한국음향학회지 제40권 제2호 pp. 176~182 (2021)
The Journal of the Acoustical Society of Korea Vol.40, No.2 (2021)
https://doi.org/10.7776/ASK.2021.40.2.176

pISSN : 1225-4428
eISSN : 2287-3775

다양한 손실 함수를 이용한 음성 향상 성능 비교 평가

Performance comparison evaluation of speech enhancement using various loss functions

황서림,[1] 변준,[1] 박영철[1†]

(Seo-Rim Hwang,[1] Joon Byun,[1] and Young-Cheol Park[1†])

[1]연세대학교 컴퓨터정보통신공학부

(Received January 19, 2021; accepted February 23, 2021)

초 록: 본 논문은 다양한 손실 함수에 따른 Deep Nerual Network(DNN) 기반 음성 향상 모델의 성능을 비교 평가한다. 베이스라인 모델로는 음성의 위상 정보를 고려할 수 있는 복소 네트워크를 사용하였다. 손실 함수는 두 가지 유형의 기본 손실 함수, Mean Squared Error(MSE)와 Scale-Invariant Source-to-Noise Ratio(SI-SNR)를 사용하였으며 두 가지 유형의 지각 기반 손실 함수 Perceptual Metric for Speech Quality Evaluation(PMSQE)과 Log Mel Spectra(LMS)를 사용한다. 성능은 각 손실 함수의 다양한 조합을 사용하여 얻은 출력을 객관적인 평가와 청취 테스트를 통해 측정하였다. 실험 결과, 지각기반 손실 함수를 MSE 또는 SI-SNR과 결합하였을 때 전반적으로 성능이 향상되며, 지각기반 손실함수를 사용하면 객관적 지표에서 약세를 보이는 경우라도 청취 테스트에서 우수한 성능을 보임을 확인하였다.
핵심용어: 음성 향상, 손실 함수, 복소 네트워크, 지각 최적화, 공동 학습

ABSTRACT: This paper evaluates and compares the performance of the Deep Nerual Network (DNN)-based speech enhancement models according to various loss functions. We used a complex network that can consider the phase information of speech as a baseline model. As the loss function, we consider two types of basic loss functions; the Mean Squared Error (MSE) and the Scale-Invariant Source-to-Noise Ratio (SI-SNR), and two types of perceptual-based loss functions, including the Perceptual Metric for Speech Quality Evaluation (PMSQE) and the Log Mel Spectra (LMS). The performance comparison was performed through objective evaluation and listening tests with outputs obtained using various combinations of the loss functions. Test results show that when a perceptual-based loss function was combined with MSE or SI-SNR, the overall performance is improved, and the perceptual-based loss functions, even exhibiting lower objective scores showed better performance in the listening test.
Keywords: Speech enhancement, Loss function, Complex network, Perception optimization, Joint learning

PACS numbers: 43.60.Uv, 43.72.Ar

I. 서 론

음성 향상은 의사전달에 방해되는 잡음을 제거하여 기존에 목표로 하였던 깨끗한 음성을 복원하는 기법이다.[1-3] 딥러닝이 발전하면서 음성 향상 또한 많은 발전을 이루고 있으며 음성 인식 인공지능, 보청기 등 다양한 응용 분야에 사용되고 있다.

그러나 딥러닝 기반 음성 향상은 음질과 음성의 명료도 측면에서 우수한 성능을 보이고 있지만 아직 보완해야하는 문제점들을 가지고 있다. 첫째로, 딥러닝 기반의 기존 음성 향상은 학습 단계에서 복소수 연산이 필요한 위상 정보를 고려하지 못한다. 이는 음질이나 음성의 명료도를 향상시키는 데에 한계를 지니게 한다. 최근에는 위상을 고려할 수 있는 복

†**Corresponding author:** Young-Cheol Park (young00@yonsei.ac.kr)
Division of Computer and Telecommunication Engineering, Yonsei University, Chang jo room 269, 1 Yonseidae-gil, Wonju, Gangwon-do 26493, Republic of Korea
(Tel: 82-33-760-2744, Fax: 82-33-763-4323)

소 네트워크들이 주목을 받고 있으며,[3] 이러한 모델들은 위상을 고려하지 않거나 위상을 고려하지만 복소 네트워크를 사용하지 않는 기존의 모델[1,2]과 비교하여 우수한 성능을 보이고 있다.[3]

두 번째로 음성 향상 성능을 높이기 위해서는 딥러닝 네트워크의 훈련을 위한 손실 함수를 최적화해야 한다. 딥러닝 모델을 학습할 때, 손실 함수는 학습 방향을 결정하는 중요한 역할을 한다. 음성 향상에서 모델의 성능을 높이기 위한 손실 함수로 평균 제곱 오차(Mean Squared Error, MSE) 함수가 일반적으로 사용된다. 그러나 평균 제곱 오차 함수는 모델의 성능을 높이는 데에 한계를 지니고 있다.[4] MSE를 대체하기 위한 손실 함수로서 Scale-Invariant Source-to-Noise Ratio(SI-SNR)가 사용되기도 한다.[3] 그리고 이외에도 음성 향상에서도 다양한 손실 함수를 통한 학습의 중요성이 관심을 받고 있으며[4,5] 특히, 사람의 청각 지각 능력에 기반을 둔 손실 함수들이 제안된 바 있다.[6,7] 이러한 지각 기반 손실 함수는 객관적인 지표에서는 다소 열세에 있을 수 있으나 지각적으로 우수한 출력을 만들어 내는 것으로 알려져 있다.

본 논문에서는 복소 네트워크 모델을 기반으로 다양한 손실함수에 따른 음성 향상의 성능을 비교 평가함으로써 음성 향상을 위한 딥러닝 네트워크 최적화 가이드라인을 제시하고자 한다. 이를 위해 MSE와 SI-SNR 두 종류의 손실 함수를 기본으로 사용하고, 이에 사람의 지각 특성에 기반을 둔 두 종류의 손실함수[7,8]를 추가적으로 결합하여 성능을 평가하였다. 성능평가는 객관적인 지표와 청취실험을 통해 비교 분석하였다.

본 논문의 구성은 다음과 같다. II장에서 음성 향상을 위한 네트워크 모델에 대해 설명하고, III장에서는 네트워크 최적화를 위한 손실 함수에 대해 설명한다. IV장에서 실험 데이터 구성과 실험 결과를 평가한 후, V장에서 결론을 맺는다.

II. 성능 평가를 위한 음성 향상 네트워크

딥러닝 기반의 음성 향상에서 가장 일반적으로 사용되는 방법은 시간-주파수 마스킹 기법이다.[2,3] 딥러닝 모델에 잡음이 섞인 음성이 입력으로 들어가면 해당 모델은 마스크를 추정하고 추정된 마스크가 입력으로 들어온 음성과 곱해져 향상된 음성을 얻는다. 이때, 시간 영역에서의 음성은 각각 Short-Time Fourier Transform(STFT)과 Inverse STFT(ISFTF)을 통해 시간-주파수 영역으로 전환되며 수식은 다음과 같다.

$$\hat{X}_{t,f} = \hat{m}_{t,f} \odot Y_{t,f}. \tag{1}$$

$Y_{t,f}$는 잡음이 섞인 음성 신호를 주파수 영역으로 변환한 것이며, $\hat{m}_{t,f}$는 모델을 통해 추정된 마스크이다. 위 수식을 통하여 얻어진 $\hat{X}_{t,f}$를 ISTFT하여 향상된 음성을 얻는다.

기존의 시간-주파수 마스킹 기법은 모델 추정 과정에서 위상 정보를 고려하지 않는다. 이는 향상된 음성을 추정하는 데 오차를 발생시키며 음질과 음성의 명료도를 높이는 데에 한계를 갖는다. 최근에는 이러한 오차를 보완하기 위하여 위상 정보를 고려하는 복소 마스크를 추정하는 방법이 제안되었으며[2,3] 복소 마스크를 효과적으로 계산할 수 있는 복소 네트워크가 제안되었다.[3]

본 논문에서는 최근 제안되어 우수한 성능을 보이는 것으로 알려진 Deep Complex Convolutional Recurrent Network(DCCRN)[3]를 베이스라인 네트워크로 사용하여 서로 다른 손실 함수에 따른 음질을 비교 평가한다. DCCRN은 합성곱 신경망과 순환 신경망을 결합한 복소 네트워크이며, 구조는 Fig. 1과 같다. 잡음 신호를 합성곱 STFT을 통해서 복소 스펙트럼을 추출한 뒤 복소 인코더와 복소 Long Short-Term Memory(LSTM), 복소 디코더를 순차적으로 거쳐 복소 마스크를 추정한다. 그리고 복소 마스크를 추출한 복소 스펙트럼에 곱한 뒤 합성곱 ISTFT을 통해서 향상된 신호를 만들어낸다.

III. 음성 향상을 위한 손실 함수

3.1 기본 손실 함수

Fig. 1의 복소 네트워크를 훈련하기 위한 기본 손

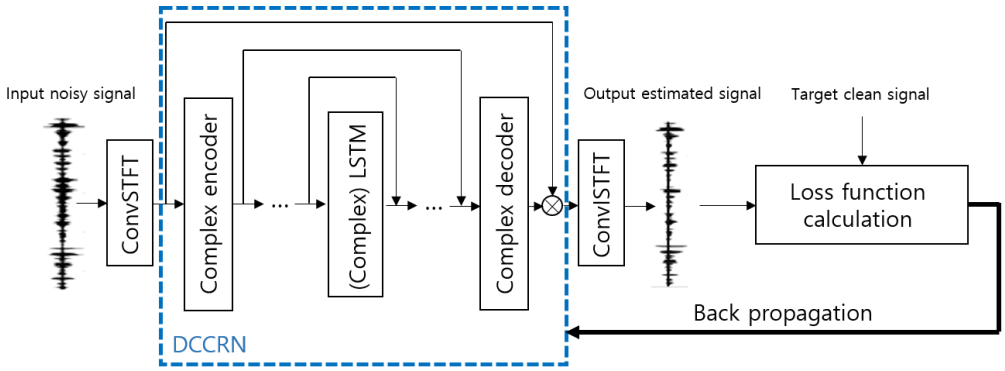

Fig. 1. (Color available online) The architecture of DCCRN.[3]

실함수로 MSE와 SI-SNR 두 가지를 고려하였다. 먼저 MSE 손실 함수는 음성을 포함하여 다양한 응용 분야에서 널리 사용되는 손실 함수이다. 수식은 다음과 같다.

$$L_{MSE} = \frac{1}{L}\|\hat{s} - s\|_2^2. \tag{2}$$

위 식에서 $\hat{s}$ 는 모델을 통해 향상된 음성을 의미하며, s 는 잡음이 없는 표적 음성을 의미한다. MSE는 단순하면서도 효과적인 손실 함수이나 앞서 언급한 것처럼 MSE만으로는 음질이나 음성의 명료도를 높이는데 한계가 있다.

SI-SNR 손실 함수는 SNR을 보완한 것으로 사용된 신호의 스케일에 따라 변하지 않는다는 특징을 가지고 있다. Reference [3]에서는 SI-SNR을 손실함수로 사용하여 DCCRN을 훈련함으로써 기존 방법보다 우수한 음질 향상 결과를 얻을 수 있음을 보였다. SI-SNR 수식은 다음과 같다.

$$L_{SI-SNR} = 10\log_{10}\left(\frac{\|s_{target}\|_2^2}{\|e_{noise}\|_2^2}\right). \tag{3}$$

이때, 타겟 음성 s_{target} 과 추정 잡음 e_{noise} 는 다음과 같이 계산된다.

$$s_{target} = <\hat{s}, s> \frac{s}{\|\hat{s}\|_2^2}, \tag{4}$$

$$e_{noise} = \hat{s} - s_{target}, \tag{5}$$

s 와 $\hat{s}$ 은 각각 깨끗한 음성과 모델을 통해 향상된 음성이다. $<\bullet, \bullet>$ 는 벡터 간의 스칼라 곱을 의미하며 $\|\bullet\|_2$ 는 L2 norm을 의미한다.

3.2 지각 기반 손실함수

Perceptual Metric for Speech Quality Evaluation (PMSQE)[7]는 음성 품질 평가를 위해 가장 널리 사용되는 측정 지표인 Perceptual Evaluation of Speech Quality (PESQ)에서 정하고 있는 라우드니스 기반 왜곡항을 단순화하여 미분할 수 있도록 변형함으로써 딥러닝 네트워크 훈련을 위해 제안된 손실함수이다. PMSQE는 단시간 스펙트럼 진폭에 대한 평균 오차 제곱 함수에 인간의 지각 정보를 함께 고려한다. PMSQE 손실 함수는 PESQ 라우드니스 기반 왜곡항으로부터 다음과 같이 구성된다.

$$L_{PMSQE} = \frac{1}{T}\sum_t (\alpha D_t^{(s)} + \beta D_t^{(a)}). \tag{6}$$

위 식에서 $D_t^{(s)}$ 는 모델을 통해 향상된 음성과 실제 타겟 음성 간의 대칭 교란[7]을 의미하고 $D_t^{(a)}$ 는 비대칭 교란[7]을 의미한다. α, β 는 각각 $D_t^{(s)}$ 와 $D_t^{(a)}$ 간의 가중치이다.

한편, Reference [8]에서는 오디오 부호화 과정에서 발생하는 양자화 잡음을 지각적으로 조절하기 위해 Log Mel Spectra(LMS)를 기반으로 하는 손실함수가

제시되었다.[8] LMS는 파워 스펙트로그램에 인간의 청각 구조를 기반으로 하는 Mel 필터뱅크를 적용하였으며, 해상도를 달리하는 다수의 Mel 필터뱅크를 결합하여 사용하였다. 수식은 다음과 같다.

$$L_{LMS} = \frac{1}{3}\sum_{i=1}^{3} \| M_i(\hat{s}),\ M_i(s) \|_2, \quad (7)$$

위 식에서 M_i는 i번째 Mel 밴드의 log 스펙트럼을 나타내며,[8] 본 논문에서는 각각 16, 32, 64개의 밴드를 갖는 3종류의 필터뱅크를 사용하였다. 결과적으로 Eq. (7)의 손실 함수는 $\hat{s}$과 s LMS 간의 L2 norm을 최소화하도록 한다.

3.3 손실 함수 결합

MSE와 SI-SNR은 시간 영역에서 손실 함수를 계산한다. 반면, PMSQE와 LMS는 주파수 영역에서 추정된 파워를 기반으로 계산된다.[7,8] 본 논문에서는 손실 함수 결합에 의한 성능 향상을 평가하기 위해 기본 손실함수(L_n)와 지각 기반 손실함수(L_p)를 결합하여 각 손실함수를 공동으로 학습하도록 하였다.

$$L_{joint} = \frac{\gamma_1 L_n + \gamma_2 L_p}{\gamma}. \quad (8)$$

위 식에서 $\gamma = \gamma_1 + \gamma_2$이며 γ_1과 γ_2는 각각 L_n과 L_p의 결합계수(coupling coefficients)를 의미하며, 각 손실함수의 동적영역을 고려하여 실험적으로 결정하였다.

IV. 실험 및 결과

각각의 손실 함수가 음성 향상 성능 변화에 어떤 영향을 주는지 확인하기 위해서, Fig. 1의 DCCRN 모델에 서로 다른 손실 함수를 사용하여 실험하였다.

4.1 실험 데이터 구성

동일한 딥러닝 모델에 다양한 손실 함수를 적용하는 경우, 공정한 성능 비교를 위해서는 각 손실 함수에 따라 최적화 조건을 달리해야 한다.[4] 이를 위해 본 논문에서는 각 손실 함수에 대해 최상의 성능을 보이는 학습률과 결합계수비를 실험적으로 결정하였다. 그 결과는 Table 1에 정리되어 있다.

Table 1. Learning rate and coupling coefficients ratio for the performance evaluation according to each loss function.

Loss function	learning rate	coupling constant ratio ($\gamma_1 : \gamma_2$)
MSE	10^{-3}	
MSE + LMS	10^{-3}	$10^3 : 1$
MSE + PMSQE	$5 \cdot 10^{-4}$	88 : 1
SI-SNR	10^{-3}	
SI-SNR + LMS	$5 \cdot 10^{-4}$	1 : 2
SI-SNR + PMSQE	$5 \cdot 10^{-4}$	1 : 10

실험 데이터는 16 kHz로 샘플링된 TIMIT 음성과 NoiseX-92, CHiME-2, CHiME-3, ETSI 잡음 데이터셋을 사용하여 생성하였다. 훈련은 3,696개의 잡음이 없는 음성에 11종류의 잡음 신호를 각각 SNR −10 dB에서 20 dB까지 총 25,872개의 데이터를 생성한 뒤 사용하였다. 검증 데이터는 훈련에 사용되지 않은 1,152개의 잡음이 없는 음성에 훈련에 사용한 것과 같은 잡음 신호를 SNR −10 dB에서 20 dB까지 무작위로 섞어 사용하였다. 테스트에는 훈련에 사용되지 않은 193개의 잡음이 없는 음성에 훈련에 사용된 잡음 신호와 훈련에 사용되지 않은 6종류의 생활 잡음 신호를 각각 검증 데이터와 같은 방식으로 만들어 사용하였다.

4.2 실험 결과 및 평가

모델의 성능을 평가하기 위해, 객관적 평가와 주관적 평가를 수행하였다. 객관적 평가 지표로는 음질 평가에서 가장 많이 사용되는 PESQ와 Short-Time Objective Intelligibility(STOI)를 사용하였다. 주관적 평가로는 성인 7명을 대상으로 청취 테스트를 수행하였다.

객관적 지표에 따른 성능 평가 결과는 Tables 2, 3과 같다. Table 2는 학습에 사용한(seen) 잡음 신호를 사용하여 측정한 결과이며, Table 3는 학습에 사용되지 않은(unseen) 잡음 신호를 사용하여 측정한 결과이다. 가장 높은 PESQ와 STOI값을 굵은 글씨로 나타내

Table 2. Performance evaluation of various loss functions using seen noise.

SNR	Metric	Loss function					
		MSE	MSE + LMS	MSE + PMSQE	SI-SNR	SI-SNR + LMS	SI-SNR + PMSQE
-10 dB	PESQ	1.3867	1.4080	**1.4808**	1.4286	1.4567	1.4520
	STOI	0.6907	0.6927	**0.6958**	0.6865	0.6879	0.6675
-5 dB	PESQ	1.6104	1.6563	**1.7445**	1.6679	1.6932	1.7085
	STOI	0.8002	0.8022	0.8031	0.8072	**0.8110**	0.7860
0 dB	PESQ	1.9767	2.0612	**2.2071**	2.0634	2.1065	2.1661
	STOI	0.8802	0.8824	0.8830	0.8905	**0.8933**	0.8746
5 dB	PESQ	2.3770	2.4883	**2.6487**	2.4905	2.5600	2.6114
	STOI	0.9261	0.9267	0.9282	0.9316	**0.9378**	0.9253
10 dB	PESQ	2.8741	3.0011	**3.1430**	3.0128	3.1180	3.1423
	STOI	0.9580	0.9596	0.9587	0.9610	**0.9675**	0.9581
15 dB	PESQ	3.2970	3.4051	3.4964	3.4420	**3.5277**	3.5126
	STOI	0.9759	0.9761	0.9756	0.9808	**0.9820**	0.9764
20 dB	PESQ	3.6440	3.7691	3.8167	3.8221	**3.8878**	3.8555
	STOI	0.9881	0.9882	0.9877	0.9899	**0.9912**	0.9886

Table 3. Evaluation of various loss functions using unseen noise.

SNR	Metric	Loss function					
		MSE	MSE + LMS	MSE + PMSQE	SI-SNR	SI-SNR + LMS	SI-SNR + PMSQE
-10 dB	PESQ	1.3793	1.3865	**1.4686**	1.4223	1.4492	1.4299
	STOI	0.7193	0.7364	0.7333	0.7371	**0.7463**	0.7013
-5 dB	PESQ	1.7561	1.7248	**1.8731**	1.7744	1.8298	1.8241
	STOI	0.8362	0.8405	0.8381	0.8427	**0.8538**	0.8255
0 dB	PESQ	2.1339	2.1852	**2.3646**	2.2316	2.3343	2.3030
	STOI	0.9085	0.9074	0.9097	0.9113	**0.9232**	0.9036
5 dB	PESQ	2.4861	2.5769	**2.7649**	2.6507	2.6973	2.6930
	STOI	0.9414	0.9429	0.9427	0.9497	**0.9531**	0.9394
10 dB	PESQ	2.9340	3.0625	3.2372	3.1417	**3.2507**	3.1907
	STOI	0.9689	0.9690	0.9684	0.9749	**0.9754**	0.9683
15 dB	PESQ	3.2863	3.3959	3.5487	3.4913	**3.5926**	3.5241
	STOI	0.9808	0.9813	0.9802	0.9840	**0.9851**	0.9812
20 dB	PESQ	3.6087	3.6980	3.8501	3.8143	**3.9000**	3.8302
	STOI	0.9890	0.9900	0.9892	0.9901	**0.9925**	0.9903

었다.

먼저 SI-SNR을 손실 함수로 사용하였을 때 MSE를 사용한 경우보다 더 높은 PESQ와 STOI를 보인다. 또한 MSE 혹은 SI-SNR에 지각 기반 손실 함수인 LMS 또는 PMSQE를 결합하면, PESQ가 개선됨을 확인할 수 있다. Table 2(seen 데이터)에서 MSE와 LMS를 결합하면 STOI 값은 큰 차이를 보이지 않으나 PESQ 값은 평균 0.07 이상 향상됨을 확인할 수 있다. MSE와 PMSQE를 결합하는 경우, PESQ 값은 평균적으로 0.21 이상 크게 향상되어 비교된 손실함수 중 가장 좋은 수치를 보인다. 한편 SI-SNR와 LMS를 결합하면 SI-SNR에 비해 PESQ 값이 평균 0.05 이상 향상되며, SI-SNR과 PMSQE를 결합하면 PESQ 값이 평균적으로 0.07 이상 향상됨을 볼 수 있다.

결과적으로 PESQ 측면에서 볼 때, MSE 혹은 SI-SNR과 PMSQE를 결합하여 사용할 때 다른 경우보다 평균적으로 더 높은 점수를 얻음을 확인할 수 있다. 그러나 출력음성을 좀 더 세밀히 관찰하였을 때 결과는 이와 상이하였다. Fig. 2에 여러 가지 조합의 손실함수로 얻어진 출력의 스펙트로그램을 비교하였다. 먼저, MSE 혹은 SI-SNR과 PMSQE를 공동으로 학습한 경우, 높은 PESQ에도 불구하고 표적 음성 이후 잡음(박스로 표시한 부분)의 일부를 완전히 제거되지 못하고 있다. 한편 SI-SNR만 사용한 경우에는 음성

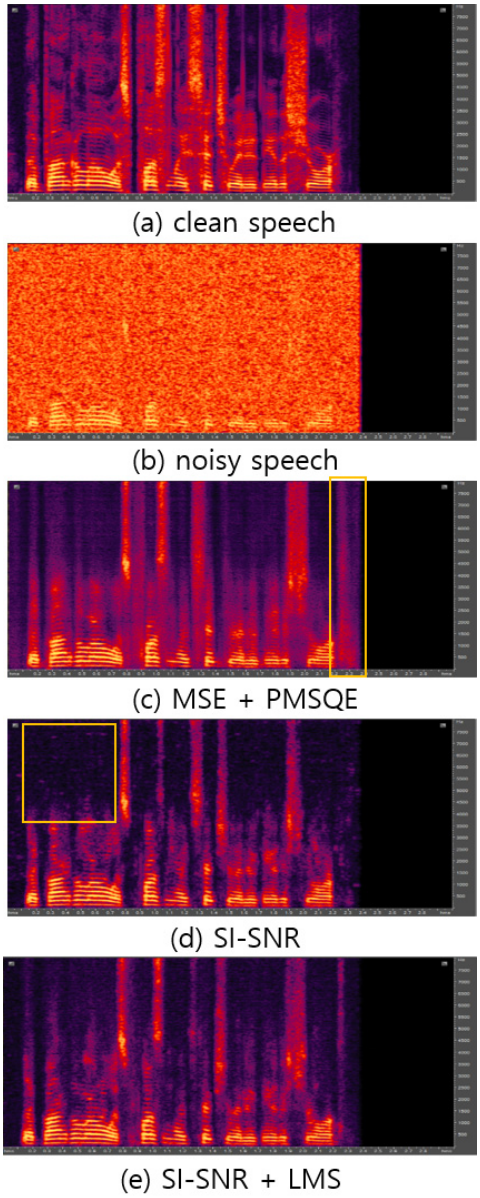

Fig. 2. (Color available online) The spectrograms of (a) clean speech, (b) noisy speech at 0 dB SNR, estimated speeches using (c) MSE and PMSQE, (d) SI-SNR, (e) SI-SNR and PMSQE, (f) SI-SNR and LMS.

초기 고주파 성분이 충실히 복원되지 않고 있다. 이러한 결과가 전반적인 음질 차이로 반영되는지를 확인하기 위해 Comparative Mean Opinion Score(CMOS) 청취 실험[9]을 수행하였다.

청취 실험에는 7명의 경험자가 참여하였으며, 실험 참가자들은 총 35쌍의 향상된 음성 한 쌍씩 듣고 −3에서 3까지 점수를 매겼다. 이때, −3은 첫 번째 음성이 두 번째 음성보다 월등히 우수, 0은 비슷, +3은

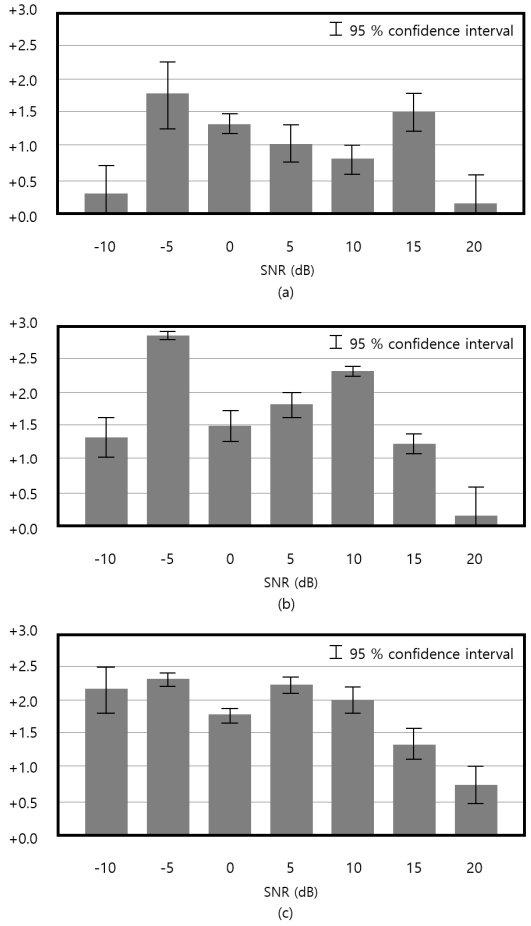

Fig. 3. The listening test results: CMOS scores of SI-SNR + LMS with respect to (a) SI-SNR, (b) MSE + PMSQE, (c) SI-SNR + PMSQE. Positive values indicate SI-SNR+LMS is better than the compared loss functions.

두 번째 음성이 월등히 우수함을 의미한다. 모든 실험에서 두 번째 음성은 SI-SNR+LMS 손실함수를 사용하여 얻은 음성이다. Fig. 3(a)는 SI-SNR과 (SI-SNR + LMS)를 비교한 결과이다. 모든 SNR에서 (SI-SNR + LMS)가 더 우수한 평가를 얻었으며, 특히 −5 dB와 0 dB SNR에서는 1.5 이상의 높은 점수를 얻었다. Fig. 3(b)는 가장 높은 PESQ를 얻었던 (MSE + PMSQE)와 (SI-SNR + LMS)를 비교한 결과이다. 20 dB SNR 경우를 제외하면 (SI-SNR + LMS)이 1.0점 이상의 높은 점수를 얻었다. 이는 PESQ 수치와 배치되는 결과이다. Fig. 3(c)는 (SI-SNR + PMSQE)와 비교한 결과로서, 마찬가지로 (SI-SNR + LMS)가 대부분의 SNR에서 1.5 이상의 높은 점수를 얻었으며, −5 dB SNR의 경우 2.8

점의 큰 차이를 확인할 수 있다. Tables 2와 3에서는 (SI-SNR+PMSQE)가 (SI-SNR+LMS)에 비해 더 높은 PESQ를 보이는 경우가 있었던 점에 비추어보면 이 또한 PESQ값과 배치되는 결과이다.

결과적으로 PMSQE를 기본 손실함수와 결합하였을 때 가장 높은 PESQ 점수를 보인 반면, 실제 주관적인 음질 평가에서는 SI-SNR과 LMS를 결합한 손실함수로 얻은 음질이 가장 우수함을 확인하였다. 이는 PESQ 점수가 주관적인 음질을 정확히 반영하지 못한다는 최근의 연구 사례[10]와 일치하는 결과이다.

V. 결 론

본 논문은 복소 네트워크를 사용하여 다양한 손실 함수들에 대한 성능을 비교 평가 하였다. 또한, 기존 손실 함수가 가지는 한계를 보완하기 위하여 기존 손실 함수에 지각 기반 손실 함수를 결합하여 그 성능을 비교 분석하였다. 지각 기반 손실 함수를 결합함으로써 PESQ 점수가 전반적으로 향상됨을 확인할 수 있었으며, CMOS 청취 실험 결과 SI-SNR과 LMS를 결합한 손실함수의 성능이 가장 우수함을 확인할 수 있었다.

References

1. H. Zhao, S. Zarar, I. Tashev, and C. Lee, “Convolutional recurrent neural networks for speech enhancement,” Proc. IEEE ICASSP. 2401-2405 (2018).
2. D. S. Williamson, Y. Wang, and D. Wang, “Complex ratio masking for monaural speech separation,” IEEE/ACM Trans. on audio, speech, and Lang. Pross. **24**, 483-492 (2015).
3. Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, “Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement,” arXiv:2008.00264 (2020).
4. M. Kolbk, Z. Tan, S. H. Jensen, and J. Jensen, “On loss functions for supervised monaural time-domain speech enhancement,” IEEE/ACM Trans. on Audio, Speech, and Lang. Pross. **28**, 825- 838 (2020).
5. S. Braun and I. Tashev, “A consolidated view of loss functions for supervised deep learning-based speech enhancement,” arXiv:2009.12286 (2020).
6. S. Fu, C. Liao, and Y. Tsao, “Learning with learned loss function: Speech enhancement with quality-net to improve perceptual evaluation of speech quality,” IEEE Signal Processing Letters, **27**, 26-30 (2020).
7. J. M. Mart´ın-Donas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, “A deep learning loss function based on the perceptual evaluation of the speech quality,” IEEE Signal Processing Letters, **25**, 1680-1684 (2018).
8. S. Kankanahalli, “End-to-end optimized speech coding with deep neural networks,” Proc. IEEE ICASSP. 2521-2525 (2018).
9. ITU-T. Rec. P.800, *Methods for Subjective Determination of Transmission Quality*, E 9713, 1996.
10. W. A. Jassim, J. Skoglund, M. Chinen, and A. Hines, “Speech quality factors for traditional and neural-based low bit rate vocoders,” arXiv:2003.11882 (2020).

저자 약력

▸ 황 서 림 (Seo-Rim Hwang)

2017년 3월 ~ 현재 : 연세대학교 컴퓨터정보통신공학부 학사 과정

▸ 변 준 (Joon Byun)

2017년 2월 : 연세대학교 컴퓨터정보통신공학부 학사
2017년 3월 ~ 현재 : 연세대학교 전산학과 석박사통합과정

▸ 박 영 철 (Young-Cheol Park)

1986년 2월 : 연세대학교 전자공학과 학사
1988년 2월 : 연세대학교 전자공학과 석사
1993년 2월 : 연세대학교 전자공학과 박사
2002년 3월 ~ 현재 : 연세대학교 컴퓨터정보통신공학부 교수