Speech Enhancement with Topology-enhanced Generative Adversarial Networks (GANs)

Xudong Zhang[1], *Liang Zhao*[2], *Feng Gu*[3]

[1]The Graduate Center, CUNY, NY, USA
[2]Lehman College, CUNY, NY, USA
[3]College of Staten Island, CUNY, NY, USA

xzhang5@gradcenter.cuny.edu, Liang.Zhao1@lehman.cuny.edu, Feng.Gu@csi.cuny.edu

Abstract

Speech enhancement is one of the effective approaches in improving speech quality. Neural network models have been widely used in speech enhancement, such as recurrent neural networks (RNN), long short-term memory (LSTM), and generative adversarial networks (GANs). However, some of them either handle the speech noise removal tasks in the spectral domain or lack the waveform recovery capability. As a result, the processed speeches still include noisy signals. In this study, we propose a topology-enhanced GAN model to tackle noisy speech in an end-to-end structure. We use the topology features of speech waves as additional constraints and modify the objective function of GAN by adding a penalty term. The penalty term is a Wasserstein distance of topology features between the generated speech and the corresponding clean speech in GAN. We evaluate the proposed speech-enhanced model on the public speech data set with 56 speakers and 20 different types of noisy conditions. The experimental results indicate that the topology features improve the performance of GAN on speech enhancement in metrics of PESQ, CBAK, COVL, and SSNR.

Index Terms: speech enhancement, topology, persistent homology, generative adversarial networks, convolutional neural networks.

1. Introduction

Speech enhancement aims to improve the speech quality including clarity, intelligibility, and compatibility in speech processing [1]. Enhancing the audio of a speech typically involves the removal of background noise, echo suppression and addition of certain frequencies [2]. Many researches have been focusing on removing the background noise in speech [3], as it is a particularly important task in speech recognition (ASR) systems [4]. It has vast applications in practice due to the fact that noise naturally occurs in speeches generated during telephone conversations, online meetings, and television programs.

One of the most popular methods to remove the background noise from the speech signals is a technique called spectral subtraction [5]. The amplitude spectra of speech indicate the amplitude of speech waves in dB scale. The speech can also be enhanced by using Wiener filter, which is a linear estimator and minimizes the mean-squared error between the original and enhanced speech [6]. The Wiener filtering algorithm is implemented in the frequency domain and depends on the filter transfer function from speech samples. However, these methods cannot differentiate the speeches from complex noises.

Recently, many researchers have obtained promising results in speech enhancement using deep learning models, such as the Long Short Term Memory (LSTM) model [7] and various deep neural networks (DNN) models [8, 9]. Some models weaken the short-time phase because the spectra of speech with short-time phase are considered to be unimportant in Fourier analysis [10]. Speech Enhancement Generative Adversarial Network (SEGAN) [11] was proposed to provide end-to-end speech enhancement approach. SEGAN learns from different speakers and noise types, and incorporates them into the same shared parametrization. However, the SEGAN cannot fully recover the clean speech from noisy speech due to the lack of morphology features of speech waves. Figure 1 shows the clean and corresponding GAN-enhanced speech waves from the same noisy speech waves. The Generative Adversarial Network (GAN) produces relative clean speech wave but also generates additional waves compared with the clean speech.

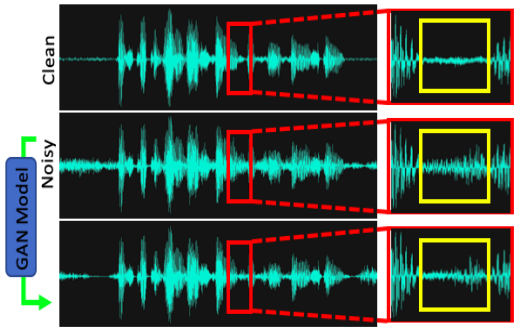

Figure 1: *The clean, noisy and GAN-generated speech waves and the corresponding differences in yellow squares.*

To overcome this draw-back of existing deep learning approaches on speech enhancement, we propose a topology-enhanced GAN model that can remove noises from speeches more effectively. This GAN model considers not only the frequency information of the speech waves, but also analyzes their corresponding topological features. We propose to use persistent homology to extract the wave features from the clean speech and the speech synthesized by GAN. Persistent homology is a method for computing topological features of a space at different spatial resolutions in arithmetic topology. The Wasserstein distance of the persistence of the extracted features are calculated as a penalty term for the objective function of the generator of GAN. By optimizing the new objective function, the generator is encouraged to synthesize a speech wave that is more similar to the clean speech wave, effectively avoiding those unnecessary noisy speech signals and keeping critical speech signals. We validate this approach with extensive experiments on several released benchmark datasets, and the results are highly promising in metrics of PESQ, CSIG, CBAK, COVL, and SSNR.

The rest of the paper is organized as follows. Section 2 in-

troduces the related work on speech enhancement. Section 3 shows the basics background of the persistent homology and its applications on feature extraction of acoustic waves. Section 4 presents the proposed speech-enhanced GAN model. Section 5 presents the experimental results. Section 6 provides discussions on related issues.

2. Related work

With the development of deep neural networks (DNN) over the last years, the speech enhancement approaches have been converted from the spectral analysis [5] to DNN-based enhancement. Pascual et al. proposed an end-to-end Speech Enhancement Generative Adversarial Networks (SEGAN) to tackle the speech enhancement from the raw speech waves [11]. The experimental results indicated the proposed SEGAN model enhanced speech in both objective and subjective evaluations. Rethage et al. introduced an end-to-end learning method, namely Wavenet, for speech denoising [12]. The model made use of non-causal, dilated convolutions, and predicted target fields instead of a single target sample. Tan and Wang proposed a speech enhancement model by incorporating a convolutional encoder-decoder (CED) and LSTM [13]. The experiments suggested that the proposed model led to consistently better objective intelligibility and perceptual quality than the existing LSTM-based model.

The persistent homology had been successfully used to segment the trabecula from 3D CT images [14]. The experimental results indicated that the persistent homology is able to capture the fine tissue in a noisy environment . Emrani et al. proposed to use persistent homology for robust analysis of point clouds of delay-coordinate embeddings in detection [15]. The quantitative and qualitative evaluations showed that the proposed method is preferred over Wiener filtering.

3. Preliminary Results

This section introduces notations needed for the definition of persistence homology. A d-dimensional simplex σ is the convex hull of $d+1$ affinely independent points. For instance, the 0-simplex, 1-simplex, and 2-simplex are a vertex, edge, and triangle in topology space, respectively. A face of a d-simplex is the convex hull of a nonempty subset of its $d+1$ vertices, i.e., an edge has two 0-dimensional (vertex) faces. A simplicial complex $K = \{\sigma_1, \sigma_2, ..., \sigma_n\}$ is a set of simplices. The dimension of a simplicial complex is the highest dimension of its simplices.

A d-chain c is the formal sum of d-simplices, $c = \sum_{\sigma \in K} \alpha_\sigma \sigma$, $\alpha_\sigma \in \mathbb{Z}_2$, where, $\mathbb{Z}_2$ is the binary space. A d-chain c is also a subset of the simplicial complex K, $c \subseteq K$. A d-chain can be denoted by a binary vector with length N_c, in which N_c is the number of d-simplices in K. The i-th entry in this binary vector is 1 if and only if $\sigma_i \in c$.

A boundary of a d-simplex is the formal sum of its $(d-1)$ faces. The boundary of a d-chain, represented by c, is the formal sum of boundaries of all d-simplices in c, $\partial_d(c) = \sum_{\sigma \in c} \partial_d(\sigma)$. When the chain is represented by a N_d-dimensional vector, the boundary operator is a $N_{d-1} \times N_d$ binary matrix, whose columns are the boundaries of d-simplices. Figure 2a and Figure 2b show two audio waves with peaks as the vertices. A edge is formed between two adjacent vertices, i.e., edges AC, AE, BC, and BD in Figure 2a and edges AD and AE in Figure 2b. Figure 2c and 2d are the boundary matrices of of 1-simplex (edge) in Figure 2a and Figure 2b, respectively.

Topology features of speech waves are analyzed with persistent homology under a filter function. A filter function is a function f: $X \to R$, where X is the topology space and R is the real coordinate space. The filtering function f induces a nested sequence of simplicial complexes, which are called filtration and denoted by $\emptyset = K_0 \subset K_1 \cdots K_p \cdots K_{m-1} \subset K_m = K$. For each simplical complex p, $K_p = \{\sigma \in K \mid f(\sigma) \leq p\}$. The sublevel set σ grows from an empty set to the whole image domain, Ω, when a threshold t increases continuously from $-\infty$ to $+\infty$. A homology class is born when a non-boundary cycle is added into the sublevel set. A homology class is dead when another homology class merges with it.

The persistence of a homology class is the difference between its death and birth times. The persistence can be calculated by reducing the boundary matrix. In the reduced matirx as shown in Figure 2d, the birth time and death time for 1-simplex edge BC is -2 and 2 because edge BC is born when vertex B appears at -2 and dead when edge BC formed at 2, and the corresponding persistence is 4. The homology class with relatively large persistence indicates the longevity feature in Ω. Persistence is able to quantify the significance of the features. In Figure 2c and Figure 2d, the numerical values with cyan background are the persistence ϕ of the homology class in the same column.

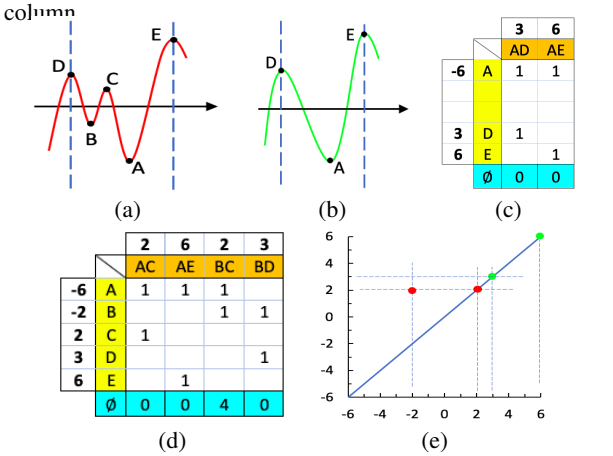

Figure 2: *(a) and (b) waves with several peaks. (c) and (d) are the reduced boundary matrices of simplical complex in (b) and (a), respectively. The numerical values besides the vertices or edges are the corresponding birth times. (e) a persistent diagram corresponding to the acoustic waves in (a) and (b) with persistence $\phi \geqslant 0$.*

Persistent homology can be visualized with a persistence diagram, where the horizontal and vertical axes represent the birth and death time, respectively. Each homology class corresponds to one dot in persistence diagram. Figure 2e shows the persistence diagram of the simplicial complex in Figure 2a and Figure 2b, where total four dots are corresponding to four non-zero columns in Figure 2d and two green dots are corresponding to two non-zero columns in Figure 2c. By comparing the distribution of homology classes in persistence diagram of two acoustic signals, the similarity of these two signal can be evaluated based on the topological features.

4. The proposed speech enhancement model

In this section, we introduce the proposed topology-enhanced GAN model. The GANs are generative neural networks where a generative model G and a discriminative model D are trained simultaneously via an adversarial process. The generator G is

expected to learn an effective mapping from real data distribution by training against the discriminator model D. D is designed as a binary classifier, which provides the feedback of classification results to G for updating. The adversarial learning process follows a mini-max algorithm, in which the GAN optimizes D by maximizing the probability of differentiate the samples from the real data and samples from G, and optimizes G by minimizing the difference between real data and data produced by G. The objective function of GAN is a cross-entropy loss as shown in Equation (1) and the training process is a mini-max game. $D(x)$ is the probability of x came from the real data rather than the G. $D(G(z))$ represents the probability of $G(z)$ came from G.

$$\begin{aligned} &\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{data}(x)}[logD(x)] \\ &+\mathbb{E}_{z\sim P_z(z)}[log(1-D(G(z)))] \end{aligned} \tag{1}$$

The network structure of generator G used in this study is an encoder-decoder structure [11] as shown in Figure 3. The input speech wave is handled and compressed through five convolutional blocks in the encoder. Each block includes a 1D convolution layer with a kernel size of 31 by 31, a stride of 4, and a PReLU [16] layer. The sizes of filters are 64, 128, 256, 512, 1,024 in these five blocks. The encoding vector c is concatenated with the noise vector z sampled from the normal distribution and presented to the decoder. In the decoder, five convolutional blocks with the same structures as the blocks in the encoder recover the speech signal to a speech wave. In order to prevent the loss of low-dimension speech information, some skip connections are created between the encoder and the decoder in G. The skip connections link each encoder layer to a decoder with the same dimensions for bypassing the signal compression in the following encoders, and thus the skip connections make the number of feature maps in every layer of the decoder to be doubled.

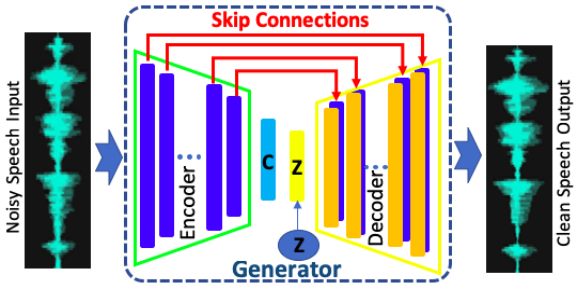

Figure 3: *The generator of GAN with an encoder and a decoder. Red arrows are skip connections between the encoder and the decoder.*

The discriminator D is a generally convolutional classifier with five convolutional blocks followed by three fully connected layers. Each convolutional block consists of one 1D convolution layer followed by one batch norm layer and PReLU layer. The input speech wave is narrower and narrower when it goes through the blocks. Finally, these three fully connected layers output the classification results to indicate whether the input speech wave is a clean speech or not.

The structure of the proposed model is shown in Figure 4. we measure the similarity of input wave and output wave from the generator of GAN. The similarity is used as part of penalized term in the objective function when training the generator. Our motivation is to force the generator to imitate the input noise speech while getting rid of the noise to generate clean speech. Firstly, the persistent diagrams of input wave and output waves from the generator of GAN are calculated by following the process shown in Section 2. The persistent diagrams include the global features as peak-to-peak distribution of speech wave. Secondly, we use the Wasserstein distance to measure [17] the similarity between two persistence diagrams. The Wasserstein distance measures the minimal distance achieved by a perfect matching between the points of the two persistent diagrams. Finally, we define the penalty term on the objective function of GAN as shown in Equation 2. The Algorithm 1 shows the details of the proposed model.

$$\begin{aligned} &\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{data}(x)}[logD(x)] \\ &+\mathbb{E}_{z\sim P_z(z)}[log(1-D(G(z)))] \\ &+\eta\Psi(P(x),P(G(z)))_{z\sim P_z(z)} \end{aligned} \tag{2}$$

Where, $P(x)$ and $P(G(z))$ represent the persistent diagrams of clean speech and generated speech from corresponding noisy speech. $\Psi()$ is the Wasserstein distance function. η is the weight of persistent penalty, which is needed to be tuned.

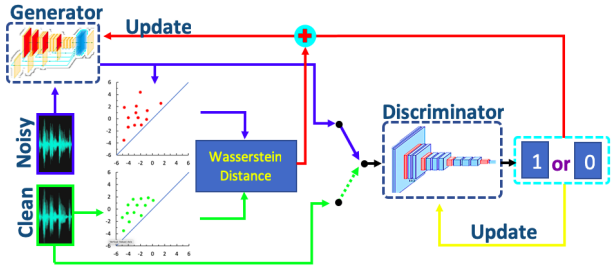

Figure 4: *The structure of the proposed topology-enhanced GAN for speech enhancement.*

Algorithm 1 The proposed topology-enhanced GAN model training with minibatch.

for $epoch = 0$ **do:**
  while $i < I$ **do:**
    Sample the minibatch:
      Sample minibatch of M noise samples from noisy speech as: $z \sim \{z|z \in \vartheta\}$.
      Sample minibatch of M samples from clean speech as: $x \sim \{x|x \in \chi\}$.
      Update the weights of the discriminator by maximizing the loss function:
      $\frac{1}{m}\sum_{i=1}^{m}\left[logD(x^{(i)}) + log(1-D(G^{(i)}))\right]$
    $i = i + 1$
  end while
    Sample minibatch of M noise samples from noisy speech as: $z \sim \{z|z \in \vartheta\}$.
    Update the weights of the generator by minimizing the loss function:
    $\frac{1}{m}\sum_{i=1}^{m}\left[log(1-D(G(z^{(i)}))) + \eta\Psi(P(x^{(i)}),P(G^{(i)})\right]$
  $epoch = epoch + 1$
end for
Where, ϑ and χ are the datasets of noisy and clean speech, respectively. The $G^{(i)}$ represents the $G(z^{(i)})$.

5. Experiments and results

5.1. Data set and parameter settings

We train and evaluate the proposed model on a public speech data set [18], including Voice Bank corpus [19] with 28 speakers (14 males and 14 females) from the same accent region

(England) and another 56 speakers (28 male and 28 female) from different accent regions (Scotland and United States). Each speaker provides roughly 400 sentences. The dataset includes general speech samples with 4 signal-to-noise ratios (SNRs) (15 dB, 10 dB, 5 dB, and 0 dB) and 10 types of noises (2 artificial and 8 from the Demand database [20]. A sliding window with a size of one second is used to segment the speeches continuously with 500 milliseconds overlap for feeding to the discriminator and generator. The proposed model is trained by using Adam optimization with $b_1 = 0.9, b_2 = 0.99, \epsilon = 1e^{-8}$ for 100 epochs. The mini-batch size is 100.

5.2. Objective evaluation

We evaluate the topology-enhanced GAN model on the testing data set and compare its performance with several existing speech enhancement models including ISEGAN [21], DSEGAN [21], SEGAN [11], DNN [8]. The evaluation metrics are (1) PESQ: Perceptual evaluation of speech quality, using the wide-band version recommended in ITU-T P.862.2 (from -0.5 to 4.5) [22]. (2) CSIG: Mean opinion score (MOS) prediction of the signal distortion attending only to the speech signal (from 1 to 5) [23]. (3) CBAK: MOS prediction of the intrusiveness of background noise (from 1 to 5) [23]. (4) COVL: MOS prediction of the overall effect (from 1 to 5) [23]. (5) SSNR: Segmental signal-to-noise ratio (from 0 to ∞) [24]. Evaluation results summaried in Table 1 indicates that the proposed model achieves the best performance in terms of PESQ, CBAK, SSNR, and maintains strong performance in CSIG and COVL. If we assign the ranks from 1 to 6 (6 models including **Noisy**) for each metric, the total ranks for **Noisy**, **ISEGAN**, **DSEGAN**, **SEGAN**, **DNN**, and the **proposed model** are 28, 21, 11, 20, 14, and 10. Our proposed model has the first rank of 10, which indicates the proposed model has a better comprehensive performance. Figure 5 shows the spectrum of noisy, clean, and enhanced speeches. Compared with the available model SEGAN, the proposed model weakens the amplitude in noisy areas (areas with blue color). The objective evaluation results indicate our proposed topology-enhanced GAN model clearly outperforms existing models in this experiment.

Table 1: *The objective evaluation results from different speech-enhanced models. Red and blue colors represent the first and second ranks, respectively.*

Metric	Noisy	ISEGAN	DSEGAN	SEGAN	DNN	Ours
PESQ	1.97	2.24	2.39	2.19	2.45	2.56
CSIG	3.35	3.23	3.46	3.39	3.73	3.25
CBAK	2.44	2.95	3.11	2.90	2.89	3.26
COVL	2.63	2.69	2.90	2.76	3.09	2.90
SSNR	1.68	8.17	8.72	7.36	3.64	9.76

5.3. Subjective evaluation

We randomly selected 30 sentences with different environmental background noises from the testing data set for subjective evaluation. Each sentence has three versions, including a noisy speeche and enhanced speeches generated by SEGAN and the proposed topology-enhanced GAN. We do not include other models due to the unavailability of the enhanced speeches. We use MOS (Mean opinion score) to evaluate the quality of speeches. A total of 90 speeches are evaluated by 5 candidates with scores from 1 to 5. For the scores, 1 represents the speech has a heavy degradation and apparent noise and 5 means the speech has no degradation and no noticeable noise. Figure 6 shows the evaluation results. Our proposed model has better performance than the SEGAN in speech-enhancement tasks because the proposed model achieves a higher MOS with narrower confidence interval.

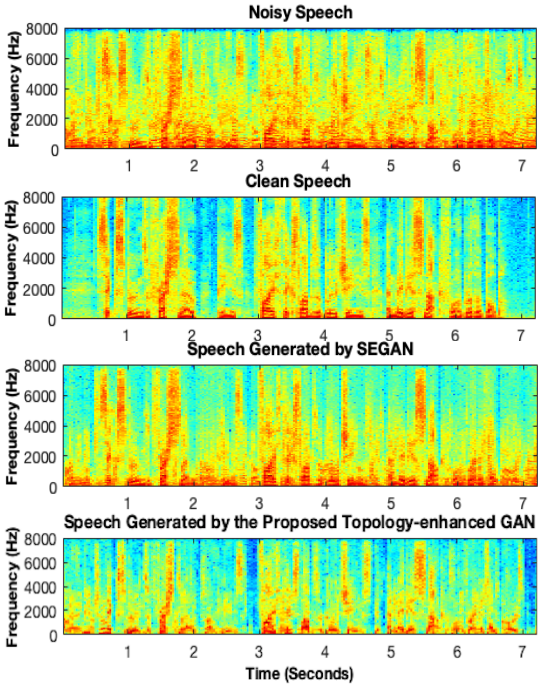

Figure 5: *The spectrum of noisy speeches, clean speech, SEGAN-enhanced speech, and enhanced speech by proposed model.*

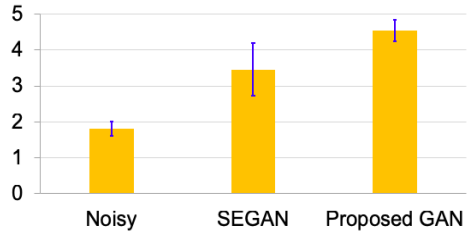

Figure 6: *The MOS of the noisy speeches and speeches generated by SEGAN and the proposed topology-enhanced GAN with 95% confidence interval.*

6. Conclusion

In this study, we propose a topology-enhanced GAN model to improve speech enhancement. Persistence homology is used to extract the features of speech waves for improving the generator of GAN, and the Wasserstein distance of topology features is used as an additional penalty term in the modified objective function of GAN. Both objective and subjective evaluations indicate that our proposed model can identify and weaken the noise amplitude in speeches significantly better than existing models. The experimental results also indicate the Wasserstein distance between the distribution of topology features of acoustic waves can be used to evaluate the similarity of wave morphologies and improve the generative process in GAN for speech enhancement.

7. References

[1] H.-M. Liu, P. K. Kuhl, and F.-M. Tsao, "An association between mothers' speech clarity and infants' speech discrimination skills," *Developmental science*, vol. 6, no. 3, pp. F1–F10, 2003.

[2] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment*. John Wiley & Sons, 2006.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[4] C. Zorilă, C. Boeddeker, R. Doddipatla, and R. Haeb-Umbach, "An investigation into the effectiveness of enhancement in asr training and test for chime-5 dinner party transcription," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 47–53.

[5] R. Martin, "Spectral subtraction based on minimum statistics," *power*, vol. 6, no. 8, 1994.

[6] N. Upadhyay and R. K. Jaiswal, "Single channel speech enhancement: using wiener filtering with recursive noise estimation," *Procedia Computer Science*, vol. 84, pp. 22–30, 2016.

[7] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *International conference on latent variable analysis and signal separation*. Springer, 2015, pp. 91–99.

[8] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[9] A. Kumar and D. Florencio, "Speech enhancement in multiple-noise conditions using deep neural networks," *arXiv preprint arXiv:1605.02427*, 2016.

[10] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[11] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[12] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[13] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement." in *Interspeech*, 2018, pp. 3229–3233.

[14] X. Zhang and P. Wu, "Heuristic search for homology localization problem and its application in cardiac trabeculae reconstruction," in *IJCAI*, 2019.

[15] S. Emrani, T. Gentimis, and H. Krim, "Persistent homology of delay embeddings and its application to wheeze detection," *IEEE Signal Processing Letters*, vol. 21, no. 4, pp. 459–463, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[17] N. Bonneel, M. Van De Panne, S. Paris, and W. Heidrich, "Displacement interpolation using lagrangian mass transport," in *Proceedings of the 2011 SIGGRAPH Asia Conference*, 2011, pp. 1–12.

[18] C. Valentini-Botinhao *et al.*, "Noisy reverberant speech database for training speech enhancement algorithms and tts models," 2017.

[19] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*. IEEE, 2013, pp. 1–4.

[20] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, no. 1. Acoustical Society of America, 2013, p. 035081.

[21] H. Phan, I. V. McLoughlin, L. Pham, O. Y. Chén, P. Koch, M. De Vos, and A. Mertins, "Improving gans for speech enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 1700–1704, 2020.

[22] I.-T. Recommendation, "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.

[23] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.

[24] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*. Prentice-Hall, 1988.