mathbf{b}} = \mathbf{b}/\|\mathbf{b}\| + \mathbf{c}$ 가 된다. 벡터 $\mathbf{a}$ 및 $\mathbf{b}$ 는 길이가 64인 중첩되지 않는 프레임들로 나눈다. 나누어진 각 64샘플 프레임에 영을 192개 보태어 영첨가(zero-padding) 한다. 이렇게 하여 생긴 256샘플의 매 프레임의 DFT를 계산한다. $\widetilde{A}_p(k)$ 및 $\widetilde{B}_p(k)$ 를 각각 벡터 $\mathbf{a}$ 및 $\mathbf{b}$ 의 DFT의 P번째 프레임의 k번째 주파수 성분이라 한다.

벡터의 $\mathbf{a}$ 및 $\mathbf{b}$ 간의 SSD는

$$S(\boldsymbol{a},\boldsymbol{b}) = \frac{1}{p}\frac{1}{256}\sum_{i=1}^{p}\sum_{k=0}^{255} 20\left|\log|\widetilde{A}_p(k)| - \log|\widetilde{B}_p(k)|\right| \tag{31}$$

로 정의 된다. 여기서 $\boldsymbol{p}$는 관측된 음성신호의 총 프레임 수이다.

제안하는 추정자는 선형이기 때문에 식 (3) 및 식 (14)로부터

$$\mathbf{y} = \mathrm{H}_{\mathbf{y}} = \mathrm{H}_{\mathbf{x}} + \mathrm{H}_{\mathbf{n}} \tag{32}$$

로 나누어 질 수 있다. 여기서 $\mathrm{H}_{\mathbf{x}}$ 는 신호 부분이고 $\mathrm{H}_{\mathbf{n}}$ 은 잡음 부분이다. 이상적으로 볼때 $\mathrm{S}(\mathbf{x},\mathrm{H}_{\mathbf{x}}) = 0$ 이 되도록 하는 선형 추정자 H 가 필요하다. 이 $\mathrm{S}(\mathbf{x},\mathrm{H}_{\mathbf{x}})$ 을 음성 스펙트럼 왜곡(SSD)이라 정의한다.

잡음에 오염된 신호 y 의 신호대잡음비(SNR)은

$$SNR = 10\log_{10}\frac{\sum_{t=1}^{N}\sum_{k=1}^{K}\mathrm{x}_t^2(k)}{\sum_{t=1}^{N}\sum_{k=1}^{K}(\mathrm{y}_{t(k)} - \mathrm{x}_t(k))^2} \tag{33}$$

로 정의 되었고, 여기서 N은 관측된 신호 y 의 총 프레임 수이고 $\mathrm{y}_t(k)$ 는 t 번째 프레임의 k번째 샘플이다. 음성 향상 시스템의 출력 SNR은 식 (32)로부터 $\mathrm{H}_{\mathbf{x}}$ 및 $\mathrm{H}_{\mathbf{y}}$ 을 사용하여 식 (33)으로 계산되었다.

4.2.2. 실험결과 및 성능평가

성능평가를 위해 5명의 남자와 5명의 여자가 발음한 문장 "She had your dark suit in greasy wash water all year"을 TIMIT 데이터베이스로부터 발췌하여 입력 음성 데이터로 사용하였다. 음성 데이터의 시작 부분에서 0.05초 미만 구간의 큰 잡음이 있어 전처리로 제거한 후 사용하였으며, 샘플링율은 16 ㎑이다. 깨끗한 음성에 자동차 잡음과 다중화자 배블(multi-talker babble) 잡음은 AURORA 데이터베이스로부터 취하여 SNR이 5dB이 되도록 음성 데이터에 각각 더하여 입력으로 사용하였

다. 음성 출력 신호의 SNR은 식 (33)을 사용하여 계산하였고, 음성 스펙트럼 왜곡은 식 (31)을 사용하여 계산하였다. 이러한 실험환경에 따라 남성 화자 5명과 여성화자 5명에 대해 각각 실험하고 평균한 결과를 표 1에서 표 4까지 나타내었다.

표 1 및 표 2 는 깨끗한 음성에 자동차 잡음을 SNR이 5dB가 되도록 첨가한 경우의 실험결과이다. 실험결과에서 남성 화자 항은 5명의 남자가 발음한 문장을 실험하여 평균하였다. 같은 방법으로 여성 화자 항은 여자 5명이 같은 문장을 발음한 것을 실험하여 평균하였다.

표 1. 5dB 자동차 소음 첨가한 경우의 성능 비교 (전체 평균 SNR)
Table. 1 Comparison for adding 5dB car noise (total average SNR)

알고리즘 종류	남성 화자	여성 화자
[9]에 사용된 알고리즘	7.67	7.32
제안한 알고리즘	8.42	8.33
제안한 알고리즘 +VAD	9.34	9.29

표 2. 5dB 자동차 소음 첨가한 경우의 성능 비교 (평균 SSD)
Table. 2 Comparison for adding 5dB car noise (average SSD)

알고리즘 종류	남성 화자	여성 화자
[9]에 사용된 알고리즘	0.68	0.65
제안한 알고리즘	0.74	0.68
제안한 알고리즘 +VAD	0.68	0.65

표 1에서 볼 수 있듯이 자동차 잡음일 경우 출력 SNR이 VAD가 적용된 제안한 알고리즘이 기존의 알고리즘보다 남성음성의 경우 약 1.67dB, 여성음성일 경우 약 1.97dB 정도 향상되었다. 또, VAD가 적용되지 않은 제안한 알고리즘이 기존의 알고리즘보다 남성음성일 경우 약 0.75dB, 여성음성일 경우 약 1.01dB정도 향상되었다. 그러나 음성 스펙트럼 왜곡의 경우 자동차 잡음일 때 기존 알고리즘에 추가되는 백색화 변환에도 불구하고, 기존 알고리즘과 제안하는 두 알고리즘이 비슷하였다. 그러나 제안한 알고리즘이 기존의 알고리즘보다 남성음성일 경우에는 0.06dB정도, 여성음성일 경우 0.03dB 정도 높았다.

표3 및 표4는 깨끗한 음성에 다중화자 배블을 SNR이 5dB이 되게 첨가하였을 경우의 실험결과이다. 실험방법은 자동차 잡음일 경우와 동일하게 실행하였다.

표 3. 5dB 다중화자 배블 잡음 첨가한 경우의 성능 비교(전체 평균 SNR)
Table. 3 Comparison for adding 5dB multi-talker babble(total average SNR)

알고리즘 종류	남성 화자	여성 화자
[9]에 사용된 알고리즘	8.28	8.06
제안한 알고리즘	7.67	7.33
제안한 알고리즘 +VAD	8.37	8.04

표 4. 5dB 다중화자 배블 잡음 첨가한 경우의 성능 비교(평균 SSD)
Table. 4 Comparison for adding 5dB multi-talker babble(average SSD)

알고리즘 종류	남성 화자	여성 화자
[9]에 사용된 알고리즘	0.62	0.58
제안한 알고리즘	0.69	0.66
제안한 알고리즘 +VAD	0.67	0.62

표 3에 언급한 바와 같이 다중 화자 배블인 경우 입력 SNR이 5dB일 때 기존 알고리즘과 VAD가 적용된 제안한 알고리즘의 출력 SNR은 거의 비슷하였으나, 오히려 VAD가 적용되지 않은 제안한 알고리즘이 남자의 경우 약 0.6dB 낮았으며 여자의 경우는 약 0.73dB가 낮았다. 그러나 표 4에서 보는 바와 같이 다중 화자 배블의 경우 음성 스펙트럼 왜곡은 기존 알고리즘이 가장 나았다.

표 1 및 표 3에서 나타난 바와 같이 SNR 성능은 VAD를 가지는 제안한 알고리즘이 가장 좋고, 표2 및 표4에서와 같이 음성 스펙트럼 왜곡은 대체로 VAD 없이 제안한 알고리즘 성능이 좋지 않았는 데, 이는 알고리즘 전단부에서 처리하는 유색잡음의 백색화 과정이 있기 때문인 것으로 분석된다.

V. 결 론

본 논문에서는 유색잡음에 오염된 음성신호의 음성 향상을 위해, 잡음에 오염된 음성신호의 전처리로 백색화 변환을 이용하는 알고리즘을 제안하였다. 음성신호와 유색잡음이 서로 상관이 없다면 음성 향상 처리 전에 유색잡음으로 처리할 수 있다. 백색화 전처리로 발생되는 음성왜곡은 음성 향상 후에 역 백색화에 의해 원상복구 할 수 있었다. 제안한 알고리즘의 컴퓨터 시뮬레이션 결과는 다음과 같다.

1) 유색잡음이 자동차 잡음인 경우 (입력 SNR은 5dB)
   ① 출력 SNR은 남성음성의 경우 기존 알고리즘보다 제안한 알고리즘이 0.75dB향상되었고, VAD가 적용된 제안한 알고리즘은 1.67dB향상되었다. 여성음성의 경우는 기존 알고리즘보다 제안한 알고리즘이 1.01dB향상되었고, VAD가 적용된 제안한 알고리즘은 1.97dB향상되었다.
   ② 음성 스펙트럼 왜곡은 남성음성과 여성음성의 경우 기존 알고리즘이나 제안한 두 알고리즘이 모두 거의 유사하였다.
2) 유색잡음이 multi-talker babble인 경우 (입력 SNR은 5dB)
   ① 출력 SNR은 기존 알고리즘과 제안한 VAD를 적용한 알고리즘은 남성과 여성음성에 관계없이 거의 비슷하였으나, 제안한 알고리즘은 기존의 알고리즘과 비교했을 경우 남성음성일 때 0.61dB, 여성음성일 때 0.73dB정도 더 감소하였다.
   ② 음성 스펙트럼 왜곡은 남성음성일 때 기존 알고리즘이 제안한 알고리즘보다 0.07dB정도 더 나았고, VAD가 적용된 알고리즘의 경우 0.05dB정도 나았고 여성음성일 때도 이와 비슷하였다.

감사의 글

"이 논문은 부산대학교 자유과제 학술연구비(2년)에 의하여 연구되었음."

참고문헌

[1] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* 1979, pp. 208 - 211.

[2] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars," *Speech Commun.,* vol. 11, no. 2, pp.215 - 228, 1992.

[3] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing,* vol. 3, pp.251 - 266, 1995.

[4] U. Mittal and N. Phamdo, "Signal/noise KLT based approach for enhancing speech degraded by colored noise," *IEEE Trans. Speech Audio Processing,* vol. 8, pp. 159 - 167, Mar. 2000.

[5] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Trans. Speech Audio Processing,* vol. 9, pp. 87 - 95, Feb. 2001.

[6] Y. Hu and P. C. Loizou, "A subspace approach for enhancing speech corrupted by colored noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* vol. 1, Orlando, FL, May 2002, pp. 573 - 576.

[7] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Commun.,* vol. 10, pp. 45 - 57, 1991.

[8] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech Audio Processing,* vol. 3, pp. 439 - 448, Nov. 1995.

[9] Yi Hu and Philipos C. Loizou, A Generalized Subspace Approach for Enhancing Speech Corrupted by Colored Noise *IEEE Trans. Speech Audio Processing,* vol. 11, no. 4, pp. 334-341. July 2003.

[10] S. B. Searle, *Matrix Algebra Useful for Statistics.* New York: Wiley, 1982.

[11] *G. Strang, Linear Algebra and Its Applications,* 3rd ed. New York: Harcourt Brace Jovanonich, 1988.

저자소개

이정욱(Jeong-wook Lee)

2009년 3월 ~ 현재 부산대학교 전자공학과 석사과정
※관심분야 : 적응신호처리, 음향신호처리

손경식(Kyung-sik Son)

1973년 2월 부산대학교
전자공학과(학사)
1977년 8월 부산대학교
전자공학과(석사)

1991년 8월 경북대학교 전자공학과(박사)
1986년 10월 ~ 현재 부산대학교 전자공학과 교수
※관심분야 : 적응신호처리, 음향신호처리

박장식(Jang-Sik Park)

1992년 2월 부산대학교
전자공학과(학사)
1994년 2월 부산대학교
전자공학과(석사)

1999년 2월 부산대학교 전자공학과(박사)
1997년 3월 ~ 2011년 2월 동의과학대학 전자과 교수
2011년 3월~현재 경성대학교 전자공학과 부교수
※관심분야 : 음성 및 음향신호처리, 멀티미디어통신, 입체음향

김현태(Hyun-Tae Kim)

1989년 2월 부산대학교
전자공학과(학사)
1995년 2월 부산대학교
전자공학과(석사)

2000년 2월 부산대학교 전자공학과(박사)
2002년 3월 ~ 현재 동의대학교 멀티미디어공학과
부교수
※관심분야 : 음성 및 음향신호처리, 멀티미디어신호처리, 입체음향