한국음향학회지 제40권 제3호 pp. 234~240 (2021)
The Journal of the Acoustical Society of Korea Vol.40, No.3 (2021)
https://doi.org/10.7776/ASK.2021.40.3.234

pISSN : 1225-4428
eISSN : 2287-3775

잡음 환경에 효과적인 마스크 기반 음성 향상을 위한 손실함수 조합에 관한 연구

A study on combination of loss functions for effective mask-based speech enhancement in noisy environments

정재희,[1] 김우일[1†]

(Jaehee Jung[1] and Wooil Kim[1†])

[1]인천대학교 컴퓨터공학부

(Received March 19, 2021; accepted April 22, 2021)

초 록: 본 논문에서는 잡음 환경에서 효과적인 음성 인식을 위해 마스크 기반의 음성 향상 기법을 개선한다. 마스크 기반의 음성 향상 기법에서는 심층 신경망을 기반으로 추정한 마스크를 잡음 오염 음성에 곱하여 향상된 음성을 얻는다. 마스크 추정 모델로 VoiceFilter(VF) 모델을 사용하고 추정된 마스크로 얻은 음성으로부터 잔여 잡음을 보다 확실히 제거하기 위해 Spectrogram Inpainting(SI)기법을 적용한다. 본 논문에서는 음성 향상 결과를 보다 개선하기 위해 마스크 추정을 위한 모델 학습 과정에 사용되는 조합된 손실함수를 제안한다. 음성 구간에 남아 있는 잡음을 보다 효과적으로 제거하기 위해 잡음 오염 음성에 마스크를 적용한 Triplet 손실함수의 Positive 부분을 컴포넌트 손실함수와 조합하여 사용한다. 실험 평가를 위한 잡음 음성 데이터는 TIMIT 데이터베이스와 NOISEX92, 배경음악 잡음을 다양한 Signal to Noise Ratio(SNR) 조건으로 합성하여 만들어 사용한다. 음성 향상의 성능 평가는 Source to Distortion Ratio(SDR), Perceptual Evaluation of Speech Quality(PESQ), Short-Time Objective Intelligibility(STOI)를 이용한다. 실험을 통해 평균 제곱 오차로만 훈련된 기존 시스템과 비교하여, VF 모델은 평균 제곱 오차로 훈련하고 SI 모델은 조합된 손실함수를 사용하였을 때 SDR은 평균 0.5dB, PESQ는 평균 0.06, STOI는 평균 0.002만큼 성능이 향상된 것을 확인했다.

핵심용어: 음성 향상, 마스크, 잡음 환경, 손실함수, 심층 신경망

ABSTRACT: In this paper, the mask-based speech enhancement is improved for effective speech recognition in noise environments. In the mask-based speech enhancement, enhanced spectrum is obtained by multiplying the noisy speech spectrum by the mask. The VoiceFilter (VF) model is used as the mask estimation, and the Spectrogram Inpainting (SI) technique is used to remove residual noise of enhanced spectrum. In this paper, we propose a combined loss to further improve speech enhancement. In order to effectively remove the residual noise in the speech, the positive part of the Triplet loss is used with the component loss. For the experiment TIMIT database is re-constructed using NOISEX92 noise and background music samples with various Signal to Noise Ratio (SNR) conditions. Source to Distortion Ratio (SDR), Perceptual Evaluation of Speech Quality (PESQ), and Short-Time Objective Intelligibility (STOI) are used as the metrics of performance evaluation. When the VF was trained with the mean squared error and the SI model was trained with the combined loss, SDR, PESQ, and STOI were improved by 0.5, 0.06, and 0.002 respectively compared to the system trained only with the mean squared error.

Keywords: Speech enhancement, Mask, Noisy environments, Loss function, Deep neural network

PACS numbers: 43.72.Bs, 43.72.Ne

†**Corresponding author:** Wooil Kim (wikim@inu.ac.kr)
Department of Computer Science and Engineering, Incheon National University, 119 Academy-ro, Yeonsu-gu, Incheon 22012, Republic of Korea
(Tel: 82-32-835-8459, Fax: 82-32-835-0780)

I. 서 론

음성 인식 시스템의 성능이 저하되는 가장 큰 원인 중 하나는 인식 시스템에 장착되는 음향 모델을 학습하는 환경과 실제 시스템을 적용하는 환경이 음향학적 측면에서 불일치한다는 점이다. 이러한 음향학적 불일치를 최소화함으로써 음성 인식 성능을 향상하기 위한 다양한 연구가 진행되어 왔다[1-8]. 본 논문에서는 음성 인식 시스템의 전처리 단계에서 잡음을 제거하고 음성의 품질을 향상시킴으로써 음성 인식 성능을 향상시키기 위한 음성 향상 기법에 관한 연구를 소개한다.

전통적인 음성 향성 기법으로 통계적 접근방법을 이용한 위너 필터,[1] 스펙트럼 차감법[2,4] 등의 방법이 오랜 기간 연구되어 왔으나 이와 같은 기법은 시간에 따라 변하는 잡음에 대해서는 성능이 좋지 못하다. 심층 신경망 기술 발전에 따라 최근에는 심층 신경망을 이용한 마스크 기반 음성 향상 기법에 대한 많은 연구가 진행되어 시간에 따라 변하는 잡음에 대해서도 비교적 높은 성능을 보이고 있다.[9,10] 하지만 아직 배경 잡음이 극심한 낮은 Signal-to-Noise Ratio(SNR) 환경에서는 성능 향상이 필요하다.

본 논문에서는 마스크 기반 음성 향상을 위한 마스크 추정 모델로 VoiceFilter(VF)[11] 모델을 사용하고 후처리 방법으로 Spectrogram Inpainting(SI)[12] 기법을 사용한다. 음성 품질 및 명료도의 성능을 높이기 위해 기존에 사용되던 컴포넌트 손실함수[13]를 개선하는 방법에 관해 연구를 진행하였다. 이를 위해 잡음 음성 스펙트럼에 마스크가 적용된 후 복원된 음성은 깨끗한 음성에 가깝도록 최대한 남기고 잡음과는 최대한 멀어지도록 하는 경우를 동시에 반영할 수 있도록 Triplet 손실함수[14]를 조합하였다.

2장에서 마스크 기반 음성 향상을 위해 VF 기반의 마스크 추정 모델과 SI 기법을 도입한 후처리 기법에 관해 설명하고, 3장에서 기존에 사용되고 있는 손실함수와 제안한 조합된 손실함수 기법에 관해 설명한다. 4장에서 실험 방법 및 손실함수에 따른 음성 품질 및 명료도에 대한 성능을 비교 평가한 후, 5장에서 결론을 맺는다.

II. 마스크 기반 음성 향상

본 논문에서는 낮은 SNR로 잡음에 오염된 음성에서 최대한 잡음을 제거하고 음성의 품질을 향상하는 것을 목표로 두고 있다. 배경 잡음에 부가적으로 시간 n에서의 오염된 음성 $y(n)$는 Eq. (1)과 같이 나타낼 수 있다. 이 식에서 $s(n)$과 $d(n)$은 각각 깨끗한 음성과 잡음 신호를 나타낸다.

$$y(n) = s(n) + d(n). \tag{1}$$

본 논문에서 사용하는 마스크 기반 음성 향상 기법은 시간-주파수 도메인에서 적용되므로 오염된 음성과 깨끗한 음성, 잡음을 단시간 푸리에 변환(Short Time Fourier Transform, STFT)을 통해 각각 크기 스펙트럼으로 변환하여 사용하였고, Eq. (2)와 같이 스펙트럼 상에서도 부가적인 관계가 유지된다. Eq. (2)에서 $Y_l(k)$, $S_l(k)$, $D_l(k)$는 각각 $y(n)$, $s(n)$, $d(n)$의 STFT 결과이며, 잡음에 오염된 음성, 깨끗한 음성, 잡음 신호의 스펙트럼을 나타낸다. 이때 l은 프레임 인덱스를 나타내고 k는 주파수 간격 인덱스를 나타낸다.

$$Y_l(k) = S_l(k) + D_l(k). \tag{2}$$

마스크 기반 음성 향상을 위해 마스크를 추정하기 위한 심층 신경망 기반 모델 구조로 VF 모델을 사용하였다. 본 논문에서는 VF 모델을 통해 추정된 마스크 적용의 결과로 남아있는 잔여 잡음들을 제거하고 잡음을 제거하는 과정에서 제거된 음성 구간을 복구하기 위해 부분 컨볼루션을 이용한 SI 기법을 후처리 단계에서 적용하였다.

2.1 마스크 추정 모델

마스크 추정 모델로 사용한 VF 모델은 Fig. 1에 나타낸 것과 같이 8개의 컨볼루션 신경망(Convolution Neural Network, CNN) 층, 1개의 양방향 장단기 메모리(Long-Short Term Memory, LSTM) 층, 1개의 완전 연결 (Fully-Connected) 계층의 모델 구조를 가진다.

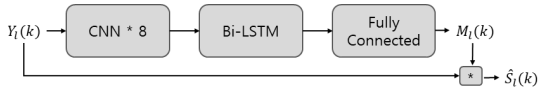

Fig. 1. Architecture of voicefilter model.

VF 모델의 입력으로는 Eq. (2) 식에서 얻은 오염된 음성의 스펙트럼을 사용하고 깨끗한 음성의 스펙트럼을 목표로 하여 이에 가까워지도록 훈련을 진행한다. 잡음 오염 음성의 스펙트럼에 제로 평균 및 단위 분산 정규화를 적용한다. 학습을 통해 추정된 마스크 $M_l(k)$는 최종 출력 단계에서 시그모이드 활성화 함수를 거쳐 [0, 1] 사이의 값을 가지며 Eq. (3)와 같이 요소 사이의 곱 연산에 의해 향상된 음성의 스펙트럼 $\hat{S}_l(k)$을 얻어낼 수 있다.

$$\hat{S}_l(k) = Y_l(k) * M_l(k). \tag{3}$$

2.2 Spectrogram Inpainting(SI) 기반의 후처리 기법

VF 모델을 이용하여 추정한 마스크로 얻은 향상된 음성 스펙트럼에서 아직 제거되지 않은 잔여 잡음을 제거하고 잡음을 제거하는 과정에서 손상된 음성 요소를 복구하기 위해서 후처리 방법으로 SI 기법을 사용하였다.

SI 기법을 위한 모델은 Fig. 2와 같이 2개의 다운 샘플링 블록, 8개의 Residual 블록, 2개의 업 샘플링 블록으로 구성되어있다. 각 Residual 블록은 2개의 CNN으로 구성되어 블록의 입력과 두 번째 CNN의 출력을 더해 해당 블록의 출력을 얻을 수 있다.

모델의 입력으로는 마스크 기반 기법에 의해 향상된 스펙트럼과 음성/비음성 구간을 구분하기 위한 이진 마스크를 사용한다. VF 모델 기법에서 추정한 마스크는 [0, 1] 사이의 연속적인 값을 가지므로 이를 이진 마스크로 변형하기 위해 Eq. (4) 와 같이 마스크 값과 문턱값 T와의 비교를 통해 0 또는 1의 값을 가지는 이진 마스크로 변경하여 사용한다. 본 논문에서는 문턱값 T를 0.35로 설정하여 사용하였다.

$$BM_l(k) = \begin{cases} 1, M_l(k) > T \\ 0, otherwise \end{cases}. \tag{4}$$

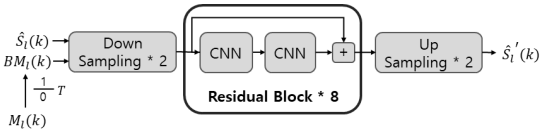

Fig. 2. Architecture of spectrogram inpainting model.

SI 기법은 기존의 컨볼루션 대신 음성 구간이 존재하는 부분에 대해서만 컨볼루션 연산을 진행하는 부분 컨볼루션을 사용한다. 부분 컨볼루션 연산은 Eqs. (5)와 (6)으로 계산되며 해당 식은 하나의 컨볼루션 필터에 해당하는 연산이다. 이때 w^T는 컨볼루션 필터의 가중치이고 b는 컨볼루션 연산의 bias 값으로 학습 과정을 통해 갱신된다. $\hat{s}$는 현재 컨볼루션 필터에 해당되는 스펙트럼 값이고 bm은 이에 대응되는 이진 마스크 값이다.

$$\hat{S}_l'(k) = \begin{cases} w^T(\hat{s} \odot bm)\frac{1}{\sum bm} + b, \sum bm > 0 \\ 0, otherwise \end{cases}. \tag{5}$$

$$BM_l'(k) = \begin{cases} 1, \sum bm > 0 \\ 0, otherwise \end{cases}. \tag{6}$$

Eq. (5)와 같이 현재 컨볼루션의 필터에 해당되는 이진 마스크의 합이 0보다 큰 경우에만 향상된 스펙트럼과 이진마스크를 요소 곱을 취한 뒤 컨볼루션 연산을 적용한다. 이 때 스펙트럼 값을 조정해주기 위한 스케일링 값을 곱해준다. Eq. (6)과 같이 해당 영역의 이진마스크 값은 1로 갱신한다.

이진 마스크로 결정된 음성 구역 정보를 바탕으로 위에서 설명한 부분 컨볼루션 과정을 통해 주변부의 음성 요소를 복구하거나 잡음으로 판단된 구역은 스펙트럼 값을 0으로 설정함으로써 잡음 요소를 억제하게 된다. 이러한 과정을 반복하게 되면서 음성 스펙트럼은 향상되고 잡음 스펙트럼은 제거된다.

최종 모델의 출력으로는 SI기법이 적용된 향상된 스펙트럼을 얻을 수 있고 얻은 스펙트럼에 역 STFT을 적용함으로써 시간 축의 음성 신호 파형으로 변환한다.

III. 손실함수

3.1 평균 제곱 오차(Mean Squared Error)

평균 제곱 오차는 마스크 기반 음성 향상 기법에서 대중적으로 사용하는 손실함수로서 음성 향상 처리 과정을 통해 얻은 향상된 음성 스펙트럼과 깨끗한 음성의 스펙트럼의 거리를 줄이도록 하는 손실함수를 정의한다. 기존의 VF 모델 기반 음성 향상 기법과 SI 기법에서는 손실함수로 평균 제곱 오차를 사용하였다. 평균 제곱 오차 손실함수는 한 프레임에 대해 다음과 같이 계산된다.

$$L_l^{mse} = \sum_{k=1}^{K} (\hat{S}_l(k) - S_l(k))^2. \quad (7)$$

3.2 컴포넌트 손실함수(Component Loss)[13]

컴포넌트 손실함수는 평균 제곱 오차를 사용했을 때 음성의 일부 요소가 제거되는 것을 최소화하고 잔여 잡음도 함께 고려하기 위해 제안된 손실함수로 추정된 마스크를 잡음 오염 음성에 곱하지 않고 깨끗한 음성의 스펙트럼과 잡음 스펙트럼에 곱하여 손실함수 계산에 이용한다.

컴포넌트 손실함수는 깨끗한 음성의 스펙트럼에 마스크를 곱한 여과된 깨끗한 스펙트럼 $\widetilde{S}_l(k)$과 깨끗한 스펙트럼의 차이를 줄이는 동시에 잡음에 마스크를 곱한 여과된 잡음 스펙트럼 $\widetilde{D}_l(k)$의 값을 최소화하는 함수로 정의된다. 한 프레임에 대한 손실함수는 다음의 식으로 계산되며 α는 가중치 요소로서 [0, 1] 사이의 범위를 갖는다. 본 논문에서 α는 0.5로 설정하였다.

$$\widetilde{S}_l(k) = S_l(k) * M_l(k). \quad (8)$$

$$\widetilde{D}_l(k) = D_l(k) * M_l(k). \quad (9)$$

$$L_l^{CL} = (1-\alpha)\sum_{k=1}^{K} (\widetilde{S}_l(k) - S_l(k))^2 + \alpha \sum_{k=1}^{K} (\widetilde{D}_l(k))^2. \quad (10)$$

3.3 제안하는 손실함수 조합

본 논문에서는 향상된 스펙트럼의 음성 부분에 남아있는 잡음을 최대한 제거하기 위해 여과된 깨끗한 음성과 가까워지도록 컴포넌트 손실함수를 개선했다.

컴포넌트 손실함수의 경우 여과된 깨끗한 스펙트럼과 깨끗한 스펙트럼을 비교하여 음성이 사라지는 것을 방지하고 여과된 잡음 값을 작게 만들어 잡음을 제거했다고 볼 수 있다. 잡음 음성 스펙트럼에서 추정된 마스크를 적용하여 얻는 향상된 스펙트럼은 음성 구간이 많이 포함되어 있다. 따라서 향상된 스펙트럼을 여과된 깨끗한 스펙트럼과 가깝도록 하여 여과된 음성에 남아있는 잡음을 집중적으로 제거하고 동시에 여과된 잡음과는 최대한 멀어지도록 하여 향상된 스펙트럼 전체에 있는 잡음도 함께 줄이고자 한다. 이 경우를 모두 한 번에 고려할 수 있도록 Triplet 손실함수[14]를 사용하여 조합하는 방법에 관해 연구를 진행하였다. Triplet 손실함수는 Anchor 입력을 Positive 입력과 Negative 입력과의 각각 비교를 수행하여 Anchor와 Positive의 거리는 최소화되도록, Anchor와 Negative의 거리는 최대화되도록 계산한다.

향상된 음성 스펙트럼은 마스크를 적용한 깨끗한 음성 부분과 최대한 가까워지도록 하고 마스크가 적용된 잡음과는 최대한 멀어지도록, Triplet 손실함수의 Anchor 입력은 향상된 스펙트럼, Positive 입력은 여과된 깨끗한 스펙트럼, Negative 입력은 여과된 잡음 스펙트럼으로 정의하여 컴포넌트 손실함수와 조합하였다. 본 논문에서 대상으로 하는 낮은 SNR 조건에서 오염된 음성 스펙트럼은 잡음 성분이 모든 시간-주파수 영역에 분포되어 있기 때문에 향상된 스펙트럼을 마스크가 적용된 잡음과 멀어지도록 하면 음성 부분도 같이 삭제되어 결과 음성에 상당한 왜곡이 발생하는 것을 관찰하였다. 이에 따라 본 논문에서는 Triplet 손실함수의 Anchor 입력과 Positive 입력과의 거리만 가까워지도록 하는 부분만 사용하여 컴포넌트 손실함수와 조합하여 사용하였다. 한 프레임에 대한 Triplet 손실함수의 Positive 부분과 조합된 손실함수는 다음과 같이 계산한다. 여기서 β는 가중치 요소로서 본 논문에서는 0.3으로 설정하였다.

$$L_l^{pos} = \sum_{k=1}^{K} (\hat{S}_l(k) - \widetilde{S}_l(k))^2. \quad (11)$$

$$L_l = L_l^{CL} + \beta L_l^{pos}. \quad (12)$$

Triplet 손실함수의 경우 모든 입력이 마스크에 곱해져 있어 마스크의 상태에 따라 영향을 많이 받는다. 이에 따라 본 논문에서는 마스크에 대해 처음 학습하는 VF 모델 훈련 시 초반 훈련에는 컴포넌트 손실함수만 사용 후 20epoch 후에는 Eq. (12)과 같이 손실함수를 조합하여 사용하였다.

IV. 실험 및 결과

깨끗한 음성 데이터는 TIMIT Data[15]를 사용하였고 배경 잡음 데이터는 비교적 Stationary한 특성을 갖는 Factory, Car 잡음과 시간에 따라 변하는 특성을 갖는 Babble, Music 잡음을 사용하였다. Factory, Car, Babble데이터는 NOISEX92[16]에 포함된 데이터를 사용하고, Music 샘플은 한국 가요 전주 부분에서 추출했다. 잡음에 오염된 음성 데이터는 깨끗한 음성 데이터에 각 잡음 샘플을 SNR이 5 dB, 0 dB, -5 dB가 되도록 생성하여 훈련용 데이터는 총 54,264개 발화, 모의 테스트용 데이터는 총 120개 발화, 테스트용 데이터는 2,268개의 발화를 사용하였다.

실험에 사용한 음성 데이터는 샘플링비율은 8 KHz, STFT를 위해 윈도우 길이는 50 ms, 이동 길이는 20 ms로 설정하였다. 고속 푸리에 변환 개수는 512로 설정해 주파수 요소 256차원과 에너지 값을 포함하여 총 257차원으로 사용하였다. 훈련에서 학습률은 0.001, 최적화 알고리즘은 'adam'을 이용하였다.

실험은 VF 모델과 SI 모델의 훈련에 사용되는 손실함수를 평균 제곱 오차, 컴포넌트 손실함수, 조합한 손실함수로 변경하면서 실험을 진행하였다. Tables 1 ~ 3의 실험은 VF 모델 훈련에 사용된 손실함수와 SI 모델 훈련에 사용된 손실함수의 조합 별로 성능을 관찰하였고, 성능 지표로는 SDR, PESQ, STOI 총 3가지를 사용하였다. 모든 표의 수치는 실험에 사용한 모든 잡음 종류의 같은 SNR 조건에 대한 평균을 나타낸 것이다.

Table 1. SDR results of model trained with three losses (loss to train voicefilter model– loss to train spectrogram inpainting model).

Loss combinations (VF-SI)	SNR 5 dB	SNR 0 dB	SNR -5 dB	Avg
Noisy speech	5.4	0.5	-4.3	0.5
MSE-MSE	13.8	11.3	8.4	11.2
CL-MSE	13.5	11.1	8.3	11.0
Proposed-MSE	13.9	11.4	8.6	11.3
MSE-CL	13.5	11.0	8.2	10.9
CL-CL	13.6	11.1	8.3	11.0
Proposed-CL	13.4	10.9	8.1	10.8
MSE-Proposed	**14.3**	**11.8**	**8.9**	**11.7**
CL-Proposed	13.9	11.4	8.6	11.3
Proposed-Proposed	**14.0**	**11.5**	**8.6**	**11.4**

Table 2. PESQ results of model trained with three losses (loss to train voicefilter model– loss to train spectrogram inpainting model).

Loss combinations (VF-SI)	SNR 5 dB	SNR 0 dB	SNR -5 dB	Avg
Noisy speech	1.89	1.60	1.40	1.63
MSE-MSE	3.02	2.68	2.26	2.65
CL-MSE	2.96	2.63	2.24	2.61
Proposed-MSE	3.02	2.69	2.29	2.67
MSE-CL	2.98	2.63	2.22	2.61
CL-CL	2.98	2.64	2.26	2.63
Proposed-CL	2.98	2.64	2.24	2.62
MSE-Proposed	**3.10**	**2.74**	**2.28**	**2.71**
CL-Proposed	3.02	2.67	2.26	2.65
Proposed-Proposed	**3.07**	**2.71**	**2.27**	**2.68**

Table 3. STOI results of model trained with three losses (loss to train voicefilter model– loss to train spectrogram inpainting model).

Loss combinations (VF-SI)	SNR 5dB	SNR 0dB	SNR -5dB	Avg
Noisy speech	0.762	0.663	0.554	0.660
MSE-MSE	0.902	0.857	0.786	0.848
CL-MSE	0.901	0.856	0.785	0.847
Proposed-MSE	0.903	0.860	0.792	0.852
MSE-CL	0.902	0.858	0.787	0.849
CL-CL	0.900	0.856	0.786	0.848
Proposed-CL	0.901	0.859	0.790	0.850
MSE-Proposed	0.902	0.857	0.782	0.847
CL-Proposed	0.902	0.858	0.786	0.849
Proposed-Proposed	**0.904**	**0.859**	**0.787**	**0.850**

4.1 Source-to-Distortion Ratio(SDR)[17]

SDR은 소스-대-왜곡 비율로 목표 음성과 원하지 않는 소스의 간섭, 잡음, 뮤지컬 잡음 같은 인공적으로 만들어진 요소들의 비율로 단위는 dB를 사용한다.

4.2 Perceptual Evaluation of Speech Quality (PESQ)[18]

PESQ는 원신호와 향상된 신호를 비교 평가하는 통화 품질 평가 방법으로 ITU-T에서 표준화하였고 0.5 ~ 4.5 사이의 값의 범위를 가진다.

4.3 Short-Time Objective Intelligibility(STOI)[19]

STOI는 표준 음성 명료도 측정 도구로 0 ~ 1 사이 값의 범위를 가진다.

Tables 1, 2, 3의 결과는 평균 제곱 오차나 컴포넌트 손실함수를 사용하여 훈련한 모델보다 제안한 Triplet 손실함수와 조합한 손실함수를 사용하여 훈련된 모델이 대부분 더 높은 성능을 가지는 것을 나타낸다. 특히 SDR과 PESQ지수는 평균 제곱 오차로 VF 모델을 훈련하고 제안한 조합 손실함수로 SI 모델을 훈련한 결과가 가장 좋은 성능을 보였다.

이러한 결과는 본 논문에서 제안하는 것과 같이 향상된 스펙트럼과 여과된 깨끗한 스펙트럼의 차이를 손실함수에 적용하는 것이 향상된 음성 스펙트럼을 얻는데 효과적임을 입증한다. 즉, 향상된 음성 스펙트럼과 여과된 깨끗한 음성 스펙트럼의 차이를 줄임으로써 향상된 음성 구간에 남아 있는 잡음을 효과적으로 제거할 수 있는 것으로 판단된다.

V. 결 론

본 논문에서는 낮은 SNR 환경에서 효과적인 음성 인식을 위해 마스크 기반의 음성 향상 기법을 제안했다. 마스크 기반의 음성 향상 기법의 성능을 높이기 위해 손실함수를 조합하는 방법에 대해 연구를 진행했다. 기존에 사용되었던 평균 제곱 오차에서 손실되는 음성 요소를 복구하고 잔여 잡음을 제거하기 위해 컴포넌트 손실함수를 도입하였다. 음성 구간에 남아 있는 잡음을 보다 효과적으로 제거하기 위해 잡음 오염 음성에 마스크를 적용한 Triplet 손실함수의 Positive 부분을 조합하여 사용하였다. 실험을 통해 VF 모델에서는 평균 제곱 오차, SI 모델에서 조합한 손실함수를 사용한 경우가 기존 두 모델 모두 평균 제곱 오차로만 사용한 경우보다 SDR은 평균 0.5, PESQ는 평균 0.06, STOI는 평균 0.002의 음질 향상 결과를 나타내는 것을 확인했다.

감사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2019R1F1A106299513).

References

1. J. Lim and A. Oppenheim. “All-pole modeling of degraded speech,” IEEE Trans. on Acoustics, Speech, and Signal Process. **26**, 197-210 (1978).
2. S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. on Acoustics, Speech and Signal Proccess. **27**, 113-120 (1979).
3. Y. Ephraim and D. Malah, “Speech enhancement using minimum mean square error short time spectral amplitude estimator,” IEEE Trans. on Acoustics, Speech and Signal Process. **32**, 1109-1121 (1984).
4. R. Martin, “Spectral subtraction based on minimum statistics,” Proc. EUSIPCO. 1182-1185 (1994).
5. P. J. Moreno, B. Raj, and R. M. Stern, “Data-driven environmental compensation for speech recognition: a unified approach,” Speech Communication, **24**, 267-285 (1998).
6. W. Kim and J. H. L. Hansen, “Feature compensation in the cepstral domain employing model combination,” Speech Communication, **51**, 83-96 (2009).
7. J. Du, L.-R. Dai, and Q. Huo, “Synthesized stereo mapping via deep neural networks for noisy speech recognition,” Proc. ICASSP. 1764-1768 (2014).
8. K. Han, Y. He, D. Bagchi, E. F. -Luissier, and D. L. Wang, “Deep neural network based spectral feature mapping for robust speech recognition,” Proc. Interspeech, 2484-2488 (2015).
9. K. Tan and D. L. Wang, “A convolutional recurrent neural network for real-time speech enhancement,” Proc. Interspeech, 3229-3233 (2018).
10. Y. Xu, J. Du; L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural

networks," IEEE/ACM. Trans. on Audio, Speech, and Lang. Process. **23**, 7-19 (2014).
11. Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," arXiv preprint arXiv:1810.04826 (2018).
12. X. Hao, X. Su, S. Wen, Z. Wang, Y. Pan, F. Bao, and W. Chen, "Masking and inpainting: A two-stage speech enhancement approach for low SNR and non-stationary noise," Proc. ICASSP. 6959-6963 (2020).
13. Z. Xu, S. Elshamy, and T. Fingscheidt, "Using separate losses for speech and noise in mask-based speech enhancement," Proc. ICASSP. 7519-7523 (2020).
14. F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," Proc. IEEE conference on CVPR. 815-823 (2015).
15. J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," NASA STI/Recon Tech. Rep., 1993.
16. 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**, 247-251 (1993).
17. E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," IEEE Trans. on audio, speech, and lang. process. **14**, 1462-1469 (2006).
18. A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," Proc. IEEE ICASSP. 01CH37221 (2001).
19. C. H. Taal, R. C. Hendriks, R, Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time- frequency weighted noisy speech," Proc. IEEE ICASSP. 4214-4217 (2010).

저자 약력

▸ **정 재 희 (Jaehee Jung)**

2021년 2월 : 인천대학교 컴퓨터공학부 공학사
2021년 3월 ~ 현재 : 인천대학교 컴퓨터 공학과 석사과정

▸ **김 우 일 (Wooil Kim)**

1996년, 1998년, 2003년 : 고려대학교 전자공학과 공학사/공학석사/공학박사
2004년 ~ 2005년 : 미국 Carnegie Mellon University 박사후연구원
2005년 ~ 2012년 : 미국 University of Texas at Dallas, 연구원, 연구교수
2012년 ~ 현재 : 인천대학교 컴퓨터공학부 조교수, 부교수