논문 13-38C-06-02

한국통신학회논문지 '13-06 Vol.38C No.06
http://dx.doi.org/10.7840/kics.2013.38C.6.503

시간 변화에 따른 사전 정보와 이득 함수를 적용한 NMF 기반 음성 향상 기법

권 기 수•, 진 유 광*, 배 수 현*, 김 남 수°

A NMF-Based Speech Enhancement Method Using a Prior Time Varying Information and Gain Function

Kisoo Kwon•, Yu Gwang Jin*, Soo Hyun Bae*, 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. 이 때 훈련으로부터, 시간 흐름에 따른 기저 사용량의 변화량을 각기 독립적인 가우시안 모델들로 만들고, 이를 이용하여 매 시간 프레임에서 주어진 모델들에 일정 가중치만큼 가까워지는 방향으로 최적화를 수행하였다. 또한 매 시간 얻은 NMF의 부호화 행렬의 결과를 이전 시간 프레임의 부호화 행렬 값과 평활화(smoothing) 과정을 수행하였다. 향상 과정에서는 Log-spectral Amplitude를 이용하여 이득(gain) 함수를 구하였다. 실험 결과에서는 PESQ 값을 지표로 사용하였고, 기존의 NMF를 이용한 음성 향상 보다 이 두 과정을 적용한 방법이 뛰어남을 확인 했다.

Key Words : speech enhancement, NMF, Gaussian distribution model, smoothing, Log-Spectral Amplitude

ABSTRACT

This paper presents a speech enhancement method using non-negative matrix factorization. In training phase, we can obtain each basis matrix from speech and specific noise database. After training phase, the noisy signal is separated from the speech and noise estimate using basis matrix in enhancement phase. In order to improve the performance, we model the change of encoding matrix from training phase to enhancement phase using independent Gaussian distribution models, and then use the constraint of the objective function almost same as that of the above Gaussian models. Also, we perform a smoothing operation to the encoding matrix by taking into account previous value. Last, we apply the Log-Spectral Amplitude type algorithm as gain function.

Ⅰ. 서 론

1990년 대 이전부터 지금까지도 음성 향상 분야에서는 통계적 방법을 이용한 음성 향상이 많이 사용되고 있다[1]. 이러한 음성 향상에서는 목표가 되는 음성과 그 외의 잡음을 각기 다른 가우시안 통계 모델로 만들고 매 시간 프레임 마다 음성 존재 검출(Voice Activity Detection, VAD) 방법 등을 결합해서 향상 과정을 수행 한다[2,3]. 이러한 방법의 향상은 잡음이 시간에 흐름에 따라 천천히 변한다는 가정을 한다. 이는 실제 잡음이 급격히 변한다거나 크기가 매우 커질 때는 높은 성능을 기대 할 수 없다. 또한 향상을 위해

※ 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2012R1A2A2A01045874).
• 주저자 : 서울대학교 전기·정보공학부 및 뉴미디어통신공동연구소, kskwon@hi.snu.ac.kr, 학생회원
° 교신저자 : 서울대학교 전기·정보공학부 및 뉴미디어통신공동연구소, nkim@snu.ac.kr, 종신회원
* 서울대학교 전기·정보공학부 및 뉴미디어통신공동연구소, ygjin@hi.snu.ac.kr, shbaenkim@snu.ac.kr
논문번호 : KICS2013-03-145, 접수일자 : 2013년 3월 28일, 최종논문접수일자 : 2013년 5월 15일

VAD 등의 과정이 필요하다는 특징도 있다. 이때 통계적 특성을 활용한 방법 등이 사용되고 있다[4].

1999년에 발표된 비음수 행렬 인수분해 (Non-negative Matrix Factorization, NMF) 는 음수가 없는 하나의 행렬을 두 개의 음수가 없는 행렬로 분해하는 알고리즘이다[5]. 해당 논문에서 저자는 얼굴 인식을 예로 들고 있다. 기존의 Principle Component Analysis (PCA) 는 NMF와 유사하지만, 비음수라는 조건이 없다. 이러한 차이로 인해, NMF는 분해된 두 행렬 중에 기저 행렬의 원소들을 단순히 더하는 방법으로 원래의 행렬과 같게 표현할 수 있다. 이러한 NMF 는 이미지 벡터의 기저를 찾거나 의미 단위로 문서를 분류하는 등 여러 분야에서 활발히 사용되었고, 지금은 데이터의 차수를 줄이는 하나의 방식으로 각광 받고 있다[6].

NMF는 초기에 제안된 배수사(multiplicative) 방식이 많이 사용되고 있다[5]. 그 이후에는 최적화 방법을 달리하여 경사 투영법(projected gradient)을 사용한 연구[6], 뉴턴 방법(Newton method)을 사용한 연구[7] 등이 있고, 목적함수로 유클리디안(Euclidean) 거리함수와 Kull-back Liebler 발산(divergence) 등을 사용한 논문 외에 다른 방법의 발산 등을 목적함수로 사용하기도 했다[7].

소리 또한 NMF를 쉽게 적용할 수 있다. 기존의 통계 특성을 이용한 음성 향상을 보면 음향 영역에서 하나의 복합된 음향은 여러 기본 음향들이 단순히 더해진 것으로 볼 수 있다. 이는 시간축의 값 뿐 아니라, Short Time Fourier Transform (STFT)를 통과한 주파수 축의 값에서도 유효하다. 이러한 성질 때문에 NMF는 처음에 음원 분리 분야에서 활발하게 연구되었다[8]. 여러 음원들을 미리 훈련을 통해 각각의 기저 행렬을 구하고 이를 이용하여 여러 음원이 섞인 소리에서 각각의 음원을 분리하는 것이다. 이와 같은 개념으로 음성향상에 적용할 수 있다. 음성과 여러 잡음이 섞인 소리에서 음성 및 잡음의 기저 행렬을 이용하여 음성과 잡음을 분리해내는 것이다. 이외에도 소리의 특성을 살리기 위해, 시간에 따른 연속성을 보장하는 방법[9]과 convolutive 개념을 적용한 cNMF[10] 등의 알고리즘도 연구되었다.

본 논문에서는 NMF를 이용한 훈련을 통해 얻은 부호화(encoding) 행렬의 시간에 따른 정보를 통계모델로 만들어 이 모델에 좀 더 가까워지는 방향으로 음성향상을 수행하는 것과, 매 시간 프레임에서 얻은 부호화 행렬을 부드럽게 이어주는 평활화 방법을 다루고 있다. 실험 과정에서는 이 두 과정을 적용하지 않은 기본적인 방법과의 성능 차이를 나타냈고, 지표로는 perceptual evaluation of speech quality (PESQ)를 사용하였다.

Ⅱ. 본 론

2.1. 비음수 행렬 인수분해

NMF 란 쉽게 말해 어떠한 정보 집단에서 개개의 정보들이 가지고 있는 공통된 부분(basis)들을 분리해내는 것이다. 실제의 정보 집단을 V, 분리하고자 하는 행렬을 W, H 라고 하면

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

를 만족하게 된다. W 는 기저 행렬을, H 는 부호화 행렬을 나타낸다. V 는 W 의 기저들의 합으로 구성되어 있다. V 는 $(n \times m)$ 크기를, W 는 $(n \times r)$, H 는 $(r \times m)$ 크기의 행렬이다. n 은 주파수 축의 개수, r 은 기저의 개수, m 은 시간 프레임의 개수를 의미한다. 이를 위해 [5]에서는 특정 거리함수를 목적함수로 정하고 이를 경사 하강법(gradient descent) 으로 W 와 H 에 대해 최적화를 하였다. 본 논문에서는 Kull-back Leibler 발산(KL Divergence) 거리함수와 경사 하강법을 사용하였다. [11] 논문을 참고하여 최적화 과정을 알아보겠다. 우리가 목적으로 하는 W 와 H 를 구하기 위해서 목적함수를 정하면 아래와 같다.

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

이 목적 함수의 값이 최소가 되는 방향으로 최적화를 진행한다. 위의 목적함수는 다음과 같이 나타낼 수 있다. 이 목적함수를 최적화 하는 과정은 convexity를 만족하지 않아 진행할 수 없다. 하지만 최적화 과정을 둘로 나누면 convexity를 만족하게 된다. 즉 W 와 H 중 하나를 고정 시키고 나머지 하나에 대해 최적화를 진행하게 되면 우리가 원하는 결과를 얻을 수 있다. 우선 H 를 구하는 과정이다.

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

이 목적 함수를 경사 하강법을 사용하기 위해 H_{ab} 에 따라 미분을 수행 하면 다음과 같다.

$$\frac{\partial}{\partial H_{ab}} D(V \| WH) = \sum_{i}^{n} W_{ia} - \sum_{i}^{n} \frac{V_{ib} W_{ia}}{\sum_{k}^{r} W_{ik} H_{kj}} \tag{4}$$

이를 경사 하강법에 적용하면

$$H_{ab} \leftarrow H_{ab} - \eta_{ab} \frac{\partial}{\partial H_{ab}} D(V \| WH) \tag{5}$$

처럼 수식이 나온다. 여기서 고정 소수점(fixed - point) 방법을 사용하기 위해 위의 증분량(step size)을 없애주어야 한다. 여기서 고정 소수점 방법이란 multiplicative 방법을 의미한다. 즉 아래와 같이 증분량을

$$\eta_{ab} = \frac{H_{ab}}{\sum_{i=1}^{n} W_{ia}} \tag{6}$$

로 치환을 해주면 아래와 같은 최종 식이 나온다.

$$H_{ab} \leftarrow H_{ab} \frac{\sum_{i=1}^{n} (W_{ia} V_{ib}) / \sum_{k=1}^{r} W_{ik} H_{jb}}{\sum_{i=1}^{n} W_{ia}} \tag{7}$$

같은 방법으로 H 를 고정하고 진행을 하면

$$W_{cd} \leftarrow W_{cd} \frac{\sum_{j=1}^{m} (H_{dj} V_{cj}) / \sum_{k=1}^{r} W_{ck} H_{kj}}{\sum_{j=1}^{m} H_{dj}} \tag{8}$$

의 W 에 관한 최적화 식을 얻을 수 있다. 위의 두 식을 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 최솟값을 정해주어야 한다[6]. 이 과정에서 중요한 것은 후에 부호화 행렬, H 를 통계적으로 사용하기 위해서 W 의 각 열을 기준으로 최댓값이 1 이 되도록 정규화 해주는 것이다.

이러한 방법 외에도 특정 증분량을 사용하는 방법 등과 성능 향상을 위해 최적화 하는 목적 함수에 sparseness와 관련된 부분을 추가해주는 여러 연구들이 있다[12]. 실제로도 본 논문을 위해 실험을 해 본 결과, sparseness 부분을 추가해주면 음성향상에도 약간의 성능 향상이 있음을 확인 하였다.

2.2. 향상 과정

향상 과정은 하나의 복원 과정으로 생각 할 수 있다. 우선 이 논문에서는 훈련 과정을 통해 얻은 기저 행렬이 향상 과정에서 별도의 업데이트 과정이 필요 없다고 가정한다.

훈련에 앞서 우리가 사용하고자 하는 신호의 데이터는 비음수를 만족해야한다. 그래서 소리 신호를 Sort Time Fourier Transform (STFT)를 통해 각 시간 프레임에서 주파수 별로 나타내고 이를 NMF에 사용하기 위해 절댓값으로 바꿔 준다. 훈련을 통해 얻은 음성 기저 행렬을 W_s, 잡음 기저 행렬을 W_n이라고 하자.

향상 과정에서도 마찬가지로 잡음이 섞인 신호를 STFT를 통해 복소수 값으로 변환 시킨다. 이렇게 변환 된 잡음이 섞인 신호를 Y 라고 하자. Y 의 절댓값 V 는 우리가 구한 기저 행렬을 통해 아래와 같이 나타 낼 수 있다.

$$V \approx [W_s, W_n] H \tag{9}$$

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. 이를 실시간으로 매 시간 프레임에서 수행하기 위해서 시간 t 에서의 복원 수식을 보면

$$V_t \approx [W_s, W_n][H_s; H_n] \tag{9}$$

로 나타 낼 수 있다. 이 때 각각의 H_s 와 H_n 은 $(r \times 1)$ 행렬이고, $[H_s; H_n]$은 $(2r \times 1)$ 행렬이 된다. 즉 $H_t = [H_s; H_n]$ 으로 볼 수 있다. 이렇게 한

프레임에서 추정 과정을 마치면 아래와 같이 음성과 잡음 크기의 추정값이 나오게 된다.

$$\hat{s} = W_s H_s \qquad (11)$$
$$\hat{n} = W_n H_n$$

이 때 매 프레임에서 부호화 행렬, H_t 의 초기값을 무작위로 준다. 하지만 이 때 음성과 잡음의 차이로 인해 초기값을 주는 형태가 다르다. 음성은 주파수 영역에서 비교적 이전 프레임의 값과 유사하다. 이러한 성질은 H_s 영역에서도 마찬가지다. 그래서 $H_{s,t}$의 초기값은 이전 프레임의 값인 $H_{s,t-1}$ 으로 주는 것이 성능 향상에 좋다. 반대로 잡음의 경우에는 이전 값으로 주게 되면 성능이 저하되는 현상을 보인다. 그래서 $H_{n,t}$ 만은 무작위로 초기값을 매 프레임 설정해 준다.

이러한 과정을 거치고 나면, 실제 잡음에는 어느 정도 $\hat{s}$ 에 존재하게 된다. 그래서 이를 바로 결과로 사용하기에는 성능 상 문제가 있다. 그래서 보통 간단한 형태로 아래와 같은 이득 함수를 구하여 사용한다.

$$G_t = \frac{\hat{s}}{\hat{s}+\hat{n}} \qquad (12)$$

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[13]. 이렇게 구한 복소수 값을 Inverse STFT 해주면 최종적으로 향상된 결과가 나오게 된다.

2.3. 제안하는 방법

실제로 위의 방법으로 향상을 하고 나면 남아있는 잡음의 존재하고, 음성 손상 또한 존재한다. 이는 시간에 따른 H_t 를 분석해 보면 원인을 찾을 수 있다. H_t 는 음성향상 과정에서 각 시간 프레임에서의 음성과 잡음의 부호화 행렬을 나타낸다. 우리가 구한 W_s 와 W_n 은 직교성(orthogonality)을 만족하지 못하기 때문에 실제 음성이 W_n 을 사용하기도 하고, 실제 잡음이 W_s 을 사용하기도 한다. 즉 H_n 의 크기가 음성이 커질수록 커지는 경향이 있다. 이를 방지하고자 아래의 방법을 제안했다.

2.3.1. 훈련 과정의 H 변화량 활용

W 에서 각각의 열은 하나의 기저를 나타낸다. 그렇다면 H 의 한 행은 하나의 기저가 시간의 흐름에 따라 사용된 정도를 나타낸다. 훈련 중에 얻은 H 를 이용하여, 인접한 프레임 값과의 차이를 절댓값으로 구하고 이를 H_d 라고 하자. 이는 각 기저가 사용된 정도의 변화를 나타낸다. 각 행의 H_d 의 평균과 분산을 이용하여 각 행의 가우시안 통계 모델을 만든다. 즉 r 개의 모델이 생기게 된다. 이 모델에 가까울수록 목적 함수에는 0 에 가까운 값이, 모델에 멀수록 0 보다 큰 값이 더해진다. 즉 해당 잡음이 가지고 있는 각 기저의 사용 정도의 변화를 매 시간 특정 값으로 가중치만큼 강제해주는 것이다. 너무 적게 변했다면 좀 더 크게, 너무 크게 변했다면 적게 모델에 근접하게 H_t 를 얻게 된다. 향상 과정에서는 기저 행렬을 고정으로 사용하기에 목적함수는 다음과 같이 H_t 로 표현 할 수 있다. 여기서 추가되는 조건은 잡음만을 대상으로 한다.

$$f(H_t) = D(V_t \parallel WH_t) - \alpha L(H_{n,t}) \qquad (13)$$

뒤의 $L(H_t)$ 함수는 훈련을 통해 얻은 H 의 가우시안 모델과 H_t 의 차이를 의미 한다. 이 목적 함수를 이용하여 다음과 같이 H_t 를 구하게 된다.

$$H_t = \arg\min\left(D(V_t \parallel WH_t) - \alpha L(H_{n,t})\right) \qquad (14)$$

$H_{n,t}$ 는 각 시간대에서 구한 부호화 행렬 중에 잡음 부분을 의미한다. 그리고 이때 2.에서 설명한 향상 방법에서는 $H_{n,t}$ 의 초기값을 무작위로 설정했지만, 이 방법을 사용할 때는 $H_{n,t}$의 초기값을 이전 프레임의 값인 $H_{n,t-1}$ 로 설정하는 것이 성능 향상에 좋다. 이는 변화량을 각 잡음의 특징에 맞게 강제하기 때문에 각 프레임에서의 초기값을 이전 프레임의 값으로 하는 것이 적절하다. 목적함수에 추가된 $L(H_t)$ 함수는 아래와 같다.

$$L(H_t) = \frac{1}{2\pi\sigma^2} exp\left(-\frac{1}{2} \frac{(H_{d,t} - \mu_d)^2}{\sigma_d^2} \right) \quad (15)$$
$$H_{d,t} = |H_t - H_{t-1}|$$

여기서 μ_d 는 훈련과정에 나온 H 의 각 행의 평균 변화율을, σ_d 는 분산을 의미한다. 이 값들은 $(r \times 1)$ 크기의 행렬이다. r 은 기저의 개수이다. 실험결과, 통계 모델을 데이터의 차수가 r 인 하나의 가우시안 모델로 보고 H 의 공분산을 이용하여 계산하는 것 보다 위와 같이 각 행 마다 독립적으로 보고 r 개의 통계 모델로 계산하는 것이 높은 성능을 나타냈다. 즉 잡음이 기저행렬을 사용하는 정도의 변화량은 서로 연관성이 없다고 가정 하는 것이다. 계산의 편의를 위해 위의 식을 log 형태로 취하고 경사 하강법과 고정 소수점 방법을 취하면 아래와 같은 반복 식이 나온다.

$$H_{ab} \leftarrow H_{ab} \left(\frac{\left(\sum_{i=1}^{n} \left(W_{ia} V_{ib} / \sum_{k=1}^{r} W_{ik} H_{ik} \right) \right)}{\sum_{i=1}^{n} W_{ia} - \alpha \left(\frac{H_{d_a} - \mu_{d_a}}{\sigma_d^2} \right)} \right) \quad (16)$$

향상과정에서 W 는 고정이기에 H_t 만 수렴할 때까지 위의 식을 반복하여 각 시간 프레임에서 H_t 를 찾는다. 음성 부분에서는 기존의 최적화 식을 사용한다. 그리고 이 제한조건과 관련해서 한 가지 조건을 추가해 준다. 변화량만 만족하는 방향으로 최적화를 하다 보면 정해진 정답에서 벗어날 수 있다. 이를 방지하고자 [14] 논문과 유사한 방법을 사용하였다. 훈련과정에서 나온 H 값으로 서로 독립적인 가우시안 모델을 만들고 목적 함수에 같은 방법으로 제한 조건을 추가해 준다.(이 조건의 가중치는 α_m이라 하고 값은 0.03을 주었다.)

그리고 여기서 크기를 맞추어 주는 작업이 중요하다. 훈련에 사용된 잡음의 크기와 실제 향상 시키는 신호의 잡음의 크기는 다르다. 크기가 다른 상태에서 훈련을 통해 얻은 통계 모델을 그대로 사용하면 높은 성능을 기대할 수 없다. 또한 일반적으로 NMF를 사용하여 향상을 하면 잡음만 있는 구간에서는 음성 기저 행렬로 인해 잡음 기저 행렬만 사용했을 때 보다 $H_{n,t}$ 의 크기가 작아진다. 이를 보완하기 위해 초반 20 프레임까지는 잡음만 존재한다고 가정하고 잡음 기저 행렬만 이용해서 H_n 를 구한다. 20 프레임 동안의 평균 크기와 훈련을 통해 얻은 H 의 평균 크기 비를 이용해서 통계 모델을 수정해 준다. 크기 비에 따라 가우시안 모델의 크기와 분산을 수정하는 것이다. 이러한 과정을 거치면 향상하고자 하는 신호에 적절한 통계 모델을 적용할 수 있고, 음성 기저 행렬과 같이 사용하는 환경에서 실제 잡음이 작게 추정되는 것을 어느 정도 보완하게 되어 성능 향상을 기대할 수 있다.

2.3.2. H_t 의 평활화

또 한 가지 추가되는 방법은 H의 평활화(smoothing)이다. 일반적으로 잡음은 짧은 시간 구간에서 크게 변하지 않는다는 가정을 H 영역에서도 적용하여 H_{t-1} 값에 일정한 가중치를 주어 현재의 H_t 에 반영해주는 것이다.

$$H_t = (1-\beta) H_t + \beta H_{t-1} \quad (17)$$

이러한 방법을 사용하는 이유는 처음에 제안한 H 의 통계를 이용하는 이유와 같다. 실제 향상 과정을 보면 H_n이 음성이 클수록 커지는 현상이 있다. 이는 음성 기저 행렬과 잡음 기저 행렁이 어느 정도의 유사도가 존재하고, 이에 따라 실제 음성이 잡음 기저 행렬을 사용하게 된 결과이다. 즉 이를 방지하기 위해 1에 가까운 크기로 이전 값을 현재 값에 반영해 준다면 음성이 큰 부분에서 H_n이 커지는 현상을 억제 할 수 있다. 또한 이 논문의 향상 과정에서는 마지막에 이득 함수로 wiener 필터 형태의 LSA를 사용하고 있다. 이러한 이득 함수에서는 잡음이 평활화가 되어도 성능이 보장 된다. 실제로 잡음에 사용되는 β_n 은 1에 가까울수록 성능이 좋다.

이 방법은 음성의 H_t 에도 가능하다. H_s 또한 H_n 과 마찬가지로, 실제 잡음이 H_s 를 조금씩 사용한 결과 약간의 잡음이 남게 된다. 이를 평활화를 통해 약간은 보상을 해줄 수 있다. 하지만 음성 부분의 β_s 값은 0에 가까워야 한다. β_s 값이 크다면 음성손실을 불러일으킬 수 있기 때문이다.

이러한 평활화 방법은 첫 번째에 제안한 방법과 별개로 적용하여도 성능 향상을 보인다. 마찬가지로 같이 적용하면 좀 더 높은 성능 향상을 보인다. 이는 변화량의 기준이 되는 이전 시간의 H 가 추정이 잘못

됐을 시에 현재 시간의 H 또한 잘못 추정될 수 있다. 하지만 평활화를 사용하여 최종적으로 이전 시간의 H 를 정함으로서, 잘못 추정되는 정도를 줄일 수 있고 이는 변화량 통계특성이 적용될 때 안정적으로 현재 시간의 변화된 H 를 추정 할 수 있게 된다. 실험 결과에서도 이와 같은 성능 향상을 볼 수 있다.

2.3.3. LSA를 이용한 이득함수

앞서 언급했듯이 안정적인 성능을 위해 추정된 음성 크기를 그대로 사용하지 않고 위너 형태의 간단한 이득함수를 사용하기도 한다. 하지만 이 방법 대신 Log-spectral Amplitude (LSA) 와 같은 방법을 이용하여 이득 함수를 사용하면 (12) 수식의 결과에 비해 높은 성능을 얻을 수 있다. 실제 이 논문에서 사용한 이득 함수는 아래와 같다[1].

$$G_t = \frac{\xi_t}{1+\xi_t} exp\left(\frac{1}{2}\int_{\nu_t}^{\infty}\frac{e^{-t}}{t}dt\right) \tag{18}$$

$$\xi_t = \frac{\hat{s_t}^2}{\hat{n_t}^2} \quad \gamma_t = \frac{V_t^2}{\hat{n_t^2}} \quad \nu_t = \frac{\gamma_t \xi_t}{1+\xi_t} \tag{19}$$

ξ_t는 사전 신호 대 잡음비(priori SNR), γ_t 는 사후 신호 대 잡음비(posteriori SNR) 이다. 원래 이 수식에서 사전 신호 대 잡음비 또는 사후 신호 대 잡음비에 사용되는 파워는 특정 시간 구간에서의 기대값 (expectation) 이다. 그렇기 때문에 (13) 식처럼 추정된 값 그대로 사용하는 것 보다 다음과 같이 망각율 (forgetting factor)를 사용하여 ξ_t 와 γ_t 를 구한다.

$$P_s(t) = \tau P_s(t-1) + (1-\tau)\hat{s^2} \tag{20}$$
$$P_n(t) = \tau P_n(t-1) + (1-\tau)\hat{n^2}$$

$$\xi_t = \frac{\widehat{P_s}}{\widehat{P_s}} \qquad \gamma_t = \frac{V_t^2}{\widehat{P}_n} \tag{21}$$

이 때 τ 의 값이 0 이라면 (19) 식과 같게 되고, 1에 가까울수록 이전 값과 거의 같게 된다. 실험적으로 τ 이 0 보다 클수록 성능이 향상됨을 보였지만, 0.6 이상이 되면 성능이 떨어지는 것을 볼 수 있다. 이 LSA 의 성능 향상 정도는 실험에서 따로 다루겠다.

Ⅲ. 실험 결과

음성은 TIMIT DB에서, 잡음은 NOISEX-92 DB에 있는, F-16, factory2, m109, leopard, white 를 사용하였다. 푸리에 변환에서의 윈도우 크기는 512개 이고, 각 윈도우 크기에서 구한 값이 75% 겹치도록 하여 주파수-시간 영역으로 변환하였다. 즉 이때의 데이터의 차수는 256 이 된다. 음성의 경우 테스트 하는 화자와 다른 화자들로, 남녀 각각 13명의 음성 파일을 사용하였다. 또한 비슷한 음절이 실험 결과에 영향을 끼칠 수 있기 때문에 테스트에 사용된 대본과 다른 대본으로 발음된 파일을 사용하였다. 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 있다. 잡음의 경우는 각기 실제 섞인 잡음과 다른 부분으로 각기 16초 정도의 wav 파일을 이용하여 잡음의 기저 행렬을 구하였다. 또한 훈련에 사용된 잡음은 실제 실험에서 음성과 섞이는 잡음과는 다른 부분이다. 이 때 각 기저의 개수는 40으로 구하였다.

각 결과의 성능은 perceptual evaluation of speech quality (PESQ)를 사용하였다. 이는 원본 음성과 가까울수록 4.5, 다를수록 0 의 값을 나타낸다. 각 잡음이 섞인 파일은 PESQ 값이 평균적으로 비슷한 수준이 되도록 선택하였다. F-16과 white 의 경우 input SNR 은 5 dB, factory2 와 m109는 0 dB, 마지막으로 leopard 는 $-5dB$인 잡음이 섞인 파일을 선택하였다. 그리고 테스트에 쓰인 파일의 남, 여 성비는 1 : 1 로 하여 진행하였다.

각 잡음과 음성을 섞은 음원을 준비하고 각각의 화자에 따라 여러 번 수행한 결과의 평균을 구하였다.

향상

과정 수행 전의 PESQ 값은 각 잡음에서 1.7493에서 1.8661 사이였다. [표 1.]을 보면 NMF를 기본적으로 사용했을 때와 여기서 제안한 변화량 통계특성과 평활화를 각기 또는 동시에 적용했을 때의 PESQ 값을 수행 전보다 상승한 정도로 나타내었다. [표 2.]에서는 기존과 달리 새롭게 추가한 LSA 가 이득함수로서 적절한지 알아보기 위한 실험 결과를 보여주고 있다.

표 1. PESQ의 상승량
(변화량 통계특성 및 평활화 적용, LSA 미적용)
Table 1. A amount of enhancement in PESQ (apply statistic of change and smoothing of H)

noise	basic NMF	+smoothing H	+statistic of H	+smoothing H +statistic of H
F-16	+0.2352	+0.2993	+0.4342	+0.4723
factory2	+0.3686	+0.4298	+0.5486	+0.6258
m109	+0.4361	+0.4157	+0.5224	+0.5718
leopard	+0.5875	+0.6156	+0.6435	+0.7438
white	+0.5331	+0.5849	+0.7428	+0.7484
Ave.	+0.4321	+0.4691	+0.5782	**+0.6324**

표 2. PESQ의 상승량(LSA 적용)
Table 2. A amount of enhancement in PESQ (LSA)

noise	basic NMF	+ LSA	+smoothing H +statistic of H	+smoothing H +statistic of H + LSA
F-16	+0.2352	+0.2629	+0.4723	+0.5366
factory2	+0.3686	+0.4001	+0.6258	+0.7065
m109	+0.4361	+0.4622	+0.5718	+0.6450
leopard	+0.5875	+0.6128	+0.7438	+0.8181
white	+0.5331	+0.6357	+0.7484	+0.7694
average	+0.4321	+0.4747	+0.6324	**+0.6951**

[표 1.] 에서는 LSA를 이득함수로 적용하지 않고 (12)의 간단한 이득함수를 공통적으로 적용한 실험 결과이다. 즉 [표 1.] 은 변화량 통계특성과 평활화의 성능을 평가한 결과이다. 첫 번째 실험(basic NMF)은 [1]에서 KL 분산 함수와 경사 하강법을 이용한 NMF를 사용했다. 두 번째 실험(+smoothing H)은 첫 번째 제안된 방법인 H 의 변화량 통계특성을 반영하지 않고 평활화 과정만을 적용한 결과이다. 세 번째 실험(+statistic H)은 H 의 변화량 통계특성을 적용한 결과이다. 이때 변화량이 반영된 정도인 α 값은 0.07 이다. 이 때 α_m 값은 경험적으로 0.03 로 정하였다. 마지막 실험(+smoothing H + statistic H)은 제안한 두 가지 방법을 모두 적용한 결과이다.

두 번째 실험 결과를 보면 평활화 된 H_t 를 이용한 방법이 기본적인 NMF 사용했을 때 보다 약 0.0370 정도 상승했음을 확인할 수 있다. 하지만 유일하게 m109 잡음에서는 성능이 약간 저하됨을 확인 할 수 있다. 그 외에는 모든 잡음에서 성능이 향상 되었다. H 통계특성을 활용한 방법은 0.1461 정도 상승하였다. 하지만 이 둘을 결합했을 때는 향상 정도가 둘의 합인 0.1831 보다 큰 0.2003 정도 상승 하였다. 이는 H_t 의 변화량을 측정 할 때 기준이 되는 H_{t-1}이 평활화 된 값이기 때문에 좀 더 안정적인 값으로 다음 프레임의 값을 얻을 수 있다고 추측 된다. 이는 기본적인 방법과 비교하여 1.46 배의 높은 성능을 나타낸다.

[표 2.] 에서는 이득함수로 제안한 LSA 가 기존의 (12) 식에 비해 얼마나 성능향상을 보이는지 알아보기 위한 실험 결과이다. 본 실험에서 망각율 τ 는 0.3의 값을 주었다. 첫 번째와 두 번째의 결과를 보면 기본적인 NMF 에서 수정된 LSA로 이득함수를 바꿈으로서 0.0426 정도 평균적으로 성능 향상이 되었음을 확인할 수 있다. 더욱이 세 번째와 네 번째 실험을 비교하면 성능 향상 폭이 더 큼을 볼 수 있다. 즉 변화량 통계특성 과 평활화를 적용한 경우에는 LSA 를 적용하면 0.0627 정도의 성능 향상을 보인다.

결과적으로 이 논문에서 제안한 세 가지, 변화량 통계특성 반영, H_t 영역에서의 평활화 그리고 이득함수로서의 LSA를 각기 적용해도 각각 모두 성능이 향상됨을 볼 수 있고, 나아가 이 세 가지를 모두 적용 했을 때 각기 적용했을 때 보다 높은 성능 향상이 있음을 확인하였다. 수치상으로 기존의 방법에 비해서 1.61 배 정도의 성능 향상이 있다.

Ⅳ. 결 론

본 논문은 NMF를 음성향상에 적용하였다. NMF를 이용한 음성향상은 기존의 통계모델을 이용한 방법에 비해 직관적이다. VAD 또는 음성 존재 확률(speech presence probability, SPP) 등의 과정이 필요하지 않다. 기저 행렬을 구하기 위한 데이터베이스, V 는 시간의 흐름에 따라 표현되어 있다. 그렇기에 부호화 행렬, H 에는 각 음원의 시간 정보를 포함하고 있다. 특히 잡음은 음성에 비해 특정 기저를 사용하는 정도의 변화를 통계 모델로 적용하기 적절하다. 훈련 과정 나온 H 의 각 행에서 열에 따른 변화 정도와 크기를 각각 기저의 개수만큼의 독립적인 통계 모델로 만들고

H_t 를 구할 때 가중치를 주어 반영해주었다. 그 결과 기존에 비해 성능이 향상함을 확인하였다. 기존 통계 특성을 이용한 음성향상에서는 추정한 잡음의 안정성을 위해 평활화 작업을 한다. 이와 같은 목적으로 매 시간 얻은 H_t 를 이전 값 H_{t-1}과 평활화 함으로써 실제 음성과 실제 잡음이 적절한 기저 행렬을 사용하도록 하였다. 이 결과 음성 향상에 있어서 성능이 높게 향상됨을 확인하였다. 마지막으로 기존의 간단한 이득함수 대신에 LSA를 이용한 이득함수와 적절한 망각율을 이용한 음성 및 잡음 파워 추정으로 높은 성능을 이끌어 낼 수 있었다.

나아가 위의 통계 모델을 음성 향상에 있어서 실시간으로 업데이트 해준다면 더 높은 성능 향상과 잡음의 크기가 변하는 상황에서도 높은 성능을 얻을 수 있을 것으로 기대 된다.

References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.,* vol. 33 no. 2, pp. 443-445, Apr. 1985.

[2] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Process.* vol. 81, no. 11, pp. 2403-2418, Nov. 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] J.-H. Chang and N.S. Kim, "Noisy speech enhancement based on multiple statistical models," *Telecommun. Review*, vol. 16, no. 4, pp.731-747, Aug. 2006.

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

[6] C.-J. Lin, "Projected gradient methods for non-negative matrix factorization," *Neural Computation*. vol. 19, no. 10, pp. 2756-2779, Oct. 2007.

[7] R. Zdunek and A. Cichocki, "Non-negative matrix factorization with quasi-Newton optimization," in *Proc. 8th Int. Conf. Artificial Intell. Soft Comput. (ICAISC 2006)*, pp. 870-879, Zakopane, Poland, June 2006.

[8] A. Cichocki, R. Zdunek, and S. Amari, "New algorithms for non-negative matrix factorization in application to blind source separation," *IEEE Acoust. Speech Signal Process.,* vol. 5, pp. 14-19, May 2006.

[9] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization With temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 1066-1074, Mar. 2007.

[10] P. D. O'Grady and B. A. Pearlmutter, "Convolutive non-negative matrix factorization with a sparseness constraint," in *Proc. 16th IEEE Signal Process. Soc. Workshop Machine Learning Signal Process.,* pp. 427-432, Maynooth, Ireland, Sep. 2006.

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

[12] P. O. Hoyer, "Non-negative sparse coding," in *Proc. IEEE Workshop Neural Networks for Signal Process.*, pp. 557-565, Martigny, Switzerland, Sep. 2002.

[13] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 30, no. 4, pp. 679-681, Aug. 1982.

[14] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. 2008*, pp. 4029-4032, Las Vegas, U.S.A., Apr. 2008.

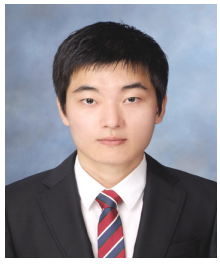
권 기 수 (Kisoo Kwon)

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

진 유 광 (Yu Gwang Jin)

2007년 2월 서울대학교 전기공학부 졸업
2007년 3월~현재 서울대학교 전기·정보공학부 석박통합과정
<관심분야> 음질 향상, 음성 신호처리, 통계적 신호 처리

배 수 현 (Soo Hyun Bae)



2012년 2월 인하대학교 정보통신공학부 졸업
2012년 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