Gender-Specific Speech Enhancement Architecture for Improving Deep Neural Networks Learning

Soha A. Nossier
Biomedical Engineering Department
Medical Research Institute, Alexandria University
Alexandria, Egypt
soha.nossier@alexu.edu.eg

Mhd Saeed Sharif
Intelligent Technologies Research Group
School of Architecture, Computing and Engineering, UEL
London, United Kingdom
s.sharif@uel.ac.uk

***Abstract*—Deep learning techniques for speech enhancement rely on training a deep neural network to process noisy speech, regardless the gender of the speaker. However, research shows that the speech of male and female stimulates different parts in human brain, and that female speech requires more complex analysis. This implies that different processing is applied on the speech, based on the speaker gender. In this work, we argue that male and female speeches have different features that can help in the learning process of speech enhancement deep neural networks if the training is performed on male and female speech data, independently, and using two different deep neural networks, specifically implemented for enhancing the clean speech signal of the target gender. This work presents a gender-specific speech enhancement architecture, which consists of a front-end binary classifier to detect the speaker gender. Based on the classifier decision, the noisy speech is enhanced using either a male or female speech enhancement model. One-stage and two-stage speech enhancement approaches are used to process male and female speeches, respectively. The results reveal that gender-specific speech enhancement has positive impact on the enhanced speech by deep neural networks. Additionally, the developed architecture achieved classifier accuracy 96.9% and 0.11 increase in Covl speech quality metric for the test data, in comparison to other best-performing networks.**

***Index Terms*—Denoising autoencoders, deep learning, gender recognition, signal processing, speech enhancement**

I. Introduction

Removing background noise that accompanies the speech signal is known as speech enhancement, and it is a common signal processing technique that has lots of applications, including mobile communications and Voice over Internet Protocol (VoIP) [1].

The implementation of deep neural networks (DNNs) for speech processing is showing a massive progression due to the availability of huge datasets and deep learning libraries, which has enabled the implementation of well-trained networks. These DNNs generate enhanced speech of better quality and intelligibility, in comparison to other speech enhancement approaches [2], [3].

Despite the effectiveness of deep learning techniques in mitigating background noise, some scenarios are still very challenging for the DNNs to deal with. Examples of these challenging scenarios include the ability to remove babble noise with low speech distortion, and processing speech in environments of high-level and diverse noise [4]. Although deep learning techniques are data-driven, the DNN often requires more information about the input data, to guide the network during the learning process [5], [6]. This becomes clear when investigating deep learning approaches for speech enhancement, as it can be noticed that powerful speech enhancement models are those that better describe the clean speech signal using the most important speech features. These features facilitate the learning process of the non-linear noisy to clean speech mapping function [7], [8].

Important speech features for network training include: framing and windowing for time domain signal, Short-Time Fourier Transform (STFT) for frequency domain signal, spectral decomposition for wavelet domain signal, spectrogram and cochleagram for image domain signal, Mel-Frequency Cepstral Coefficients (MFCC) for cepstral domain signal [9].

Research shows that there are specific properties for male and female speeches related to the different biological structure of males and females [10], [11]. Among these differences, the length and size of the vocal folds, where males usually have longer and thicker vocal folds than females, resulting in lower sound frequency and pitch for males [12].

Another interesting difference between male and female speeches is the human brain response to them. A study [13] has shown that male voices activate specific region in the brain; while, female voices activate different and more complex regions that are responsible for speech processing. An important finding of the study is that this brain behaviour could not be explained by the known speech features, such as frequency and pitch, or the behavioural response of the study subjects. This is because the study was performed on ”gender-ambiguous” voices, defined as the range of fundamental frequencies in which the speech of males and females overlaps [14].

In this paper, we argue that as long as the human brain processes the speech of males and females differently, employing two different DNNs for eliminating background noise from speech based on speaker gender will result in better network performance.

Fig. 1 clarifies the suggested idea, where a gender-specific speech enhancement architecture is proposed, which first classifies the input noisy speech as male or female speech. Afterwards, the noisy speech is processed by one of two speech enhancement models, trained on either male speech (network

a) or female speech (network b). Male speech is processed by a Deep Encoder - Convolutional Autoencoder DEnoiser (DE-CADE) model [15], shown in Fig. 2. While, female speech is enhanced by a two-stage network, in which a Deep Denoising Autoencoder network (DDAE) and DE-CADE models are employed. Details about architecture implementation will be described in Section III.

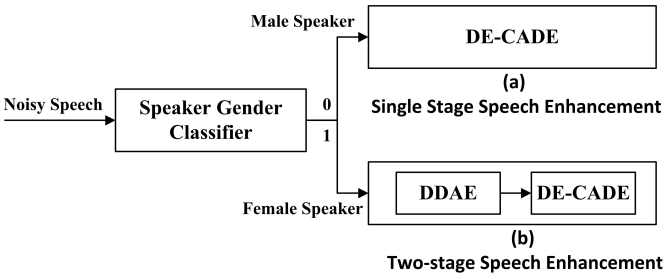

Fig. 1. Gender-specific speech enhancement architecture: a gender classifier, a single stage speech enhancement model for male speech (a), and a two-stage speech enhancement model for female speech (b)

This paper makes the following contributions:

- Showing speaker gender effect on the performance speech enhancement DNNs
- Proposing a gender-specific network for speech enhancement deep learning approach

The following sections are structured as follows. An explanation is given to the research problem of investigation in Section II; while, Section III illustrates the presented gender-specific speech enhancement architecture. Datasets for network training and testing are described in Section IV. The obtained results are explained and discussed in Section V. Paper conclusion is given in section VI.

II. Problem Statement

A. Male versus Female Speech Enhancement

As a way to analyse the learning process of speech enhancement DNNs for male and female speeches, the single stage DE-CADE frequency domain speech enhancement model [15] was first tested using the Valentini test dataset [16], after manually dividing the test set into male and female noisy speech sets. Details about the dataset are presented in Section IV. Perceptual Evaluation of Speech Quality (PESQ) [17] was used for measuring generated speech quality. While, evaluating the clarity of the output speech was performed using the Short-Time Objective Intelligibility (STOI) [18].

The outcome of this analysis is shown in Fig. 3. PESQ and STOI scores clearly show that the network generate better estimate in the case of male speech, as both scores are higher than the case of female speech. It should be noted here that the pre-trained network has no bias to any gender, as the training and testing sets have equal number of speakers for both genders. Moreover, all the speech samples for both genders were corrupted with the same noise environments at the same noise level, Signal to Noise Ratio (SNR), which means that the only factor affecting the results is the speaker gender.

To provide further evidence that the gender of the speaker affects network performance, the frequency domain-based single stage DE-CADE speech enhancement model was retrained twice: one time using male speech samples only, and another time using female speech samples only. The samples were taken from the Valentini train set [16]. Fig 4 shows the training Mean Square Error (MSE) loss curves in both cases, where the loss in the case of male noisy speech training is significantly lower than female noisy speech loss. This means that the model was able to better understand male speech features, and generated a closer estimate than that of the case of female speech data.

This analysis proves that female speech is more complex than male speech and requires more processing by the DNN, same as the human brain behaviour reported in the literature. Additionally, a speech enhancement model which improves male noisy speech may not be suitable for enhancing female noisy speech with the same performance. In other words, male and female noisy speeches should be enhanced using two separate and different DNNs.

B. The Proposed Gender-Specific Speech Enhancement Procedure

The following equation defines the input signal, *y*, to the DNN. This signal is defined in Equation (1) as the additive mixing of speech and noise signals (*s* and *n*):

$$y = s + n \tag{1}$$

The proposed speech enhancement architecture in this work extracts more information about the input speech signal using a binary classifier. This classifier acts as a first feature extraction stage that identifies the input signal as either male speech, s_m or female speech s_f. MFCCs features of the input noisy speech were obtained, denoted by *C*. Equation (2) defines a Binary Cross Entropy (BCE) loss function that the classifier uses, L_C, defined as follows:

$$L_C = \frac{1}{M}\sum_{i=1}^{M}\left[Z_i \log \hat{Z}_i + (1 - Z_i)\log\left(1 - \hat{Z}_i\right)\right], \tag{2}$$

where *i* denotes the sample index. *M* represents total sample number. *Z* denotes the binary value of the target (0 for male speaker and 1 for female speaker). $\hat{Z}$ is the speaker gender, estimated by the classifier.

Based on classifier's decision, the input signal is then sent to a single stage or two-stage speech enhancement model in the case of male or female speech, respectively. Both models operate in the frequency domain using signal spectrogram, which can be obtained by applying STFT to *y*. This creates a time-frequency form, *Y(t,f)*, for the noisy speech, calculated using the following equation:

$$Y(t,f) = \sum_{f=0}^{F-1} y(a+t)h(a)e^{-j2\pi fa/F}, \tag{3}$$

where *t* and *T* represent the time frames and the total number of frames, respectively. *a* denotes the time samples index.

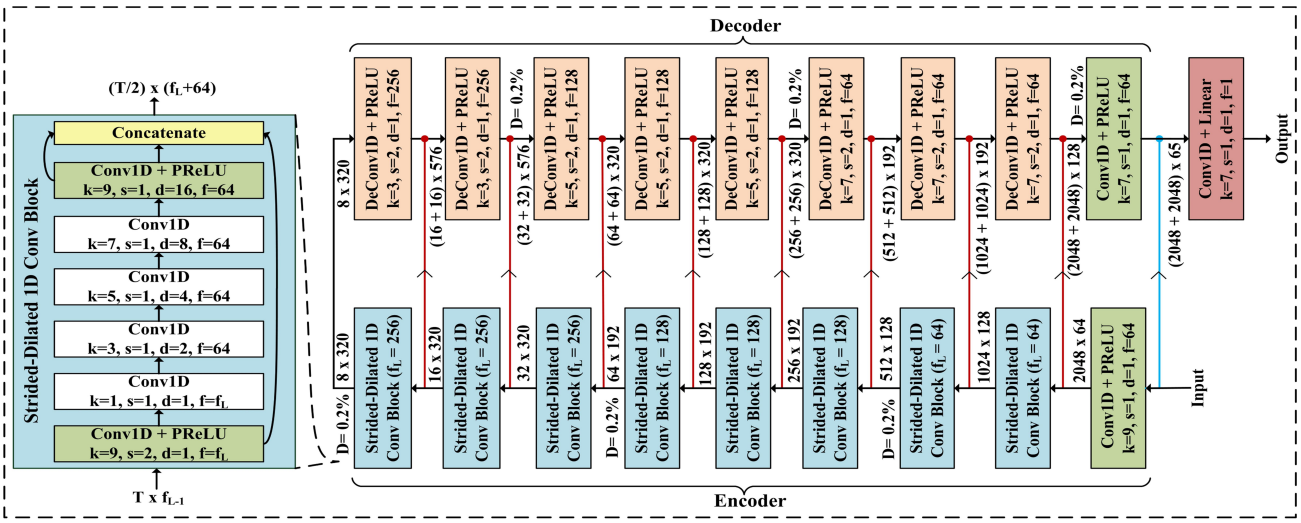

Fig. 2. The DE-CADE speech enhancement network [15].

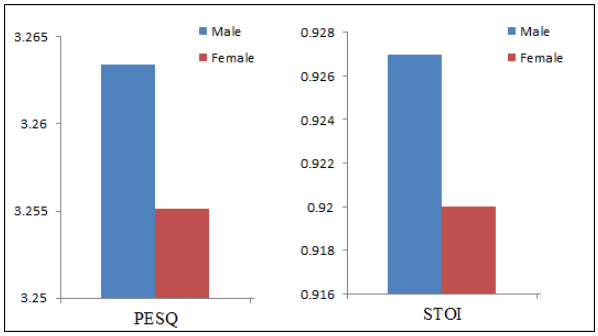

Fig. 3. PESQ and STOI scores for male and female enhanced speech from the Valentini Voice Bank test set [16], using DE-CADE speech enhancement model [15]

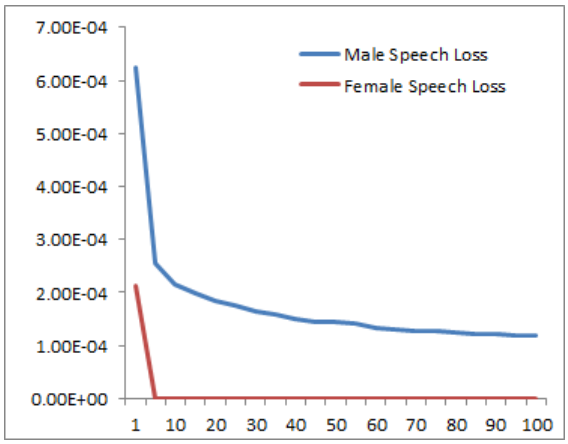

Fig. 4. Training loss curves for the DE-CADE speech enhancement model [15], trained on male speech only (blue curve) and female speech only (red curve). Vertical and horizontal axes represent the MSE and epoch number, respectively.

All the implemented speech enhancement networks use time frame of size 256 with 50% overlap. Windowing was applied to the time frames using Hamming window function, denoted as h. F is the total number of frequency bins. f is the frequency bin index.

Spectrograms of the input samples were then calculated using the magnitude STFT $|Y(t,f)|$, so Equation (1) is redefined by the below equation:

$$|Y(t,f)| = |S(t,f)| + |N(t,f)|, \tag{4}$$

where $|S(t,f)|$ represents speech spectrogram, and $|N(t,f)|$ represents the spectrogram of the noise environment.

In the case that the classifier detects male speaker, the DE-CADE speech enhancement model enhances the noisy speech, to predict male clean speech spectrogram, $|\hat{S}_m(t,f)|$, where:

$$|\hat{S}_m(t,f)| = DE-CADE(|Y(t,f)|) \tag{5}$$

If female speech was detected, two speech enhancement networks process the noisy speech, using two-stage enhancement approach. The first stage is a DDAE network and the second stage is the DE-CADE network; details about networks implantation will be given in Section III. The DDAE network generates a first estimate to the spectrogram of the female speech, $|\hat{S1}_f(t,f)|$, and then this estimate is concatenated with the first noisy speech spectrogram, $|Y(t,f)|$, to create a second input signal, $|Y_2(t,f)|$.

The second-stage DE-CADE network performs further denoising on $|Y_2(t,f)|$, to generate the final estimate, $|\hat{S2}_f(t,f)|$.

The estimated time speech signal is then recovered by calculating the Inverse STFT (ISTFT). This procedure can be described by the following equations.

$$|\hat{S1}_f(t,f)| = DDAE(|Y(t,f)|) \tag{6}$$

$$|Y_2(t,f)| = (|Y(t,f)|, |\hat{S1}_f(t,f)|), \tag{7}$$

$$|\hat{S2}_f(t,f)| = DE-CADE(Y_2(t,f)) \tag{8}$$

III. Gender-Specific Speech enhancement Architecture

The architecture contains a binary classifier that detects the gender of the speaker. Depending on the detected gender, the input signal is then enhanced by one of two speech enhancement networks, trained to process either male or female speech. The following subsections describe the structure of the architecture in details.

A. Speaker Gender Classifier

The classifier network is shown in Fig. 5. First, feature extraction is performed to get the MFFCs of the input noisy speech. These features were then processed by three fully connected layers, using 128 nodes in the first layer, 64 nodes in the second, and 32 nodes in the third layer. A final fully connected layer with sigmoid activation function and dropout layer of rate 10% were used to generate the prediction. BCE is the used loss function. Network training lasts for 100 epochs using batch size equals 64.

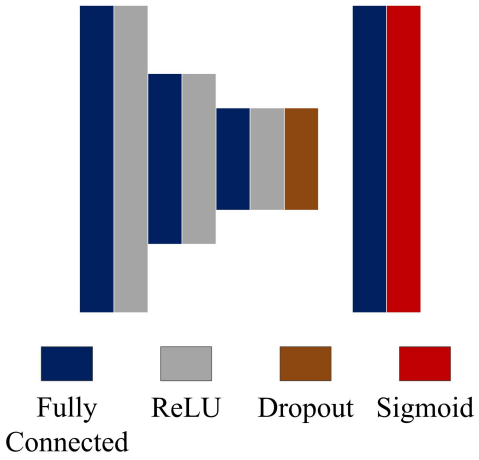

Fig. 5. Speaker gender classifier deep neural network for noisy speech

B. Male Speech Enhancement Network

The DE-CADE speech enhancement network [15] was used to process male noisy speech; network implementation is presented in Fig. 2. Convolution layers are the main building blocks of this network, in encoder-decoder structure. The network has an encoder deeper than the decoder, to allow for better feature extraction. The implementation also uses strided and dilated convolution blocks, which improve the overall performance. Furthermore, the encoder and decoder networks are connected using skip connections, to prevent information loss that is likely to happen in deep networks.

Training was performed in the frequency domain using speech spectrograms as input feature. The used loss function is MSE with Adam optimizer. The training lasts for 100 epochs with 64 batch size.

C. Female Speech Enhancement Network

Two-stage speech enhancement architecture was used to process noisy speech from female speakers, considering the complexity of the input, proved in Section II.

The input signal is first enhanced using a DDAE network, shown in Fig. 6, which was proven to be effective in noise removal [19]. The encoder network has 2 dense layers. The first and second layers use 2,048 and 500 hidden units, respectively. The network has a middle bottleneck dense layer that uses 180 nodes. The decoder network also has 2 dense layers, but the first and second layers use 500 and 2,048 nodes, respectively. All the dense layers have ReLU activations. Dropout technique was used to avoid training data overfitting, where two 10% rate dropout layers were added, represented in brown colour in Fig. 6. The final output is predicted using a dense output layer with linear activation function. Skip connections were added in the implementation.

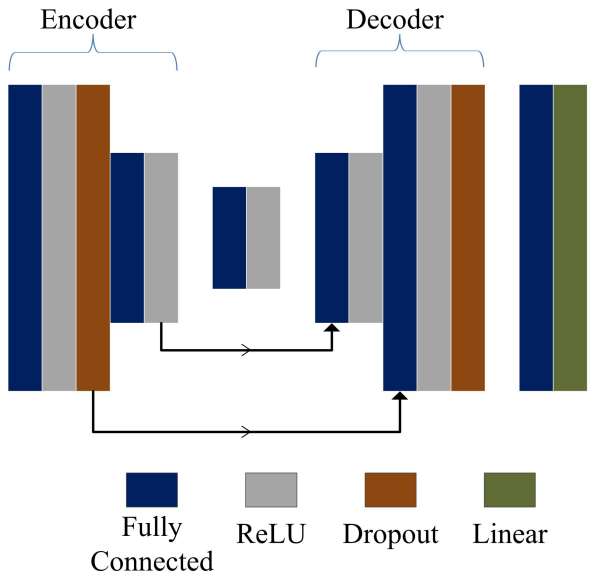

Fig. 6. The first stage female speech enhancement deep neural network, black arrows represent skip connections

The estimated speech by the first DDAE speech enhancement network was then enhanced by a second-stage frequency domain-based DE-CADE architecture, to perform further denoising. The DDAE training lasts for 100 epochs; while DE-CADE training lasts for 50 epochs. The batch size was 64 in both training processes. MSE with Adam optimizer was utilized in the entire training process.

IV. Dataset

The Valentini dataset benchmark [16] was used in evaluating the performance of our gender-specific speech enhancement network and state of the art implementations. This dataset includes 30 speakers, each reads around 400 sentences. All the speakers are English native speakers, and only English sentences are included in the dataset.

On the one hand, the speech audio samples in the train set were mixed with 10 noise types, including a mix of real noise and artificially generated noise. The real noise environments were taken from the DEMAND dataset [20]. The additive noise mixing was performed at four SNR levels: 0, 5, 10 and 15dB. The samples number in the train set is 11,572.

On the other hand, The noisy test set includes 5 noise environments, selected from the DEMAND noise set and different from the ones used in the train set. The testing data has a total of 824 noisy speech samples.

We divided the training set into two sets equal in size: male train set and female train set, to create the training data for male and female speech enhancement models, respectively. 10% of the training speech samples was used for validation during the training process.

More noisy speech samples were used in the training process of the speaker gender classifier, to improve the accuracy. A total of 2,000 speech samples for 4 males and 4 females speakers were taken at random from the Librispeech corpus [21], which has read English sentences, making around 1000 hours. These 2,000 clean speech samples were corrupted by random additive mixing of different noise environments. The noises were randomly selected from the DNS challenge dataset [22]. These noisy samples were used in addition to the Valentini train set, to create more challenging data for the training process of the speaker gender classifier.

V. Results and Discussion

The speaker gender classifier performance was evaluated using the Valentini dataset, and it achieved 97.6% and 96.9% accuracies for the train and test sets, respectively.

The proposed gender-specific speech enhancement model was evaluated using six speech evaluation metrics. PESQ [17] and STOI [18] scores were calculated for the enhanced speech; scores definitions were given in Section II. To measure speech distortion, Log Spectral Distortion (LSD) was used [23], where low LSD value indicates low speech distortion.

The results also include the three composite mean opinion score predictions, speech signal quality (Csig), residual noise (Cbak), and overall quality of the enhanced speech (Covl) [24]. The three predictions generate a number that ranges from 0 to 5, and high values relate to better network performance.

Two experiments were carried out to evaluate gender-specific speech enhancement. First, the performance of the single-stage DE-CADE speech enhancement model was evaluated when trained on male and female speech, independently. This model was then compared with the same speech enhancement network, but when trained using both male and female speech samples; half of the training data was used here to perform fair comparison. This comparison shows the effect of gender-specific training on the generated speech by the model. Second, the presented gender-specific speech enhancement architecture, demonstrated in Section III, was evaluated and compared with well-known speech enhancement models.

Table I shows first experiment results, where $DE-CADE_{gs}$ denotes independent male and female speeches training of the frequency domain-based DE-CADE model.

In comparison to the original DE-CADE network, $DE-CADE$, gender-specific training has a positive impact on network performance with reference to all the measures, expect the Cbak metric that is lower for the $DE-CADE_{gs}$ network, indicating more residual noise.

Although Cbak score is higher for $DE-CADE$ network, the enhanced speech signal quality is better in the case of $DE-CADE_{gs}$, which means less speech distortion, resulting in better overall enhanced speech quality, Covl.

Table II provides the Csig, Cbak, and Covl outputs for the presented gender-specific architecture. The reported results of high-performance speech enhancement models are also included in the Table, for comparison.

The comparison includes 12 speech enhancement models: a classical approach; Wiener filter approach [25], and 11 deep learning models: [26]–[35], including the DE-CADE model [15]. Noisy test set evaluation is also included in Table II.

Our architecture achieved the highest overall speech quality score, Covl, in comparison to the other models. As discussed above, although Koizumi et al. [35] architecture is the best-performing in terms of removing background noise (the highest Cbak score), this degrades speech signal quality, Csig, leading to lower Covl. This means that the estimated speech by our architecture is of the best overall quality.

Referring to Table I and II, it can be concluded from the results that gender-specific training improves the learning process of speech enhancement DNNs, resulting in better enhanced speech overall quality.

VI. Conclusion

This paper investigates the training of DNNs for speech enhancement on male and female speech samples, independently, and using different deep learning models. The work is based on research in the literature that shows that male and female voices are processed differently by the human brain. A gender-specific speech enhancement architecture was proposed, which employs a binary classifier to detect the speaker gender, and two speech enhancement modes, one for male noisy speech and another for female noisy speech. The results show that performing gender-specific training improves speech enhancement DNNs training, leading to better performance when comparing our network to best-performing speech enhancement models, using a variety of speech evaluation measures.

TABLE I
Effect of gender-specific speech enhancement on DNN performance

Metric	PESQ	STOI	LSD	Csig	Cbak	Covl
Noisy	1.97	91.5	1.32	3.35	2.44	2.63
DE-CADE	3.09	93.2	0.84	3.82	**3.19**	3.52
$DE-CADE_{gs}$	**3.18**	**93.5**	**0.75**	**4.05**	3.11	**3.62**

TABLE II
GENDER-SPECIFIC SPEECH ENHANCEMENT ARCHITECTURE PERFORMANCE COMPARISON WITH THE REPORTED RESULTS OF OTHER MODELS.

Metric	Csig	Cbak	Covl
Noisy	3.35	2.44	2.63
Wiener [26]	3.23	2.68	2.67
SEGAN [27]	3.48	2.94	2.80
Wave U-Net [28]	3.52	3.24	2.96
WaveNet [29]	3.62	3.23	2.98
MMSE-GAN [30]	3.80	3.12	3.14
Deep Feature Loss [31]	3.86	3.33	3.22
Deep Xi-ResLSTM [32]	4.01	3.25	3.34
Metric-GAN [33]	3.99	3.18	3.42
SEGAN-D [34]	3.46	3.11	3.50
DEMUCS [35]	4.14	3.21	3.54
Koizumi et al. [36]	4.15	**3.42**	3.57
DE-CADE(F)	4.00	3.11	3.60
Ours	4.30	3.15	**3.71**

REFERENCES

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[2] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "An experimental analysis of deep learning architectures for supervised speech enhancement," *Electronics*, vol. 10, no. 1, p. 17, 2021.

[3] N. Saleem and M. I. Khattak, "A review of supervised learning algorithms for single channel speech enhancement," *Int. J. Speech Tech.*, vol. 22, no. 4, pp. 1051–1075, 2019.

[4] S. Chhetri, M. S. Joshi, C. V. Mahamuni, R. N. Sangeetha, and T. Roy, "Speech enhancement: A survey of approaches and applications," in *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*. IEEE, 2023, pp. 848–856.

[5] H. Shi, M. Mimura, L. Wang, J. Dang, and T. Kawahara, "Time-domain speech enhancement assisted by multi-resolution frequency encoder and decoder," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[6] C. Zheng, H. Zhang, W. Liu, X. Luo, A. Li, X. Li, and B. C. Moore, "Sixty years of frequency-domain monaural speech enhancement: From traditional to deep learning methods," *Trends in Hearing*, vol. 27, p. 23312165231209913, 2023.

[7] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR," *Trans. Audio Speech Lang. Proc.*, vol. 28, pp. 1778–1787, 2020.

[8] Y. Xia, S. Braun, C. K. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *ICASSP*. IEEE, 2020, pp. 871–875.

[9] F. Alías, J. C. Socoró, and X. Sevillano, "A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds," *Applied Sciences*, vol. 6, no. 5, p. 143, 2016.

[10] L.-M. Zhang, Y. Li, Y.-T. Zhang, G. W. Ng, Y.-B. Leau, and H. Yan, "A deep learning method using gender-specific features for emotion recognition," *Sensors*, vol. 23, no. 3, p. 1355, 2023.

[11] I. Potamitis, N. Fakotakis, and G. Kokkinakis, "Gender-dependent and speaker-dependent speech enhancement," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2002, pp. I–249.

[12] A. P. Simpson, "Phonetic differences between male and female speech," *Language and linguistics compass*, vol. 3, no. 2, pp. 621–640, 2009.

[13] D. S. Sokhi, M. D. Hunter, I. D. Wilkinson, and P. W. Woodruff, "Male and female voices activate distinct regions in the male brain," *Neuroimage*, vol. 27, no. 3, pp. 572–578, 2005.

[14] C. Henton, "Pitch dynamism in female and male speech," *Language & Communication*, vol. 15, no. 1, pp. 43–61, 1995.

[15] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "Two-stage deep learning approach for speech enhancement and reconstruction in the frequency and time domains," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–10.

[16] C. Valentini-Botinhao *et al.*, "Noisy speech database for training speech enhancement algorithms and tts models," *University of Edinburgh*, 2017.

[17] 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.

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

[19] M. Tu and X. Zhang, "Speech enhancement based on deep neural networks with skip connections," in *ICASSP*. IEEE, 2017, pp. 5565–5569.

[20] J. Thiemann, N. Ito, and E. Vincent, "Diverse environments multichannel acoustic noise database DEMAND," 2013.

[21] 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.

[22] C. K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Interspeech 2021 deep noise suppression challenge," *arXiv*, 2021.

[23] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions," in *Conf. Int. Speech Comm. Assoc.*, 2008.

[24] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *Trans. Audio Speech Lang. Proc.*, vol. 16, no. 1, pp. 229–238, 2007.

[25] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *ICASSP*, vol. 2. IEEE, 1996, pp. 629–632.

[26] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *INTERSPEECH*, 2017, pp. 3642–3646.

[27] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-u-net," *arXiv*, 2018.

[28] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *ICASSP*. IEEE, 2018, pp. 5069–5073.

[29] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *ICASSP*. IEEE, 2018, pp. 5039–5043.

[30] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," in *INTERSPEECH*, 2019, pp. 2723–2727.

[31] A. Nicolson and K. K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," *Speech Comm.*, vol. 111, pp. 44–55, 2019.

[32] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Int. Conf. Mach. Learn.* PMLR, 2019, pp. 2031–2041.

[33] H. Phan, I. V. McLoughlin, L. Pham, O. Y. Chén, P. Koch, M. De Vos, and A. Mertins, "Improving GANs for speech enhancement," *Sig. Proc. Lett.*, vol. 27, pp. 1700–1704, 2020.

[34] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *INTERSPEECH*, 2020, p. 3291–3295.

[35] Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi, "Speech enhancement using self-adaptation and multi-head self-attention," in *ICASSP*. IEEE, 2020, pp. 181–185.