주파수 특성 기저벡터 학습을 통한 특정화자 음성 복원

(Target Speaker Speech Restoration via Spectral bases Learning)

박 선 호 † 유 지 호 † 최 승 진 ††
(Sunho Park) (Jiho Yoo) (Seungjin Choi)

요 약 본 논문에서는 학습이 가능한 특정화자의 발화음성이 있는 경우, 잡음과 반향이 있는 실 환경에서의 스테레오 마이크로폰을 이용한 특정화자 음성복원 알고리즘을 제안한다. 이를 위해 반향이 있는 환경에서 음원들을 분리하는 다중경로 암묵음원분리(convolutive blind source separation, CBSS)와 이의 후처리 방법을 결합함으로써, 잡음이 섞인 다중경로 신호로부터 잡음과 반향을 제거하고 특정화자의 음성만을 복원하는 시스템을 제시한다. 즉, 비음수 행렬분해(non-negative matrix factorization, NMF) 방법을 이용하여 특정화자의 학습음성으로부터 주파수 특성을 보존하는 기저벡터들을 학습하고, 이 기저벡터들에 기반 한 두 단계의 후처리 기법들을 제안한다. 먼저 본 시스템의 중간단계인 CBSS가 다중경로 신호를 입력받아 독립음원들을(두 채널) 출력하고, 이 두 채널 중 특정화자의 음성에 보다 가까운 채널을 자동적으로 선택한다(채널선택 단계). 이후 앞서 선택된 채널의 신호에 남아있는 잡음과 다른 방해음원(interference source)을 제거하여 특정화자의 음성만을 복원, 최종적으로 잡음과 반향이 제거된 특정화자의 음성을 복원한다(복원 단계). 이 두 후처리 단계 모두 특정화자 음성으로부터 학습한 기저벡터들을 이용하여 동작하므로 특정화자의 음성이 가지는 고유의 주파수 특성 정보를 효율적으로 음성복원에 이용 할 수 있다. 이로써 본 논문은 CBSS에 음원의 사전정보를 결합하는 방법을 제시하고 기존의 CBSS의 분리 결과를 향상시키는 동시에 특정화자만의 음성을 복원하는 시스템을 제안한다. 실험을 통하여 본 제안 방법이 잡음과 반향 환경에서 특정화자의 음성을 성공적으로 복원함을 확인할 수 있다.

키워드 : 특정화자 음성복원, 잡음과 반향 환경, 다중경로 암묵음원분리(CBSS), 비음수 행렬분해, 채널선택 단계, 복구 단계

Abstract This paper proposes a target speech extraction which restores speech signal of a target speaker form noisy convolutive mixture of speech and an interference source. We assume that the target speaker is known and his/her utterances are available in the training time. Incorporating the additional information extracted from the training utterances into the separation, we combine convolutive blind source separation(CBSS) and non-negative decomposition techniques, e.g., probabilistic latent variable model. The nonnegative decomposition is used to learn a set of bases from the spectrogram of the training utterances, where the bases represent the spectral information corresponding to the target speaker. Based on the learned spectral bases, our method provides two postprocessing steps for CBSS. Channel selection step finds a desirable output channel from CBSS, which dominantly contains the target speech. Reconstruct step recovers the original spectrogram of the target speech from the selected output channel so that the remained interference source and

· 이 논문은 제35회 추계학술대회에서 '주파수 특성 기저벡터 학습을 통한 특정화자 음성 복원'의 제목으로 발표된 논문을 확장한 것임
· 본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 지원 사업의 연구결과로 수행되었음(IITA-2008-C1090-0801-0045)

† 학생회원 : 포항공과대학교 컴퓨터공학과
titan@postech.ac.kr
zentasis@postech.ac.kr
†† 종신회원 : 포항공과대학교 컴퓨터공학과 교수
seungjin@postech.ac.kr

논문접수 : 2008년 12월 15일
심사완료 : 2009년 1월 29일

정보과학회논문지: 소프트웨어 및 응용 제36권 제3호(2009.3)

background noise are suppressed. Experimental results show that our method substantially improves the separation results of CBSS and, as a result, successfully recovers the target speech.

Key words : target speech extraction, convolutive blind source separation(CBSS), training utterances, non-negative decomposition techniques, postprocessing steps for the CBSS

1. 서 론

실 환경에서 특정화자만의 음성을 추출, 복원하는 기술은 이론적 연구가치 뿐 아니라 실무적 유용성을 동시에 지닌다. 예를 들어, 사무실에서 음성인식을 실행 할 경우 여러 잡음과 다른 방해음원신호(interference source)에 의해 인식률이 크게 저하된다. 이때 잡음과 방해음원을 제거하고 화자의 명령 음성신호만을 추출함으로써 음성 인식률을 크게 향상시킬 수 있다. 그러나 실 환경에서의 특정화자 음성복원은 아직 뚜렷한 방법론이 제시되지 않아 이에 대한 많은 연구가 필요한 영역이다. 특히, 특정화자의 음성에서 음원분리에 유용한 정보를 추출하고 이를 잡음과 반향의 제거에 적용하는 효과적인 방법론이 제시되지 못하고 있다. 다만 이와 비슷한 목적을 가지는 여러 다중경로 암묵음원분리(convolutive blind source separation: CBSS)[1-5]들에서 특정화자 음성복원의 실마리를 찾을 수 있다. 이는 CBSS가 반향 환경에서 독립적인 음원들을 효과적으로 분리할 뿐 아니라 활발한 연구를 통해 수많은 알고리즘들이 개발되었기 때문이다. 다만 CBSS의 암묵적 성질(blindness)에 의해, 음원에 대한 사전정보(prior information)를 음원분리에 활용하지 못하므로 논문의 최종 목표인 특정화자 음성복원에 바로 적용하기 어렵다. 이에 본 논문은 CBSS와 학습 가능한 특정화자의 음성으로부터 얻은 사전정보를 결합하여 특정화자의 음성만을 복원하는 시스템을 제시하고자 한다.

음원을 분리하는데 사전지식을 활용하여 성능을 향상하려는 여러 시도가 있어왔다[6-8]. Low와 Togneri 외 1인의 연구[8]에서는 누화간섭 잡음(babble noise)과 음성신호가 선형적으로 섞인 경우에, 음성신호를 분리해내는 방법을 제시하였다. 누화간섭 잡음에 비해 음성신호가 보다 큰 kutosis를 가진다는 사실에 착안하여, 분리된 독립음원에서 음성신호를 찾아냄으로써 교환 모호성(permutation ambiguity)을 해결하였다. 또한 Sawada와 Araki 외 2인의 연구[7]에서는 특정 음원이 우세한 파워를 가질 경우, time-frequency masking 방법을 이용하여 교환 모호성을 해결하였다. 그러나 앞서의 방법들은 음성신호의 일반적인 통계적 특성이나 음원이 섞이는 환경에 대한 지식을 일시적으로 이용할 뿐, 특정화자의 음성이 가지는 고유 정보를 이용하지 못한다. 이에 본 논문은 비음수 행렬분해(non-negative matrix factorization: NMF)의 여러 접근방법들을[9-14] 살펴보고 이 방법들을 이용하여 특정화자 음성의 주파수 특징을 잘 나타내는 기저벡터를 학습 한 후 이를 특정화자 음성복원에 적용하는 시스템을 개발하고자 한다.

본 논문에서 제안하는 특정화자 음성복원 시스템은 잡음과 반향이 있는 환경에서 스테레오 마이크로폰으로부터(2 입력채널) 음성 신호와 방해신호를 입력받아, 최종적으로 특정화자의 음성만을 결과로 출력한다(그림 1 참고). 본 제안 방법은 기존의 CBSS의 후처리 방법을 재시하여 성능을 향상시키고자 한 방법들과[15-17] 유사한 방법으로 문제에 접근한다. 즉, CBSS의 출력물을 가공하여 우리가 원하는 특정화자의 음성만을 복원하고자 한다. 다만 제안방법이 기존 접근방법들과 차별화 되는 점은 학습 가능한 특정화자의 음성으로부터 추출한 주파수특성 기저벡터들을 이용하여 특정화자 음성의 고유정보를 직접적으로 음성복원에 적용할 수 있다는 점이다. 본 제안방법은 기저벡터 학습단계와 다중경로 신호로부터 독립음원들을 분리하는 CBSS 단계, CBSS의 두 출력채널로부터 특정화자의 음성이 우세한 채널을 찾고(channel selection: 채널선택 단계) 이로부터 특정화자의 음성만을 복원해내는(reconstruction: 복원 단계) 두 후처리 단계로 나눌 수 있다. 실제 잡음과 반향 환경에서 특정화자의 음성을 복원하는 실험을 통하여 본 제안 방법의 우수성을 정량적으로 확인할 수 있다.

2. 문제 공식화

잡음과 반향이 존재하는 환경에서 t시간의 관측 신호 $\boldsymbol{x}_t=[x_{1,t},\ldots,x_{m,t}]^{\top}\in\mathrm{R}^{\mathrm{m}}$를 아래와 같이 공식화 한다:

$$\boldsymbol{x}_t=\sum_{\tau=0}^{P}\boldsymbol{A}_{\tau}\boldsymbol{s}_{t-\tau}+\boldsymbol{n}_t. \tag{1}$$

위 식에서 $\boldsymbol{s}_t=[s_{1,t},\ldots,s_{m,t}]^{\top}\in\mathrm{R}^{\mathrm{n}}$은 n개의 음원들이며 $\boldsymbol{A}_{\tau}$은 τ 지연 시간의 혼합행렬, P는 τ의 최대 값, 마지막으로 $\boldsymbol{n}_t$는 백색잡음이다. 본 논문에서는 잡음과 반향환경에서 특정화자와 방해음원을 스테레오 마이크로폰으로 입력받아(즉, $n=m=2$), 최종적으로 특정화자만의 깨끗한 음성을 복원하는 것을 목표로 한다.

본 제안방법 그림 1과 같이 세 단계로 나누어 생각할 수 있다: 아래의 목록은 각 단계를 그림과 대응하여 간단히 설명하고 있으며, 이를 통해 본 알고리즘의 전체 구성을 한눈에 알 수 있다.

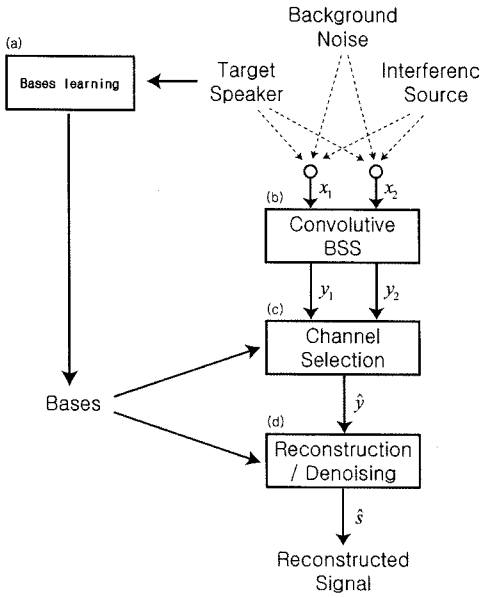

그림 1 본 제안 방법의 도식도

- ***기저벡터학습 단계***(그림 1(a)): 특정화자의 발화 음성을 이용하여 주파수 특성을 잘 나타내는 주파수특성 기저벡터를 학습한다. 3장에서는 NMF의 일종인 확률적 해석이 가능한 probabilistic latent variable model 기법들을[13,18,20] 이용하여 기저벡터를 학습한다.
- ***음원분리 단계***(그림 1(b)): 잡음과 반향 환경에서 신호를 입력받아 CBSS를 이용하여 중간 결과물로 y_1과 y_2을 출력한다. 여기서 우리는 y_1과 y_2 중 어느 채널이 특정화자의 음성에 해당한지 알지 못하며, 아직 잡음과 방해음원이 남아있기 때문에 이에 대한 후처리가 필요하다.
- ***후처리 단계***(그림 1(c),(d)): 채널선택단계(c)를 통하여 특정화자의 음성에 해당하는 채널을 선택하고, 복원단계(d)를 통하여 선택된 채널에 남아있는 잡음과 방해음원을 제거하고 특정화자만의 음성을 복원한다.

위의 모든 단계는 주파수 영역에서 실행되며 다음 장들은 각 단계에 대한 자세한 설명을 제시한다.

3. 주파수 특성 기저벡터 학습

본 제안 방법에서 중요한 요소 중 하나는 학습 가능한 특정화자의 발화음성을 통해 특정화자의 음성이 가지는 고유한 주파수 특성정보를 추출하는 것이다. 이를 위해 short-term Fourier transformation(STFT)를 통한 시간-주파수 표현을 고려한다. $\boldsymbol{M}(F\times T$ 행렬, F: 주파수 대역, T: 전체 프레임 수)을 특정화자 음성의 시간-주파수 표현이라 할 때, $\boldsymbol{M}$(비음수 2차원 행렬)에 NMF를 적용하여 $\boldsymbol{M}$을 두 행렬로 분해할 수 있다. 즉,

$$\boldsymbol{M}\approx\boldsymbol{UV}. \tag{2}$$

위 식에서, $\boldsymbol{U}$는 $F\times K$(K: 기저벡터의 수를 결정하는 인자)행렬로써 각 열(column)이 기저벡터에 해당하며, $\boldsymbol{V}$는 $K\times T$행렬로써 각 열이 인코딩에 해당한다. 즉, 식(2)에서 $\boldsymbol{U}$의 각 열은 시간의 변화에 불변한 주파수 영역의 기저벡터이므로 $\boldsymbol{U}$를 학습함으로써 특정화자의 음성이 가지는 주파수 특성정보를 쉽게 추출할 수 있다.

본 연구진은 기본 NMF 방법[9]외에도 다른 제한 조건들이 추가된 다양한 NMF 방법들을 고려해 보았다. 확률적인 해석이 가능한 probabilistic latent variable model 기법들[13,18,20], 모델 (2)에 직교제한조건(orthogonality)을 추가 한 orthogonal NMF[14], 희소성 제한조건(sparsity)을 추가한 sparse NMF[10], convoulutive NMF[11] 등을 고려하였으며, 각 방법론에 따라 서로 다른 성질의 기저벡터들이 학습된다. 그러나 본 논문에서는 CBSS의 후처리 단계들이 확률모델에 기반을 두어 동작하므로, 확률적 해석이 가능한 probabilistic latent variable model을 사용하고자 한다. 또한 probabilistic latent variable model은 entropic 사전확률을 [21] 활용하여 overcomplete representation ($K\gg F$)이 적용 가능하다. 이 방법은 entropic 사전확률을 이용하여 다른 NMF방법들 보다 더 많은 수의 기저벡터들을 학습할 수 있기 때문에, 특정화자 음성정보를 더 효과적으로 추출 할 수 있다. 본문에는 자세히 언급되지 않았지만 실험적으로도 overcomplete representation이 적용했을 때 그렇지 않은 경우보다 성능이 더 향상됨을 관찰했다. 이에 entropic 사전확률에 대한 자세한 설명 및 구현은 [18,20]을 참조할 수 있다.

4. 다중경로 암묵음원분리(Blind Source Separation of Convolutive Mixture)

다중경로 혼합신호를 분리하기 위해서 [22-25]에 소개된 CBSS를 사용한다. 여기서 다중경로 혼합신호는 다음과 같이 정의된다:

$$\boldsymbol{x}(t)=\sum_{\tau=0}^{P}\boldsymbol{A}(\tau)\boldsymbol{s}(t-\tau). \tag{3}$$

음원분리의 목표는 혼합신호로부터 원래의 음원들을 찾는 것이며, 이를 위해 아래의 식에서 정의된 혼합행렬의 역인 역혼합행렬(demixing matrix)$\{\mathbf{W}(\tau)\}$를 구한다:

$$\boldsymbol{s}(t)=\sum_{\tau=0}^{Q}\boldsymbol{W}(\tau)\boldsymbol{x}(t-\tau). \tag{4}$$

위 식에서 Q는 역혼합행렬의 최대 길이이다. 반향환경에서의 음원분리 문제는 주파수 영역에서 각 주파수

대에서의 곱으로 간단히 표현되기 때문에, 식 (3)을 주파수 영역으로 변환하면 문제해결이 용이해진다. 즉, f가 주파수일 때, 식 (3)은 각 주파수대의 곱으로 표현된다:

$$X(f,t)=\boldsymbol{A}(f,t)S(f,t).$$

그러므로 아래와 같은 관계를 이용하여($\hat{\boldsymbol{R}}_x$와 $\hat{\boldsymbol{R}}_s$은 각각 관측신호와 음원의 분산행렬)

$$\boldsymbol{W}(f,t)\hat{\boldsymbol{R}}_x(f,t)\boldsymbol{W}(f,t)^H=\hat{\boldsymbol{R}}_s(f,t)$$

최종적인 목적식을 유도할 수 있다. 각 음원들이 서로 통계적으로 독립이라는 점으로부터 $\hat{\boldsymbol{R}}_s$이 대각행렬이 된다는 것을 알 수 있다. 그러므로 $\hat{\boldsymbol{R}}_s$의 비 대각성분들을 0으로 만드는 역혼합행렬을 구하는 목적식 (5)을

$$\sum_t \frac{1}{2}\log\det\left(\boldsymbol{W}(f,t)\hat{\boldsymbol{R}}_x(f,t)\boldsymbol{W}(f,t)^H\right)-\log\det|\boldsymbol{W}(f)|. \quad (5)$$

얻을 수 있고, Pham의 연구에서[26] 제시된 joint approximate diagonalizaton을 통해 각 주파수 대역에서의 역혼합행렬들을 구할 수 있다. 그러나 교환모호성 문제가 남아 있는데, 이는 혼합행렬의 연속성(continuity)을 이용하여 풀 수 있고 최종적으로 반향 상황에서 다중경로 신호로부터 원래의 음원들을 분리하는 알고리즘을 얻을 수 있다.

5. 학습된 기저벡터를 이용한 CBSS의 후처리 기법들

CBSS는 각 음원들이 '서로 통계적으로 독립'이라는 정보만을 이용하여 음원들을 분리한다. 그러므로 CBSS만으로는 특정화자의 정보를 미리 가지고 있는 상황에서 이 추가적인 정보를 활용하여 CBSS의 분리성능을 향상시킬 수 없다. 이에, 본 논문에서는 CBSS의 후처리 단계를 제안함으로써 기존 CBSS의 성능을 향상시킴과 동시에 최종적으로 특정화자의 음성을 복원하고자 한다. 다음 5.1장은 확률적 분해 모델에 미리 학습한 기저벡터들을 적용하여 특정화자의 음성에 해당하는 채널을 선택하는 단계를 설명하고 있으며, 5.2장은 마찬가지로 앞서의 기저벡터들을 이용하여 특정화자만의 음성을 복원하는 단계를 설명하고 있다.

5.1 채널선택 단계(Channel Selection)

보다 효율적인 채널선택 판단 식을 유도하기 위해 확률적 해석이 가능한 모델을 도입한다. 아래의 식 (6)은 이를 위해 제시된 확률모델로써 [18]에서는 단일 마이크로폰에서 여러 음성을 분리하기 위해 사용되었다:

$$P_t(f)=\sum_s P_t(s)\sum_{z\in z_s}P_t(z|s)P_s(f|z). \quad (6)$$

위 식에서 $s\in\{s^*,s^\ddagger\}$(s^*=특정화자, $s^\ddagger$=방해음원)는 음원변수이고 $P_t(s)$는 각 음원에 대한 사전확률, $P_s(f|z)$는 각 음원에 해당하는 기저함수, $P_t(z|s)$는 각 음원이 주어졌을 경우의 시간에 따른 가중치이다. 식 (6)의 모델을 확률-그래프 모델로 나타내면 아래의 그림 2와 같다:

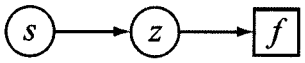

그림 2 식 (6)에 대한 확률 그래프 모델

이제 식 (6)의 모델을 CBSS의 출력채널에서 얻은 신호 y_1과 y_2의 시간-주파수 표현에(3장 참조) 각각 적용한다. 다만 여기서 $P_{s^*}(f|z)$은 3장에서 미리 학습해놓은 기저벡터들을 그대로 사용하기 때문에($P_{s^*}(f|z)=[U]_{f,z}$, $[A]_{f,z}$는 A행렬의 (f,z)번째 원소), 특정화자의 음성정보를 채널선택에 효과적으로 적용시킬 수 있다. CBSS를 통해 얻은 출력신호들은 특정화자의 음성과 방해 음원 및 잡음이 완벽히 분리가 안 되어 섞어 있기 때문에, (6)와 같은 음원변수에 대한 확률모델을 세움으로써 이를 확률적으로 기술하고, 이를 통해 채널선택의 판단기준을 마련하고자 한다.

우리가 식 (6)에서 구해야 하는 요소들은 $P_{s^*}(f|z)$를 제외한

$$\theta=\left[\{P_t(s),P_t(z|s)\}_{s\in\{s^*,s^\ddagger\}},P_{s^\ddagger}(f|z)\right]$$

이다. 이 값들은 잠재변수가 있을 경우, 유사도(likelihood)를 최대화하는 expectation and maximization (EM) 알고리즘으로 구할 수 있다. EM 알고리즘은 두 단계, E 단계와 M 단계로 나누어져 있으며 이 두 단계를 수렴 때까지 반복적으로 수행한다. 식 (6)에 대한 상세한 식은 아래와 같다[13,18,20]:

• E-step

$$P_t(s,z|f)=\frac{P_t(s)P_t(z|s)P_s(f|z)}{\sum_{s'}P_t(s')\sum_{z'\in z_s}P_t(z'|s')P_{s'}(f|z')}, \quad (7)$$

• M-step

$$P_t(s)=\frac{\sum_{z\in z_s}\sum_f P_t(s,z|f)\boldsymbol{M}_{f,t}}{\sum_{s'}\sum_{z\in z_s}\sum_f P_t(s,z|f)\boldsymbol{M}_{f,t}}, \quad (8\text{-}10)$$

$$P_t(z|s)=\frac{\sum_f P_t(s,z|f)\boldsymbol{M}_{f,t}}{\sum_{z'\in z_s}\sum_t P_t(s,z'|f)\boldsymbol{M}_{f,t}},$$

$$P_{s^\ddagger}(f|z)=\frac{\sum_t P_t(s^\ddagger,z|f)\boldsymbol{M}_{f,t}}{\sum_{f'}\sum_t P_t(s^\ddagger,z|f')\boldsymbol{M}_{f',t}}.$$

이제 채널선택에 대한 판단 기준식은 특정화자의 사전 확률을 이용하여 구해진다. 직관적으로 $P_t(s^*)$가 클수록 주어진 시간-주파수 표현에 특정화자(s^*)의 기여

도가 크다는 것을 알 수 있다. 그러므로 판단 기준식 $\hat{P}(s^*)$은 아래와 같이 정의된다:

$$\hat{P}(s^*)=\frac{1}{T}\sum_{t=1}^{T}P_t(s^*). \tag{11}$$

즉, 두 시간-주파수 표현 $\boldsymbol{M}^{(1)}$과 $\boldsymbol{M}^{(2)}$에 앞서의 확률모델을 적용하여 θ들을 구하고 최종적으로 식 (11)을 서로 비교함으로써 우리는 특정화자의 음성이 보다 우세하게 포함되어 있는 채널을 선택 할 수 있다:

$$\{\hat{\boldsymbol{M}},\hat{\theta}\}=\begin{cases}\{\boldsymbol{M}^{(1)},\theta^{(1)}\} & \text{if } \hat{P}^{(1)}(s^*)\geq\hat{P}^{(2)}(s^*),\\ \{\boldsymbol{M}^{(2)},\theta^{(2)}\} & otherwise\end{cases} \tag{12}$$

5.2 복원 단계(Reconstruction)

앞서 구한 $\{\hat{\mathbf{M}},\hat{\theta}\}$를 이용하여 최종적으로 특정화자의 음성을 복원하고자 한다. 이를 위해 먼저 특정화자의 음성만을 담고 있는 파워 스펙트로그램 $\tilde{\boldsymbol{M}}$을 복원한다. 이를 위해 $\hat{\boldsymbol{M}}$의 (f,t)의 원소 $\hat{\boldsymbol{M}}_{f,t}$는 다음과 같고:

$$\hat{\boldsymbol{M}}_{f,t}=\sum_{s\in\{s^*,s^\ddagger\}}\hat{\boldsymbol{M}}_{f,t}(s), \tag{13}$$

$\hat{\boldsymbol{M}}_{t,f}(s)$는 s 음원에서 특정 확률분포로(식 (6) 참고) 뽑힌 수라는 점을 이용한다[13]. 그러므로 특정화자의 파워 스펙트로그램 $\widetilde{\boldsymbol{M}_{f,t}}(s^*)$는 $\{s^*,s^\ddagger\}$가 이루는 이항 분포에서 총 뽑힌 횟수 $\hat{\boldsymbol{M}}_{f,t}$가 주어졌을 때의 평균으로 정의될 수 있다[13]:

$$\widetilde{\boldsymbol{M}_{f,t}}(s^*)=\frac{\hat{P}^*_{t(s)}\hat{P}_t(f|s^*)}{\sum_{s'}\hat{P}_t(s')\widehat{P_{t(f|s')}}}\hat{\boldsymbol{M}}_{f,t} \tag{14}$$

여기서

$$\hat{P}_t(f|s^*)=\sum_{z\in z_{s^*}}\hat{P}_t(z|s^*)P_s(f|z). \tag{15}$$

앞서 구한 시간-주파수 표현 $\tilde{\boldsymbol{M}}$과 해당 채널의 신호에서 얻은 위상정보를 이용, 이를 역 STFT를 통하여 최종적으로 특정화자의 음성신호만을 복원한다.

6. 실험 결과

제안 방법의 성능을 평가하기 위해, 실제 잡음과 반향이 있는 환경에서 잡음과 반향 제거정도를 평가한다. 직접 녹음을 할 경우 관측 신호만을 얻을 수 있어, 본래의 음성신호와 제안방법의 복원결과와의 차이를 분석할 수 없다. 그러므로 정량적인 분석을 위해 roomsim[27]을 이용하여 실제 잡음과 반향 상황을 시뮬레이션하고, 이 가상 환경에서 제안방법의 성능을 평가하는 방식을 취한다(이 경우 원 음성신호와 제안방법이 최종적으로 복원한 음성신호와의 비교가 용이하다). 그림 3에 제시 되었듯이, 방의 크기는 6.75m × 3.75m × 2.5m(가로 × 세로 × 높이)이며 반향 정도를 나타내는 충격함수 반응은 그림 4와 같다. 또한 그림 3에서 s1과 s2는 각 음원이 자리 할 수 있는 위치를 나타내는데, ①은 s1이 s2 보다 가까운 곳에 위치함을(d1), ②는 같은 거리(eq), ③은 s1이 s2보다 먼 위치에 있음을(d2) 의미한다. 특정화자가 s1에 위치하게 되는데 ①~③의 위치로 자리를 바꾸어 가면서 녹음을 해, 다양한 위치에서의 제안방법의 성능 평가를 하게 된다. 본 실험에서는 특정화자 s1은 여성화자 또는 남성화자의 음성이고, 방해음원 s2는 프린터 출력 소리 또는 트럼펫 연주소리로 가정한다.

주파수 특성 기저벡터 학습을 위해 각 30초 길이의 여성 또는 남성의 음성신호를 사용한다. 이는 뒤의 분리 실험에 사용되는 음성신호와는 별개의 학습 데이타이다. 이 여성 및 남성의 음성신호를 STFT(윈도우 크기: 1024, hop 크기: 256, hamming window 사용) 변환을 통하여 시간-주파수 표현 $\boldsymbol{M}$을 얻는다. 3장에서 소개한 대로 $\boldsymbol{M}$을 비음수 분해하기 위해 probabilistic latent variable model을 적용하고, 이때 K를 1000($\gg F=512$)으로 하여 over-representation이 가능토록 하였다. 이 밖에도 entropic 사전확률을 정의할 때 요구되는 사용자 계수들이 있는데, 이는 Shashanka와 Smaragdis 연구[18]에서 제시된 수치를 따른다. 이를 통해 얻어진 여성과 남성의 주파수 특성 기저벡터들을 그림 5,6에서 확인할 수 있다.

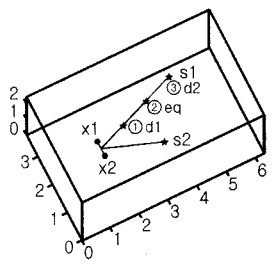

그림 3 실험 방의 모형도

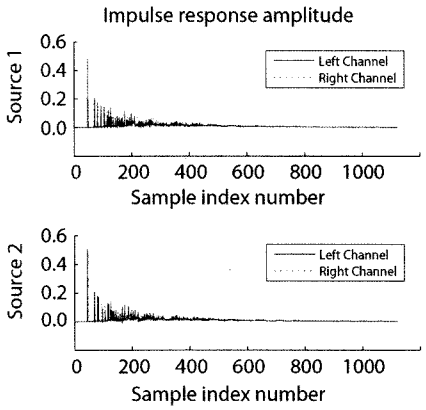

그림 4 실험 방의 충격함수(impulse) 반응

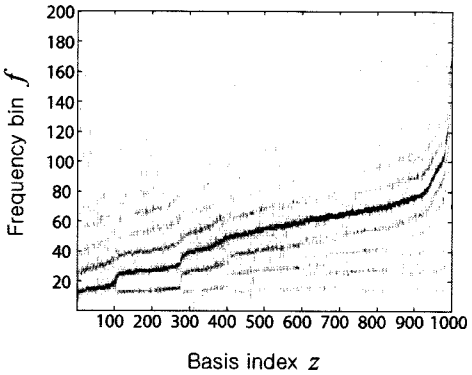

그림 5 남성의 학습발화음성으로부터 얻은 기저벡터들을 정렬한 모습

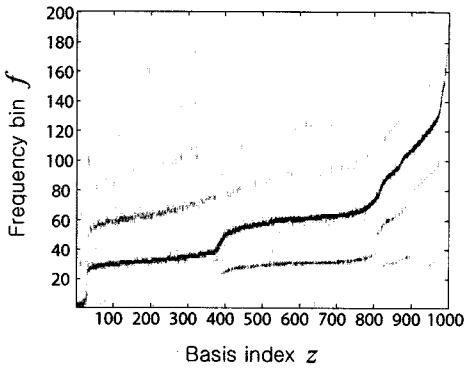

그림 6 여성의 학습발화음성으로부터 얻은 기저벡터들을 정렬한 모습

복원결과의 정량적 평가를 위하여 source to distortion ration(SDR)[28]을 도입한다. SDR은 잡음이 있는 환경에서 음원분리 방법들의 품질을 평가하는 측정단위로써, 그 값이 클수록 알고리즘이 출력한 결과신호가 목표 신호에 근접하다는 것을 의미한다. 이 계산을 위해, 주어진 신호 $\hat{s}_t$를 아래의 분리모델 [28]을 이용하여 분리한다.

$$\hat{s}(t) = s_{\text{target}}(t) + s_{\text{interf}}(t) + s_{\text{noise}}(t) + s_{\text{artif}}(t).$$

위 분리모델을 바탕으로 SDR을 다음과 같다[28]:

$$SDR = 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|s_{\text{interf}} + s_{\text{noise}} + s_{\text{artif}}\|^2}. \quad (16)$$

본 방법의 성능을 평가하기 위해 각 단계별로 SDR를 측정하고 이의 증가폭을 기존 CBSS와 비교하고자 한다. 즉, 그림 1에서와 같이 크게 세 부분에서 매 실험마다 SDR을 측정한다: (ㄱ) 마이크로폰으로 얻어진 관측신호의 SDR값($\text{SDR}_x = \max(SDR_{x_1}, SDR_{x_2})$, 두 입력 중 SDR이 높은 쪽 선택); (ㄴ) 음원분리만을 통해 얻은 결과의 SDR값 (SDR_y); (ㄷ) 본 방법의 복원 단계를 거친 최종 결과신호의 SDR값($\text{SDR}_{\hat{s}}$). 여기서 주지하여야 할 사실은 SDR_y는 CBSS의 두 출력 채널 중 채널선택 단계가 특정화자의 음성에 해당하는 신호로 선택한 채널의 SDR값이라는 점이다. 즉, 본 방법은 채널선택 단계가 1차적으로 CBSS의 교환 모호성 문제를 해결하여 SDR값을 향상시키고, 복원 단계를 통해 2차적으로 SDR값을 향상시킨다. 표 1과 표 2는 각각 여성과 남성이 특정화자일 경우, 여러 상황에서 본 방법의 복원결과를 평가한 것이다. 특히 SDR 값의 증가분을 통해(각 표의 마지막 줄), 본 방법이 반향과 잡음에 의해서 생기는 왜곡을(식 (16) 참고)을 상당 부분 제거하고, CBSS의 분리 성능을 큰 폭으로 다시 향상시키는 것을 알 수 있다. 즉, 본 제안방법이 잡음과 반향 환경에서 잡음 및 반향을 제거하고 본 특정화자의 음성을 복원하는데 적합함을 입증한다.

표 1 여성 화자인 경우의 각 단계별 결과(SDR): pr(프린터 잡음)이나 tr(트럼펫 음악)이 방해음원일 경우 여러 가지 조합의(예, eq: 특정화자와 방해음원이 마이크에 대해 비슷한 위치에 존재함, 그림 3 참조) 위치에 있을 경우 각 단계별(그림 1 참고) SDR 값 테이블

위치	SDR_x (그림 1 참고)	SDR_y	$\text{SDR}_{\hat{s}}$
f-pr-eq	0.53	12.55	16.76
f-pr-d1	3.39	5.08	17.50
f-pr-d2	-1.38	9.14	13.80
f-tr-eq	0.36	3.89	16.66
f-tr-d1	2.83	10.30	18.39
f-tr-d2	-1.93	-0.94	6.78
평균	0.63	6.67	14.98
증가	-	6.04	8.31

표 2 남성 화자인 경우의 각 단계별 결과(SDR): pr(프린터 잡음)이나 tr(트럼펫 음악)이 방해음원일 경우 여러 가지 조합의 위치에 있을 경우 각 단계별 (그림 1 참고) SDR 값 테이블

위치	SDR_x	SDR_y	$\text{SDR}_{\hat{s}}$
m-pr-eq	-1.59	1.08	5.68
m-pr-d1	1.09	2.17	6.85
m-pr-d2	-3.16	0.85	4.08
m-tr-eq	-1.13	-3.14	2.87
m-tr-d1	1.18	1.73	7.77
m-tr-d2	-3.32	-3.81	2.85
평균	-1.16	-0.19	5.02
증가	-	0.97	5.20

7. 결 론

본 논문은 잡음과 반향이 있는 실 환경에서의 특정화자음성을 복원하는 시스템을 제안하였다. NMF 방법 중 하나인 probabilistic latent variable model를 주어진 특정화자의 발화음성에 적용하여 주파수 특성을 잘 표현하는 기저벡터들을 학습하였다. 이와 같이 얻어진 기저벡터들을 CBSS의 후처리 부분에 유기적으로 이용함으로써 특정화자의 음성만을 복원할 수 있었다. 즉, CBSS의 두 출력 채널로부터 특정화자의 음성에 해당하는 채널을 선택하고(채널선택 단계), 미리 학습한 기저벡터들을 활용하여 선택된 채널의 신호로부터 남아 있는 잡음과 방해음원을 제거, 최종적으로 특정화자의 음성만을 복원할 수 있었다(복원 단계).

본 제안방법은 암묵적 속성을 지니는 CBSS에 효율적으로 특정음원의 사전정보를 적용시키는 방법을 제시하였다. 또한 잡음 및 반향환경에서의 실험을 통하여 본 제안 방법이 특정화자의 음성에 가까운 음성신호를 복원함을 확인 할 수 있었다. 다만 본 방법이 신뢰성 있는 복원성능을 보이기 위해서는 특정화자의 주파수 특성이 방해음원의 그것과는 달라야 한다는 한계점이 있다. 즉, 아직 본 방법은 2명의 여성화자나 2명의 남성화자 간의 대화에서 특정화자의 음성만을 선택하여 복원하지 못한다. 이는 주파수 특성이 비슷할 경우 채널 선택단계에서 주파수 특성 기저벡터에 의한 선별이 제대로 이루어지지 않기 때문이다. 이처럼 주파수 특성이 비슷할 경우에 음원분리나 복원 문제는 음원분리 연구영역에서도 해결되지 않은 과제로써 현재 이에 대한 연구가 진행 중이다.

참 고 문 헌

[1] L. C. Parra, C. Spence, "Convolutive blind source separation of non-stationary sources," IEEE Trans. Speech and Audio Processing 320-327, 2000.

[2] D. Pham, C. Serviere, H. Boumaraf, "Blind separation of convolutive audio mixtures using nonstationarity," in: Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation, pp. 107-110, 2003.

[3] K. Torkkola, "Blind separation of convolved sources based on information maximization," in: Proceedings of the IEEE Workshop on Neural Networks for Signal Processing, pp. 423-432, 1996.

[4] S. Amari, S. C. Douglas, A. Cichocki, H. H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," in: Proceedings of the IEEE International Conference on Signal Processing Advances in Wireless Communications, Paris, France, pp. 101-104, 1997.

[5] P. Smaragdis, "Information-theoretic approaches to source separation," Master's thesis, Massachusetts Institute of Technology, 1997.

[6] extraction from interferences in real environment using bank of lters and blind source separation, in:Proceedings Third AustralianWorkshop on Signal Processing and Applications, 2000.

[7] H. Sawada, S. Araki, R. Mukai, S. Makino, "Blind extraction of a dominant source from mixtures of many sources using ica and time-frequency masking," in: Proceedings of the IEEE International Symposium on Circuits and Systems, pp. 5882-5885, 2007.

[8] S. Y. Low, R. Togneri, S. Nordholm, "Spatio-temporal processing for distant speech recognition," in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004.

[9] D. D. Lee, H. S. Seung, "Algorithms for non-negative matrix factorization," in: Advances in Neural Information Processing Systems, Vol. 13, MIT Press, 2001.

[10] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," Journal of Machine Learning Research 5, 1457-1469, 2004.

[11] P. D. O. Grady, B. A. Pearlmutter, "Convolutive non-negative matrix factorisation with sparseness constraint," in: Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing, 2006.

[12] T. Hofmann, "Probablistic latent semantic indexing," in: Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval, 1999.

[13] B. Raj, P. Smaragdis, "Latent variable decomposition of spectrograms for single channel speaker separation," in: IEEE Workshop of Applications of Signal Processing to Audio and Acoustics, pp. 17-20, 2005.

[14] Jiho Yoo and Seungjin Choi (2008), "Orthogonal nonnegative matrix factorization: Multiplicative updates on Stiefel manifolds," in Proceedings of the 9th International Conference on Intelligent Data Engineering and Automated Learning, IDEAL-2008.

[15] E. Visser, M. Otsuka, T.-W. Lee, "A spatio-temporal speech enhancement scheme for robust speech recognition in noisy environments," Speech Communication 41(15), 393-407, 2003.

[16] C. Choi, G. Jang, Y. Lee, S. R. Kim, "Adaptive cross-channel interference cancellation on blind source separation outputs," in: Proceedings of International Conference on Independent Component Analysis and Blind Signal Separation, 2004.

[17] J. Kocinski, "Speech intelligibility improvement using convolutive blind source separation assisted by denoising algorithms," Speech Communication 50, 29-37, 2008.

[18] M. V. S. Shashanka, P. Smaragdis, "Sparse overcomplete decomposition for single channel speaker separation," in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 641-644, 2007.
[19] P. Smaragdis, B. Raj, M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in: Proceedings of International Conference on Independent Component Analysis and Signal Separation, 2007.
[20] M. V. S. Shashanka, "Latent variable framework for modeling and separating single channel acoustic sources," Ph.D. thesis, Department of Cognitive and Neural Systems, Boston University, 2007.
[21] M. Brand, "Structure learning in conditional probability models via an entropic prior and parameter extinction, Neural Computation," 11(5), 1155-1182, 1999.
[22] J. F. Cardoso, A. "Souloumiac, Blind beamforming for non Gaussian signals," IEE Proceedings-F 140 (6), 362-370, 1993.
[23] A. Belouchrani, K. Abed-Merain, J. F. Cardoso, E. Moulines, "A blind source separation technique using second order statistics," IEEE Trans. Signal Processing 45, 434-444, 1997.
[24] S. Choi, A. Cichocki, A. Belouchrani, "Blind separation of second-order nonstationary and temporally colored sources," in: Proceedings of IEEE Workshop on Statistical Signal Processing, Singapore, pp. 444-447, 2001.
[25] A. Ziehe, P. Laskov, G. Nolte, K. R. Muller, "A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation," Journal of Machine Learning Research 5, 777-800, 2004.
[26] D. T. Pham, "Joint approximate diagonalization of positive denite matrices," 22(4), 1163-1152, 2001.
[27] D. R. Campbell, K. J. Palomaki, G. J. Brown, A "matlab simulation of shoebox room acoustics for use in research and teaching," Computing and Information Systems Journal 9(3), 1352-1404, 2005.
[28] E. Vincent, C. Fevotte, R. Gribonval, "Performance measurement in blind audio source separation," IEEE Trans. on Audio, Speech and Language Processing 14(4), 1462-1469, 2006.

박 선 호
2004년 고려대학교 전자공학부 학사. 2006년 포항공과대학교 컴퓨터공학과 석사 2006년~현재 포항공과대학교 컴퓨터공학과 박사과정. 관심분야는 기계학습 - kernel, Gaussian process 등

유 지 호
2006년 포항공과대학교 컴퓨터공학과 학사. 2006년~현재 포항공과대학교 컴퓨터공학과 통합과정. 관심분야는 Document Clustering with Nonnegative Matrix Factorization, Music Transcription

최 승 진
1987년 서울대학교 전기공학과 학사. 1989년 서울대학교 전기공학과 석사. 1996년 University of Notre Dame 전기공학과 박사. 1997년 일본 RIKEN Frontier Researcher. 1997년~2001년 충북대학교 전기전자공학부 교수. 2001년~현재 포항공과대학교 컴퓨터공학과교수. 관심분야는 통계적 기계학습, 확률 그래프 모델