Mapping and Masking Targets Comparison using Different Deep Learning based Speech Enhancement Architectures

Soha A. Nossier
Dept. of Engineering and Computing
University of East London
London, UK
soha.abdallah.nossier@gmail.com

Julie Wall
Dept. of Engineering and Computing
University of East London
London, UK
j.wall@uel.ac.uk

Mansour Moniri
Dept. of Engineering and Computing
University of East London
London, UK
m.moniri@uel.ac.uk

Cornelius Glackin
Intelligent Voice Ltd
London, UK
neil.glackin@intelligentvoice.com

Nigel Cannings
Intelligent Voice Ltd
London, UK
nigel.cannings@intelligentvoice.com

***Abstract*—Mapping and Masking targets are both widely used in recent Deep Neural Network (DNN) based supervised speech enhancement. Masking targets are proved to have a positive impact on the intelligibility of the output speech, while mapping targets are found, in other studies, to generate speech with better quality. However, most of the studies are based on comparing the two approaches using the Multilayer Perceptron (MLP) architecture only. With the emergence of new architectures that outperform the MLP, a more generalized comparison is needed between mapping and masking approaches. In this paper, a complete comparison will be conducted between mapping and masking targets using four different DNN based speech enhancement architectures, to work out how the performance of the networks changes with the chosen training target. The results show that there is no perfect training target with respect to all the different speech quality evaluation metrics, and that there is a tradeoff between the denoising process and the intelligibility of the output speech. Furthermore, the generalization ability of the networks was evaluated, and it is concluded that the design of the architecture restricts the choice of the training target, because masking targets result in significant performance degradation for deep convolutional autoencoder architecture.**

***Index Terms*—Deep Learning, Speech Enhancement, Training Targets, Time-Frequency Mapping, Time-Frequency Masking**

I. Introduction

Speech enhancement is one of signal processing's most challenging tasks that has shown great improvement through deep learning [1]. It is the process of separating clean speech from background noise in order to improve speech perception. The idea of supervised deep learning based speech enhancement is to develop an algorithm that can learn the mapping function that maps noisy speech to clean speech. This is achieved by feeding the algorithm with a huge dataset of pairs of noisy and clean speech during the training process, and then the trained network is expected to output an enhanced clean speech signal with better quality and intelligibility [2].

The time domain audio signal is usually transferred to the frequency domain before being fed to the DNN. This transformation is useful in obtaining more meaningful features about the speech signal, such as the harmonics and its relative amplitudes, which make the learning process easier and more generalized [3]. This transformation results in a Time-Frequency (T-F) representation of the signal in the form of a spectrogram or cochleagram [4]. Many features can be extracted from these two forms to act as input to the DNN, while the output or the target in the case of the speech denoising process can be one of two forms: a spectrogram or cochleagram of the clean speech signal; or a spectrographic mask, which will be discussed later. Hence the training targets are divided into two types: mapping and masking targets [5]. These two training target types arise from the fact that the supervised speech enhancement problem can be seen from two perspectives: it can be defined as a regression problem if our target is directly mapping to a clean speech T-F representation, or a classification problem if our target is to produce a matrix, known as a mask, that classifies every portion of the signal either as speech or noise, and then by weighting, or filtering the noisy speech with this mask, the enhanced clean speech signal can be generated [2].

Although masking targets were proved to produce more intelligible speech for MLP architectures [5]–[7], other studies [8], [9] claimed that mapping targets outperform masking targets, especially at low Signal to Noise Ratio (SNR). For that reason, both mapping and masking approaches have both been

This research is sponsored by Intelligent Voice Ltd.

used in recent DNN based speech enhancement research. However, no comprehensive comparison was made between the two approaches using different speech enhancement architectures to show how the performance is affected by the used training target. The aim of this paper is to give an overview of the two approaches for speech enhancement, and then make a complete comparison between them using four different, recent and best performing DNN based speech enhancement architectures.

The rest of this paper is organized as follows. Section II and III cover mapping and masking approaches, and the different training target types that are based on each approach. The experimental work is presented in Section IV. In Section V, the results are illustrated and discussed. Finally the conclusion is given in Section VI.

II. Time-Frequency Mapping (T-F Mapping)

Mapping based targets are based on either the spectrogram of the speech signal, which is created by performing Short Time Fourier Transform (STFT) to the time domain audio signal [10], or the cochleagram of the speech signal, which is created by time windowing responses of a filterbank representing the frequency analysis of the cochlea [11]. A study [9] that compares masking and mapping targets reported that the mapping approach is less sensitive to SNR variations, so it will be useful for applications where a wide range of SNRs are expected. In the following subsections, the two mapping based targets will be discussed.

A. Spectrogram based T-F Mapping Targets

When using this approach, the target is to map to the clean speech spectrogram, which is obtained using STFT. The STFT operation is applied to the noisy speech, and then the magnitude of the STFT is the feature used in the training process. In order to reconstruct the speech signal back to the time domain, the inverse operation (ISTFT) is performed [12]. In most research that is based on this approach, the phase of the noisy speech is kept, so as to be used in the reconstruction process of the audio based on the assumption that the phase is not sensitive to the noise [13], so the phase of the noisy speech is approximately the same as that of the clean speech. However, some studies managed to develop techniques to retrieve the clean phase, as they conversely believe that it will have a positive impact on the general quality of the output speech [14]. Some of these techniques are based on separate algorithms for retrieving the clean phase from the noisy speech [15], while others are based on using the complex spectrogram in the training process [16], [17].

There are other features that can be extracted from the spectrogram, such as the power spectrum, which shows the distribution of the power of the frequency components of the speech; Mel spectrum, which represents the spectrum in the Mel scale; and log power spectrum, in which the log operation is performed to the power spectrum in order to decrease the dynamic range, and ease the training process [18]. Mel-Frequency Cepstral Coefficients (MFCC) is another feature extracted by applying a Discrete Cosine Transform (DCT) to the log-compressed Mel scale power spectrum. Perceptual Linear Prediction (PLP) is also a feature that can be created from the spectrogram by extracting spectral characteristics that match that of the human auditory system, and discards any information irrelevant to the speech signal [19].

Many DNN architectures for speech enhancement were found in the literature to be based on spectrogram based T-F mapping targets [20]–[24].

B. Cochleagram based T-F Mapping Targets

The target here is mapping to the clean speech cochleagram, instead of the spectrogram. Defining the signal on the cochleagram is achieved by passing the signal through a number of Gammatone filters in order to extract specific characteristics for the signal at different frequencies, leading to a time frequency representation of the signal similar to the one obtained from the STFT, called Gammatone Frequency Target Power Spectrum (GF-TPS) [25]. Eq. (1) represents the impulse response of the gammatone filter in the time domain:

$$g(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi f_c t + \phi), \tag{1}$$

where the constant *a* is the amplitude that controls the gain, *t* is the time, *n* is the order of the filter, f_c is the central frequency of the filter, and ϕ is the phase. Eq. (2) defines *b*, which is the decay factor determining the filter bandwidth:

$$b = 1.019 * 24.7(4.73\frac{f_c}{1000} + 1). \tag{2}$$

The gammatone filterbank is created by changing the center frequency f_c of the filter in the above equations. There are many features that can be extracted based on the Gammatone Frequency (GF) feature, such as Gammatone Frequency Cepstral Coefficients (GFCC) [26], which is calculated by applying the DCT to the GF feature. Gammatone Frequency Modulation Coefficients (GFMC) [27] is another feature based on GFCC. Multiresolution Cochleagram (MRCG) [28] and Pitch-Based Feature (PITCH) [2] are other features that can be created based on the cochleagram. Some features are also found to integrate both cochleagram and spectrogram, such as Gabor Filterbank Feature (GFB) [29] and Power-Normalized Cepstral Coefficients (PNCC) [30].

There are many speech enhancement networks that are based on the use of cochleagrams in representing the speech signal [5], [31]–[33]. Although the use of a cochleagram might not be as accurate as STFT in reconstructing the time domain speech signal due to the absence of a direct inverse process, research in the literature proves that it will result in better performance because of its better representation of the speech characteristics [4], [5]. The reconstruction of the time domain speech signal when using a cochleagram is done indirectly using an idea dating from 1983 [34]. Researchers first used cochleagrams in the field of Computational Auditory Scene Analysis (CASA) in order to separate sound sources by segmenting the cochleagram into regions belonging to each sound source. These regions are then grouped into streams to form a binary matrix of 1 or 0 weights for different sound

sources, and then this matrix of weights is applied to the mixture to separate the target sound, which corresponds to the 1s weights in the matrix [35]. Based on this idea, the matrix of weights can be extracted from the noisy speech cochleagram and the estimated clean one, and then the speech signal can be re-synthesized by weighting any T-F representation of the mixture signal with this matrix. This matrix of weights is actually the spectrographic or T-F mask, which will be discussed in the following section, so when using cochleagram based mapping targets, speech re-synthesis is done indirectly through a spectrographic mask.

III. Time-Frequency Masking (T-F Masking)

As discussed in the previous section, the idea of T-F masking is not new and it has been applied in the CASA field. Recent research in DNN based speech enhancement used this approach to deal with the speech de-noising process as a supervised deep learning classification problem [36]–[38]. There are two basic types of T-F masks: Binary Masking and Soft Masking [39]. In Binary Masking, the frequency bins that are likely to belong to the target signal are set to 1, while other bins are set to 0, assuming sparseness and disjointness of the two signals in the mixture. Sparseness means that most of the T-F bins have low energy, while disjointness means that the T-F bins of the two signals in the audio mixture do not overlap [40]. For soft masking, each bin is set to a probability value between 0 and 1, based on how much it is likely to belong to the target signal. Soft masking is used for mixtures in which the earlier discussed assumptions for binary masks are not fulfilled [41]. In the following subsections, an overview of different masking targets will be presented.

A. Ideal Binary Mask (IBM)

This was one of the first binary masks used in supervised speech separation. In this mask, portions of the spectrogram that have a high noise intensity are set to 0, while others with higher speech amplitude are set to 1 [42]. Eq. (3) defines the IBM:

$$IBM(t,f) = \begin{cases} 1, & if\, SNR(t,f) > LC \\ 0, & otherwise \end{cases}, \tag{3}$$

where t and f denote time and frequency, respectively. *LC* is the local criterion or threshold that the classification to 1 or 0 is based on. This value should be chosen based on practical trials; but in the literature, it is kept to 5 dB lower than the SNR of the mixture so as to preserve enough speech information [5].

Another version of IBM can be defined, which is independent of the noise in the mixture, as in this type of mask the *SNR(t, f)*, in Eq. (3), is redefined using the target speech energy in each T-F unit and the average spectral energy of a reference Speech-Shaped Noise (SSN) instead of the local noise energy. This mask is known as Target Binary Mask (TBM) [43], and it has been also used in many speech enhancement approaches.

B. Ideal Ratio Mask (IRM)

Ideal Ratio Mask is a soft masking target that has proved to be very efficient in the speech enhancement process, as it results in higher speech intelligibility [44]. A reason that this type outperforms IBM in the speech enhancement process is the complexity of the noise speech mixture, in which the assumptions for binary masking are not always fulfilled. The IRM is presented below in Eq. (4):

$$IRM(t,f) = \left(\frac{S(t,f)^2}{S(t,f)^2 + N(t,f)^2}\right)^{\beta}, \tag{4}$$

where $S(t,f)^2$ and $N(t,f)^2$ denote the speech and noise energy, respectively, in a particular T-F unit. β is a tunable parameter to scale the mask.

C. Complex Ideal Ratio Mask (cIRM)

With the introduction of research that has shown the importance of retrieving the clean phase instead of using the noisy one, the idea of cIRM was proposed in order to be used as a target for supervised speech enhancement [38]. The speech enhancement network in this case is supposed to enhance both the magnitude and the phase during the training process, which results in a better clean speech reconstruction, although it might be less effective in removing noise than the normal IRM [45]. The STFT in this masking target type is expressed in the Cartesian coordinates so as to give a meaningful phase representation that can be used in the training process. Eq. (5) defines the cIRM so that when it is applied to the noisy complex spectrum, it produces a clean complex spectrum:

$$cIRM = \frac{Y_r S_r + Y_i S_i}{Y_r^2 + Y_i^2} + i\frac{Y_r S_i - Y_i S_r}{Y_r^2 + Y_i^2}, \tag{5}$$

where Y_r and Y_i are the real and imaginary parts of the noisy speech, respectively, and S_r and S_i are the real and imaginary parts of the clean speech, respectively. In practice, cIRM is expressed in a compressed format in order to be bounded to ensure training stability, and the work in [38] and [45] defined these compression techniques. Afterwards, the estimated compressed mask is decompressed and multiplied by the noisy spectrum to produce the clean complex spectrum.

D. Spectral Magnitude Mask (SMM)

This mask, which is also called FFT-Mask, takes advantage of mapping and masking targets together by applying a mapping target in the form of a masking approach. In this type of masking target, the STFT magnitude of the clean speech is divided by the STFT of the noisy speech so as to generate a mask, when multiplied with the noisy speech signal, the result will be clean speech only [5]. Like cIRM, this mask is not bounded by 0 and 1 so a truncation to the high values should be applied for the stability of the training process [2], [5]. The definition of SMM is expressed in Eq. (6):

$$SMM(t,f) = \frac{|S(t,f)|}{|N(t,f)|}, \tag{6}$$

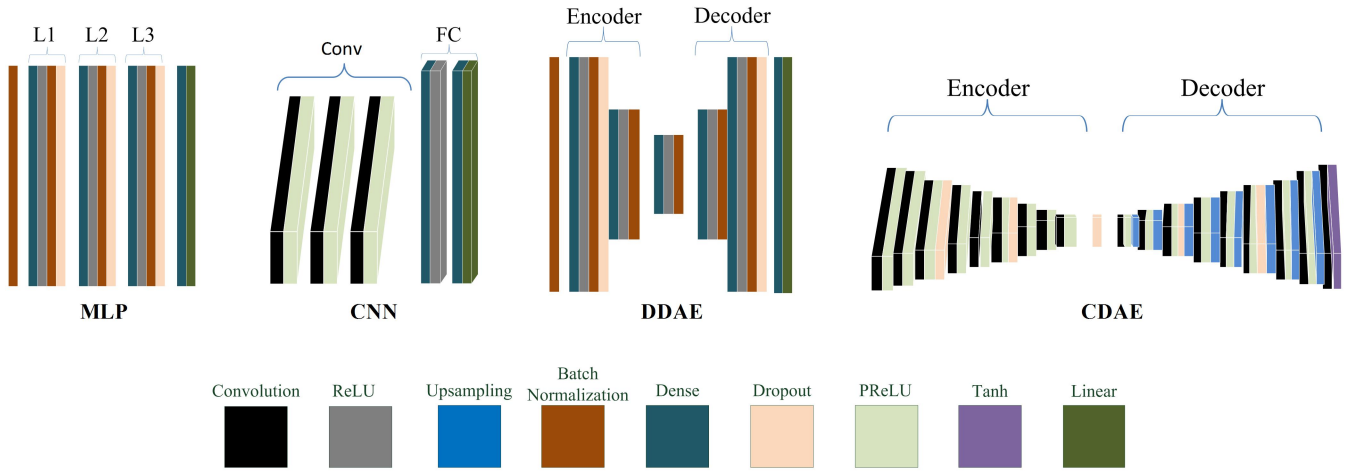

Fig. 1: The four implemented Deep Neural Network architectures

where $|S(t, f)|$ and $|N(t, f)|$ are the magnitude spectrum of the clean and noisy speech, respectively. There is another version of this type of mask named Phase-Sensitive Mask (PSM) [46], in which the SMM mask is multiplied by the cosine of the phase difference between noisy and clean speech, and this was reported to have a positive impact on the overall performance.

IV. Experimental Work

In this section, details of the technical work done will be presented, by illustrating the followed procedure to re-implement and compare different DNN speech enhancement architectures using mapping and masking targets. This illustration is given in the following subsections.

A. Datasets

Two clean speech datasets were used in this work to train and test the DNNs. The first one is the Voice Bank corpus [47], which consists of 400 English sentences for each of 28 English speakers, 14 male and 14 female, and another 56 different accent speakers, 28 male and 28 female, from Scotland and the United States. Five hours of clean speech were randomly selected from this dataset for training purposes, and another 30 minutes clean speech for testing purposes. The other clean speech dataset is the LibriSpeech corpus [48], which consists of 1,000 hours of English speech with various accents, derived from audio books. The LibriSpeech corpus was not used in the training, but 30 minutes of clean speech were randomly selected from this dataset to test the networks' generalization ability. On the other hand, a variety of noise environments were used to generate the noisy audio. For training, a total of 105 noise environments were used, 90 were collected from the 100 Noise Environment dataset [49] and 15 from the NOISEX-92 corpus [50]. While for testing, 20 noise environments were used, half seen and the other half unseen during the training, taken from the 100 Noise Environment dataset. The seen noise environments are used to test the network's ability to remove the noise, while the unseen noise environments are used to evaluate the generalization ability of the network. The noise environments used in the testing process are a mix of human generated noise, such as crying sounds, yawning sounds, and human crowd sounds, and other non human generated noise, such as AWGN, phone dialing, shower noise, tooth brushing, and wood creaks.

B. Audio Preprocessing

Before feeding the audio signals to the DNN, an 8 kHz down-sampling operation was applied to all audio, and the noise and speech intensity were adjusted to be the same, for the training to be done at 0 dB SNR. A transformation to the frequency domain was then performed by calculating the STFT of the audio signals, using a Hamming window with 256 frame size and 50% overlap, and FFT size of 256. The results of this signal preprocessing are spectrogram based T-F representations to be used as input features to the DNN. It should be mentioned here that the choice of spectrogram features was made so as to make a fair comparison between mapping and masking, due to the fact that cochleagram mapping involves the use of a masking target for speech re-synthesis.

C. Speech Enhancement Architectures

This comparison is based on the evaluation of four different speech enhancement architectures, shown in Fig. 1. The basic MLP architecture is used as the most commonly used architecture in speech enhancement research [21]. The architecture is based on three hidden layers with Rectified Linear Unit (ReLU) activation and an output layer with linear activation for prediction. All the hidden layers are followed by a Batch normalization layer for training stability, and a dropout layer of 20% rate to avoid overfitting to the training data. A Convolutional Neural Network (CNN) architecture is also used, as it is proved to be very powerful in speech enhancement [37]. The architecture was modified in our work to have three 1D convolutional layers with Parametric Rectified Linear Unit

(PReLU) activation, and two other fully connected layers at the end with ReLU and linear activation for predicting the output. Another two autoencoder based architectures were used in this work. The first is a Deep Denoising Autoencoder (DDAE) architecture [22] with an encoder and decoder network, each has 2 fully connected layers with ReLU activation functions and batch normalization, and a bottleneck middle fully connected layer with ReLU activation and 180 hidden units. 2,048 and 500 hidden units were used in the first and second hidden layers of the encoder, respectively, and vice versa for the hidden layers of the decoder. A dropout technique of 10% rate was applied to the first and last layers of the encoder and decoder networks, respectively. The final fully connected layer with linear activation was used to predict the output. The second autoencoder architecture is a Convolutional Denoising Autoencoder (CDAE) [51] with nine 1D convolutional layers with PReLU activations. Strided convolution was used in the encoder network, while upsampling was used in the decoder. Skip connections were included in this network to prevent the loss of important information, which may happen due to the deep nature of the architecture.

D. Training Targets

For the mapping based approach, the magnitude spectrogram was used as the training target. Although a cochleagram mapping was proved to be better, the use of spectrogram based mapping will result in a fair comparison, because the speech reconstruction operation for cochleagram is done using masking targets. While for the masking based approach, IRM and SMM were both used, because they were proved to be the two best performing masking targets. β is set to 0.5 for the IRM, as the default used value in most practice, in Eq. (4). Due to the fact that the SMM is not bounded, the SMM values were truncated to 10 to make the training process more stable, as suggested by [5].

E. Evaluation Metrics

Four different evaluation metrics were used in this comparison: Perceptual Evaluation of Speech Quality (PESQ) [52], Short Time Objective Intelligibility (STOI) [53], Log Spectral Distortion (LSD), and Segmental Signal to Noise Ratio increase (ΔSSNR). All these metrics are based on comparing the clean speech audio with the processed one coming from the network. PESQ is an objective method of measuring speech quality, its score ranges from -0.5 to 4.5 and the higher the score, the better the speech quality. STOI is another measure that evaluates the intelligibility of the enhanced speech after removing the noise, which means how many words could be interpreted from the processed speech. It ranges from 0 to 1, and the higher the value, the better the speech intelligibility. LSD is a measure of the distortion in the processed speech so it should be kept as minimum as possible; while ΔSSNR increase shows the ability of the network in removing the noise. Spectrograms for the noisy, clean, and processed speech were also used to visually compare between the two approaches.

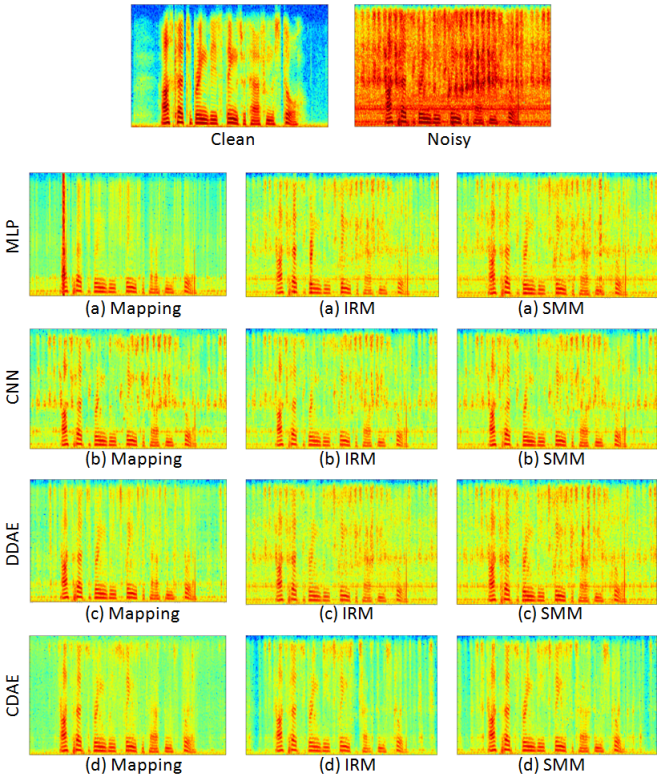

Fig. 2: Spectrograms of the clean speech, noisy speech with tooth brushing noise at 0dB, and output processed speech by MLP (a), CNN (b), DDAE (c), and CDAE (d) using spectrogram mapping, IRM, and SMM.

V. Results and Discussion

The results of the experiments are shown in Table I to IV, which represent the networks' performance with respect to the four evaluation metrics: PESQ, STOI, LSD, and ΔSSNR, respectively. Regarding the speech quality (PESQ score), at very high SNR, 20 and 15 dB, the masking based approaches are generating speech with better quality for all architectures. However, mapping based approaches managed to output better quality speech at low SNR, and this is significant in the DDAE architecture. Moreover, the standard deviation (SD) for the mapping approach in all architectures is lower, which means that this approach is more sustainable. Masking based targets outperform mapping based targets with respect to speech intelligibility (STOI score) for all architectures. Furthermore, the SMM, specifically, generates enhanced speech with the least distortion. The increase in SSNR is relatively high for both approaches. By comparing the architectures' output with the input noisy speech, the fully connected networks (MLP and DDAE) generated worse speech quality at high SNRs (20 and 15 dB) in the case of a mapping target. At the same time, speech intelligibility shows no improvement in the case of both mapping and masking approaches at high SNRs for all architectures, except the CDAE.

The results also prove that the speech quality evaluation metrics are very sensitive to any change, because there is no

TABLE I: Speech Quality Results (PESQ)

SNR		20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	AVG	SD
Noisy		2.92	2.62	2.32	2.04	1.81	1.60	2.219	0.498
MLP	MAP	2.41	2.34	2.25	2.16	2.02	1.70	2.147	0.258
	SMM	3.04	2.79	2.57	2.35	2.09	1.75	**2.433**	0.469
	IRM	2.97	2.75	2.54	2.33	2.05	1.70	2.388	0.465
CNN	MAP	3.09	2.90	2.68	2.46	2.21	1.87	2.537	0.449
	SMM	3.12	2.92	2.71	2.47	2.19	1.87	2.546	0.469
	IRM	3.15	2.94	2.72	2.48	2.21	1.88	**2.564**	0.470
DDAE	MAP	2.82	2.72	2.58	2.41	2.19	1.83	2.424	0.368
	SMM	3.03	2.78	2.55	2.32	2.05	1.73	2.411	0.477
	IRM	3.06	2.80	2.56	2.34	2.08	1.75	**2.430**	0.478
CDAE	MAP	2.93	2.81	2.68	2.52	2.32	2.01	2.543	0.339
	SMM	3.19	3.01	2.83	2.62	2.38	2.04	**2.680**	0.422
	IRM	3.19	3.00	2.80	2.61	2.38	2.03	2.667	0.424

TABLE II: Speech Intelligibility Results (STOI)

SNR		20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	AVG	SD
Noisy		0.91	0.88	0.83	0.78	0.71	0.64	0.790	0.101
MLP	MAP	0.82	0.81	0.79	0.77	0.73	0.65	0.760	0.063
	SMM	0.89	0.87	0.83	0.80	0.75	0.68	**0.804**	0.078
	IRM	0.89	0.86	0.83	0.80	0.75	0.68	0.801	0.078
CNN	MAP	0.88	0.86	0.83	0.79	0.74	0.67	0.795	0.078
	SMM	0.89	0.86	0.83	0.80	0.75	0.67	0.800	0.079
	IRM	0.89	0.87	0.84	0.80	0.76	0.68	**0.808**	0.077
DDAE	MAP	0.85	0.83	0.81	0.79	0.75	0.68	0.785	0.062
	SMM	0.90	0.87	0.83	0.80	0.75	0.68	0.804	0.080
	IRM	0.90	0.88	0.85	0.81	0.76	0.69	**0.814**	0.078
CDAE	MAP	0.89	0.87	0.85	0.82	0.78	0.72	0.820	0.064
	SMM	0.91	0.89	0.86	0.83	0.79	0.72	0.832	0.071
	IRM	0.91	0.89	0.87	0.83	0.79	0.72	**0.834**	0.071

TABLE III: Log Spectral Distortion Results (LSD)

SNR		20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	AVG	SD
Noisy		1.36	1.62	1.92	2.21	2.46	2.62	2.032	0.489
MLP	MAP	1.05	1.12	1.18	1.22	1.32	1.68	**1.261**	0.225
	SMM	0.96	1.10	1.20	1.30	1.49	1.82	1.312	0.306
	IRM	1.05	1.18	1.26	1.33	1.51	1.85	1.362	0.285
CNN	MAP	1.09	1.18	1.30	1.44	1.64	1.98	1.438	0.330
	SMM	0.97	1.10	1.25	1.42	1.64	1.95	**1.389**	0.363
	IRM	0.97	1.11	1.27	1.44	1.67	2.00	1.411	0.378
DDAE	MAP	1.23	1.28	1.32	1.40	1.54	1.85	1.437	0.230
	SMM	1.01	1.15	1.26	1.35	1.53	1.82	**1.354**	0.288
	IRM	1.04	1.21	1.34	1.46	1.64	1.93	1.437	0.316
CDAE	MAP	1.37	1.41	1.44	1.51	1.62	1.82	1.529	0.168
	SMM	0.86	0.93	1.02	1.13	1.30	1.54	**1.129**	0.252
	IRM	0.87	0.94	1.03	1.14	1.29	1.53	1.133	0.242

TABLE IV: Segmental SNR Increase Results (ΔSSNR)

SNR		20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	AVG	SD
MLP	MAP	6.44	7.12	7.56	7.77	7.65	7.03	7.262	0.502
	SMM	6.91	7.48	7.80	7.81	7.42	6.68	**7.350**	0.465
	IRM	6.12	6.79	7.22	7.36	7.15	6.72	6.894	0.452
CNN	MAP	6.98	7.60	7.92	7.94	7.46	6.43	7.388	0.586
	SMM	7.08	7.70	8.03	7.96	7.37	6.49	**7.437**	0.588
	IRM	6.20	6.91	7.38	7.52	7.17	6.53	6.952	0.509
DDAE	MAP	6.73	7.44	7.85	7.86	7.63	7.02	**7.422**	0.459
	SMM	6.96	7.53	7.85	7.85	7.41	6.80	7.400	0.442
	IRM	6.08	6.76	7.17	7.29	7.06	6.58	6.823	0.448
CDAE	MAP	7.07	7.77	8.23	8.33	7.98	7.51	**7.814**	0.473
	SMM	7.10	7.77	8.19	8.29	7.93	7.37	7.773	0.463
	IRM	6.22	6.98	7.50	7.73	7.58	7.31	7.222	0.554

TABLE V: PESQ Results Considering Generalization Ability

SNR		20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	AVG	SD
Noisy		2.69	2.36	2.03	1.75	1.51	1.27	1.935	0.532
MLP	MAP	2.16	2.09	2.02	1.93	1.75	1.37	1.887	0.293
	SMM	2.79	2.54	2.33	2.11	1.80	1.42	**2.166**	0.501
	IRM	2.74	2.52	2.30	2.07	1.73	1.36	2.119	0.513
CNN	MAP	2.36	2.34	2.28	2.16	1.89	1.43	2.074	0.361
	SMM	3.01	2.80	2.57	2.30	1.95	1.55	**2.364**	0.545
	IRM	3.03	2.80	2.56	2.28	1.94	1.56	2.361	0.548
DDAE	MAP	2.68	2.57	2.42	2.22	1.91	1.48	**2.215**	0.451
	SMM	2.85	2.60	2.34	2.08	1.76	1.40	2.172	0.539
	IRM	2.89	2.63	2.38	2.12	1.79	1.41	2.201	0.545
CDAE	MAP	2.81	2.68	2.54	2.34	2.06	1.68	**2.352**	0.424
	SMM	1.64	1.62	1.60	1.54	1.45	1.30	1.526	0.130
	IRM	1.65	1.62	1.59	1.53	1.45	1.29	1.523	0.133

TABLE VI: STOI Results Considering Generalization Ability

SNR		20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	AVG	SD
Noisy		0.95	0.92	0.87	0.81	0.73	0.65	0.823	0.117
MLP	MAP	0.84	0.82	0.81	0.79	0.74	0.63	0.772	0.076
	SMM	0.92	0.89	0.86	0.83	0.77	0.68	**0.825**	0.088
	IRM	0.91	0.89	0.86	0.82	0.76	0.67	0.817	0.091
CNN	MAP	0.77	0.77	0.76	0.73	0.69	0.60	0.719	0.067
	SMM	0.93	0.91	0.88	0.84	0.78	0.69	0.836	0.090
	IRM	0.93	0.92	0.89	0.85	0.79	0.70	**0.844**	0.090
DDAE	MAP	0.89	0.88	0.86	0.83	0.78	0.68	0.821	0.080
	SMM	0.93	0.91	0.87	0.82	0.76	0.68	0.829	0.094
	IRM	0.94	0.92	0.88	0.84	0.78	0.69	**0.842**	0.093
CDAE	MAP	0.94	0.92	0.90	0.86	0.81	0.73	**0.859**	0.080
	SMM	0.70	0.70	0.69	0.67	0.65	0.60	0.667	0.040
	IRM	0.70	0.70	0.69	0.67	0.65	0.59	0.667	0.041

specific target that works better with respect to all metrics. This is clear in the visual comparison shown in Fig. 2, which represents the spectrogram of a noisy speech signal and its processed speech from the four implemented architectures. Each row represents a different architecture, using the two approaches. It can be noticed that the mapping based approach for the two fully connected architectures, MLP and DDAE shown in sub figures a and c, is doing very well in removing the background noise; however, the approach is not efficient in reconstructing the entire speech signal, especially the high frequency components. On the other hand, masking based approaches are better at representing the clean speech signal at the expense of the ability to remove the noise. This is why masking based targets produce more intelligible speech. Consequently, the choice between a masking and mapping target, in this case, is a speech denoising and speech intelligibility tradeoff.

It is also clear in Fig. 2 that the convolutional based architectures, CNN and CDAE shown in sub figures b and d, are less affected by the used training targets, because for these architectures masking and mapping approaches are approximately giving the same performance. This introduces another factor, which is that the architecture design may compensate the negative effects of the chosen target. Fig. 2 also shows a comparison between different architectures performance, where the CDAE architecture, sub figure (d),

is the best performing one. Finally, when looking into the intensity of the clean and processed speech, it is clear that the processing causes attenuation to the intensity of output speech in all cases, which is a loss in the strength of the signal. This is mainly due to the transformation process between the frequency and time domain.

Considering the effect of the training target on the network generalization ability, Table V and VI show the results of the PESQ and STOI scores when testing the networks using a different speech dataset from the one used in the training process, mixed with the same seen and unseen noise environments used in the previous evaluation. It is clear that there is a degradation in the performance for all architectures; however, it can be noticed that masking training targets are not suitable for any architecture design. For example, the autoencoder based architectures (DDAE and CDAE) output speech with better quality in the case of a mapping target, although masking targets showed better performance previously when the networks' generalization ability was not considered. Additionally, there is a significant negative effect on the performance of the CDAE architecture when using masking targets (SMM and IRM), and the output speech is unintelligible at low SNR. As a conclusion, the choice of the training target is bounded by the type of the used architecture, because autoencoder based architectures are proved to better generalize when using mapping targets.

VI. Conclusion

In this paper, a comprehensive comparison was presented between masking and mapping targets for speech enhancement using four different, state of the art, DNN architectures. The comparison covers how the networks' performance change with respect to the chosen target. The results show that there is no training target that is considered to be the best with respect to the four used evaluation metrics. The performance of masking targets was shown to be much better at high SNR for all architectures, leading to a higher variance for all the evaluation metrics than that of the mapping targets, which makes mapping targets less affected by SNR changes. It is also shown that for the fully connected architectures, MLP and DDAE, there is always a tradeoff between the ability of the network to remove the noise, where the mapping targets outperform, and the reconstruction of a more intelligible clean speech signal, where the masking targets surpass. It can be concluded that the choice of the training target when using these architectures will be based on the application in which the speech enhancement process is applied, which will define the metric with the highest priority to improve. Applications such as mobile communications will be more interested in removing noise due to the noisy nature of the wireless communication medium. Conversely, speech intelligibility is a more important factor for applications such as hearing aids and Automatic Speech Recognition systems. On the other hand, convolution based architectures, such as CNN and CDAE, are proved to be less affected by the training target when tested using unseen noise environments and unseen speech from the same training dataset. However, when considering the generalization ability of the networks by testing the performance using a different dataset from the one used in the training process, the results show that masking targets are not recommended for autoencoder architectures, because there is a significant performance degradation in the case of using masking targets, especially for the CDAE architecture.

References

[1] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *ICASSP*. IEEE, 2014, pp. 1562–1566.

[2] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE Trans. Audio Speech Lang. Proc.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[3] L. Hertel, H. Phan, and A. Mertins, "Comparing time and frequency domain for audio event recognition using deep learning," in *IJCNN*. IEEE, 2016, pp. 3407–3411.

[4] Y. K. Muthusamy, R. A. Cole, and M. Slaney, "Speaker-independent vowel recognition: Spectrograms versus cochleagrams," in *ICASSP*. IEEE, 1990, pp. 533–536.

[5] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE Trans. Audio Speech Lang. Proc.*, vol. 22, no. 12, pp. 1849–1858, 2014.

[6] B. O. Odelowo and D. V. Anderson, "A study of training targets for deep neural network-based speech enhancement using noise prediction," in *ICASSP*. IEEE, 2018, pp. 5409–5413.

[7] Y. Zhao, D. Wang, I. Merks, and T. Zhang, "DNN-based enhancement of noisy and reverberant speech," in *ICASSP*. IEEE, 2016, pp. 6525–6529.

[8] T. Kounovsky and J. Malek, "Single channel speech enhancement using convolutional neural network," in *ECMSM*. IEEE, 2017, pp. 1–5.

[9] X.-L. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE Trans. Audio Speech Lang. Proc.*, vol. 24, no. 5, pp. 967–977, 2016.

[10] M. Portnoff, "Time-frequency representation of digital signals and systems based on short-time fourier analysis," *IEEE Trans. Acoust. Speech Sig. Proc.*, vol. 28, no. 1, pp. 55–69, 1980.

[11] M. Slaney and R. F. Lyon, "On the importance of time-a temporal representation of sound," *Visual Representations of Speech Signals*, vol. 95116, 1993.

[12] D. Yu and J. Li, "Recent progresses in deep learning based acoustic models," *IEEE/CAA J. Automatica Sinica*, vol. 4, no. 3, pp. 396–409, 2017.

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

[14] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Comm.*, vol. 53, no. 4, pp. 465–494, 2011.

[15] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Proc. Mag.*, vol. 32, no. 2, pp. 55–66, 2015.

[16] S.-W. Fu, T.-Y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *MLSP*. IEEE, 2017, pp. 1–6.

[17] Z. Ouyang, H. Yu, W.-P. Zhu, and B. Champagne, "A fully convolutional neural network for complex spectrogram processing in speech enhancement," in *ICASSP*. IEEE, 2019, pp. 5756–5760.

[18] S. Pirhosseinloo and J. S. Brumberg, "A new feature set for masking-based monaural speech separation," in *52nd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2018, pp. 828–832.

[19] N. Dave, "Feature extraction methods LPC, PLP and MFCC in speech recognition," *Int. j. Advance Research in Engineering and Technology*, vol. 1, no. 6, pp. 1–4, 2013.

[20] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Proc. Letters*, vol. 21, no. 1, pp. 65–68, 2013.

[21] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Trans. Audio Speech Lang. Proc.*, vol. 23, no. 1, pp. 7–19, 2015.

[22] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *INTERSPEECH*, 2013, pp. 436–440.

[23] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *arXiv preprint arXiv:1609.07132*, 2016.

[24] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE Trans. Audio Speech Lang. Proc.*, vol. 23, no. 6, pp. 982–992, 2015.

[25] J. Qi, D. Wang, Y. Jiang, and R. Liu, "Auditory features based on gammatone filters for robust speech recognition," in *ISCAS*. IEEE, 2013, pp. 305–308.

[26] Y. Shao and D. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *ICASSP*. IEEE, 2008, pp. 1589–1592.

[27] H. K. Maganti and M. Matassoni, "An auditory based modulation spectral feature for reverberant speech recognition," in *Eleventh Annual Conf. Int. Speech Comm. Assoc.*, 2010.

[28] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE Trans. Audio Speech Lang. Proc.*, vol. 22, no. 12, pp. 1993–2002, 2014.

[29] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 131, no. 5, pp. 4134–4151, 2012.

[30] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE Trans. Audio Speech Lang. Proc.*, vol. 24, no. 7, pp. 1315–1329, 2016.

[31] K. Han, Y. Wang, and D. Wang, "Learning spectral mapping for speech dereverberation," in *ICASSP*. IEEE, 2014, pp. 4628–4632.

[32] Y. Jiang and R. Liu, "Binaural deep neural network for robust speech enhancement," in *ICSPCC*. IEEE, 2014, pp. 692–695.

[33] T. Goehring, F. Bolner, J. J. Monaghan, B. van Dijk, A. Zarowski, and S. Bleeck, "Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users," *Hearing Research*, vol. 344, pp. 183–194, 2017.

[34] R. Lyon, "A computational model of binaural localization and separation," in *ICASSP*, vol. 8. IEEE, 1983, pp. 1148–1151.

[35] M. Weintraub, "A theory and computational model of auditory monaural sound separation," Ph.D. dissertation, Stanford University, 1985.

[36] D. Wang, "Deep learning reinvents the hearing aid," *IEEE Spectrum*, vol. 54, no. 3, pp. 32–37, 2017.

[37] S. Chakrabarty, D. Wang, and E. A. Habets, "Time-frequency masking based online speech enhancement with multi-channel data using convolutional neural networks," in *IWAENC*. IEEE, 2018, pp. 476–480.

[38] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE Trans. Audio Speech Lang. Proc.*, vol. 24, no. 3, pp. 483–492, 2016.

[39] S. Samui, I. Chakrabarti, and S. K. Ghosh, "Time–frequency masking based supervised speech enhancement framework using fuzzy deep belief network," *Applied Soft Computing*, vol. 74, pp. 583–602, 2019.

[40] G. S. Alberti and H. Ammari, "Disjoint sparsity for signal separation and applications to hybrid inverse problems in medical imaging," *Applied and Computational Harmonic Analysis*, vol. 42, no. 2, pp. 319–349, 2017.

[41] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. Audio Speech Lang. Proc.*, vol. 15, no. 6, pp. 1766–1776, 2007.

[42] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*. Springer, 2005, pp. 181–197.

[43] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, 2009.

[44] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Comm.*, vol. 48, no. 11, pp. 1486–1501, 2006.

[45] Z. Wang, X. Wang, X. Li, Q. Fu, and Y. Yan, "Oracle performance investigation of the ideal masks," in *IWAENC*. IEEE, 2016, pp. 1–5.

[46] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *ICASSP*. IEEE, 2015, pp. 708–712.

[47] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *O-COCOSDA/CASLRE*. IEEE, 2013, pp. 1–4.

[48] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.

[49] G. Hu, "100 nonspeech environmental sounds." [Online]. Available: http://www.cse.ohio-state.edu/pnl/corpus/HuCor- pus.html, 2014.

[50] A. Varga and H. 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 Comm.*, vol. 12, no. 3, pp. 247–251, 1993.

[51] A. Pandey and D. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE Trans. Audio Speech Lang. Proc.*, vol. 27, no. 7, pp. 1179–1188, 2019.

[52] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Recommendation, p. 862.*, 2001.

[53] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio Speech Lang. Proc.*, vol. 19, no. 7, pp. 2125–2136, 2011.