SWAR: A LONGFORMER-BASED GAN VOCODER FOR GUJARATI LANGUAGE

Ravindrakumar M. Purohit, Hemant A. Patil
Dhirubhai Ambani University (DAU), Gandhinagar (GJ), India
{202321002, hemant_patil}@daiict.ac.in

***Abstract*—Recently, Longformer has been proposed to efficiently handle long-range sequence text data and has demonstrated state-of-the-art performance on various Natural Language Processing (NLP) tasks. On the other hand, existing neural vocoders struggle to balance computational efficiency and long-dependency modeling. While Generative Adversarial Network (GAN)-based vocoders are capable of generating high-quality speech quickly, they face challenges in capturing long-range dependencies. Especially in Indian languages, such as Gujarati, where 14 vowel sounds can be spoken in various regional dialects. In this context, we propose a novel vocoder, *Swar*, which uses sliding window attention for high-quality speech generation in Gujarati. The attention and diffusion-based vocoders are unsuitable due to training and inference complexity, respectively. *Swar* achieves a 4.51 Subjective Mean Opinion Score (SMOS) on NVIDIA GTX 1080 8GB GPU. It generates 22.05 kHz audio 833.33 $\times$ times faster than real-time on an NVIDIA GTX 1650. Since SMOS has reproducibility limitations, we evaluated using the Pearson Correlation Coefficient (PCC), Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI), Modulation Spectra Distance (MSD), and Mel Cepstral Distortion (MCD). to evaluate a comprehensive and robust evaluation of Swar's performance.**

***Index Terms*—Longformer, Neural Vocoder, Generative Adversarial Networks (GANs), Speech Prosody, Gujarati Language.**

I. Introduction

Text-to-Speech (TTS) is used in various applications, such as virtual assistants, content creation, and human-computer interaction. Renowned TTS methods work in two phases: (1) Text-to-Mel Spectrogram Prediction [1], [2] and (2) Mel spectrogram-to-waveform generation (vocoder) [3]–[5]. In second phase, traditional vocoders, such as WORLD [6], analyze speech signals and extract key acoustic parameters e.g., pitch (i.e., fundamental frequency, F_0), formants, and spectral envelope. These vocoder often produce robotic and unnatural speech. In 2016, Oord *et al.* [7] proposed the autoregressive (AR)-based vocoder WaveNet, which model the waveform from a Mel spectrogram by capturing complex characteristics in the speech wave. Successively, non-autoregressive vocoders (NAR), flow-based [8], [9], GAN-based [10]–[13], and diffusion-based models [14], [15] gives improved speech generation quality while achieving speeds hundreds of times faster than the real-time. However, each approach presents its unique set of difficulties (or limitations). In particular, AR models suffer from slow inference speed due to sequential nature. while NAR models were faster but demonstrated reduced speech quality. Flow-based models improved quality but required high computational resources. Similarly, diffusion-based vocoders [14], [15], capable to generate high-quality speech samples, however, require high-end GPU resources for real-time inference. In contrast, the GAN-based vocoder becomes a preferred choice due to its real-time capabilities w.r.t. the speech quality and inference speed [10]–[13].

In 2017, the work proposed by Vaswani *et al.* [16] proposed the attention mechanism in the transformer, which revolutionized the NLP field. The self-attention mechanism allows models to dynamically focus on different parts of an input sequence, capturing *contextual* dependencies more effectively than the traditional RNNs or LSTMs. However, the use of self-attention increases complexity quadratically with a sequence length of the ($O(n^2d)$), makes it complex for long sequences and limits real-world applications, and still struggle with long sequences due to vanishing attention weights over distant tokens. To address these challenges, *Longformer* [17] was introduced as an efficient alternative that reduces the second-order complexity of the self-attention computation to the lower-order ($O(nd)$) using the local sliding windows and global attention mechanism. Here, local attention mechanism ensures that each token only attends to a fixed number of nearby tokens, reducing computations while global attention selectively allows key tokens to attend across the entire sequence.

In this work, we introduce ***Swar***- a Longformer-based GAN vocoder designed especially for Gujarati (an Indian language) while highlighting the following key contributions:

1) We propose a Longformer-based GAN vocoder, which uses the sliding window mechanism, to focus on local *phonetic* details as well as *global prosodic* features.
2) We successfully trained the ***Swar*** on the Gujarati language (55.5 million speakers worldwide), without explicit regularization in either the generator or the discriminator of GAN [18].
3) The earlier works were done heavily using subjective measures, such as MOS. In the context of this work, we believe the perceptual evaluation is not sufficient for most noisy conditions and phonetic information preservation and hence, we evaluated our samples using objective measures, such as PCC, PESQ, STOI, MSD, and MCD. These measures allow us to establish a strong correlation with subjective evaluations while providing a detailed assessment of speech quality. Samples available

at website [1].

II. Related Work

A. Why Gujarati Language?

Most speech vocoders are trained in English, French, and Mandarin, which is challenging due to notable differences in phonetic structures and *prosody* of these languages. As shown in Table I, the Gujarati language contains 57 phonemes, 32 consonants, and 25 vowels, as compared to other languages, e.g., 44 phonemes in English, 35 in Mandarin, and 35 in French, which means many Gujarati phonemes, such as retroflex, consonants, and nasalized vowels are missing in pre-trained models, leads to pronunciation errors and phoneme mismapping (i.e., wrong assignment), as model attempt to replace unavailable sounds with the closest matches from their training language phonemes. However, the phonotactic rules and *prosody* patterns of Gujarati differ from those in English and Mandarin, making speech synthesis audio unnatural without proper training with Gujarati training data. The grapheme-to-phoneme (G2P) is also a challenge due to the Gujarati Abugida script, leading to the inaccurate phonetic transcript in another language. The proposed vocoder is trained entirely on the Gujarati dataset I, which covers the recording samples from Kathiawar, North Gujarat, Kutch, and Central Gujarat, where Gujarati is spoken with different *accents* and unique pronunciation styles.

TABLE I
Comparative Analysis of Phoneme, Consonant, and Vowel Counts Across Languages. After [19].

Language	Phonemes	Consonants	Vowels
Mandarin	35	26	9
French	35	21	14
English	44	24	20
Gujarati	**57**	**32**	**25**

The word 'Swar' originates from Sanskrit, where the term 'Svara' refers to vowels, that are self-sounding. It originates from the root, referring to vowels and intonations. Which plays a fundamental role in speech production and perception. In the context of signal processing, it refers to the pitch (F_0) of pronunciation and determines the meaning of words [20].

B. GANs Vocoders

The explicit density models demand high computational costs due to explicit probability modeling, which made the real-time inference challenging [21]. In contrast, implicit density models offered high quality and fast inference with limited resource conditions, making the GANs-based vocoder a perfect choice for various real-world applications. Among GAN vocoders [4], [10], [11], [22] several vocoders can be considered due to their effectiveness on evaluation criteria, such as speech quality, computational efficiency, robustness, stability, *accents*, and *prosody* control. In 2019, MelGAN [4] and Parallel WaveGAN [10] optimised speed over speech quality for waveform generation. HiFi-GAN [11] performed 167.9 $\times$ times faster than the real-time factor without diminishing speech quality and naturalness. Another work, Multi-band MelGAN [23], uses multiple frequency bands to significantly improve the generation speed with high-quality output. At the same time, UnivNet [24] proposed resolving the generalization across different datasets while offering stability and reliability as compared to other models. In 2022, BigVGAN set a benchmark by proposing a zero-shot universal vocoder, trained at the scale of 112M parameters [22]. This large-scale training helps it to generate high-frequency sounds, such as birds and audio effects, which contain electronic music with loud drums even with out-of-distribution scenarios.

C. Attention Mechanism

The classic work on "*Attention Is All You Need*" introduced the transformer architecture, which replaced the Recurrent Neural Network (RNN) with a self-attention structure for the sequential processing of text [16]. The transformer revolutionized the NLP, by making deep learning more scalable and effective through parallel training. Subsequently, this architecture inspired authors, such as Li *et al.* [25] to apply it to audio tasks, specifically in the first phase of the TTS systems to generate a Mel spectrogram. Wu *et al.* [26] were the first to address long-range dependencies of input text by developing a transformer *prosody* model for expressive synthesized speech. However, this model struggled to maintain control over *accents* and *prosody*.

III. Swar: Proposed Architecture

To the best of our understanding and belief, no prior work integrates Longformers into the vocoder models and removes the challenges regarding accents and linguistic influences. In the proposed work, sparse attention (as shown in Fig. 1) uses character embeddings to process phoneme sequences and combines local sliding windows and global attention to handle long sequences. Given an input character embedding X, the

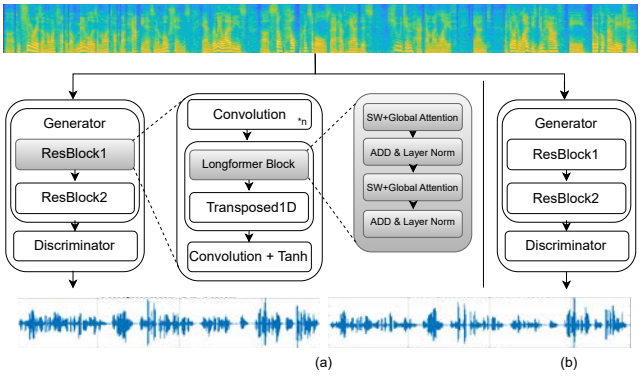

Fig. 1. Comparison of raw speech waveform generation between (a) The proposed *Swar*, and (b) the traditional GAN architecture.

Longformer attention can be defined as follows: each character embedding attends to a fixed window size w around it instead

[1] https://iamshreeji-copy1.github.io/SWAR_Gujarati_Vocoder/

of all character embeddings in the sequence. The attention for each head i is computed as,

$$\text{Head}_i = \text{Attention}(XW_i^Q, XW_i^K, XW_i^V, w), \quad (1)$$

where W_i^Q, W_i^K, W_i^V are the learnable projection matrices for Query, Key, and Value. Certain preselected embeddings have global attention, meaning they attend to all possible characters, and all embeddings can attend back to them. Thus global attention becomes,

$$\text{Head}_j^{\text{global}} = \text{Attention}(XW_j^Q, XW_j^K, XW_j^V). \quad (2)$$

In the case of multi-head self-attention consists of both local (sliding window) and global attention. So final attention computation is given by:

$$Y = \text{Concat}(\text{Head}_1, ..., \text{Head}_h)W^O + X, \quad (3)$$

where W^O is the learned transformation matrix applied after concatenating all attention heads. The attention scores are computed as follows:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V, \quad (4)$$

where d_k is the dimensionality of the key vector.

A. Sliding Windows

For an input sequence $S = \{S_1, S_2, \ldots, S_n\}$ (where S_t represents a speech frame or phoneme embedding at time step t), then attention mechanism is modified to restrict each element to attend only within a fixed window size w. So, the single attention head is computed by:

$$\text{Head}_i = \text{Attention}(S_t W_i^Q, S_{t-w:t+w} W_i^K, S_{t-w:t+w} W_i^V), \quad (5)$$

where, S_t is the current frame or phoneme embedding, $S_{t-w:t+w}$ represents the local neighbourhood around S_t, and W_i^Q, W_i^K, and W_i^V are learned projection matrices for queries, keys, and values. Each speech frame computes attention only within its window instead of the entire sequence, which means scaled dot product attention becomes:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V, \quad (6)$$

where, $Q = S_t W^Q$ (query at time t), $K = S_{t-w:t+w} W^K$, $V = S_{t-w:t+w} W^V$, d_k is the key dimensionality. So each speech frame attends only to its local context, avoiding unwanted computational overhead. Indirectly, Longformer uses multiple attention heads to capture different aspects of local dependencies, So, Multi-Head Attention Output will become,

$$Y = \text{Concat}(\text{Head}_1, ..., \text{Head}_h)W^O + S, \quad (7)$$

where h is the number of attention heads, W^O is the learned output transformation matrix, and S is the residual connection to preserve original speech features.

IV. Experiments

A. Dataset Used

For this study, we used the IndicVoices-R Gujarati dataset, which includes a total of 8.94 hours of speech data from 45 speakers with an average duration of 9.74 seconds per utterance. The dataset was collected from the various districts of Gujarat State, India. The primary reason behind choosing this dataset is its ability to capture ground-level linguistic diversity. Unlike scripted speech datasets, it includes extempore conversations, making it more representative of natural spoken language rather than text being read aloud (i.e., read speech). Initially, the datasets' raw audio was sampled at 44 KHz. We pre-processed (e.g., resampling) and converted paired representations of Mel spectrograms and corresponding audio waveforms using Wav2Mel [2].

B. Model Parameter Configuration

Proposed models use residual blocks, to preserve essential features and reduce the artifacts in the generated waveform, where it increases the resolutions of the generated spectrogram using a multi-stage transposed convolution with upsample rates [8,8,2,2], and a corresponding kernel size of [16,16,4,4] progressively. The training was conducted with 128 Mel bands spectrograms with a hop size of 256 and a Windows size of 1024. The **Swar** model is capable of capturing a frequency range of 0 Hz to 8000 Hz. The convolution layers were designed with kernel sizes [3,7,11], each paired with a dilation rate of [1,3,5] for each kernel. The batch size is limited to 8 on GTX 1080 ARMOR, learning rate of 0.0002 with a decay factor of 0.999, and momentum parameters β_1 and β_2 were 0.8 and 0.99, respectively.

C. Setup

All experiments were executed on an ubuntu-powered workstation equipped with a 12^{th}generation Intel Core i7 processor, 16GB of RAM, and an MSI GTX 1080 ARMOR 8GB graphics card with 2TB external SSD for storage.

V. Results and Discussion

To assess the synthesized quality of synthesized samples, we conducted the following experiments. Specifically, 20 generated utterances were randomly selected from the out-of-distribution dataset. The previously proposed methods were generally trained in languages other than Gujarati, mainly in English. This remains the primary constraint of our evaluation paradigm. In this setting, we observed that some models show a generalized ability to an avoidable rate, when trained on a multi-speaker dataset. However, they still fail to generate perfect pronunciation per Gujarati phonemes, which should be acceptable.

[2] https://github.com/rhasspy/wav2mel.git

A. Subjective Measures

To analyze this difference, we conducted a SMOS evaluation, according to ITU-T P.800 standard, by ensuring sufficient test duration and phonetic balance in test material [27]. All 10 subjects were naïve listeners with no prior training in speech or audio processing (e.g., age group of 19-28 years, gender distribution: 8 male, 2 female). Notably, all were native speakers of Gujarati, ensuring they could reliably judge the naturalness and intelligibility of synthesised Gujarati speech. None of the participants reported any hearing impairments or known biases toward the speech samples. They were asked to rate the best-fit option in comparison to the reference audio on a scale of 1 to 5 via *TestVox* web app [28], where 5 represents excellent (clear and natural), 4 is good (acceptable), 3 is fair (noticeable degradation), 2 is poor (notable distortion), and 1 is bad (unacceptable or unintelligible). Each listener was assigned a score based on the perceived quality, and we calculated the arithmetic mean (average) score of all ratings the listeners gave as per eq. 8. As shown in Table II, 95% Confidence Interval (CI) means 'true SMOS score is expected to lie with 95% certainty.'

$$\text{SMOS} = \frac{1}{N}\sum_{i=1}^{N} S_i. \tag{8}$$

B. Objective Measures

We evaluated speech quality using four objective measures, such as PCC [29], PESQ [3] [30], STOI [31], MSD [32], and MCD [33]. Higher PESQ and STOI scores demonstrate better speech quality and intelligibility, and a lower MCD and MSD score shows reduced distortion and a closer match to the Ground Truth (GT) speech sample.

TABLE II
COMPARISON OF OBJECTIVE AND SUBJECTIVE MEASURES (REPORTED WITH 95% CI), ON SYNTHETIC AND GT FOR INDICVOICES-R DATASET

Method	**PCC** (↑)	**PESQ** (↑)	**STOI** (↑)	**MSD** (↓)	**MCD** (↓)	**SMOS** (↑)
GT	-	-	-	-	-	4.62
MelGAN [4]	0.18	0.59	0.12	76.82	267.12	1.86
P.WaveGAN [10]	0.01	1.98	0.56	45.62	143.61	2.02
HiFi-GAN [11]	-0.00	2.22	0.84	37.09	93.61	3.09
BigVGAN [22]	-0.34	4.38	0.99	11.33	10.72	4.39
IndicVoices-R						
Swar (Proposed)	0.01	2.64	0.93	25.20	35.66	4.51 (±0.12)

1) PCC: It measures the linear association between two signals x and y, with values ranging from 1 (perfect positive) to -1 (perfect negative) [29]. It is computed as:

$$r = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}}. \tag{9}$$

2) PESQ: It measures speech quality. Higher PESQ ratings indicate high speech quality. In particular,

$$\text{PESQ} = 4.5 - 0.1d_{\text{sym}} - 0.0309d_{\text{asm}}. \tag{10}$$

[3]https://pypi.org/project/pesq/

3) STOI: It is widely used to evaluate speech intelligibility in noisy or degraded conditions. i.e.,

$$\text{STOI} = \frac{1}{N}\sum_{t=1}^{N}\frac{\sum_{f=1}^{F}(x_f(t)-\bar{x}_f)(y_f(t)-\bar{y}_f)}{\sqrt{\sum_{f=1}^{F}(x_f(t)-\bar{x}_f)^2}\sqrt{\sum_{f=1}^{F}(y_f(t)-\bar{y}_f)^2}}. \tag{11}$$

4) MSD: It calculates the likeness or disparity between two signals through modulation spectra. As given in Eq.(12),

$$\text{MSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(s(y)_i^t - s(y)_i^{\hat{t}}\right)^2}. \tag{12}$$

5) MCD: It is used to quantify the variation among two sets of Mel cepstral coefficients, i.e.,

$$MCD = \frac{\sqrt{\sum_{t=1}^{N}\left(\sqrt{\sum_{i=1}^{D}(A(t,i)-B(t,i))^2}\right)^2}}{N}. \tag{13}$$

Except for the PCC score, As mentioned in Table II, *Swar* remarkably surpasses the performance of traditional vocoders, such as MelGAN [4], Parallel WaveGAN [10], and HiFi-GAN [11] across all the metrics, by capturing local and long-range speech dependencies in waveform prediction. We noted that BigVGAN [22] achieves the best overall performance in objective evaluations, closely followed by *Swar*, which achieves a PESQ of 2.64 (as per eq. 10), STOI of 0.93 (as per eq. 11), an MSD of 25.20 (as per eq. 12), and an MCD of 35.66 (as per eq. 13) but as per the SMOS results, the listers rated BigVGAN [22] samples lower regardless of its strong objective performance, this variation is possible due to perceived excessive smoothing, and unconditioned *accents* in BigVGAN [22], which leads listeners to choose the generated samples from the *Swar* model, as it produces more expressive and natural-sounding speech outputs compared to BigVGAN. In short, *Swar* maintains a balance between objective and subjective quality by giving importance to perceptual evaluation, leading to a more natural and intelligible speech synthesis experience as compared to existing vocoders.

VI. FUTURE DIRECTIONS

Our future improvements will focus on enhancing *prosody* control, optimizing computational efficiency for real-time application, and integrating mode-diverse linguistic datasets, especially for Gujarati. Without any doubt, this research lays a strong foundation for future advancements in Indic language vocoders, to address challenges of low-resource language, where existing non-Indic pre-trained models struggle with phoneme mismapping and lack of naturalness. This work provides the robust foundation for future advancements in Indic language vocoders, which are low-resource and complex to synthesize from the existing non-Indic pre-trained models, which miss the important phonemes and mispredict the Indic phonemes with need more naturalness and intelligible low-resource and phonetically rich languages, we evaluated with subjective and objective evaluations for better judgement in terms of naturalness and expressiveness.

VII. Closing Remarks

In this paper, we introduced *Swar*, a longformer-based GAN vocoder for Gujarati speech synthesis. By integrating the sliding windows mechanism with a GANs-based vocoder, Swar captures local phonetic details and long-range dependencies, in order to improve speech quality and intelligibility for Gujarati. The subjective and objective evaluation shows that the proposed work surpasses traditional vocoders by achieving comparable results to BigVGAN in terms of subjective metrics. Also, listeners preferred *Swar* synthesized speech due to its expressiveness and naturalness, highlighting the importance of perceptual quality in vocoder.

Acknowledgment

This work is supported by the Ministry of Electronics and Information Technology (MeitY), Government of India, under Project Grant ID: 11(1)/2022-HCC(TDIL).

References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, and S. Bengio, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017 {Last Accessed: March 16^{th}, 2025}.

[2] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *Advances in Neural Information Processing Systems (NIPS), vol. 32, pp. 3171–3180*, 2019, Vancouver Convention Centre, Canada.

[3] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning wavenet on Mel spectrogram predictions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4779–4783*, 2018, Calgary, Canada.

[4] K. Kumar, R. Kumar, T. De Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. De Brebisson, Y. Bengio, and A. C. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," *Advances in Neural Information Processing Systems (NIPS), Vancouver, Canada,*, vol. 32, pp. 2672–2680, 2019.

[5] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3617–3621, Brighton, United Kingdom*, 2019.

[6] M. MORISE, F. YOKOMORI, and K. OZAWA, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[7] A. v. d. Oord, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016, {Last Accessed: March 16^{th}, 2025}.

[8] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *International Conference on Machine Learning (ICML), pp. 3918–3926*, 2018, Stockholm, Sweden.

[9] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," *arXiv preprint arXiv:1807.07281*, 2018.

[10] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6199–6203*, 2020, Barcelona, Spain.

[11] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems (NIPS), Vol. 33, pp. 17022–17033*, 2020, Virtual-only Conference.

[12] W. Jang, D. Lim, and J. Yoon, "Universal melgan: A robust neural vocoder for high-fidelity waveform generation in multiple domains," *arXiv preprint arXiv:2011.09631*, 2020 {Last Accessed: March 16^{th}, 2025}.

[13] M. He, T. Guo, Z. Lu, R. Zhang, C. Gong, and D. Chuxing, "Improving gan-based vocoder for fast and high-quality speech synthesis." in *INTERSPEECH, pp. 1601–1605*, 2022, Incheon, South Korea.

[14] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "Wavegrad: Estimating gradients for waveform generation," *arXiv preprint arXiv:2009.00713*, 2020 {Last Accessed: March 16^{th}, 2025}.

[15] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," *arXiv preprint arXiv:2009.09761*, 2020 {Last Accessed: March 16^{th}, 2025}.

[16] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems (NIPS), pp. 6000-6010*, 2017, Long Beach, USA.

[17] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020 {Last Accessed: March 16^{th}, 2025}.

[18] Wikipedia contributors, "Gujarati language — Wikipedia, the free encyclopedia," 2024, [Online; {Last Accessed: March 16^{th}, 2025}]. [Online]. Available: https://en.wikipedia.org/wiki/Gujarati_language

[19] W. contributors, "List of languages by number of phonemes," 2025 {Last Accessed: March 16^{th}, 2025}. [Online]. Available: https://en.wikipedia.org/wiki/List_of_languages_by_number_of_phonemes

[20] Svara: Wikipedia, "Svara: Wikipedia," 2025, {Last Accessed: March 16^{th}, 2025}. [Online]. Available: https://en.wikipedia.org/wiki/Svara

[21] R. Valle, K. Shih, R. Prenger, and B. Catanzaro, "Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis," *arXiv preprint arXiv:2005.05957*, 2020, {Last Accessed: March 16^{th}, 2025}.

[22] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigVgan: A universal neural vocoder with large-scale training," *arXiv preprint arXiv:2206.04658*, 2022 {Last Accessed: March 16^{th}, 2025}.

[23] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multi-band melgan: Faster waveform generation for high-quality text-to-speech," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 492–498.

[24] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, "Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation," *arXiv preprint arXiv:2106.07889*, 2021 {Last Accessed: March 16^{th}, 2025}.

[25] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *AAAI*. AAAI Press, 2019 {Last Accessed: March 16^{th}, 2025}.

[26] C. Wu, Z. Xiu, Y. Shi, O. Kalinli, C. Fuegen, T. Koehler, and Q. He, "Transformer-based acoustic modeling for streaming speech synthesis." in *Interspeech*, 2021, pp. 146–150.

[27] International Telecommunication Union, "Itu-t recommendation p.800: Methods for subjective determination of transmission quality," https://www.itu.int/rec/T-REC-P.800-199608-I, Aug. 1996, {Last Accessed: March 16^{th}, 2025}.

[28] A. Parlikar, "Testvox: web-based framework for subjective evaluation of speech synthesis," *Opensource Software, pp. 13*, 2012.

[29] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," *Noise Reduction in Speech Processing*, pp. 1–4, 2009.

[30] 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 *IEEE International Conference on Acoustics, Speech, and Signal processing. Proceedings (Cat. No. 01CH37221) (ICASSP)*, vol. 2, Salt Lake City, USA, 2001, pp. 749–752.

[31] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[32] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[33] R. Kubichek, "Mel cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing, Vol. 1, pp. 125-128*, 1993.