http://dx.doi.org/10.7840/kics.2015.40.4.619

잡음 데이터를 활용한 음성 기저 행렬과 NMF 기반 음성 향상 기법

권 기 수◆, 김 형 용*, 김 남 수°

Speech Basis Matrix Using Noise Data and NMF-Based Speech Enhancement Scheme

Kisoo Kwon◆, Hyung Young Kim*, Nam Soo Kim°

요 약

본 논문은 비음수 행렬 인수분해(NMF)를 이용한 음성향상 기법을 다루고 있다. 음성과 잡음에서 적절한 훈련을 통해 각각의 기저(basis) 행렬을 구하고 이 행렬들을 이용하여 두 음원을 분리 하는 것이다. 그 중에서도 음성 향상의 성능은 사용하게 되는 기저 행렬에 따라 크게 달라짐을 보인다. 기존의 독립적으로 구한 음성 기저 행렬에 비해서, 잡음 데이터를 복원하는데 부적합한 방향으로 최적화시킨 음성 기저 행렬을 사용하였을 때 더 높은 음성 향상 성능을 보임을 실험으로 확인하였다. 이 때 잡음 데이터의 복원 오차 자체를 크게 해주는 방향과 해당 인코딩 행렬(encoding matrix) 원소의 값을 작게 해주는 두 가지 방법을 적용하여 비교하였다. 좀 더 음성 복원에만 특화된 기저 행렬을 구함으로서 음성 기저 행렬이 잡음 데이터 복원에 사용되는 것을 최소화 하였다. 실험 결과에서는 perceptual evaluation speech quality값과 signal to distortion ratio를 지표로 사용하였고, 기존 기법에서 사용하는 기저 행렬 보다 더 높은 성능을 보임을 확인 하였다.

Key Words : communication, signal processing, Neutral systems, Communication Sciences, Network

ABSTRACT

This paper presents a speech enhancement method using non-negative matrix factorization (NMF). In the training phase, each basis matrix of source signal is obtained from a proper database, and these basis matrices are utilized for the source separation. In this case, the performance of speech enhancement relies heavily on the basis matrix. The proposed method for which speech basis matrix is made a high reconstruction error for noise signal shows a better performance than the standard NMF which basis matrix is trained independently. For comparison, we propose another method, and evaluate one of previous method. In the experiment result, the performance is evaluated by perceptual evaluation speech quality and signal to distortion ratio, and the proposed method outperformed the other methods.

※ 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2012R1A2A2A01045874).
※ 이 논문은 2014년도 대검찰청 지원을 받아 수행된 연구임.
◆ First Author : Department of Electrical and Computer Engineering and the Institute of New Media and Communications, Seoul National University, kskwon@hi.snu.ac.kr, 학생회원
° Corresponding Author : Department of Electrical and Computer Engineering and the Institute of New Media and Communications, Seoul National University, nkim@snu.ac.kr, 종신회원
* 서울대학교 전기·정보공학부 및 뉴미디어통신공동연구소, hykim@hi.snu.ac.kr
논문번호 : KICS2014-12-478, Received December 4, 2014; Revised March 13, 2015; Accepted April 15, 2015

Ⅰ. 서 론

1980년대부터 통계모델 기반 음성 향상 기법이 제안되고 활발히 연구되었다[1-3]. 이러한 통계모델 기반 기법의 경우 음성과 잡음을 각기 다른 하나의 통계 모델로 근사 시켜 음성 향상 방법에 접근을 한다. 즉 음성은 상대적으로 크고 빠르게 변하며, 반대로 잡음의 경우에는 변화의 폭이 작다고 가정을 하고 통계모델을 사용하게 된다. 이러한 통계모델을 업데이트 해주기 위한 목적 등으로 통계모델 기반 음성 향상의 경우에는 대부분 음성 활성화 구간 검출(voice activity detection)이라는 기법을 동시에 사용하고 있고 그 성능에 크게 의존하고 있다. 이러한 기법은 계산량이 상대적으로 적고 또한 음성 또는 잡음에 대해 특별한 사전 정보가 필요 없다는 큰 장점이 있다. 하지만 애초에 잡음에 대한 가정 때문에 잡음의 소리 크기 변화가 큰 경우에는 성능 저하가 뚜렷하다는 단점이 있다.

1990년대 후반부터 제안되고 연구가 활발히 이루어진 분야는 템플릿 기반 음성 향상이다[4-8]. 이 기법의 경우에는 사전 정보가 매우 중요하다. 음성 또는 잡음의 사전정보가 없는 경우 높은 성능을 기대할 수 없다. 하지만 사전 정보를 이용하기 때문에 잡음의 크기 변화가 큰 경우에도 통계모델 기반 음성 향상에 비해 높은 음성 향상 성능을 보인다. 그 중에 최근 많이 연구 되고 있는 분야는 비음수 행렬 인수분해(nonnegative matrix factorization, NMF)와 사전 학습(dictionary learning)이다[9,10]. NMF의 경우 데이터 차수를 줄이는 알고리즘이며, 부분 기반의 방법으로 설명할 수 있다. 이러한 NMF를 수행하기 위해 여러 가지 최적화 기법 등이 제안되었고 각기 단일 음원 복원에서 좋은 성능을 보이고 있다.

이러한 NMF는 복원하고자 하는 데이터가 가진 랭크(rank)문제로 낮은 개수의 기저(basis)에서는 어느 정도 크기의 복원 오차(reconstruction error)를 보일 수밖에 없다. 또한 음원에 사용할 경우 음원이 한 시간 프레임에서 가지는 정보의 한계로 인해 각 음원을 표현하는 기저들이 유사성을 보인다. 이를 해결하고자 orthogonal NMF, discriminative NMF, convolutive NMF, group sparsity NMF 등이 제안되었다[11-14]. 이 중 주로 데이터의 차수를 늘려서 데이터 간의 차별성을 두려는 기법의 경우 어느 정도의 성능향상을 보이지만, 계산량의 문제와 단체화(grouping)하게 되는 시간 프레임의 정확한 길이를 알기 어렵기 때문에 실제 음성향상 적용에는 어려운 점이 있다. 또한 group sparstiy NMF의 경우 그룹(gropu)의 개수가 적을 경우에는 효과적일 수 있지만, 그룹의 개수가 많아질 경우 효과가 떨어지거나 성능 저하를 불러일으킬 수 있다.

본 논문에서는 각 음원의 기저가 다른 음원을 복원하는데 사용되지 않도록 기저의 모양을 최대한 조절해주는 알고리즘을 제안한다. 음성 향상에 사용하기 위해서, 음성 기저행렬을 구할 때 해당 기저 행렬이 특정 잡음 데이터를 복원하는데 부적절하도록 NMF 목적 함수에 특정한 제한 조건을 추가하였다. 실험 부분에서 분석을 한 결과 제안된 방법은 기존 NMF에 비해서 2배 정도의 높은 음성 향상 성능을 보였다.

본론에서는 기본적인 NMF 알고리즘의 설명과 NMF 기반 음성 향상 기법에 대해 설명을 하고, 제안하고자 하는 알고리즘에 대한 설명을 하겠다. 그 후 실험 부분에서 기존의 독립적으로 구한 기저를 사용하는 NMF 기반 음성 향상과 제안한 두 가지 알고리즘의 기저 행렬을 이용한 음성 향상 성능을 PESQ (perceptual evaluation speech quality)[15]와 SDR (signal to distortion ratio)[16]로 분석하겠다.

Ⅱ. 비음수 행렬 기반 음성 향상

2.1 비음수 행렬 인수분해

NMF는 앞선 서론에서 설명 했듯이, 여러 샘플이 모인 데이터 세트를 특정 몇 개의 기저들로 근사하는 것이다. 당연히 이 데이터 세트는 같은 종류로 지칭되는 샘플들의 집단이어야 한다. NMF는 일반적으로 데이터의 차수를 낮추는 역할을 한다. 즉 일반적인 상황에서는 데이터의 차수 보다 기저의 개수를 작게 정해서 데이터를 표현하는 차수가 작아지는 결과를 보여 준다. 당연히 반대로 기저의 개수를 데이터의 차수보다 크게 정하여 비음수 행렬 분해를 수행할 수는 있지만, 이러한 환경에서는 NMF가 효율적으로 동작하지 않을 수 있다. 우선 일반적인 NMF 알고리즘에 대해 알아보자. NMF는 아래의 수식을 목표로 한다.

$$V \approx WH \tag{1}$$

W 는 기저 행렬을, H 는 부호화 행렬을 나타낸다. V는 W의 기저들의 선형 합(linear combination)으로 구성되어 있다. V 는 $(n \times m)$ 크기, W 는 $(n \times r)$, H 는 $(r \times m)$ 크기의 행렬이다. n 은 주파수 축을 나타내는 데이터의 개수, r 은 기저의 개수, m 은 시간 프레임의 개수를 의미한다. 이를 위해 [9]에서는 특정 거리함수를 목적 함수로 정하고 이를 경

사 하강법(gradient descent)으로 W 와 H 에 대해 최적화를 하였다. (1)의 수식을 만족하는 W 와 H 를 구하기 위해서 일반적으로 유클리디언 거리 함수(euclidean distance function)와 Kull-back Leibler divergence (KL 발산) 거리 함수 등을 택하고 있다. 이러한 거리 함수를 정하게 되면 multiplicative 방법을 이용하여 W 와 H 를 최적화 하게 된다. 우선 KL 발산을 이용한 최적화 수식을 [17]을 참고하여 알아보겠다. 우리가 목적으로 하는 W 와 H 를 구하기 위해서 목적함수를 정하면 아래와 같다.

$$D(V \| WH) = \sum_{i,j}(V_{ij}\log\frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}) \quad (2)$$

이 목적 함수의 값이 최소가 되는 경우의 W 와 H 가 구하고자 하는 결과이다. 이 목적 함수의 경우에는 구하고자 하는 변수 또는 행렬이 W 와 H 두 개로 convexity를 만족하지 않기 때문에 최적화 수식을 한 번 수행해서는 답을 구할 수 없다. 이를 위해 반복적인 과정을 통해서 W 와 H 의 최적화를 수행하게 된다. 즉, 한 번은 V 와 W 를 고정 상태로 보고 H 에 대해서만 최적화를 수행하고, 다음 작업에서는 반대로 V 와 H 를 고정 상태로 보고 W 에 대해서 최적화를 수행하게 된다. 이러한 두 번의 수행 과정을 하나의 작업으로 보고, 이 작업을 반복적으로 수행 하게 된다. 또한 이 때 이 반복적인 최적화 수행 과정을 언제 마치느냐가 중요한 문제가 된다. 일반적으로 오차의 제곱에 기반한 방법 등을 사용하기도 하지만, 그 오차의 문턱값을 정하는 것 또한 문제가 되기도 한다[9]. 그렇기 때문에 이 논문에서 수행한 NMF 과정에서는 반복 횟수를 특정 값으로 정하였다.

우선 H 를 구하는 과정에 대해서 설명하겠다. 이 논문에서는 KL 발산을 이용하여 수식을 전개하고 실험하였다. (2) 식을 항이 두 개가 되도록 변형을 하면 아래와 같다.

$$D(V \| WH) = \sum_{i,j}(\sum_{k}^{r} W_{ik}H_{kj} - V_{ij}\ln\sum_{k}^{r} W_{ik}H_{kj}) \quad (3)$$

이 목적 함수를 시작으로 하여 일종의 경사 하강법인 multiplicative 방법을 사용하여 각 변수의 업데이트 수식을 구하면 아래와 같다.

$$H_{ab} \leftarrow H_{ab}\frac{\sum_{i=1}^{n}(W_{ia}V_{ib})/\sum_{k=1}^{r}W_{ik}H_{jb}}{\sum_{i=1}^{n}W_{ia}} \quad (4)$$

$$W_{cd} \leftarrow W_{cd}\frac{\sum_{j=1}^{m}(H_{dj}V_{cj})/\sum_{k=1}^{r}W_{ck}H_{kj}}{\sum_{j=1}^{m}H_{dj}} \quad (5)$$

위의 두 식, (4)과 (5)을 적용하여 업데이트 하는 것을 한 번의 과정으로 보고, W 와 H 가 수렴할 때까지 한 과정을 반복해 준다. 하지만 고정 소수점 방법의 특성상 최적화식을 매 번 사용할 때마다 분모 부분이 0 이 되지 않도록 최소값을 정해주어야 한다[9]. 이 과정에서 각 원소의 최소값을 0 보다 크게 정해주는 것 이외에 중요한 작업은 각 기저의 $L2-norm$이 1이 되게 정규화 해주는 것이다. 일반적인 NMF 기반 음성 향상의 경우에는 이러한 작업이 없어도 크게 문제가 되지 않지만, H 에 어떠한 별도의 알고리즘 및 작업을 추가하는 경우 문제가 될 수 있다. 각 기저의 파워가 1이 되게 해줌으로써 H 의 각 원소는 해당 프레임에서의 해당 기저가 쓰인 정도를 의미하게 된다.

이러한 방법 외에도 특정 증분량을 사용하는 방법 등과 성능 향상을 위해 최적화 하는 목적 함수에 sparseness와 관련된 부분을 추가해주는 여러 연구들이 있다[16]. 또한 multiplicative 방식은 아직 완벽하게 수렴 증명이 되지 않았기 때문에 projected gradient descent 기법을 이용하여 W 와 H 를 적용하기도 한다[18]. 본 논문에서는 비교적 구현이 쉽고, 실제 논문들에서 많이 구현이 된 KL 발산 기반의 multiplicative 방법을 이용하여 NMF 과정을 수행하였다[17].

2.2 향상 과정

위의 설명한 NMF는 단순히 한 종류의 음원만을 복원할 때 적절하다고 생각되기 싶다. 하지만 둘 이상의 음원이 동시에 존재하는 경우에도 적절히 사용할 수가 있다. 각 음원의 기저 행렬을 적절히 사전에 구해 놓는다면, 각 기저행렬로 부터 복원되는 결과가 해당 음원 종류의 추정값이 될 것이다. 즉 향상 과정을 수행하기 앞서서 필수적으로 수행되어야 하는 것은 훈련 과정이다. 적절하고 충분한 각 음원의 훈련용 데이터베이스가 준비가 되어야 하고, 이를 통해 각 음원의 기저 행렬을 향상 과정 전에 완성해야 한다. 훈련

에 앞서 우리가 사용하고자 하는 신호의 데이터는 비음수를 만족해야한다. 그래서 소리 신호를 Short Time Fourier Transform (STFT)을 통해 각 시간 프레임에서 주파수 별로 나타내고 이를 비음수 값으로 사용하기 위해 절대값으로 바꿔 준다. 훈련을 통해 얻은 음성 기저 행렬을 W_S, 잡음 기저 행렬을 W_N 이라고 하자. 즉 각 기저의 차수는 FFT 크기에 따라 정해진다.

훈련 과정을 통해 각 음원에 적절한 기저 행렬을 구한 후 향상 과정을 진행하게 된다. 전체적인 향상 과정은 그림-1. 에 표현되어 있다. 이는 [7] 논문에서 좀 더 자세히 설명되어 있다. 향상 과정에서도 마찬가지로 잡음이 섞인 신호를 STFT를 통해 복소수 값으로 변환 시킨다. 이렇게 변환 된 잡음이 섞인 신호를 Y 라고 하자. 음성 신호, S 와 잡음 신호 N 이 독립적이란 가정을 하면 Y 는 $S+N$ 과 같이 생각 할 수 있다. Y 의 절대값 V 는 훈련 과정에서 구한 기저 행렬을 통해 아래와 같이 나타 낼 수 있다.

$$V \approx [W_S, W_N]H \tag{6}$$

이 때 V 는 $(n \times m)$ 행렬이고, H 는 $(2r \times m)$ 이 된다. 그리고 각각의 기저 행렬은 $(n \times r)$이다. $[W_S, W_N]$은 $(n \times 2r)$ 크기의 행렬이 된다. 당연히 상황에 맞게 음성과 잡음 기저의 개수는 달라도 된다. 위 수식을 시간 t 에 따른 수식으로 보면 다음과 같다.

$$V_t \approx [W_S, W_N]H_t = [W_S, W_N][H_{S,t};H_{N,t}] \tag{7}$$

로 나타 낼 수 있다. 이 때 각각의 $H_{S,t}$ 와 $H_{N,t}$ 은 $(r \times 1)$ 행렬이고, $H_t = [H_{S,t};H_{N,t}]$은 $(2r \times 1)$ 행렬이 된다. 이렇게 한 프레임에서 데이터의 추정 과정을 마치면 아래와 같이 음성과 잡음 크기의 추정값이 나오게 된다.

$$\hat{s_t} = W_S H_{S,t}$$
$$\hat{n_t} = W_N H_{N,t} \tag{8}$$

이 때 매 프레임에서 부호화 행렬, H_t 의 초기값을 무작위로 준다. 이렇게 얻은 음성과 잡음의 추정 크기에는 어느 정도의 오차가 존재한다. 이를 보완하고자 일반적으로 위너 필터(Wiener filter) 형태의 이득 함수(gain function)을 사용한다[4]. 즉 아래와 같은 형태의 이득 함수를 사용하여 최종 음성 추정 크기를 구하게 된다.

$$G_t = \frac{\hat{s_t}}{\hat{s_t} + \hat{n_t}} \tag{9}$$

각 시간 프레임에서 이득 함수를 구하고 이를 시간 t 프레임에서의 Y 와 곱해주면 우리가 구하고자 하는 값이 복소수 형태로 나오게 된다. 여기서 크기만 추정하는 이유는 신호의 위상(phase) 정보를 추정할 때는 처음의 잡음이 섞인 형태의 위상을 그대로 사용하여도 큰 문제가 없기 때문이다[20]. 이렇게 구한 복소수 값을 Inverse STFT 해주면 최종적으로 향상된 결과가 나오게 된다.

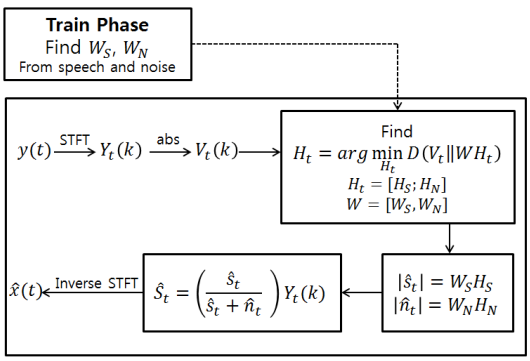

그림 1. NMF 기반 음성향상 블록 다이어그램
Fig. 1. the block diagram of speech enhancement based NMF

Ⅲ. 잡음 데이터를 이용한 음성 기저 행렬 훈련

3.1 기존의 문제점

NMF 기반 음성 향상의 경우 각 음원으로부터 얻은 기저들이 온전히 해당 음원만을 복원하는데 사용된다면 문제없이 완벽하게 수행될 것이다. 하지만 원래 데이터의 차수 보다 작은 기저들의 개수를 이용하기도 하고, 각 음원의 데이터들의 랭크 보다 작은 숫자의 기저들로 복원하기 때문에 무시하지 못 할 크기의 복원 오차도 발생하게 된다. 또한 각 음원의 한 순간, 즉 한 프레임의 데이터는 간혹 다른 음원의 한 프레임의 데이터와 유사할 수도 있고, 더욱이 이러한 상황에서 얻은 각 음원의 기저들은 유사할 수 있는 확률이 더 높다. 이러한 상황에서 아무리 각 음원의 기저를 적절하게, 많은 반복횟수를 이용하여 구했다 하여도 음원 분리 또는 음성 향상에서 복원 오차 이외에 기저의 잘못된 사용으로 인한 오차 또한 발생하게 된

다. 이는 달리 말해 각 음원에서 얻은 기저들의 선형 조합의 범위가 겹치게 된다는 것이다. 이렇게 겹치는 범위로 인해서 특정 기저들이 사용되지 말아야 되는 순간에도 사용되게 되어 큰 성능 저하를 불러일으킨다. 예를 들어 음성 기저 행렬이 섞여있는 잡음으로 인해 더 사용된다면 이는 잔여 잡음(residual noise)로 볼 수 있다. 반대로 실제 음성 성분이 잡음 기저 행렬이 사용 되어 복원이 된다면 그 부분만큼은 잡음으로 판단되어 실제 음성 부분이 사라진 결과를 보인다. 이는 음성 손상(speech distortion)으로 생각할 수 있다.

이를 해결하기 위해 크게 세 가지 접근 방법을 생각 할 수 있다. 첫 번째는 데이터를 좀 더 다른 음원의 데이터와 차별될 수 있도록 변형시키는 것이다. 예를 들어, 한 프레임의 데이터만을 사용하는 것이 아니라 두 프레임 이상의 데이터를 이어 붙여 데이터의 길이를 두 배, 세 배 이상 늘리는 것이다. 즉 한 프레임의 소리는 귀로 듣고 눈으로 본다고 해도 음원 간에 차이를 보기 어려울 때도 있지만, 여러 프레임의 데이터를 붙여서 보면 더 쉽게 음원 간의 차이를 알아 볼 수 있다는 것이다[21]. 이러한 방법의 경우 일정 크기의 성능 향상을 보이지만, 높은 성능 향상을 가져오진 못한다[20]. 두 번째로 접근하는 방식은 인코딩 행렬 영역에서 처리하는 것이다. [4]에서는 훈련 과정에서 얻은 H를 이용하여 하나의 통계 모델을 만들었다. 이 통계 모델(Gaussian distribution)을 향상 과정에서 사용하여 특정 음원의 기저들이 사용되는 정도, 즉 인코딩 행렬의 모양이 통계 모델과 큰 차이가 나지 않도록 강제하였다. 또는 [22]처럼 deep neural network를 사용하여 NMF 특성 상 발생할 수밖에 없는 기저들의 잘못된 사용을 방지하여 높은 성능을 얻기도 하였다. 접근 방법의 마지막으로 기저 행렬 자체의 모양을 최대한 적절하게 변형시키는 것이다. [23]에서는 중복된 사람의 음성을 분리하는데 NMF를 사용하고 있고, 이 때 discriminative NMF라는 것을 사용하였다. 이 논문은 시간 축의 데이터에서 NMF를 사용했다는 점이 기존 논문에 비해 특이한 점으로 볼 수 있다. 또한 훈련하고자 하는 데이터 세트를 클러스터링 등의 기법을 통해 작은 단위의 데이터 세트로 다시 나누고 해당 작은 단위의 데이터 세트에서 각기 다른 기저들을 구하여 하나의 통합된 기저 행렬로 사용하는 기법 또한 연구되었다[13]. 쉽게 말해, 여자 음성만을 이용해서 기저 행렬 하나를 만들고, 남성 음성만을 이용해서 기저 행렬을 하나 만든 후 두 행렬을 하나의 행렬로 통합하여 사용하는 것이다.

본 논문에서는 위에 설명한 세 가지 접근 방법 중 세 번째의 접근 방법을 택하였다. 인코딩 행렬의 이용과 데이터 변형 없이 단순히 기저 행렬을 적절히 만들어 기저들의 잘못된 사용을 막았다. 다음 절에는 이 논문에서 제안하는 새로운 기법을 설명한다.

3.2 잡음 데이터의 복원 오차를 이용한 음성 기저 행렬 훈련

앞서 설명했듯이, NMF 알고리즘의 경우 convexity를 만족하지 못 하는 등의 문제로 하나의 해답을 가지지 못 할 수 있다. 이러한 상황에서 NMF로 얻을 수 있는 기저 행렬은 목적 함수에 따라 또는 반복 횟수에 따라 또는 최적화 방법에 따라 다양할 수 있다. 본 논문에서는 이러한 가능한 다양한 기저 행렬들 중 음원 분리, 잡음 제거에 적절한 기저 행렬을 구하는 방법을 제안한다.

음성 데이터만을 이용하여 기저 행렬을 만든다면, 결과로 얻은 기저 행렬은 음성에만 특화된, 즉 음성 데이터를 복원할 때 가장 작은 복원 오차를 보이는 기저 행렬일 것이다. 즉 음성을 복원하는 방향으로는 적절한 기저 행렬이지만, 반대로 다른 음원과 분리할 때는 적절하지 못한 기저 행렬이 될 수도 있다. 이를 해결하고자 직관적으로 생각할 수 있는 해결 방법은 다음과 같다. 음성 음원의 기저 행렬은 음성 데이터는 잘 복원하되 그 외의 잡음 데이터는 잘 복원하지 못하도록 만드는 것이다. 이는 달리말해 음성 데이터를 대상으로 하였을 때는 복원 오차가 작고, 잡음 데이터를 대상으로 하였을 때는 복원 오차가 큰 기저 행렬을 말한다. 이를 수식으로 구현하기 위해 다음과 같이 목적 함수를 새롭게 정의하였다.

$$f(W_S, H_S, H_N) = D(V_S \| W_S, H_S) - \lambda D(V_N \| W_S, H_N) \quad (10)$$

V_S는 음성 데이터 세트의 STFT 크기값을, V_N은 잡음 데이터 세트의 STFT 크기값을 나타낸다. 이 때 λ의 경우 구하고자 하는 기저 행렬이 잡음 데이터 세트에 부적합한지와 음성 데이터 세트에 적합한지를 조절해주는 하나의 trade-off 성격을 가지는 변수이다. 즉, λ가 0 이라면 음성 데이터 세트에만 적합한 기저 행렬이 결과로 나오게 되고, 반대로 λ가 크면 클수록 음성 데이터 세트는 어느 정도 복원하되 잡음 데이터 세트를 복원하는데 부적합한 기저 행렬이 되도록 만들어 준다. 위의 수식을 Ⅱ.2.절에서 설명한 것처럼 multiplicative 방법을 이용하여 최적화 수식을 구할 수 있다. 이 때 W_S 를 업데이트 시키는 수식, H_S 를

업데이트 시키는 수식 그리고 H_N 을 업데이트 시키는 수식으로 총 세 개의 최적화 관련 수식이 나오게 된다. H_S 를 업데이트 시키는 경우 (10) 식에서 앞의 항에만 H_S 가 있기에 두 번째 항은 제외하고 생각하면 된다. 반대로 H_N 을 업데이트 시키는 경우 앞항은 제외하고 생각할 수 있다. 즉 H_S 와 H_N 을 업데이트 시키는 수식은 아래와 같다.

$$H_{S,ab} \leftarrow H_{S,ab} \frac{\sum_{i=1}^{n}(W_{S,ia}V_{S,ib}) / \sum_{k=1}^{r} W_{S,ik}H_{S,jb}}{\sum_{i=1}^{n} W_{S,ia}}$$

$$H_{N,ab} \leftarrow H_{N,ab} \frac{\sum_{i=1}^{n}(W_{S,ia}V_{N,ib}) / \sum_{k=1}^{r} W_{S,ik}H_{N,jb}}{\sum_{i=1}^{n} W_{S,ia}} \quad (11)$$

여기서 $H_{S,ab}$ 는 H_S 중 a행 b열의 원소를 나타낸다. W 에 대한 업데이트 수식은 아래와 같다.

$$W_{S,cd} \leftarrow W_{S,cd} \frac{\sum_{j=1}^{m}(H_{S,dj}V_{S,cj}) / \sum_{k=1}^{r} W_{S,ck}H_{S,kj}}{\sum_{j=1}^{m} H_{S,dj} + \lambda C}$$

$$C = \sum_{j=1}^{m}(H_{N,dj}V_{N,cj}) / \sum_{k=1}^{r} W_{S,ck}H_{N,kj} \quad (12)$$

이렇게 (11)와 (12) 세 수식을 반복적으로 수행하게 되면 음성 기저 행렬을 구할 수 있다. 이 때 λ값에 따라 잡음 데이터 복원에 부적절한 기저 행렬을 얻을 수 있다. 이 알고리즘에 대한 성능 평가는 Ⅳ. 실험 파트에서 다루도록 하겠다.

Ⅳ. 실 험

제안하는 알고리즘의 성능 평가를 위해 음성 데이터베이스는 TIMIT을 사용하였고, 잡음의 경우에는 NOISEX-92에 있는 F-16과 factory1 잡음을 선택하여 실험을 진행하였다. 이 데이터들의 샘플링 레이트는 $16kHz/s$이고 푸리에 변환에서의 윈도우 크기는 512으로 하여, 각 윈도우 크기에서 구한 값이 75% 겹치도록 하여 주파수-시간 영역으로 변환하였다. 기저 행렬을 구하기 위한 훈련 과정에서는 모든 데이터를 short time Fourier transform (STFT) 한 후 그것의 크기값만을 취해 데이터로 사용하게 된다. 음성 기저 행렬을 구하기 위한 데이터 세트는 TIMIT DB에서 남여 비율이 1 : 1이 되도록 하여 정하였다. 남자 화자 13파일과 여자 화자 13파일을 모아 하나의 데이터 행렬로 만들어주었다. 잡음의 경우에는 각각의 F-16, factory1 그리고 babble 음원 20초 정도만을 이용하였다. 이 모든 훈련용 데이터는 실험에서 쓰이는 부분과는 다른 부분을 택하였고, 음성 또한 실험에서 사용된 화자와 다른 화자의 목소리로 녹음된 파일만을 선택하였다. 이렇게 실험을 진행하는 이유는 다음과 같다. 음성의 경우 적은 화자로 훈련을 하고 그 화자 그대로 실험을 하게 되면 음성 기저 행렬이 실험 화자를 잘 표현하고 있기에 성능은 좋게 나올 수 있지만, 다른 화자를 대상으로 실험 시에는 큰 성능 저하를 불러일으킬 수 있기 때문이다.

각 결과의 성능은 perceptual evaluation of speech quality (PESQ)[15]를 사용하였다. 이는 원본 음성과 가까울수록 4.5, 다를수록 0 의 값을 나타낸다. 또 다른 성능 지표로 Signal to distortion ratio (SDR)[16]을 채택하여 결과를 얻었다. 각 잡음이 섞인 파일 입력 SNR이 $5dB$가 되도록 하였고, 남자화자 13명, 여자화자 13명으로 총 26파일로 PESQ 값을 얻고 이 값의 평균값을 최종 성능 지표로 정하였다.

비교하고자 하는 실험에 대해 설명하면 다음과 같다.

1) basic NMF: 제안하는 알고리즘과 비교를 위해 독립적으로 음성과 잡음의 기저 행렬을 구한 것을 이용한 음성 향상 기법
2) RE NMF: 잡음데이터의 복원 오차를 이용한 방법을 적용한 기저 행렬을 이용한 음성 향상 기법
3) Ortho. NMF: 기존 논문 [24]을 구현하여 얻은 기저 행렬을 사용. 음성과 잡음 기저 행렬이 직교(orthogonal)한 방향이 되도록 함.

위 비교한 실험 Ortho. NMF는 기존 다양한 DNMF 논문들 중 하나이다. 위 논문에서는 음원 분리 측면에서 각 음원의 기저 행렬이 직교한 것이 문제가 되기에 NMF 목적 함수에 각 음원의 기저 행렬이 직교하는 방향의 제한 조건을 추가해 주게 된다. 실험에 사용한 NMF 기반 음성 향상 기법은 세 조건 모두에서 동일한 조건으로 수행되었다. 각 시간 프레임에서 H 업데이트 수식이 반복된 횟수는 30번으로 동일하며 별도의 정지 조건(stop condition)을 적용하지 않았다. 단, basic NMF와 RE NMF 시에는 basic NMF에서 사용한 잡음 기저 행렬을 사용하였고, Ortho. NMF

에서는 제안된 기법을 적용하여 나온 잡음 기저 행렬을 사용하였다. Ortho. NMF의 경우 출력으로 나오게 되는 음성 기저 행렬은 잡음 기저 행렬과 함께 쓰여야 의미가 있기 때문이다. 각 음성과 잡음의 기저 개수는 동일하게 정하여 수행하였다.

음성 기저의 개수가 128개일 때의 성능은 표-1.과 같다. noisy signal 이란 잡음이 음성에 섞인 상태 그대로를 의미하며 향상 과정 전에는 1.9060의 PESQ 값을 나타내고 있다. 표-1.을 보면 잡음 데이터의 복원 오차를 키우는 방향으로 음성 행렬을 구한 RE NMF의 결과가 가장 좋음을 볼 수 있다. 기본적인 기저 행렬을 사용했을 때 보다 0.2 PESQ 점수가 올랐음을 볼 수 있다. 기존의 Ortho. NMF 또한 전체적으론 0.02 정도의 성능 향상이 있다고 볼 수 있지만 PESQ 상으로는 크게 의미가 없는 수준이라고 할 수 있다. 특히 babble 잡음 상황에서만 어느 정도 성능이 오르고 나머지 잡음의 경우에는 성능이 오히려 떨어지는 모습을 보인다. 이를 통해 기존의 직교 성질을 강제하는 방법은 적어도 음성 신호에는 적합하지 않다고 판단할 수 있다.

SDR로 향상 정도를 측정한 실험 결과는 표-2.와 같다. SDR 결과도 PESQ와 거의 같은 경향을 보인다. 제안된 RE NMF 의 실험 결과가 기존의 NMF 보다 0.51 SDR 상승을 보였다. 그리고 기존의 Ortho. NMF 의 경우도 PESQ와 동일하게 babble 상황에서만 성능이 향상 됨을 볼 수 있다. 여기서 주목할 점은 Ortho. NMF의 경우 babble 상황에서는 제안된 RE NMF 보다 높은 성능을 보인다는 것이다. 즉 기존의 직교성 강제는 적어도 babble 잡음 상황에서는 효과적이다.

위 두 실험 결과로 부터 알 수 있는 점은 다음과 같다. 음성과 잡음 즉, 분리하고자 하는 두 음원의 기저 행렬을 만들 때 서로 독립적으로 각 기저를 만들게 되면 서로 동시에 음원이 섞인 상황에서 효과적으로 분리를 잘 못 해낸다. 이는 서로의 기저가 서로의 복원 과정에서 간섭을 일으켜 기저의 잘못된 사용을 불러

표 1. 음성 향상에서의 PESQ 결과
Table 1. PESQ score in speech enhancement

$r_s = 128$	*noisy signal*	*basic NMF*	***RE NMF***	***Ortho. NMF***
F-16	1.9501	2.2081	**2.3791**	2.1980
factory2	1.8776	2.0912	**2.3619**	2.0935
babble	1.9830	2.1020	**2.2708**	2.1823
Average	1.9369	2.1337	**2.3373**	2.1579

표 2. 음성 향상에서의 SDR 결과
Table 2. SDR score in speech enhancement

$r_s = 128$	*noisy signal*	*basic NMF*	***RE NMF***	***Ortho. NMF***
F-16	5.0402	9.0334	**9.5313**	8.9235
factory2	5.0687	8.8288	**9.3896**	8.9232
babble	4.4962	7.2240	**7.7094**	8.0851
Average	5.0472	8.3620	**8.8768**	8.6439

일으킨 결과이다. 이는 실험적으로 충분히 증명 가능하다. 이 때문에 각 기저 행렬을 구할 때 서로의 데이터 정보를 활용하게 되면 좀 더 각 음원의 특징에 맞는 기저 행렬을 구해 낼 수 있다. 그 중에서도 다른 음원의 복원 오차가 큰 방향으로 기저 행렬을 구하는 것이 효과적임을 확인 할 수 있다.

V. 결 론

본 논문은 비음수 행렬 분해 과정에서 구해지는 기저 행렬에 관해 다루고 있다. 단일 음원을 복원 시에는 해당 음원 데이터만을 이용하여 충분히 적절한 기저 행렬을 얻을 수 있다. 하지만 두 개 이상의 음원이 섞인 상황에서의 복원에서는 이렇게 독립적으로 구한 기저 행렬이 효과적이지 못 할 수 있다. 이를 해결하고자 본 논문에서는 각 기저 행렬을 구할 때 다른 음원의 복원에는 큰 복원 오차를 가지게 강제하는 방법을 제안 하였다. 음성 기저 행렬을 잡음 데이터 복원에 부적절하게 만들어 준 결과 음성 향상 성능이 크게 향상됨을 확인할 수 있었다.

References

[1] G. Huang, J. Benesty, T. Long, and J. Chen, “A family of maximum SNR filters for noise reduction,” *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 22, no. 12, Dec. 2014.

[2] I. Cohen and B. Berdugo, “Speech enhancement for non-stationary noise environments,” *Signal Process.,* vol. 81, pp. 2403-2418, 2001.

[3] N. S. Kim and J.-H. Chang, “Spectral enhancement based on global soft decision,” *IEEE Signal Process. Lett*., vol. 7, no. 5, pp. 108-110, May 2000.

[4] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *IEEE Int. Conf. Acoustics, Speech and Signal Process.*, 2008.

[5] 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.

[6] M. N. Schmidt, J. Larsen, and F. T. Hsiao, "Wind noise reduction using non-negative sparse coding," *2007 IEEE Workshop Machine Learning for Signal Process.,* pp. 431-436, 2007.

[7] K. Kwon, J. W. Shin, S. Sukanya, I. Choi, and N. S. Kim, "Speech enhancement combining statistical models and NMF with update of speech and noise bases," *IEEE ICASSP*, vol. 21, no. 10, pp. 7103-7107, 2014.

[8] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 10, pp. 2140-2151, 2013.

[9] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature,* 1999.

[10] M. Julien, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," *Advances in Neural Inf. Process. Syst.,* 2009.

[11] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proc. 12th ACM SIGKDD Int. Conf. Knowledge Discovery and Data mining. ACM*, pp. 126-135, 2006.

[12] P. D. O'Grady and B. A. Pearlmutter, "Convolutive non-negative matrix factorisation with a sparseness constraint," in *Proc. 16th IEEE Signal Process. Soc. Workshop on Machine Learning for Signal Process.,* pp. 427-432, Arlington, VA, Sept. 2006.

[13] J. Huang and T. Zhang, "The benefit of group sparsity," *Annal. Statistics,* vol. 38, no. 4, pp. 1978-2004, 2010.

[14] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent," *IEEE Trans. Image Process.*, vol. 20, no. 7, Jul. 2011.

[15] *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, Tech. Rep. ITU-T P.862, 2001.

[16] E. Vincent, R. Gribonval, and C. Fvotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Language Process.,* vol. 14, no. 4, pp. 1462-1469, 2006.

[17] A. Pascual-Montano, J. M. Carazo, K. Kochi, D. Lehmann, and R. D. Pascual-Marquil, "Nonsmooth nonnegative matrix factorization (nsNMF)," *IEEE Trans. Pattern Anal. and Machine Intell.,* vol. 28, no. 3, pp. 403-415, 2006.

[18] L. Chin-Jen, "Projected gradient methods for non-negative matrix factorization," *Neural Computation,* vol. 19, no. 10, pp. 2756-2779, Oct. 2007.

[19] 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. 38, no. 6, pp. 503-511, Jun. 2013.

[20] D. Wang and J Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.,* vol. 30, No. 4, pp. 679-681, Aug. 1982.

[21] H.-T. Fan, J.-w. Hung, X. Lu, S.-S. Wang, and Y. Tsao, "Speech enhancement using segmental nonnegative matrix factorization," *IEEE ICASSP,* pp. 4516-4520, May 2014.

[22] P.-S. Huang, M. Kim, M. H-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," *IEEE ICASSP*, pp. 3433-3437, May 2014.

[23] G. Bao, Y. Xu, and Z. Ye, "Learning a discriminative dictionary for single-channel

speech separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 22, no. 7, Jul. 2014.

[24] E. M. Grais and H. Erdogan, "Discriminative nonnegative dictionary learning using cross coherence penalties for single channel source separation," *INTERSPEECH*, pp. 808-812, 2013.

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 형 용 (Hyung Yong Kim)

2014년 2월 : 광운대학교 전기공학과 졸업

2014년 3월~현재 : 서울대학교 전기·정보공학부 석박통합과정

<관심분야> 음성 신호처리, 통계적 신호처리

김 남 수 (Nam Soo Kim)

1988년 2월 : 서울대학교 전자공학과 졸업

1990년 2월 : 한국과학기술원 전기 및 전자공학과 석사

1994년 8월 : 한국과학기술원 전기 및 전자공학과 박사

1998년 3월~현재 : 서울대학교 교수

<관심분야> 음성 신호처리, 음성 인식, 통계적 신호처리, 패턴 인식, 휴먼 인터페이스