INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

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

jzha1990@gmail.com, 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 (RNNs), long short-term memory networks (LSTMs), 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 enhanced speeches still include noisy signals. In this study, we propose a topology-enhanced GAN model to tackle noisy speeches in an end-to-end structure. We use the topology features of speech waves as additional constraints and modify the objective function of the GAN by adding a penalty term. The penalty term is a Wasserstein distance of topology features measuring the difference between the generated speech and the corresponding clean speech. 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 GANs 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 certain frequencies [2]. Many studies have been focusing on removing the background noise in speeches [3], as it is a particularly important task in speech recognition (ASR) systems [4]. It has broad applications because noises naturally occur in speech generation during telephone conversations, online meetings, and television programs.

One of the popular methods to remove background noises from speech signals is a technique called spectral subtraction [5]. The amplitude spectra of speeches indicate the amplitudes of speech waves in dB scale. Speeches can also be enhanced by using Wiener filter, which minimizes the mean-squared error between the original and the 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 suffer from differentiating the speeches from complex noises.

* Corresponding author.

Recently, many researchers have obtained promising results in speech enhancement using deep learning models, such as Long Short Term Memory (LSTM) models [7] and various deep neural networks (DNN) models [8, 9]. Some models weaken short-time phases because the spectra of speeches with short-time phases are considered to be unimportant in Fourier analysis [10]. Speech Enhancement Generative Adversarial Network (SEGAN) [11] was proposed to provide an 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 clean speeches from noisy speeches due to the lack of morphological features of speech waves. Figure 1 shows the clean wave and the corresponding GAN-enhanced speech wave from the same noisy speech wave. GAN produces a relatively clean speech wave but also generates additional waves compared with the clean speech.

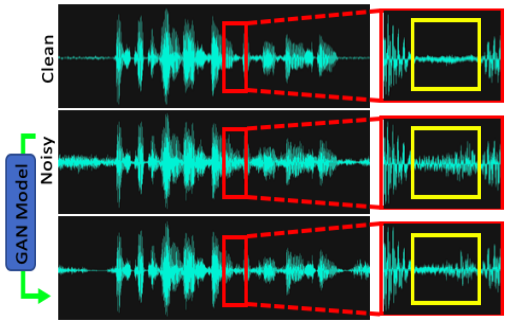

Figure 1: *The clean, noisy and GAN-generated speech waves and the detailed differences shown in yellow squares.*

To improve existing deep learning models for speech enhancements, 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 in model training. Persistent homology is a method for computing topological features of a space at different spatial resolutions in arithmetic topology [12]. The Wasserstein distance of the persistence of the extracted features is calculated as a penalty term in 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, which effectively avoids unnecessary noisy speech signals and keeps critical speech signals. We validate this proposed approach with extensive experiments on several released benchmark datasets, and the results

http://dx.doi.org/10.21437/Interspeech.2021-1411

are highly promising in metrics of PESQ, CSIG, CBAK, COVL, and SSNR [13].

The rest of the paper is organized as follows. Section 2 introduces the related work on speech enhancements. 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 conclusions of this study.

2. Related work

With the development of DNN over the last decades, the speech enhancement approaches have been converted from the spectral analysis [5] to DNN-based enhancement. Pascual et al. proposed an end-to-end SEGAN to tackle the speech enhancement from the raw speech waves [11]. The experimental results indicated the proposed SEGAN enhanced speeches in both objective evaluations and subjective evaluations. Rethage et al. introduced an end-to-end learning method, namely Wavenet, for speech denoising [14]. The model used 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 [15]. The experiments suggested that the proposed model led to consistently better objective intelligibility and perceptual quality than the existing LSTM-based model. A persistent homology is a powerful tool in capturing topological properties of information at different spatial resolutions. The persistent homology had been successfully used to segment the trabecula from 3D CT images [16]. The experimental results indicated that the persistent homology can capture the fine tissues in a noisy environment. Emrani et al. proposed to use persistent homology for robust analysis of point clouds of delay-coordinate embeddings in detection [17]. The experimental results showed that the proposed method was preferred over Wiener filtering algorithm in quantitative and qualitative evaluations.

3. Preliminaries

This section introduces notations used for definitions of persistent homology. A d-dimensional simplex σ is the convex hull of $d+1$ affinely independent points. For instance, a 0-simplex, a 1-simplex, and a 2-simplex are a vertex, an edge, and a triangle in the 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 a 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 a 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 an N_d-dimensional vector, the boundary operator is an $N_{d-1} \times N_d$ binary matrix, whose columns are the boundaries of d-simplices.

Figure 2a and Figure 2b show two audio waves with peaks as the vertices. An 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 Figure 2d are the boundary matrices of of 1-simplex (edge) in Figure 2b and Figure 2a, respectively.

Topology features of speech waves are analyzed with persistent homology under a filter function. A filter function is a function $f\colon X \to R$, 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 is 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 time and its birth time. The persistence can be calculated by reducing the boundary matrix. In the reduced matirx as shown in Figure 2d, the birth time and the death time for the 1-simplex edge BC is -2 and 2 because the edge BC is born when the vertex B appears at -2 and dead when the edge BC formed at 2, and the corresponding persistence is 4. The homology class with a relatively large persistence indicates a 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.

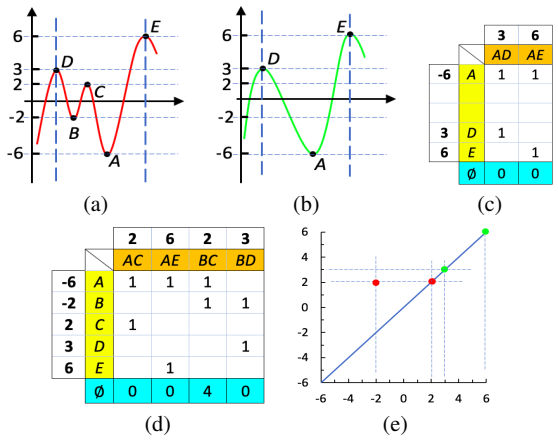

Figure 2: *(a) and (b) waves with several peaks. (c) and (d) reduced boundary matrices of simplical complex in (b) and (a), respectively. The numerical values besides the vertices or edges are the corresponding birth times. (e) a persistent diagram corresponding to the acoustic waves in (a) and (b) with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

Persistent homology can be visualized with a persistence diagram, where the horizontal and vertical axes represent the birth time and the death time, respectively. Each homology class corresponds to one dot in the persistence diagram. Figure 2e shows the persistence diagram of the simplicial complex in Figure 2a and Figure 2b, where total four dots (2 red dots and 2 green 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 signals 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 GAN includes a generative model, G, and a discriminative model, D, and both G and D are trained simultaneously via an adversarial process. G is expected to learn an effective mapping from the real data distribution by training against D. D is designed as a binary classifier, which provides feedbacks 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 differentiating the samples from the real data and samples generated 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 coming 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 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 [18] 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 doubled.

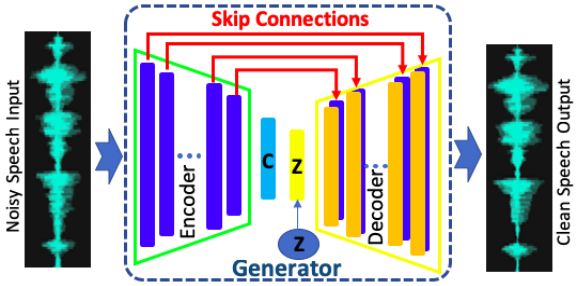

Figure 3: *The generator of GAN with an encoder and a decoder. Red arrows are skip connections between the encoder and the decoder.*

The discriminator, D, is a generally convolutional classifier with five convolutional blocks followed by three fully connected layers. Each convolutional block consists of a 1D convolution layer followed by a batch norm layer and a PReLU layer. The input speech wave becomes 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 an input wave and an 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 encourage the generator to imitate the input noise speech while getting rid of the noise to generate clean speeches. Firstly, the persistent diagrams of the input wave and output waves from the generator of GAN are calculated by following the process shown in Section 3. The persistent diagrams include the global features as peak-to-peak distribution of speech waves. Secondly, we use the Wasserstein distance [19] to measure the similarity between two persistence diagrams. The Wasserstein distance measures the minimal distance achieved by a perfect matching between the homology classes indicated by two persistent diagrams. Finally, we define a penalty term on the objective function of GAN as shown in Equation 2. Algorithm 1 shows the training 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 the clean speech and the generated speech from corresponding noisy speech, respectively. $\Psi()$ is the Wasserstein distance function. η is the weight of the persistent penalty, which needs to be tuned.

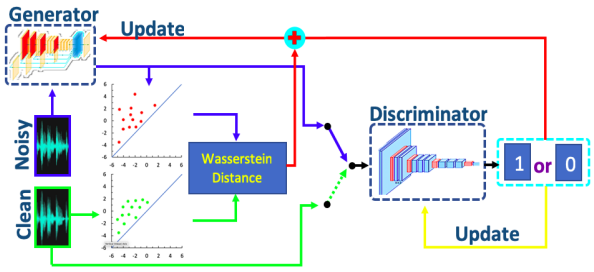

Figure 4: *The structure of the proposed topology-enhanced GAN for speech enhancement.*

Algorithm 1 Training of the proposed topology-enhanced GAN

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: $\frac{1}{m}\sum_{i=1}^{m}\left[logD(x^i) + log(1-D(G(z^i)))\right]$
    $i = i + 1$
  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: $\frac{1}{m}\sum_{i=1}^{m}\left[log(1-D(G(z^i))) + \eta\Psi(P(x^i),P(G(z^i))\right]$
  $epoch = epoch + 1$
end for

Where, i is the index of minibatch. ϑ and χ are the datasets of noisy and clean speech, respectively.

5. Experiments and results

5.1. Data set and parameter settings

We train and evaluate the proposed model on a public speech data set [20], including Voice Bank corpus [21] with 28 speakers (14 males and 14 females) from the same accent region (England) and another 56 speakers (28 males and 28 females) from different accent regions (Scotland and United States). Each speaker provides roughly 400 sentences with a sampling rate of 48KHz. 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 types and 8 natural types) from the Demand database [22]. A sliding window with a size of one second is used to segment the speeches continuously with a 500-millisecond overlap. The proposed model is trained by using the Adam optimization with $\beta_1 = 0.9, \beta_2 = 0.99, \epsilon = 1e^{-8}$ for 100 epochs [23]. The mini-batch size is 100.

5.2. Objective evaluation

We evaluate the proposed topology-enhanced GAN model on the testing data set and compare its performance with several existing speech enhancement models including ISEGAN [24], DSEGAN [24], SEGAN [11], DNN [8]. The evaluation metrics are (1) PESQ: Perceptual evaluation of speech quality $(-0.5 \sim 4.5)$ [13]. (2) CSIG: Mean opinion score (MOS) prediction of the signal distortion attending only to the speech signal $(1 \sim 5)$ [25]. (3) CBAK: MOS prediction of the intrusiveness of background noise $(1 \sim 5)$ [25]. (4) COVL: MOS prediction of the overall effect $(1 \sim 5)$ [25]. (5) SSNR: Segmental signal-to-noise ratio $(0 \sim \infty)$ [26]. Evaluation results summaried in Table 1 indicate 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) 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 model has the first rank of 10. Figure 5 shows the spectra of noisy, clean, and enhanced speeches. Compared with the available model SEGAN, the proposed model weakens the amplitude in noisy areas (areas in blue). The objective evaluation results indicate our proposed topology-enhanced GAN model has a better comprehensive performance than other models in this experiment.

Table 1: *The objective evaluation results. Areas in red and blue represent the first rank and the second rank, 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 from the testing data set for the subjective evaluation. Each sentence has three versions, including a noisy speech and two enhanced speeches generated by SEGAN and our model. Other models are not included due to the unavailability of enhanced speeches. We use MOS (Mean opinion score) to evaluate the quality of speeches. A total of 90 speeches are evaluated by 12 candidates with scores from 1 to 5.

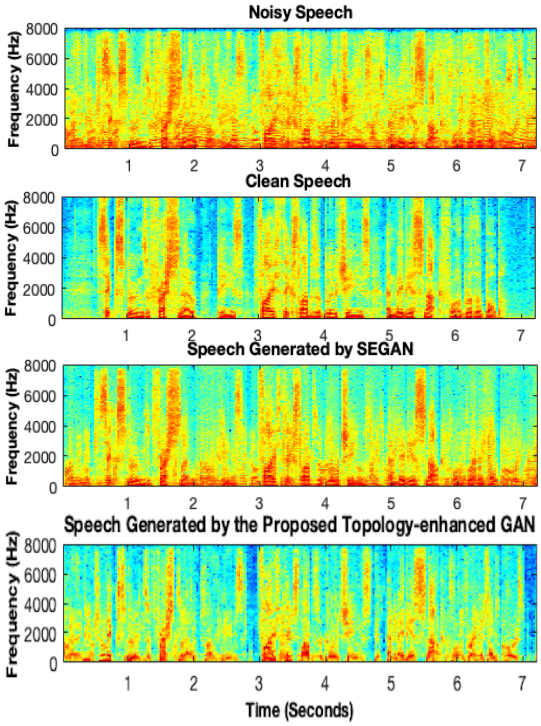

Figure 5: *The spectrum of a noisy speech, a clean speech, a SEGAN-enhanced speech, and an enhanced speech by the proposed model.*

For the scores, 1 represents the speech has a heavy degradation and apparent noise, and 5 means the speech has no degradations and no noticeable noises. Figure 6 shows the evaluation results. Our proposed model has better performance than the SEGAN in speech-enhancement tasks because our model achieves a higher MOS with a narrower 95% confidence interval.

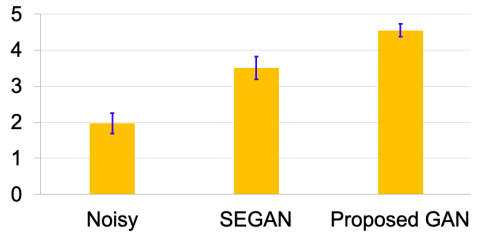

Figure 6: *The MOS of the noisy speeches and the enhanced speeches with 95% confidence intervals.*

6. Conclusions

In this study, we propose a topology-enhanced GAN model to improve speech enhancements. Persistent homology is used to extract the features of speech waves. The Wasserstein distance of topology features is used as an additional penalty term for improving the generator of the GAN. Both the objective evaluation and the subjective evaluation indicate that our proposed model can identify and weaken the noise amplitude in speeches significantly, which is better than existing models. Also, 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 enhancements.

7. References

[1] H. Liu, P. K. Kuhl, and F. 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. Dai, and C. 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] H. Edelsbrunner and J. Harer, *Computational Topology: An Introduction*. American Mathematical Soc., 2010.

[13] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 2001, pp. 749–752.

[14] 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.

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

[16] X. Zhang, P. Wu, C. Yuan, Y. Wang, D. Metaxas, and C. Chen, "Heuristic search for homology localization problem and its application in cardiac trabeculae reconstruction," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 2019, pp. 1312–1318.

[17] 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.

[18] 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.

[19] 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.

[20] Valentini-Botinhao and Cassia, "Noisy reverberant speech database for training speech enhancement algorithms and TTS models," 2017.

[21] 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.

[22] 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.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] 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.

[25] 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.

[26] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. Prentice-Hall, 1988.