Simultaneous Progressive Filtering-based Monaural Speech Enhancement

Haoran Yin[1], Hao Shi[2⋆], Longbiao Wang[1(✉)], Luya Qiang[1], Sheng Li[3], Meng Ge[1], Gaoyan Zhang[1(✉)], and Jianwu Dang[1,4]

[1] Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2] Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto, Japan
[3] National Institute of Information and Communications Technology, Kyoto, Japan
[4] Japan Advanced Institute of Science and Technology, Ishikawa, Japan
{haoran_yin, longbiao_wang, gaoyan_zhang}@tju.edu.cn

Abstract. Speech enhancement (SE) benefits from multi-stage stacking. However, this will introduce a lot of new parameters to the neural network. In this paper, we propose a simultaneous progressive filtering-based monaural SE model. Mapping-based and masking-based SE systems are simultaneously obtained with multi-target learning (MTL). Different from other MTL systems, our proposed model addresses different enhancement needs. The mapping-based SE system aims to recover speech signals from noisy features. While the masking-based SE system serves as a post-filtering to further reduce the noise that still exists after the mapping-based SE system. With the high signal-to-noise ratio inputs, noise reduction of the masking-based SE system is obvious with little speech signal loss. These two stages share one neural network which controls the parameters of the entire system with little or no increase. In addition, our approach is easy to integrate with existing methods and improve their performance significantly and stably. The experiments on Valentini-Botinhao data set show our proposed model achieves 0.12 PESQ improvement compared with directly mapping-based and masking-based SE systems both in single-target and multi-target learning. Furthermore, by comparing spectrograms, we find that our proposed models are able to recover better harmonic information.

Keywords: Speech enhancement · Multi-target learning · Simultaneous progressive filtering · Deep learning.

1 Introduction

Speech is the main mode of communication of human beings and has a wide range of applications. However, the unavoidable inclusion of a high level of undesirable noise in real scenes considerably reduces the intelligibility and quality of speech and seriously deteriorates performance in speech applications [1].

[⋆] Hao Shi is the joint first author.

Speech enhancement (SE) is the major front-processing method for recovering clean speech from noisy speech and is an indispensable technique in the speech field. Given the important role of SE, an increasing number of researchers are exploring more effective SE systems, especially monaural SE systems [2,3].

Deep learning-based methods are not based on any assumptions, resulting in an enormous improvement over traditional methods [4–7], especially in unsteady-state noisy environments [8]. Mapping and masking are two commonly used learning targets for training a deep learning-based SE system. Mapping methods [9,10,8] use the nonlinear mapping ability of neural networks to recover clean speech features from noisy features. However, limited by the capabilities of current DNN models, mapping-enhanced features contain residual noise. Masking methods [8,11–13] first learn a mask, and the estimated mask is multiplied with noisy features to reconstruct enhanced features. As a ratio mask is used to extract a speech signal from a noisy speech signal during masking, some speech signals may be lost [13]. Furthermore, some researchers have observed complementarity between mapping and masking targets in SE tasks [14,15]. Some multi-stage approaches [16,17] outperform single-stage approaches by completing more than one task during different stages. However, multi-stage approaches commonly require more parameters than single-stage approaches and thus, more training time and computing resources. Besides, although multi-target learning (MTL) with mapping and masking targets perform well with the complementarities between mapping-based and masking-based system, it is still hard to further use these two outputs.

In this paper, we propose a simultaneous progressive filtering-based monaural SE approach to eliminate the shortcomings of the above mentioned approaches. We get mapping-based and masking-based systems simultaneously with MTL. First, we use the mapping-based SE system to enhance the original noisy features. The mapping-based SE system keeps the speech information well but there are still some noise residue. Then we use the masking-based SE system to do the post-filtering. Difference from previous work, masking-based SE system recover the output of the mapping-based SE system. Although some masking-based SE systems lead to the loss of speech signal, the masking-based SE system with high signal-to-noise ratio inputs has obvious noise reduction and little speech signal loss. As mapping and masking share a common hidden layer, complementary information is available for both processes. Furthermore, the number of parameters of our entire system are not increased at all or only by a small number.

The rest of this paper is organized as follows. Section 2 presents conventional SE methods. Our proposed method is discussed in Section 3. Section 4 shows detailed experiments and results. Section 5 draws conclusions.

2 Conventional SE methods

The mean squared error (MSE) is a widely used loss function in SE systems. The MSE loss function of **direct mapping (DM)** method is represented as follows:

$$\mathcal{L}_{DM} = \frac{1}{TF} \sum || |\widetilde{X}_{DM}| - |X| ||_F^2 \tag{1}$$

$\mathcal{L}_{DM}$ is the loss for the DM approach. $|\widetilde{X}_{DM}|$ and $|X|$ denote the mapping-estimated spectrogram and the reference clean spectrogram respectively.

Signal approximation (SA) is an effective masking technique. SA trains a ratio mask [11] to approximate the spectrogram of clean speech using the product of the estimated mask and the noisy spectrogram, where the MSE loss function for the SA method is represented as follows:

$$\mathcal{L}_{SA} = \frac{1}{TF}\sum ||\widetilde{M} \odot |Y| - |X|||_F^2 = \frac{1}{TF}\sum |||\widetilde{X}_{SA}| - |X|||_F^2 \tag{2}$$

where $|Y|$ denotes the noisy speech spectrogram and $|\widetilde{X}_{SA}|$ denotes the masking-enhanced spectrogram. $\odot$ denotes point-wise matrix multiplication. $\mathcal{L}_{SA}$ denotes the loss for the SA approach and $\widetilde{M}$ denotes the estimated mask.

The principle of multi-target learning (MTL) is to learn different targets in one model. Complementary learning targets result in enhanced performance of all outputs. Therefore, MTL can be used to tarin mapping and masking targets. The MTL loss function is represented as follows:

$$\mathcal{L}_{MTL} = \alpha\mathcal{L}_{DM} + (1-\alpha)\mathcal{L}_{SA} \tag{3}$$

$\mathcal{L}_{MTL}$ is the loss for the MTL method. α is the weight coefficient of the two MSE target items. The MTL-based SE flowchart is shown in Fig. 1 (a).

3 Simultaneous progressive filtering

Simultaneous progressive filtering (SPF) contains two modules: mapping-based pre-filtering and masking-based post-filtering modules. The mapping-based pre-filtering module aims to recover speech signal from noisy features and obtain the high SNR pre-enhanced spectrogram. While the masking-based post-filtering module further reduces the noise that still exists in the pre-enhanced spectrogram. With the high SNR inputs, noise reduction of the masking-based post-filtering module is obvious with little speech signal loss.

3.1 Mapping-based pre-filtering module

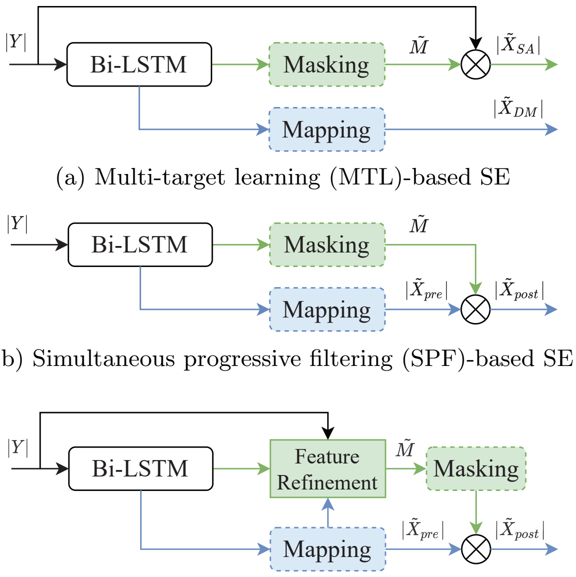
The pre-filtering module maps the noisy magnitude spectrogram to the pre-enhanced spectrogram to preserve the clean speech signals and increase the SNR of the spectrogram. We use a mapping target to train this module. The loss function of the pre-filtering module is calculated in the same way as $\mathcal{L}_{DM}$. However, we use a different symbol, $\mathcal{L}_{pre}$, to represent the loss function of this module, which is given as follows:

$$\mathcal{L}_{pre} = \frac{1}{TF}\sum |||\widetilde{X}_{pre}| - |X|||_F^2 \tag{4}$$

where $\mathcal{L}_{pre}$ is the loss of the pre-filtering module. $|\widetilde{X}_{pre}|$ is the enhanced spectrogram of the pre-filtering module.

3.2 Masking-based post-filtering module

The post-filtering module reduces the residual noise of pre-filtering enhanced spectrogram and is trained using the masking target. The use of masking targets



(a) Multi-target learning (MTL)-based SE

(b) Simultaneous progressive filtering (SPF)-based SE

(c) SPF w/ feature refinement block-based SE

Fig. 1. Multi-target learning (MTL)-based and our proposed SE flowcharts: (a) MTL-based SE have $|\widetilde{X}_{SA}|$ and $|\widetilde{X}_{DM}|$ two outputs. (b) SPF-based SE simply uses MTL to achieve simultaneous progressive filtering without introducing new parameter. (c) SPF w/ feature refinement block-based SE adding a feature refinement block to the post-filtering module, the input of the block contains three components in this configuration. The green part in the flowcharts denotes masking-based module and the blue part denotes mapping-based module.

may result in the loss of clean speech signals but greatly reduces noise. As using the pre-filtering enhanced spectrogram as the input to the post-filtering module considerably increases the SNR over that of the original noisy spectrogram, the masking target that we used dose not cause serious speech distortion and enhances performance. The loss function of this module $\mathcal{L}_{post}$ is given as follows:

$$\mathcal{L}_{post} = \frac{1}{TF}\sum ||\widetilde{M} \odot |\widetilde{X}_{pre}| - |X|||_F^2 = \frac{1}{TF}\sum |||\widetilde{X}_{post}| - |X|||_F^2 \qquad (5)$$

where $\widetilde{M}$ is the estimated mask of the post-filtering module. $|\widetilde{X}_{post}|$ is the output spectrogram of the post-filtering module and the final enhanced spectrogram.

3.3 Simultaneous progressive filtering (SPF) system

The loss function of the entire SPF system $\mathcal{L}_{SPF}$ is given as follows:

$$\mathcal{L}_{SPF} = \beta\mathcal{L}_{pre} + (1-\beta)\mathcal{L}_{post} \qquad (6)$$

We compress the pre-filtering and post-filtering modules in one bidirectional long short-term memory (Bi-LSTM) neural network and utilize the complementarity features through sharing the Bi-LSTM layers to the pre-filtering and post-filtering modules. Therefore, the Bi-LSTM output layer contains information common to both the pre-filtering and post-filtering modules. Thus, we simply use MTL to run the pre-filtering module and post-filtering modules simultaneously and do not introduce any new parameters into the SPF system. Moreover, our system fully utilizes the complementarity between the mapping and masking targets and use the masking method to filter the mapping-enhanced spectrogram again. The flowchart of SPF is shown in Fig. 1 (b).

3.4 Feature refinement Block

We design a feature refinement block to refine the shared information and supplement speech information that may be lost in the pre-filtering module. The block consists of a concatenation part and a hidden layer. The task of concatenation part is to concatenate the two or three inputs into one as the input of hidden layer. The flowchart of feature refinement block is shown in Fig.2.

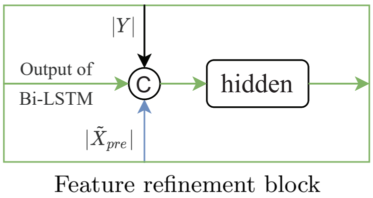


Fig. 2. Flowchart of feature refinement block: the 3 optional inputs of feature refinement block are the noisy spectrogram $|Y|$, the hidden output of the shared Bi-LSTM and the pre-filtering enhanced spectrogram $|\widetilde{X}_{pre}|$ from top to bottom.

Based on the SPF system, we add a feature refinement block into the post-filtering module to estimate a better mask. We explore 3 configurations of SPF with feature refinement block. In the first configuration, the input of the hidden layer only contains the hidden output of the shared Bi-LSTM. In the second configuration, the input of the hidden layer is the concatenation of the hidden ouputs of the shared Bi-LSTM and the pre-filtering enhanced spectrogram. In the third configuration, we concatenate the hidden output of the shared Bi-LSTM, the noisy spectrogram, and the pre-filtering enhanced spectrogram as the input of the hidden layer. The SPF with feature refinement block system are trained in the same way as the SPF system. But the addition of the hidden layer in the block introduces several new parameters into the SPF with feature refinement block system. The flowchart of SPF with feature refinement block in the third configuration is shown in Fig. 1 (c).

4 Experiments

We conduct experiments using the Valentini-Botinhao data set [18]. Some of the noise in the data set is obtained from the Demand database [19] and the speech database is obtained from the Voice Banking Corpus [20]. We adopt the validation set to control the learning rate (initialized as 0.001), which is decreased by 50% in the absence of improvement between two consecutive epochs. All speech signals are sampled at 16kHz. The Hamming window is used for framing, where the frame size is set to 512 with a 50% overlap. We use the magnitude spectrogram as the input and output features.

We implement our model using TensorFlow. All the baseline models use Bi-LSTM, where the Bi-LSTM model contains a 257-dimensional input layer and two hidden layers with 1024 nodes each. A 257-dimensional output layers are used for each of the mapping or masking targets for the single-target mapping or masking methods. Two 257-dimensional output layers are used simultaneously for both the mapping and masking outputs in the multi-target learning method. For our SPF approach, SPF has the same structure as the multi-target learning method. For SPF with feature refinement block, we add a fully connected hidden layer with 512 nodes after the Bi-LSTM layers. The input of the feature refinement block has three components: the hidden outputs of the shared Bi-LSTM, the noisy spectrogram, and the pre-filtering enhanced spectrogram. We simply concatenate the three components to form the input. The parameters of our models are randomly initialized.

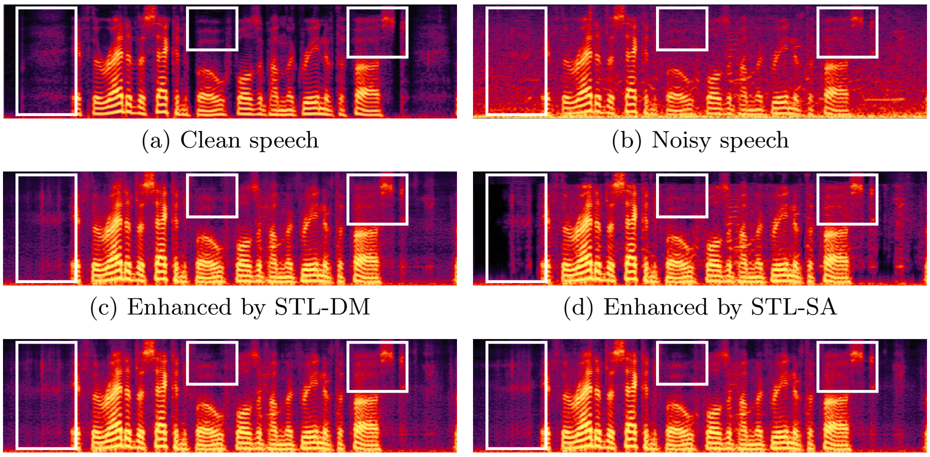
We evaluate the performance of SPF and baseline methods, by using the CSIG, CBAK and COVL [21] to measure the speech intelligibility and use the PESQ [22] to measure the speech quality.

Table 1. Results obtained using **baseline** and **our proposed methods**. The "Input of Post-filtering Config" part displays the configurations of the input of the masking-based post-filtering module, all the inputs are concatenated as the input of the post-filtering module in the feature refinement block if there are more than 1 input.

	Models	**Input of Post-filtering Config**			**Metrics**			
		output of Bi-LSTM	output of pre-filtering	noisy spectrogram	CSIG	CBAK	COVL	PESQ
Noisy	-	-	-	-	3.345	2.442	2.631	1.970
Baseline	STL-DM	-	-	-	3.849	2.547	3.226	2.604
	STL-SA	-	-	-	3.650	2.488	3.072	2.513
	MTL-DM	-	-	-	3.820	2.538	3.205	2.594
	MTL-SA	-	-	-	3.785	2.551	3.202	2.631
Ours	SPF	√	×	×	3.556	2.601	3.138	2.721
	SPF w/ Feature Refinement	√	×	×	**3.874**	2.610	**3.301**	2.729
		√	√	×	3.860	2.603	3.288	2.717
		√	√	√	3.841	**2.632**	3.297	**2.752**

4.1 The effect of multi-target learning

The upper half part of Table 1 lists the CSIG, CBAK, COVL, and PESQ performance obtained using different baseline systems with the test data sets. "Noisy" denotes the performance for untreated noisy speech. "STL-DM" and "STL-SA" denote the two single-target learning methods used for mapping and masking, as shown in Eq.(1) and Eq.(2), respectively. "MTL-DM" and "MTL-SA" denote



(a) Clean speech

(b) Noisy speech

(c) Enhanced by STL-DM

(d) Enhanced by STL-SA

(e) Enhanced by SPF

(f) Enhanced by SPF w/ feature refinement

Fig. 3. Magnitude spectrograms obtained from different types of speech (a) Clean speech; (b) Noisy speech; (c) Speech enhanced by STL-DM; (d) Speech enhanced by STL-SA; (e) Speech enhanced by SPF; (f) Speech enhanced by SPF w/ feature refinement when the input of post-filtering module consists of the hidden output of Bi-LSTM, the output of pre-filtering module and the noisy spectrogram.

the two outputs of the MTL method, as shown in Eq.(3). The hyperparameter α is set to 0.5 for the MTL. Using MTL improves the SE performance, e.g., The PESQ of "MTL-SA" is 0.118 higher than that of "STL-SA". However, using MTL causes a slight drop in the performance for some systems, e.g., the performance of "MTL-DM" was slightly lower than that of "STL-DM".

4.2 The performance of proposed methods

The bottom half part of Table 1 shows the results obtained using our SPF methods. "SPF" is described in section 3.3, and the corresponding β was set to 0.2. The 3 configurations of "SPF with feature refinement block" are described in section 3.4, and the corresponding β values were set to 0.9, 0.8 and 0.3, respectively. All of proposed methods considerably outperformed the conventional SE methods. Even without the feature refinement block, the PESQ of "SPF" was 0.09 higher than the best performance obtained using the baseline methods. These results shows that our simultaneous progressive filtering approach can be used to produce an enhancement system with superior performance and without introducing any new parameters. The main reason for the enhanced performance is that designing our mask for the mapping-enhanced spectrogram instead of the original spectrogram fully utilizes the complementarity of the two targets.

The β of "SPF" was set to 0.2. Thus, the masking-based pre-filtering module is more important than the mapping-based post-filtering module in "SPF". Unlike "STL-SA" and "MTL-SA", masking is used to recover a high-SNR spectrogram instead of a noisy spectrogram using "SPF". Thus, the masking method lowers the information loss for high-SNR spectrograms. However, for “SPF with feature refinement block”, the focus of the network gradually shifts from the pre-filtering module to the post-filtering module as more information is added to the hidden layer. Compared with the baseline methods, “SPF with feature refinement block” not only exhibited a PESQ improvement of more than 0.12 but also showed improved performance in the other three indicators. These results provide strong evidence that our proposed method can recover speech signals and remove residual noise more effectively than existing methods.

4.3 The effect on the spectrogram

Fig. 3 shows the magnitude spectrograms of clean, noisy and speech enhanced using the 4 systems. We observed clear differences among the spectrograms. Noise severely negatively impacts the speech signal for the noisy spectrogram. The STL-DM enhanced spectrogram still contains a considerable level of noise, and the details of many speech signals are not clear. The details of the STL-SA enhanced spectrogram are also not clear, and some speech signals are lost. Compared with the two STL enhanced spectrograms, some details of the SPF spectrogram are clearer, but some residual noise remains. The enhanced spectrogram by SPF with feature refinement block has the highest level of detail of the enhanced spectrograms, but there is a slight loss of some speech signals. We will address this speech distortion problem in a future study.

5 Conclusions and future work

In this paper, we proposed a simultaneous progressive filtering-based monaural speech enhancement approach. Two filtering modules were used: a mapping-based pre-filtering module and a masking-based post-filtering module. The pre-filtering module obtained a mapping-enhanced spectrogram from a noisy spectrogram to preserve clean speech signals and obtain a high-SNR spectrogram. The post-filtering module reduced the residual noise of the enhanced spectrogram obtained using the pre-filtering module. Our proposed simultaneous progressive filtering method exhibited a high SE performance; e.g., “SPF with feature refinement block” had a PESQ improvement of more than 0.12. As the post-filtering module filters the high-SNR spectrogram instead of the original noisy spectrogram, masking reduced the information loss and enhanced performance. As a multi-target learning strategy was used to develop these two modules, the number of the parameters of our proposed system was not increased or only by a small number. In addition, our SPF strategy can be easily integrated with many existing methods. In the future, we will apply this system to other speech processing tasks such as ASR.

References

1. Li, J., Deng, L., Gong, Y., Haeb-Umbach, R.: An overview of noise-robust automatic speech recognition. TASLP **22**(4), 745–777 (2014)

2. Wang, Y., Narayanan, A., Wang, D.: On training targets for supervised speech separation. TASLP **22**(12), 1849–1858 (2014)
3. Xia, B., Bao, C.: Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. Speech Communication **60**, 13–29 (2014)
4. Scalart, P., et al.: Speech enhancement based on a priori signal to noise estimation. In: ICASSP. vol. 2, pp. 629–632. IEEE (1996)
5. Boll, S.: Suppression of acoustic noise in speech using spectral subtraction. TASSP **27**(2), 113–120 (1979)
6. Deng, F., Bao, F., Bao, C.c.: Speech enhancement using generalized weighted β-order spectral amplitude estimator. Speech Communication **59**, 55–68 (2014)
7. Vihari, S., Murthy, A.S., Soni, P., Naik, D.: Comparison of speech enhancement algorithms. Procedia computer science **89**, 666–676 (2016)
8. Wang, D., Chen, J.: Supervised speech separation based on deep learning: An overview. TASLP **26**(10), 1702–1726 (2018)
9. Xu, Y., Du, J., Dai, L.R., Lee, C.H.: An experimental study on speech enhancement based on deep neural networks. IEEE Signal processing letters **21**(1), 65–68 (2013)
10. Xu, Y., Du, J., Dai, L.R., Lee, C.H.: A regression approach to speech enhancement based on deep neural networks. TASLP **23**(1), 7–19 (2014)
11. Srinivasan, S., Roman, N., Wang, D.: Binary and ratio time-frequency masks for robust speech recognition. Speech Communication **48**(11), 1486–1501 (2006)
12. Narayanan, A., Wang, D.: Ideal ratio mask estimation using deep neural networks for robust speech recognition. In: ICASSP. pp. 7092–7096. IEEE (2013)
13. Williamson, D.S., Wang, Y., Wang, D.: Complex ratio masking for monaural speech separation. TASLP **24**(3), 483–492 (2015)
14. Shi, H., Wang, L., Ge, M., Li, S., Dang, J.: Spectrograms fusion with minimum difference masks estimation for monaural speech dereverberation. In: ICASSP. pp. 7544–7548. IEEE (2020)
15. Sun, L., Du, J., Dai, L.R., Lee, C.H.: Multiple-target deep learning for lstm-rnn based speech enhancement. In: HSCMA. pp. 136–140. IEEE (2017)
16. Hao, X., Su, X., Wen, S., Wang, Z., Pan, Y., Bao, F., Chen, W.: Masking and inpainting: A two-stage speech enhancement approach for low snr and non-stationary noise. In: ICASSP. pp. 6959–6963. IEEE (2020)
17. Jin, Y.G., Lee, C.M., Cho, K., Kim, N.S.: A data-driven residual gain approach for two-stage speech enhancement. In: ICASSP. pp. 4752–4755. IEEE (2011)
18. Botinhao, C.V., Wang, X., Takaki, S., Yamagishi, J.: Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks. In: Interspeech. pp. 352–356 (2016)
19. Thiemann, J., Ito, N., Vincent, E.: The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings **19**(1), 035081 (2013)
20. Veaux, C., Yamagishi, J., King, S.: The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In: O-COCOSDA/CASLRE. pp. 1–4. IEEE (2013)
21. Hu, Y., Loizou, P.C.: Evaluation of objective quality measures for speech enhancement. TASLP **16**(1), 229–238 (2007)
22. Recommendation, I.T.: Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Rec. ITU-T P. 862 (2001)