International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Data Generation for Speech Recognition based on Generative Adversarial Networks

R. Lavanya[1], K.B. Kishore Mohan[2], G. Gomathy[3], Appana Naga Lakshmi[4], R. Salini[5]

Submitted: 14/12/2023 **Revised:** 26/01/2024 **Accepted:** 05/02/2024

Abstract: Individuals who are deaf or dumb are likely to derive extra benefit via a speech recognition system that uses GANs. However distracting outside circumstances, individuals will find it straightforward to grasp the information. The speech enhancement approaches prevalent currently work in the frequency domain and/or take the benefit of higher-level elements. Many of them utilize first-order analytics and only solve a restricted set of noise scenarios. Deep networks are being adopted increasingly to get around these drawbacks as a result of their ability to learn challenging tasks from sizable sample datasets. In this paper, a GAN-based strategy is proposed for generating synthetic data for speech emotion recognition. More specifically, we glance into using GANs for collecting the data stream. We examine the implementation of Generative Adversarial Networks (GANs) for trained data enrichment to yield samples for disproportionately represented emotions. The updated specimens demonstrate the recommended model's viability, and appraisals from both specialists and laypeople encourage its efficacy. In doing so, we start looking into generative architectures for voice enhancements, which may gradually comprise more speech-centric design choices to improve their functionality.

Keywords: *Generative Adversarial Networks; Speech Recognition; Speech Generation; SEGAN*

1. Introduction

Eliminating background noise constitutes one of the main objectives of speech augmentation technology. Regardless of whether stated or not, these are crucial elements of automated speech recognition (ASR) systems. The primary objective is to make speech easier to absorb, which will boost ASR's immunity to noise. Deep neural networks are currently frequently utilized alongside the use of standard signal processing methods like harmonic subtraction and Wiener adjustments to either instantly relive clear speech or estimate covers from the noisy information. Speech synthesis footage is typically produced in controlled circumstances with minimal interference from outside noise or reverberation. Still, there are (at least) two scenarios in which the environment cannot be managed: (a) when mobile devices are being utilized to make recordings, and (b) when speech omitted from recordings of excellent quality are conversely engaging but are unable to be captured again. As a result, there is a considerable market for a speech enhancement (SE) strategy developed for speech synthesis (SS) and speech adjusting in halfway-15–30 dB signal–to–noise (SNR) environments. Conventional methods like spectrum elimination or Wiener filtering methods work effectively in higher (30–50 dB SNR) situations. Their utility gets limited in mid-SNR levels, nonetheless. More recently, this author published a recurrent network-based SE approach for noise-robust speech synthesis. However, because this method works in the feature domain rather than the waveform field, stronger speech inadvertently incorporates vocoding fidelity. This technique has been examined in low SNR (0–15 dB) situations. On the flip side, the switch from the pattern domain to the waveform domain culminated in considerable improvements in voice transmission and speech synthesis, with spoken word quality nearly identical to that of real recordings [1].

In principle, pristine speech and reverberant speech can connect in the time-space as a room impulse response (RIR) voice dereverberation is an easy approach that entails filtering out resonance from the contaminated voice. Despite the microphone arrangement and multilingual signal conditioning in this track being extremely helpful, single-channel audio reverberation remains the choice in several real-world situations where it cannot always be feasible to use several microphones. In the arena of signal processing, single-microphone speech dereverberation has been substantially discovered, and multiple techniques have been proposed. Deep neural networks (DNNs) were originally working in voice augmentation and subsequent speech dereverberation because of their high relapse learning skills. Naturally, the complex structure can be envisioned as a dereverberation filter that, granted an array of multi-condition

[1]*Assistant Professor, Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur*
Email: lavanya27382@gmail.com
ORCID: 0000-0001-6383-6209
[2]*Professor & Head, Department of Bio Medical Engineering, Sri Shanmugha College of Engineering and Technology - [SSCET], Sankari, Salem*
Email: kishorekbmtech@yahoo.co.in
ORCID: 0009-0002-6305-1457
[3]*HOD, EEE Department, Jaya Engineering College, Chennai-24, Thiruvallur District, Tamil Nadu, Chennai*
Email: gomathypaul@gmail.com
ORCID: 0009-0007-1367-0715
[4]*Assistant Professor, Artificial Intelligence, Madanapalle Institute of Technology & Science, Madanapalle, A.P*
Email: anlassistantprofessor@gmail.com
ORCID: 0009-0008-1000-3957
[5]*Assistant Professor, Department of CSE, Panimalar Engineering College, Chennai*
Email: shalinirajendran@gmail.com
ORCID: 0000-0001-7266-3389

data, can figure out the fundamental connection between the reverberant speech and its symmetrical equivalent, or the pristine speech [2]. Recurrent neural networks (RNNs) are also selected, though. A good instance of this is the recurrent blurring autoencoder, which achieved noteworthy efficiency by taking full advantage of the time-based context knowledge contained in embedded impulses. The vast majority of current techniques solve the denoising issue using extended short-term memory networks. In the investigation papers, noise attributes are identified and contributed to DNN input traits. It was successfully shown that failure, post-filtering, and perceptually oriented measures work successfully. The short-time Fourier analysis/synthesis architecture forms the backbone of the vast majority of the algorithms in use currently. They merely alter the power of the spectrum given that it is widely accepted that the short-time portion has nothing to do on enhancing pronunciation. Further investigation, however, suggests major enhancements in speech quality are attainable, most notably in the scenario of an established phase spectrum. The creator of a widely recognized academic paper recommended a deep network that functioned directly on the unprocessed audio waveform, but they additionally created feed-forward layers that functioned frame-by-frame (60 samples) on an isolated-word collection that hinged on the speaker [3].

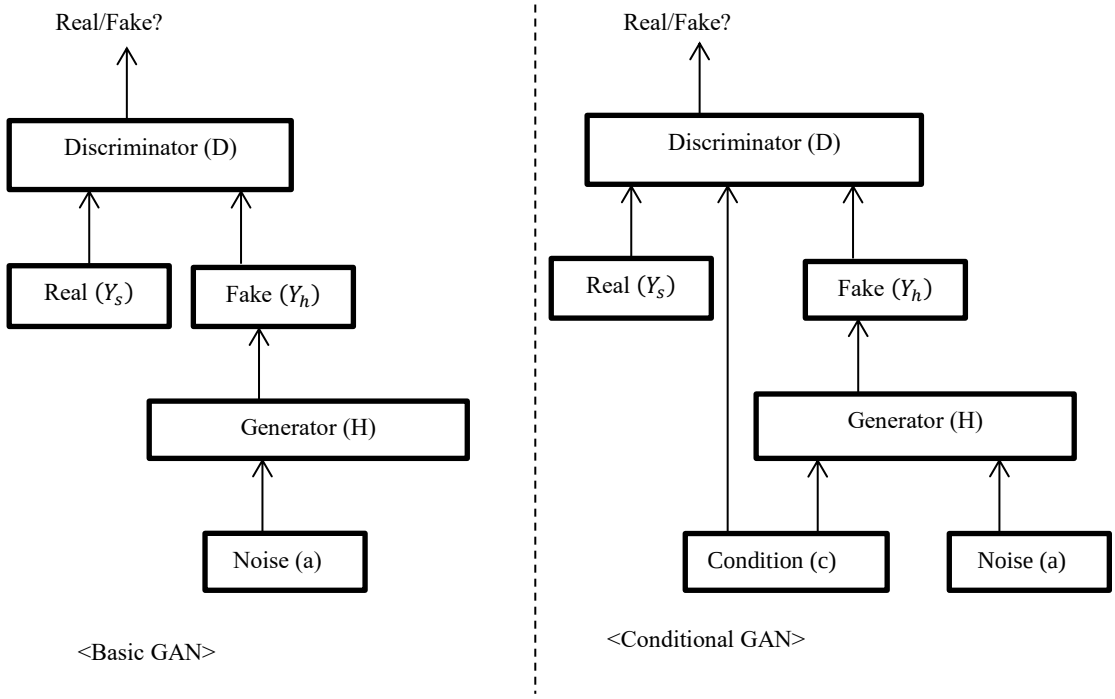


Fig.1. Architecture of basic GAN and conditional GAN

The generative adversarial network, or GAN, was initially brought in as a robust yielding model with a wide range of uses. Two neural systems network structures make up a basic GAN, illustrated in the left part Figure 1: The real instances and fake samples are categorized by a discriminator E; instances originate from information dispersion, which is regularly truncated dimensional subjective flash by a generator H. When the discriminator undergoes training to separate between actual and fraudulent specimens, the generator is programmed to confuse it. The original GAN frameworks lacked scenario knowledge, suggesting they required an ability to guide the manner of generating knowledge. As a consequence, conditional GAN is displayed by incorporating additional conditional info, which appears as the right portion in Figure 1. A desirable form of data can be produced via conditional GAN with the integrated parameter [4].

Below is an outline of the numerous elements that collectively make up the essay. A discussion of the main preceding investigations is offered in Section 2. The third portion provides a discussion of the proposed Generative Adversarial Networks, encompassing its mentioned schemes, accomplishment basis, graph-based workflow fields, and data processing. GAN is analyzed in Section 4 to figure out speech recognition based on distinct graphs and instances. Section 5, which involves the conclusion, analyzes the last concern.

2. Related works

Vijay, I. et. al [5] Following that, the discriminator performs a conscious decision to differentiate between actual information and erroneous data. Since both of these networks represent two distinct varieties of CNNs, they will consistently enhance and function based on consumer input. In this article, we'll provide an immense dataset that comprises both noise and multiple sounds together. Now, the data will be transferred to the templates that have the GAN framework. Now, the generator neural network obtains the data. The generator merges the synthetic clones it constructs of the original dataset with the genuine data. Following that, the filtering neural network collects the data.

Donahue, C. et. al [6] Unsupervised generative algorithms termed generative adversarial networks (GANs) educate from low-dimensional, arbitrary latent matrices to produce genuine examples of a specific data set. A generator and a discriminator are the double separate models (often neural systems) that collectively form a GAN. We further investigate the accuracy of an earlier ASR model educated with MTR on loud speech with and without augmenting to figure out this effect. Even after

retraining, we demonstrate that the accuracy of this model is diminished by GAN-based boosting. Productivity gets better nevertheless, when the MTR algorithm undergoes retraining with both loud and strengthened attributes in its input version.

Sriram, A. et. al [7] In the unsupervised learning paradigm referred to as GAN, the generator network develops the capacity to produce data that is progressively more convincing in an attempt to lure in a rival discriminator. Equilibrium arrives at a saddle location, keeping education extremely tricky. This approach has been enhanced extensively. For example, the Earth-Mover range can be applied by Wasserstein GAN to reduce optimization problems. It is also less vulnerable to decisions made in architecture. Lacking the goal of acquiring speech recognition as its final objective provides SEGAN, a GAN-based speech improvement methodology. In response to its dependency on untreated audio samples, SEGAN is not technically feasible for extensive study.

Benzeghiba, M. et. al [8] Lacking the goal of acquiring speech recognition as its final objective, provides SEGAN, a GAN-based speech improvement methodology. In response to its dependency on untreated audio samples, SEGAN is not technically feasible for extensive study. As an outcome, it is probable that a great deal of variability-proof ASR techniques effectively address numerous elements that can lead to reminiscent speech modification. In order to grade pronunciation correctness in foreign proficiency, a voice recognition technique is also applied. Analyses established that when there is significant speech offered, the score that is given aligns with expert judgments from individuals.

Chatziagapi, A. et. al [9] In this research, we emphasize on spectrogram formation for the minority sentimental categories, broadening the strategy of the SER sector. To boost the caliber of the spectral images that are generated, we suggest implementing alterations to both the training practices and the first system topology. The proposed technique resolves data imbalance in a better way, as illustrated by extensive research observations on the earlier stated solution and several substitute audio data augmentation techniques. To the greatest extent of our knowledge, this is the first occasion that GANs have been incorporated into SER or another type of sound categorization task to deal with an issue of data mismatch through data enhancement.

Hu, H. et. al [10] The machine intelligence teams have displayed an enormous amount of curiosity about GANs. Through adversarial conditioning, which produces observations from genuine data dispersion, it has the potential to form predictive models. Improved performance can be obtained by further enhancing the training procedure and loss function of the Wasserstein GAN (WGAN), which is built on the fundamental GAN. Provisional usages for GAN in speech analysis encompass verbal language recognition, voice transformation, speech improvement, speech generation, and perhaps even acoustic scene assessment. Still, a little work has been invested in voice recognition. In this investigation, we provide an innovative approach to noise-robust detection of speech using generative adversarial networks as a data enhancement tool. We deploy an intricate very deep convolutional neural network (VDCNN) as our primary audio paradigm.

Goodfellow, I. et. al [11] The vast majority of investigations on deep generative models focused on architectures that presented a spectrum of probability function's quantitative description. The log potential is subsequently improved to train the framework. The deep Boltzmann machinery is likely the most effective concept in this family. These representations usually consist of challenging confidence distributions, forcing many iterations of the probabilistic gradient. These difficulties prompted the establishment of "generative machines"—models that will generate instances from the appropriate dispersion even when they don't convey chances right away A generative engine that can be developed with precise backpropagation as opposed to the multiple guesses wanted for Boltzmann machines is a dynamic stochastic network. By carrying out something with the Markov chains that exist in generative stochastic networks, this research advances on the premise of a creative device.

3. Methods and Materials

3.1. Generative Adversarial Networks

A straightforward GAN comprised of a detector that seeks to separate generated data from actual ones and a device that retrieves real-like data from sample selections. The generator H, given an array of random specimens A's from a probability distribution, changes those models to copycat the pattern of actual information y, allowing the discriminator to acknowledge samples produced $H(a)$'s as real. The discriminator E operates in the background to differentiate the real instances $(y)$from the bogus signals $H(a)$. The subsequent section is a phrase of these targets:

$$\min_H \max_E W(E,H) = \mathrm{F}_{y\sim q_{data}(y)}\{\log E(y)\} + \mathrm{F}_{a\sim q_a(a)}\left\{\log\left(1-E\left(H(a)\right)\right)\right\} \quad (1)$$

In practical terms, as suggested we can train generator H to maximize log(D(G(z))), in contrast to educating it to minimize $\log\left(1-E\left(H(a)\right)\right)$. Greater variations that assist to solve the slope diminishing challenge can be generated via this goal function without impacting the balance point that is attained by the discriminator E and generator H. E and H were taught to limit the declines outlined in equations (2) and (3) utilizing these new objective tasks.

$$\mathcal{M}_E^{(GAN)} = -\mathrm{F}_{y\sim q_{data}(y)}\{\log E(y)\} - \mathrm{F}_{a\sim q_a(a)}\left\{\log\left(1-E\left(H(a)\right)\right)\right\} \quad (2)$$

$$\mathcal{M}_E^{(GAN)} = -\mathrm{F}_{a\sim q_a(a)}\left\{\log\left(E\left(H(a)\right)\right)\right\} \quad (3)$$

3.2. Conditional Generative Adversarial Networks

Conditional GAN is a refinement of the vanilla GAN that takes into account extra data, which involves class labels or additional kinds of information. The data entered for the generator is the combination of y and random vector z; the juxtaposition of actual data y and additional data z is the source for the discriminator. As a result, the cGAN's goal function is

$$\min_H \max_E W(E,H) = \mathrm{F}_{y\sim q_{data}(y)}\{\log E([y,z])\} + \mathrm{F}_{a\sim q_a(a)}\{\log(1-E([H([a,z]),z]))\} \quad (4)$$

where the combining vectors x and y are labeled by $[y,z]$. As with the vanilla GAN, the discriminator and generator are taught to lessen the associated damages:

$$\mathcal{M}_E^{(cGAN)} = -\mathrm{F}_{y\sim q_{data}(y)}\{\log E([y,z])\} - \mathrm{F}_{a\sim q_a(a)}\{\log(1-E([H([a,z]),z]))\} \quad (5)$$

$$\mathcal{M}_E^{(cGAN)} = -\mathrm{F}_{a\sim q_a(a)}\{\log(E([H([a,z]),z]))\} \quad (6)$$

3.3. Adversarial Autoencoders

An encoder, a decoder, and a discriminator are all involved in an adversarial autoencoder (AAE). The encoding unit in GANs performs out functions identical to those of the generator. Nevertheless, unlike GANs, which manufacture fake samples, the encoder in AAEs wants to pair an accumulated posterior to an arbitrary prior. The encoder F and the decoder S are both programmed to mitigate reconditioning error, as opposed to the adversarial learning target:

$$\mathcal{M}_E^{(AAE)} = -\mathrm{F}_{y\sim q_a(a)}\{\log E([a])\} - \mathrm{F}_{y\sim q_{data}(y)}\left\{\log\left(1 - E\big(F(y)\big)\right)\right\} \quad (7)$$

$$\mathcal{M}_E^{(AAE)} = -\mathrm{F}_{y\sim q_{data}(y)}\left\{\log\left(E\big(F(y)\big)\right)\right\} \quad (8)$$

$$\mathcal{M}_E^{(AAE)} = \mathrm{F}_{y\sim q_{data}(y)}\left\{\left\|y - S\big(F(y)\big)\right\|^2\right\} \quad (9)$$

where 'a' could represent either the encoder's result or a sample from the preceding dispersion $q_a(a)$, and y is the encoder's feed.

3.4. Adversarial Network for Data Augmentation

The design of the adversarial data augmentation network (ADAN) provided contains an autoencoder $S(F(y))$, an auxiliary classifier $D(F(y))$, a generator $H(a,z)$, and a discriminator $E(i)$. These three goals inspired the setting up of the ADAN. It functions by collecting a latent framework that preserves psychological input unharmed. Additionally, it strives to coordinate with the posterior dispersion $q(\hat{\imath}|a,z)$ to the posterior dispersion $q(i|y)$. Finally, it lessens the errors in the rebuilding between y and $\hat{y}$. For the achievement of those objectives, the three components are instructed adversarially. For instance, extremely emotion-discriminative N-dimensional latent depictions i's are conveyed to the classifier D and encoder F. Meanwhile, the decoder can use the implicit representations to reassemble sentiment vectors in the real space. The generator is a tool that creates samples in the space of latent events through the use of one-hot embedded emotion descriptors and samples extracted from an N-dimensional Gaussian distribution as source. The generator's purpose is to derive values in the dormant space that are similar to the genuine specimens, or $q(i|y) \approx q(\hat{\imath}|a,z)$. To figure out whether a latent vector derives from the generator or the real information, the discriminator is tuned. Developing samples in the space of latent values instead of in the genuine space has an advantage of eliminating the formation of high-dimensional variables. We minimize the losses illustrated with the goal of simulating the suggested network:

$$\mathcal{M}_E^{(ADAN)} = -\mathrm{F}_{y\sim q_{data}(y)}\{\log E\big(F(y)\big)\} - \mathrm{F}_{a\sim q_a(a)}\left\{\log\left(1 - E\big(H(a,z)\big)\right)\right\} \quad (10)$$

$$\mathcal{M}_E^{(ADAN)} = -\mathrm{F}_{y\sim q_{data}(y)}\left\{\sum_{l=1}^{L} z_{emo}^{(l)} \log D\big(F(y)\big)_l\right\} \quad (11)$$

$$\mathcal{M}_E^{(ADAN)} = \mathrm{F}_{y\sim q_{data}(y)}\left\{\left\|y - S\big(F(y)\big)\right\|^2\right\} \quad (12)$$

$$\mathcal{M}_E^{(ADAN)} = \mathrm{F}_{y\sim q_{data}(y)}\left\{\left\|y - S\big(F(y)\big)\right\|^2 - \sum_{l=1}^{L} z_{emo}^{(l)} \log D\big(F(y)\big)_l\right\} \quad (13)$$

$$\mathcal{M}_E^{(ADAN)} = \mathrm{F}_{a\sim q_a(a)}\left\{\log\left(1 - E\big(H(a,z)\big)\right) - \beta\sum_{l=1}^{L} z_{emo}^{(l)} \log D\big(H(a,z)\big)_l\right\} \quad (14)$$

where H symbolizes the generator, S is the decoder, F is the encoder, E is the discriminator, and D is the auxiliary classifier. The syntax of $()_l$ represents the l-th component of an array. The generator's loss can be calculated by the influence of the tagging mistake, expressed by β.

3.5. Wasserstein ADAN

The gradient disappearance challenge has been considered to be resolved with Wasserstein GANs. The Wasserstein separation is described below, given two likelihood distributions, $\mathbb{Q}_s$ and $\mathbb{Q}_h$:

$$X_1(\mathbb{Q}_s, \mathbb{Q}_h) = \sup_{\|g\|_M \le 1} \mathbb{F}_{y\sim\mathbb{Q}_s}\{g(y)\} - \mathbb{F}_{\tilde{y}\sim\mathbb{Q}_h}\{g(\tilde{y})\} \quad (15)$$

Where $\|g\|_M \le 1$indicates the 1-Lipschitz commitment satisfaction of g. The gradient punishment and burden clipping are dual collective slants for enforcing the 1-Lipschitz constraint. Weight reducing, however, could shrink function g's search field and yield a less-than-ideal response. The gradient fee originated in an effort to overcome the negative aspects of weight clipping. However, it can be tricky to fulfill the l-Lipschitz limitation for an entire data area when data-sparsity assumptions are met. A novel Wasserstein dispersion that may reach the Wasserstein space without employing the Lipschitz restriction is highlighted with these variables in mind. It has the following description:

$$M_{DIV} = \mathbb{F}_{y\sim\mathbb{Q}_s}\{g(y)\} - \mathbb{F}_{\tilde{y}\sim\mathbb{Q}_h}\{g(\tilde{y})\} + \sigma\,\mathbb{F}_{\tilde{y}\sim\mathbb{Q}_h}\{\|\nabla g(\tilde{y})\|^q\} \quad (16)$$

where the gradient term's effect on the desired equation is modulated by σ, the Radon risk metric $\mathbb{Q}_v$, and the M^q space for variable g is represented by q. It has also been showed in that M_{DIV} in equation (16) is a balanced split if σ and q satisfies $\sigma > 0$ and $p > 1$. Once equation (16) is implemented into ADAN, the discriminator and generator deficits grows into

$$\mathcal{M}_E^{(WADAN)} = \mathbb{F}_{q(y,a,\tilde{y},z)}\left\{E\big(F(Y)\big) - E\big(H(a,z)\big) + \sigma\left[\left\|\nabla_{\tilde{y}} E(\tilde{y})\right\|^q\right]\right\} \quad (17)$$

$$\mathcal{M}_E^{(WADAN)} = \mathbb{F}_{q(y,z,a)}\left\{E\big(H(a,z)\big) - \beta\sum_{l=1}^{L} z_{emo}^{(l)} \log D\big(H(a,z)\big)_l\right\} \quad (18)$$

They are identical to equations: (11) – (13) for extra losses. The Wasserstein ADAN (WADAN) network layout resembles that of Figure 2. The discriminator's ultimate layer in ADAN utilizes the function of sigmoid activation, whilst WADAN employs a linear activation to the discriminator's ultimate layer. This explains the main distinction between the two methods. We attached the generator H to the decoder S (illustrated by the dotted arrow in Figure 2 for statistics enhancement following ADAN or WADAN learning. Synthetic instances can be created from the decoder's output by giving the generator the one-hot mood labels and Gaussian random vectors a. In the final section, the augmentation will be described in more detail [12].

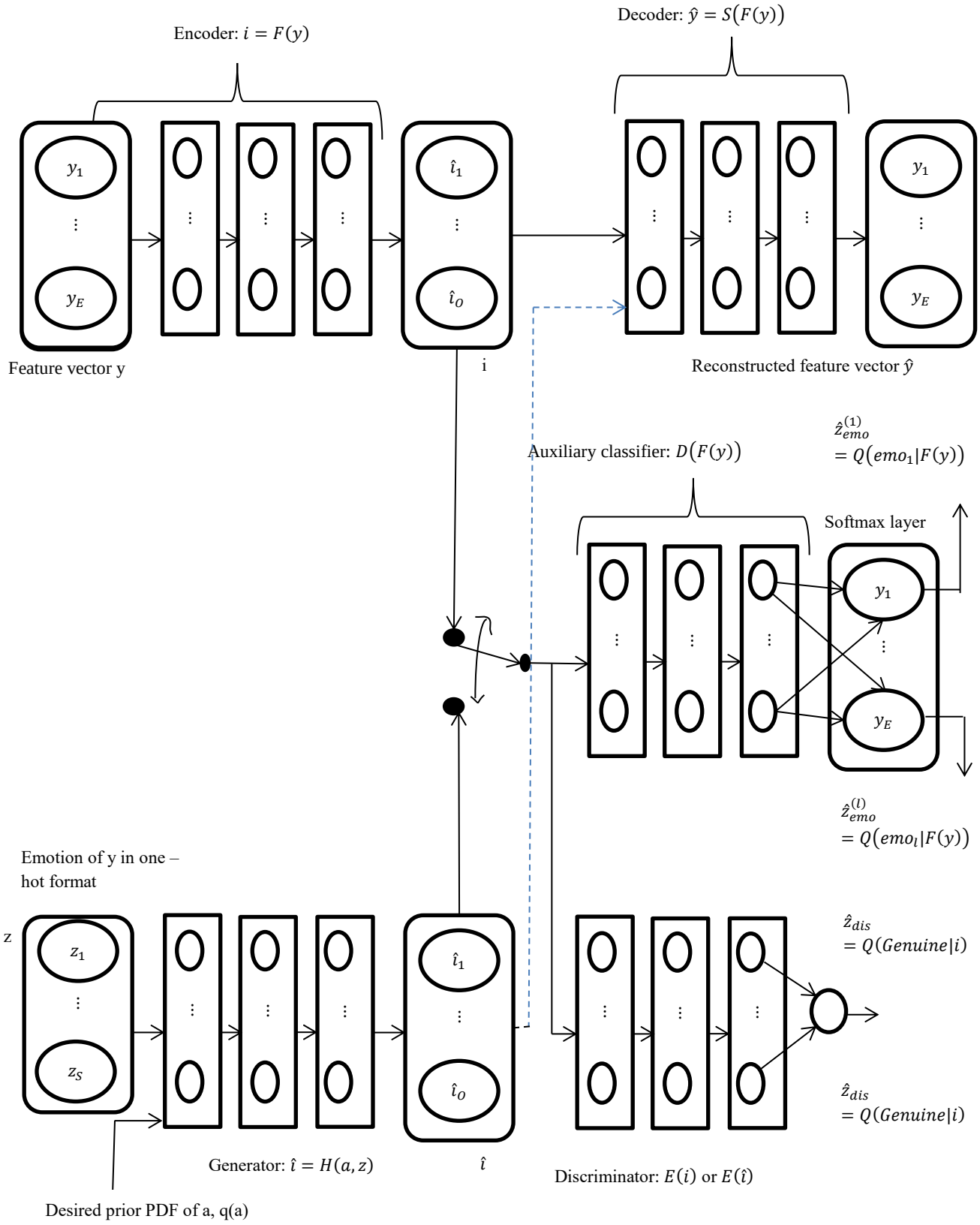


Fig.2. illustrates an adversarial data augmentation network's (ADAN) structure and data flow. The network comprises of a discriminator (lower-right), a generator (lower-left), and an autoencoder that works with an extra classifier (top). DNNs are the unique subnetworks. After preparation, the dotted line is used only for data enhancement.

4. Implementation and Results

A certain feature of current speech enhancement GAN (SEGAN) techniques is that improvement mapping is executed by a single generator H in a single stage, which might not be the most effective technique. Here, our objective is to partition the procedure for improvement into several stages, each of which will be accomplished via an alternate enhancement mapping. A generator achieves each mapping, and subsequent generators are connected to progressively and step-by-step boost a noisy input message to generate a strengthened signal. A generator's task is to upgrade or fix the results that its forerunner generates similarly. We anticipate that multi-stage advancement tracking, as contrasted with single-stage organizing, as utilized in earlier papers, might be desirable. Following that, to analyze two scenarios, we recommend two fresh SEGAN frameworks: iterated SEGAN (ISEGAN) and deep SEGAN (DSEGAN). (1) We use an integrated visualization for all improvement stages, and (2) we employ autonomous mappings at different levels of growth. In the earlier scenario, ISEGAN's generators are obligated to learn an identical mapping (i.e., they execute the same mapping continually) due to a shared parameter and linking of the generators' attributes [13].

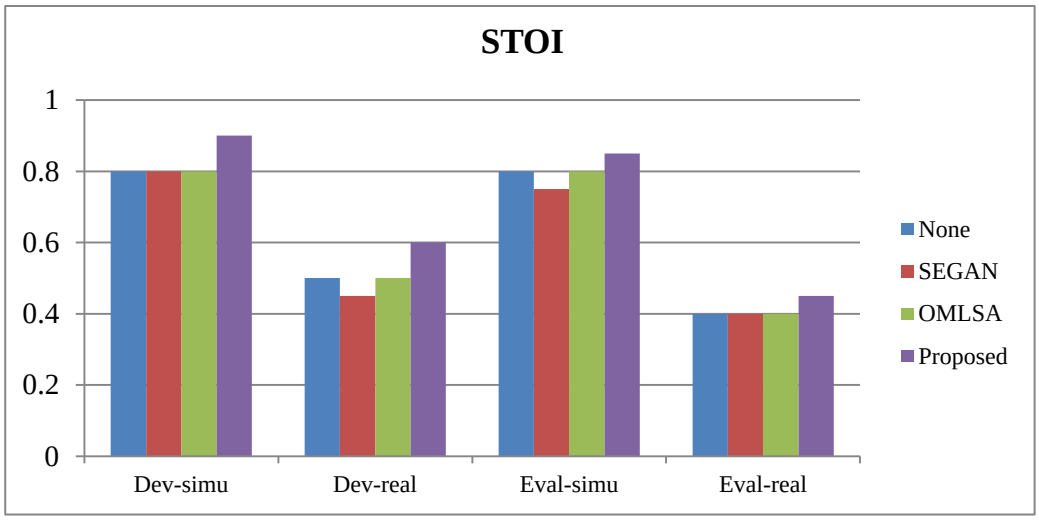
4.1. Configuration

Rather than employing the TIMII data collection and sound to produce distorted speech, we do improve the speech using the CHiME4 information library given that the TIMII data corpus lacks real noisy information for the modeled noisy speech that is produced. Real and synthetic audio recordings from the 5k WSJ0-Corpus with four distinct noise types—bus (BUS), cafe (CAF), pedestrian area (PED), and street junction (STR)—make up the corpus. In total, 8849 phrases have been set aside for training, 3391 for confirmation, and 3751 for testing. Furthermore, considering this research exclusively investigates single-channel speech improvement computations, we merely use the single-channel portion of the six-channel microphone arrangement that acquired the corpus. The network's configuration and learning variables are as follows. An auto-encoder structure with skip links between the encoder and the decoder is implemented by generator H. The decoder is an exact duplicate image of the encoder with the same set of characteristics. The encoder comprises 22 one-dimensional convolutional layers with a filter width of 32 and intervals of 2. The discriminator E and the encoder of H possess identical one-dimensional multilayer structures. The discriminator E and the encoder of H have an identical one-dimensional convolutional architecture. An Xavier initializer starts the mass of every single layer, and zeros start for every single bias. The predictive models are developed utilising an RMSprop optimizer which has a fixed rate of learning set to 0.0003. The weight attribute σ has been configured at 101 and the L1 component serves as a legalization to decrease the variance between the computer generated audio and actual pristine voice. Combining two GTX 1080ti GPUs and an Intel Xeon E5-2630 CPU, we utilize a desktop computer to manage the training and evaluation activities.

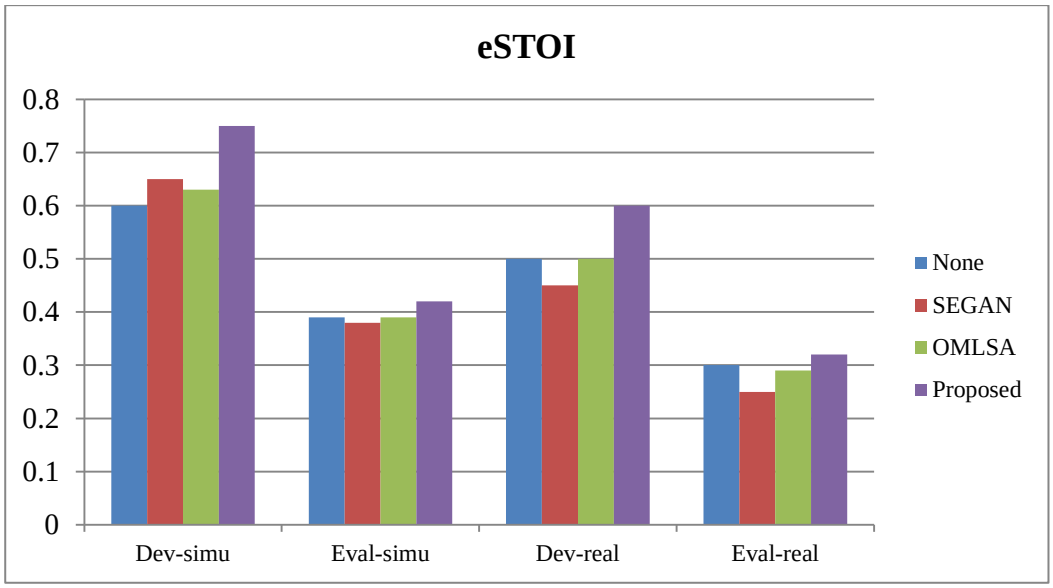
4.2. Assessment

The standard of the upgraded voice signal can be evaluated utilizing the following indicators: short-time objective intelligibility (STOI), prolonged STOI (eSTOI), channel to distortion ratios (SDR in dB), and perceptual evaluation of speech quality (PESQ). The ground-truth pure audio is not readily accessible for the real speech info, but it is obtainable for the computerized voice data. As an outcome, we rely on the channel index zero of close-talking microphone samples as the foundational clean speech. For instance, we deploy the OMLSA and the prepared SEGAN. Additionally, the cacophonous sound is put into consideration for contrast. In the artificial data of the creation set, the suggested technique attains 27.02% and 22.12% relative gain in the sense of SDR in contrast to SEGAN and OMLSA. In the actual information of the validation set, the eSTOI rating of the brought forward strategy is 0.35 however the eSTOI rating of the SEGAN and OMLSA are 0.29 and 0.32 accordingly. The SDR, STOI, and eSTOI evaluations for the speech quality evaluation on the growth set and the assessment set are depicted in Figure 4.1, in which the dev and eval are short for the development and evaluation collection, correspondingly. The findings indicate that our method significantly surpasses the SEGAN and OMLSA for the portrayed data and produces identical results as the conventional state-of-the-art OMLSA technique for the actual data. The mean scores for each circumstance in the environment are provided in Figure 3. Evaluating the individual elements of the contrasting strategies in the distinct environments—BUS, CAF, PED, and STR, for instance -- is crucial. PESQ is the measure we make use of in the current study for evaluating success. According to Table 1's PESQ accomplishments, all of the analyzed approaches receive the best PED scores and the least BUS rankings for the actual data. We've listened to and analyzed the recording's audio. This is because the city's pedestrian area is significantly calmer than the bus environment. Concerning findings, we may argue that the proposed strategy is far more accurate than SEGAN and is not prone to numerous types of distortion.

a) SDR (dB) values



b) STOI values



c) eSTOI values

Fig.3. SDR (dB), eSTOI and STOI scores for the development and evaluation collections for the audio frequency test

Table 1. Displays the PESQ outcomes for the audio standard assessment for every environment's development and evaluation collections

Technique	Environment	Dev-simu	Dev-real	Eval-simu	Eval-real
None	BUS	3.16	3.04	3.17	3.18
	CAF	1.88	3.29	2.99	3.48
	PED	3.23	3.44	2.04	3.46
	STR	2.06	3.22	2.08	3.49
SEGAN	BUS	3.19	2.07	3.21	2.09
	CAF	2.06	3.12	2.90	3.37
	PED	3.14	3.31	3.14	3.32
	STR	2.99	3.14	2.02	3.43
OMLSA	BUS	3.29	3.17	3.32	3.20
	CAF	2.02	3.32	2.03	3.43
	PED	3.43	3.51	3.43	3.53
	STR	3.13	3.38	3.13	3.57
Proposed	BUS	3.31	3.14	3.43	3.39
	CAF	3.19	3.41	3.19	3.62
	PED	3.47	3.56	3.59	3.54
	STR	3.19	3.37	3.24	3.61

Table 2. Outcomes of the juxtaposed tackles' private assessment on an evaluation set of 101 loud speech occurrences

Method	Eval-simu%	Eval-real%	Average%
SEGAN	16.4	13.8	15.1
OMLSA	13.7	17.8	15.9
Proposed	73.2	71.5	72.4

Finally, for 101 loud speech phrases that were picked at random from the testing set, we undertook an informal personal choice inspection contrast among the SEGAN enhanced, the OMLSA enhanced, and the suggested modified voice. Nine male and three female respondents were selected for involvement in the review. For each interactive word, the individual conducting the test gets a chance to nominate their preferred one. As a consequence, there is an overall of $3600\,(100 \times 3 \times 12)$ selections; Table 2 illustrates the statistical outcome. The most preferred proportions for SEGAN, OMLSA, and the recommended alternatives are 15.1%, 15.9%, and 72.4%, correspondingly, by the results. We could conclude from our research that, when thinking of the goal-oriented measurement, the suggested method succeeds better compared to the equivalent substitute [14].

In the beginning, we examined the possibility that training with both the loud and pristine forms turned out to be advantageous. Throughout 100 hours, we educated the DeepSpeech model as the recognizer utilising both fresh and noisy blends. Five hours of fresh and noisy test info was employed for assessing the machine. We used the word error rate (WER), the in-effect criterion for ASR infrastructure, for analysis. WER is the ratio of words that the Automatic Speech Recognition (ASR) algorithm erroneously classifies; the lesser more accurate. Three separate scenarios—the actual-world scenario (noisy), the optimum case (noisy+clean), and our answer (noisy+enhanced)—were brought seriously to distinguish between our results. We are capable of acquiring noisy data via open sources in the actual world. Thus, we utilized just noisy data to train the ASR system. To accomplish the best accomplishments, we educated DeepSpeech employing both its neat and messy forms of the dataset. Ultimately, we placed our scheme into reality. To generate a cleaner information set, we first trained SEGAN to cope with the distracting dataset. Following that, we employed both the upgraded DeepSpeech model and noisy datasets to refine it. We could achieve throughput which is comparable to the ideal situation scenario only if our language augmentation model works extremely well. The primary pair of instances illustrate a back-end strategy mainly because preprocessing is not performed. What counts as disturbance and what is irrelevant is left up to the simulation. Our approach utilizes an amalgamation of front-end and back-end methodologies because we screen out noise from the background and apply a finished version of the spoken word for training. A neat test set delivers superior outcomes for the DeepSpeech model versus a noisy test set, as Figure 4 demonstrates. Identical to this, the DeepSpeech model scores higher on noisy test sets following being taught on noisy data, yet leaves pathways on pristine sample sets. At last, on the identical clean and chaotic test packages, the DeepSpeech models taught on the clean+noisy blend behave more efficiently than the other instances. Undoubtedly, exercising with a clean, loud version is helpful.

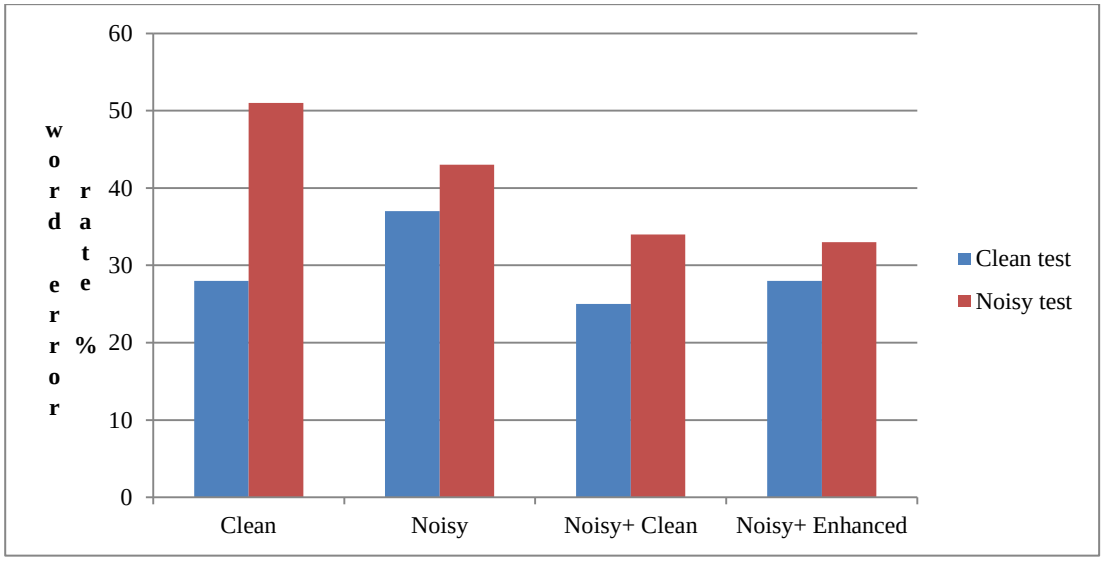


Fig.4. The DeepSpeech model's phrase error rate immediately following its training and evaluation on multiple data sets

We replaced augmented phrases with sanitized speech since we occasionally do not have availability to the clean version of data that is freely reachable to the public. For ASR systems, noisy speech blended with its SEGAN-enhanced speech performed remarkably well. When contrasting test data from noisy (43% vs. 33%) and clean (37% vs. 28%) scenarios to a real-world formulating, such as practicing with only chaotic utterance, we observed a 9.6% average decrease in the WER. For the distorted test information set, we noticed that audio filtered with SEGAN functioned comparably to the best-case situations (33% vs. 34%). The failure rate on the pure test set is slightly greater (28 % vs. 25%) in comparison to an ideal condition. The vocal improvement system's distortions on pristine voice may be at fault for this. To wrap things up, our research represents proof of theory exhibiting how data analyzed employing a speech augmentation framework can enhance the endurance and precision of ASR [15, 16].

5. Conclusion

In this research, we extracted information gathered via the statistical method and proposed an accurate end-to-end platform augmentation methodology employing the generative adversarial network. The recommended strategy proved more beneficial for the reason it reduces the necessity to train different models on the identical set of info, retrieve features from the audio data, and leverage an abundance of noise patterns to boost emulated noisy speech. As a consequence, it is free from the three main problems associated with DNN-based voice enhancement procedures: not enough extension abilities, excess fitting of the generated data, and phase discrepancy among basic clean pronunciation and predicted speech. CHiME4 functions more effectively than OMLSA and SEGAN, notably when it utilizes honest data, based on investigations carried out with data sets. As a result of its flexibility, the unique voice augmentation methodology might be adopted in real-life situations. Our article demonstrates that trustworthy artificial intelligence (ASR) systems may be developed by integrating explicitly available data with voice enhancement frameworks. We consequently intend to assess our strategy with further SE models, which include Wave-u-net and FSEGAN. Evaluating the variations between the end-to-end and back-end techniques will also be interesting. All things accounted for, we believe that the research will encourage greater research into creating cutting-edge ASR systems employing obtained or openly available information.

References

[1] Adiga, N., Pantazis, Y., Tsiaras, V., & Stylianou, Y. (2019, September). Speech Enhancement for Noise-Robust Speech Synthesis Using Wasserstein GAN. In *INTERSPEECH* (pp. 1821-1825).

[2] Wang, K., Zhang, J., Sun, S., Wang, Y., Xiang, F., & Xie, L. (2018). Investigating generative adversarial networks based speech dereverberation for robust speech recognition. *arXiv preprint arXiv:1803.10132*.

[3] Pascual, S., Bonafonte, A., & Serra, J. (2017). SEGAN: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*.

[4] Qian, Y., Hu, H., & Tan, T. (2019). Data augmentation using generative adversarial networks for robust speech recognition. *Speech Communication*, *114*, 1-9.

[5] Vijay, I., Banwari, H., Saluja, G., & Khatri, A. (2021). IMPROVING SPEECH ENHANCEMENT USING GENERATIVE ADVERSARIAL NETWORKS (SEGAN) BY USING MULTISTAGE-ENHANCEMENT. *International Research Journal of Modernization in Engineering Technology and Science*, *3*(6), 214-217.

[6] Donahue, C., Li, B., & Prabhavalkar, R. (2018, April). Exploring speech enhancement with generative adversarial networks for robust speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5024-5028). IEEE.

[7] Sriram, A., Jun, H., Gaur, Y., & Satheesh, S. (2018, April). Robust speech recognition using generative adversarial networks. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5639-5643). IEEE.

[8] Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., ... & Wellekens, C. (2007).

Automatic speech recognition and speech variability: A review. *Speech communication*, *49*(10-11), 763-786.

[9] Chatziagapi, A., Paraskevopoulos, G., Sgouropoulos, D., Pantazopoulos, G., Nikandrou, M., Giannakopoulos, T., ... & Narayanan, S. (2019, September). Data Augmentation Using GANs for Speech Emotion Recognition. In *Interspeech* (pp. 171-175).

[10] Hu, H., Tan, T., & Qian, Y. (2018, April). Generative adversarial networks based data augmentation for noise robust speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5044-5048). IEEE.

[11] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

[12] Yi, L., & Mak, M. W. (2020). Improving speech emotion recognition with adversarial data augmentation network. *IEEE transactions on neural networks and learning systems*, *33*(1), 172-184.

[13] Phan, H., McLoughlin, I. V., Pham, L., Chén, O. Y., Koch, P., De Vos, M., & Mertins, A. (2020). Improving GANs for speech enhancement. *IEEE Signal Processing Letters*, *27*, 1700-1704.

[14] Wu, J., Hua, Y., Yang, S., Qin, H., & Qin, H. (2019). Speech enhancement using generative adversarial network by distilling knowledge from statistical method. *Applied Sciences*, *9*(16), 3396.

[15] Ghai, B., Ramanan, B., & Mueller, K. (2019). Does speech enhancement of publicly available data help build robust speech recognition systems?. *arXiv preprint arXiv:1910.13488*.

[16] Victoria, D. A. H. ., Manikanthan , S. V. ., H R, D. V. ., Wildan, M. A. ., & Kishore, K. H. (2023). Radar Based Activity Recognition using CNN-LSTM Network Architecture. *International Journal of Communication Networks and Information Security (IJCNIS)*, *14*(3), 303–312. https://doi.org/10.17762/ijcnis.v14i3.5630 (Original work published December 31, 2022).