BAANI: A 296M-Parameter Neural Vocoder for End-to-End Punjabi Speech Synthesis

Siddharth Kumar*, Nisarg Trivedi*, Ravindrakumar M. Purohit*, and Hemant A. Patil
Speech Research Lab, Dhirubhai Ambani University (formerly DA-IICT), Gandhinagar (GJ), India
E-mail: {202418054, 202401129, 202321002, hemant_patil}@dau.ac.in (* Equal contribution)

***Abstract*—In this work, we present BAANI, a neural vocoder comprising 296 million parameters, designed for end-to-end speech synthesis (SS) in the Punjabi language. Recognizing the unique *phonetic* and *prosodic* characteristics of a low-resource and underrepresented Punjabi language, BAANI aims to enhance the naturalness and intelligibility of synthesized speech. The proposed model is trained on an IndicTTS Punjabi corpus and evaluated using an NVIDIA GTX 1080 GPU and an Intel Core i7 12th Gen CPU-powered system. We conducted a comparative analysis (e.g., subjective, objective, and quantitative metrics) with several state-of-the-art neural vocoders to validate the effectiveness of BAANI. It achieved 4.18 on the Mean Opinion Score (MOS). On the other side, from five different objective measures, all measures surpass the current SOTA models. Furthermore, BAANI generates 22.05 kHz speech with real-time factors (RTF) of 3.2 $\times$ and 1.5 $\times$ on the NVIDIA GTX 1080 GPU and the Intel Core i7-12700 CPU, respectively, demonstrating its practical deployment potential. Notably, the vocoder effectively preserves high-frequency harmonics and pitch (i.e., F_0) contours, aligning closely with the perceptual preferences of native Punjabi speakers.**

***Index Terms*—Speech Synthesis, Generative Adversarial Networks (GANs), Neural Vocoder, Punjabi Language.**

I. Introduction

Recent advances in Generative adversarial networks (GANs) have significantly improved speech synthesis (SS), enabling real-time, high-quality waveform generation. SOTA vocoders can take acoustic features, such as Mel spectrograms, and generate waveforms, which played a central role in the transformation of generative AI. Following the development of deep learning-based models, which made huge quality improvements, it became clear that the extensive computational resource usage made real-time speech generation infeasible. GAN-based models step forward with the ability for real-time implementation, then autoregressive [1], [2], non-autoregressive [3], [4], flow-based [5], and diffusion-based [6], [7], making them a preferred choice for many real-time SS systems. However, the majority of vocoder research and optimization has been applied to well-resourced Western languages (e.g., English, Japanese, Mandarin), which has led to less focus on Punjabi and a lack of research on how well the model will perform in those languages. Punjabi presents distinct challenges for waveform generation because of its tonality distinctions, extensive inventory of consonants and vowels, and variations in dialects. Furthermore, the lack of open access, high-quality Punjabi speech datasets poses a problem for the development of comprehensive voice technologies. Addressing the gap is crucial not only for advancements in technology but also for promoting linguistic diversity in SS.

In this paper, we propose BAANI, which is capable of representing and synthesizing the Punjabi phonetic and prosodic characteristics more accurately. BAANI is trained on the IndicTTS dataset and evaluated against several state-of-the-art (SOTA) vocoders to investigate its performance in terms of speech quality, intelligibility, and inference efficiency. Through this work, we aim to demonstrate that language-specific architectural enhancements can significantly improve vocoder performance for low-resource languages like Punjabi, improving it in the broader goal of inclusive and diverse speech synthesis. Some of our contributions include :

1) Here BAANI architecture is strategically enhanced by doubling the initial channel capacities and progressively increasing the channel capabilities of the residual blocks through the upsampling layers in order to capture richer patterns from the input speech and achieve the 4.38 and 4.18 by BAANI-V1 and BAANI-V2 with out-of-distribution samples.
2) We hypothesised that integrating ReLU6 activation in the BAANIS' generators' (G) upsampling layers can improve training stability by preserving nonlinearity. We noticed our approach produced intelligible samples in the beginning stages of training.
3) We proposed two architectures of BAANI, V1 and V2 to maintain trade-offs between SS quality and speed. The *V1* and *V2* models contain 296 M and 7.6 M parameters, respectively. The generated samples are available at url[1].

The remainder of the paper is organized as follows. Section II discusses the related work in the domain, and Section III mentions experiments along with dataset details and performance evaluation. Section IV includes the results and discussion, and finally, in Section V paper concludes with a summary and open research problems.

II. Related works and Motivation

Traditional statistical parametric speech synthesis (SPSS) relied heavily on hand-crafted rules for controlling pitch (F_0), duration, and formants. These systems, although computationally efficient, often generated robotic and

* Equal contributions
[1]https://normie27.github.io/Baani_Punjabi_Vocoder/ {Last Accessed: July 12^{th}, 2025}

unnatural speech due to their inability to capture the nonlinear and contextual transitions between phonemes. In contrast, neural vocoders have rapidly gained popularity for their ability to produce high-quality, natural-sounding speech by learning complex mappings from mel-spectrograms to waveforms. Among various neural vocoders, such as WaveNet [1], WaveGlow [5], ParallelWaveGAN [8], and HiFi-GAN [9] have demonstrated remarkable performance on large-scale English datasets. However, these models tend to perform poorly on low-resource languages, particularly in terms of intelligibility, naturalness, and prosodic fidelity. Recent initiatives like IndicTTS and IndicVoices have attempted to address these gaps across various Indian languages, yet Punjabi remains significantly underexplored.

Neural vocoders typically use a GAN architecture, where the generator (G) synthesizes waveform audio from mel-spectrograms, and the discriminator (D) attempts to distinguish real from generated audio. G and D train adversarially using a combination of adversarial, feature-matching, and reconstruction losses. At convergence, the generator produces high-quality waveforms that closely resemble real human speech. Fig. 1 illustrates this process, based on the original GAN architecture [10].

Fig. 1: Architecture of GANs-based Vocoder(s).

Among existing vocoders, HiFi-GAN stands out due to its ability to produce high-fidelity audio in real-time. It combines multiple discriminators—Multi-Scale Discriminator (MSD) and Multi-Period Discriminator (MPD)—to jointly model periodic and spectral features. HiFi-GAN V1, for example, offers up to 167.9× real-time inference speed while achieving near-parity with ground-truth recordings in MOS scores. Extensions such as BigVGAN [11], VocGAN [12], and UnivNet [13] improve training stability and generalization across speakers and languages, but they also require significantly more parameters and training data.

A. Why Punjabi Language?

Punjabi is one of the 22 scheduled languages of India and is spoken by over 100 million people globally. Despite this, Punjabi speech synthesis remains underdeveloped, largely due to the lack of high-quality, curated datasets and language-specific vocoders. Punjabi has a rich inventory of retroflex and aspirated consonants, as well as tonal variations that distinguish it from other Indic languages like Hindi or Marathi. This phonetic complexity is summarized in Table I, which compares Punjabi's phoneme set against other commonly studied languages.

TABLE I: Analysis of phoneme, consonant, and vowel counts across languages

Language	Phonemes	Consonants	Vowels
Mandarin	35	26	9
English	44	24	20
Hindi	48	33	15
Punjabi	**49**	**35**	**14**

III. Proposed Framework

A. BAANI-Generator

The BAANI-Generator (shown in Fig. 2) is a fully convolutional, one-dimensional neural network designed to convert mel-spectrograms into high-quality, time-domain audio waveforms. The model takes 80-channel mel-spectrograms as input and first processes them through a 1D convolutional layer with a kernel size of 7. This initial layer expands the input into a higher-dimensional representation using 1536 channels, laying the groundwork for the subsequent upsampling stages.

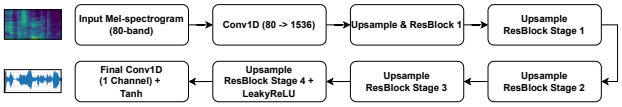

Fig. 2: Architecture of the BAANI-Generator.

To gradually increase the temporal resolution of the audio signal, the network applies four transposed convolutional layers with upsampling rates of $[8, 8, 2, 2]$ and matching kernel sizes of $[16, 16, 4, 4]$. After each upsampling step, a Multi-Receptive Field (MRF) module is introduced to enhance the model's capacity to capture both local and global acoustic patterns.

B. BAANI-Discriminator

The BAANI-Discriminator (shown in Fig. 3) is designed to critically assess the realism of generated audio waveforms by analyzing them from multiple perceptual perspectives. To do so, it combines two complementary modules: a Multi-Period Discriminator (MPD) and a Multi-Scale Discriminator (MSD). This dual-branch setup allows the model to detect both short-term periodic patterns and broader temporal inconsistencies that may arise during synthesis.

The MPD focuses on capturing periodicity in the audio signal—an essential characteristic of natural speech, especially in voiced segments. It consists of several sub-discriminators, each operating on a version of the waveform sampled at a fixed period (e.g., 2, 3, 5, 7, 11 samples apart). By examining these staggered views of the waveform, the MPD is able to highlight pitch-related features and rhythmic artifacts that are often missed by conventional discriminators.

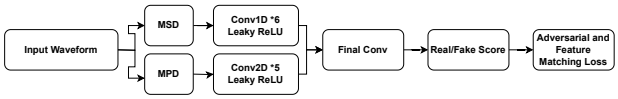

Fig. 3: Architecture of the BAANI-Discriminator, featuring both Multi-Period and Multi-Scale modules. Adapted from [9].

In parallel, the MSD examines the waveform at multiple resolutions by progressively downsampling the input audio.

This enables the discriminator to evaluate the signal at both fine-grained and global levels, helping it detect subtle artifacts as well as broader inconsistencies in prosody or structure. Each sub-discriminator within the MSD consists of a series of convolutional layers with different kernel sizes and strides, effectively building a hierarchical understanding of the signal's temporal structure.

Together, the MPD and MSD branches provide a rich and diverse feedback signal during adversarial training. By simultaneously focusing on short-term regularities and long-term coherence, the BAANI-Discriminator plays a crucial role in guiding the generator toward producing speech that not only sounds realistic, but also captures the natural rhythm, pitch, and continuity of human voice.

C. BAANI- Loss Function

BAANI trained using a combination of adversarial [10], feature matching [14], and mel-spectrogram losses [8]. The adversarial losses for the discriminator and generator are defined as:

$$L_{\text{Adv}}(D;G) = E_{(x,s)}\left[(D(x)-1)^2 + (D(G(s)))^2\right] \ , \quad (1)$$

$$L_{\text{Adv}}(G;D) = E_{s}\left[(D(G(s))-1)^2\right] \ . \quad (2)$$

The mel-spectrogram reconstruction loss is given by:

$$L_{\text{Mel}}(G) = E_{(x,s)}\left[\|\phi(x) - \phi(G(s))\|_1\right] \ , \quad (3)$$

where $\phi(\cdot)$ represents the mel-spectrogram transformation. To encourage the generator to produce perceptually similar audio, a feature matching loss is used:

$$L_{\text{FM}}(G;D) = E_{(x,s)}\left[\sum_{i=1}^{T}\frac{1}{N_i}\|D_i(x) - D_i(G(s))\|_1\right] \ , \quad (4)$$

where $D_i(\cdot)$ denotes the activation from the i^{th} layer of the discriminator, and N_i is the number of elements in that layer. The final generator and discriminator losses are computed as:

$$L_G = \sum_{k=1}^{K}\left[L_{\text{Adv}}(G;D_k) + \lambda_{\text{fm}}L_{\text{FM}}(G;D_k)\right] + \lambda_{\text{mel}}L_{\text{Mel}}(G) \ , \quad (5)$$

$$L_D = \sum_{k=1}^{K} L_{\text{Adv}}(D_k;G) \ . \quad (6)$$

IV. Experimental Analysis

A. Database Details

In all experiments, we used the monolingual IndicTTS corpus [15], which consists of paired audio and text data in the form of `<wavs, UTF-8 encoded transcripts in Gurmukhi script>` and approx. 10.5 hours of studio-quality recordings from a male and female speaker(s), sampled at 48 kHz in 16-bit mono WAV format.

B. Model Parameters

In BAANI training, we used a batch size of 16 for all experiments. The hyperparameter comparison between BAANI-V1 and BAANI-V2 is presented in Table II. The proposed models, V1 and V2, were trained for up to 1×10^5 steps to achieve stable performance and generate satisfactory quality Mel-spectrograms.

TABLE II: Comparison of Parameter Configurations Between the Proposed BAANI-V1 and BAANI-V2 Models.

Parameter	BAANI-V1	BAANI-V2
α (Learning rate)	2×10^{-4}	
β_1	0.8	
β_2	0.99	
Total parameters	296M	2.1M
Learning rate decay	0.999	
Upsample rates	[8, 8, 2, 2]	[4, 4, 2]
Initial channel	1536	128
No. of kernels	3	
Kernel sizes	[3, 7, 11]	[3, 5, 7]
No. of upsample layers	4	3
Upsampling layer	[16, 16, 4, 4]	[8, 8, 4]
Dilation sizes	[[1, 3, 5], [1, 3, 5], [1, 3, 5]]	[[1, 2], [1, 2], [1, 2]]
num_mels	80	
No. of FFT points	1024	512
Hop size	256	
Window size	1024	512

C. Experimental Setup

All experiments were performed using the 24.04 LTS-based Ubuntu workstation with an Intel Core i7-12700 processor, 16GB of RAM, an NVIDIA GTX 1080 GPU (8GB VRAM).

V. Results and Discussion

To assess the quality and intelligibility of the synthesized Punjabi speech samples, we employed a comprehensive set of evaluation metrics: Mean Opinion Score (MOS), Perceptual Evaluation of Speech Quality (PESQ), Mel Cepstral Distortion (MCD), Mel Spectral Distortion (MSD), Short-Time Objective Intelligibility (STOI), Cosine Similarity (CS), Word Error Rate (WER), and Character Error Rate (CER). These were used to compare our two model variants—PunjabiV1 (high-capacity) and PunjabiV2 (lightweight and quantized)—against SOTA vocoders such as UnivNET [16], Parallel WaveGAN [8], HiFi-GAN [9], MelGAN [17], BigVGAN [18].

A. Melspectrogram Comparison

Figure 4 presents the mel spectrograms (0–2s) for the female Punjabi utterance *train_punjabifem_00067.wav*, synthesized using various state-of-the-art vocoders. Compared to the ground truth, BAANI V1 captures the broader spectral structure and slightly finer harmonics and exhibits energy smearing in mid-frequency bands compared to HiFi-GAN [9] and BigVGAN [11], while ParallelWaveGAN and UnivNet display oversmoothing artifacts. VOCGAN struggles with high-frequency reconstruction, producing blurred formant transitions.

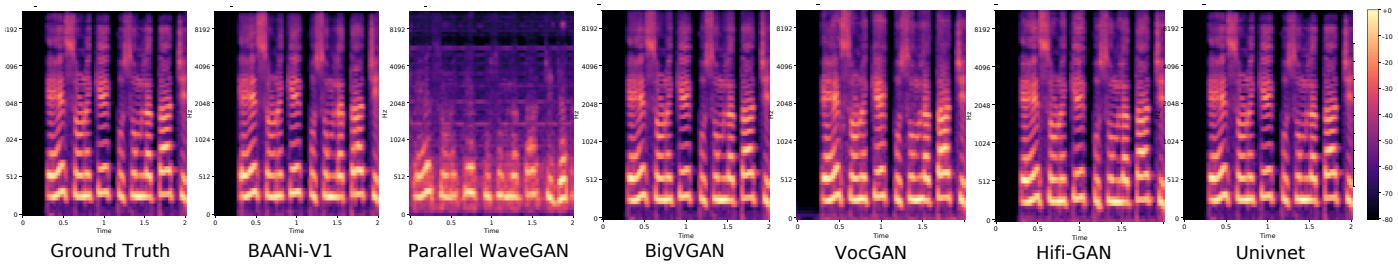

Fig. 4: Comparison of generated speech samples across various state-of-the-art (SOTA) architectures, including Parallel WaveGAN [8], BigVGAN [11], VocGAN [12], HiFi-GAN [9], and UnivNet [16].

B. Inference Speed

All experiments were conducted on a single NVIDIA GTX 1080 GPU (8 GB GDDR5X, 320 GB/s memory bandwidth, 8.87 TFLOPS FP32) [19]. BAANI-V2, with just 7.5 million parameters and 8-bit quantization, achieves real-time inference speeds of **21.4**× on CPU and **18.6**× on GPU. BAANI-V1, with approximately 296 million parameters, achieves **1.5**× CPU and **3.2**× GPU real-time speed while still fitting within the 8 GB VRAM limit.

C. Subjective Measure

As shown in Table IV, BAANI-V1 achieves the highest MOS score of 4.18 ± 0.66 by outperforming most of the existing SOTA vocoders. BAANI-V2 also performs competitively with a MOS of 4.38 ± 0.66, near to the BigVGAN MOS of 4.42 ± 0.13. BAANI-V1 and V2 perform better than HiFi-GAN (4.36 ± 0.14) and MelGAN (3.12 ± 0.18). For subjective evaluations, we took the inputs from 10 subjects (6 subjects belong to the native language as Punjabi), and the male: female ratio was 5:5. We noticed that, generated samples of the BAANI contain more natural phonemes and pleasant-sounding speech than existing models.

1) **MOS** [20]: The final MOS score was computed as the arithmetic mean of all the individual ratings. In particular,

$$\text{MOS} = \frac{1}{N}\sum_{i=1}^{N} S_i \ , \tag{7}$$

where S_i is the rating provided by the i^{th} listener and N is the total number of listeners. The 95% confidence interval for the true MOS is also reported in the Table IV.

2) **NISQA-MOS** [21]: It is a non-intrusive DL-based DL, used to predict human-perceived speech quality across multiple dimensions, such as noise, coloration, discontinuity and loudness [2]. To capture long-range dependencies, both convolutional neural networks (CNNs) and self-attention mechanisms are used. A higher NISQA-MOS score indicates better sample quality and a lower score represents degraded samples.

[2]https://github.com/gabrielmittag/NISQA {Last Accessed: June 10^{th}, 2025}

D. Objective Measures

To compare the effectiveness of the BAANI-V1 and V2 against SOTA architectures, as shown in Table V we evaluated generated samples with objective measures. BAANI-V1 achieved the highest PESQ [22] score of 3.12, with superior perceptual quality, with the lowest MCD [23] (6.65) and MSD (0.184). This indicates better spectral accuracy and reduced distortion. BAANI-V2 also closely matches BigVGAN's and performs better than VocGAN [12] and HiFi-GAN [9]. In STOI [24], BAANI-V1 and V2 achieve the highest scores (0.945 and 0.961, respectively) and CS values (0.902 and 0.926).

1) **PESQ** [22]: It measures perceptual speech quality using a standardized auditory model. In particular,

$$\text{PESQ} = f(x, \hat{x}) \ , \tag{8}$$

where x and $\hat{x}$ is the reference and synthesized speech, respectively. f is a perceptual model as defined by ITU-T P.862 [3].

2) **MCD** [23]: It is evaluated using distortion between *reference* and *generated* Mel cepstral coefficients. *i.e.,*

$$\text{MCD[dB]} = \frac{10}{\log 10}\sqrt{2\sum_{d=1}^{k}\left(\text{mcc}_d^t - \text{mcc}_d^{\hat{t}}\right)^2}, \tag{9}$$

where mcc_d^t and $\text{mcc}_d^{\hat{t}}$ indicate the d^{th} dimensional coefficient of MCC features of t and $\hat{t}$, respectively. k is the number of MCC dimensions.

3) **MSD** [25]: It estimates the deviation in Mel spectrogram shapes, to indicating naturalness in time-frequency representation. In particular,

$$\text{MSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(s(\mathbf{y})_i^t - s(\mathbf{y})_i^{\hat{t}}\right)^2}, \tag{10}$$

where i is the frame values of logarithmic modulation spectra that vary from [1, N] values.

4) **STOI** [24]: It estimate(s) intelligibility between degraded and original speech samples:

[3]https://pypi.org/project/pesq/ {Last Accessed: July 14^{th}, 2024}

TABLE III: Comparison of the proposed models with existing state-of-the-art (SOTA) architectures in terms of inference speed Real-Time Factor (RTF) on CPU and GPU both, along with their GPUs' floating-point operations capacity. '–' denotes information not disclosed by the authors.

Architecture	Params (M)	CPU (↑)	GPU (↑)	GPU Used	TFLOPS
UnivNET [16]	10.5	-	200×	NVIDIA Tesla V100 (80 GB)	125
P.WaveGAN [8]	1.44	-	28.68×	NVIDIA Tesla V100 (80 GB)	125
HiFi-GAN [9]	14.0	13.4×	167.9×	Tesla V100 (80 GB)	125
BigVGAN [11]	112	–	44.7×	RTX A8000 (48 GB)	16.3
VocGAN [12]	6.8	3.24×	416.7×	NVIDIA GTX 1080TI (11 GB)	11.3
BAANI-V2	**2.1**	**21.4**×	**18.6**×	**GTX 1080 (8 GB)**	**8.87**
BAANI-V1	**296**	**1.5**×	**3.2**×	**GTX 1080 (8 GB)**	**8.87**

TABLE IV: Subjective evaluation of BAANI against SOTA models, incorporating 95% confidence intervals (CI) from the standard deviation.

Model	MOS Score (↑)	NISQA MOS (↑)
GT	-	4.41 ± 0.31
UnivNET [16]	3.90 ± 0.09	4.4112 ± 0.3184
P.WaveGAN [8]	4.06 ± 0.10	2.3840 ± 1.0023
HiFi-GAN [9]	4.36 ± 0.14	4.42 ± 0.29
BigVGAN [11]	4.42 ± 0.13	4.40 ± 0.30
VocGAN [12]	4.20 ± 0.08	4.1239 ± 0.3160
BAANI-V2	**4.21 ± 0.23**	**4.630 ± 0.08**
BAANI-V1	**4.18 ± 0.66**	**4.3925 ± 0.30**

$$X_j(m) = \sqrt{\sum_{k=k_1(j)}^{k_2(j)-1} |\hat{x}(k,m)|^2}, \tag{11}$$

where k_1 and k_2 denote the one-third octave band edges, and rounded to the nearest Discrete Fourier Transform (DFT) bin.

5) **PCC** [26]: It evaluates linear *alignment* between ground truth and generated waveforms, capturing global waveform similarity, *i.e.,*

$$\text{PCC} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(\hat{x}_i - \bar{\hat{x}})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(\hat{x}_i - \bar{\hat{x}})^2}}, \tag{12}$$

where x_i and $\hat{x}_i$ are the i^{th} sample of the reference and synthesized signals, respectively, and $\bar{x}$, $\bar{\hat{x}}$ are their mean values.

6) **CS** [27]: It measures angular distance between feature vectors (e.g., embeddings or spectrogram frames), providing insight into geometrical closeness in the feature space. *i.e.*,

$$\text{CS} = \frac{\sum_{i=1}^{N} x_i \cdot \hat{x}_i}{\sqrt{\sum_{i=1}^{N} x_i^2} \cdot \sqrt{\sum_{i=1}^{N} \hat{x}_i^2}}, \tag{13}$$

where x_i and $\hat{x}_i$ are the components of the reference and synthesized feature vectors of dimension N., respectively.

E. Quantitative Measures

As shown in quantitative analysis (Table VI), it demonstrates the performance comparison of the proposed models, BAANI-V1 and BAANI-V2, over SOTA vocoders. BAANI-V1 achieved the lowest WER of 0.670 and CER of 0.159 by generating the most accurate speech samples. We used the Whisper (large) [28] model to calculate the performance. BAANI-V2 achieved a WER of 0.671 and CER of 0.165 by outperforming BigVGAN (WER: 0.683, CER: 0.164), MelGAN, and significantly surpassing HiFi-GAN.

1) **WER [29]**: It is used to evaluate the accuracy of transcription or synthesis by comparing the words in the reference and generated outputs. i.e.,

$$\text{WER} = \frac{S + D + I}{N}, \tag{14}$$

where S, D, and I are character-level substitutions, deletions, and insertions, respectively, and N is the total number of characters in the reference.

2) **CER[30]**: CER functions similar to WER but at the character-level, making it suitable for languages with high script complexity or ambiguous word boundaries, such as Marathi. In particular,

$$\text{CER} = \frac{S + D + I}{N}. \tag{15}$$

VI. Conclusion

Although BAANI has achieved better results, numerous hurdles still pose a way for advancing technology for Indic languages. Most of the existing vocoder systems are crafted with high-resource Western languages in mind, making them poorly compatible with the distinct characteristics of Indian languages. Additionally, the availability of labeled speech data is very limited and common across most of the regional languages, which poses a significant obstacle. Future Research must focus on extracting various intonations of different dialects and integrating them seamlessly into TTS systems. Solving these issues will not only uplift Punjabi speech synthesis but also contribute to better advancements in multilingual and low-resource languages while preserving their meaning.

It is still an open chellenge to deal with dialectal and intonation preservation with minimal supervised data. Small models struggles to capture fine-grained prosody and linguistic

TABLE V: Comparison of BAANI with different SOTA architectures using Objective measures.

Model	PESQ (↑)	MCD (↓)	MSD (↓)	STOI (↑)	CS (↑)	PCC (↑)
UnivNET[16]	3.28	15.56	1.21	0.98	0.99	0.026
P. WaveGAN[8]	1.17	432.00	23.27	0.21	0.11	-0.000
HiFi-GAN[9]	1.98	8.21	0.22	0.86	0.86	-0.010
BigVGAN[11]	2.95	6.78	0.19	0.95	0.90	0.455
VocGAN [12]	2.77	0.96	25.79	1.55	0.97	-0.016
BAANI-V2	**2.91**	**6.82**	**0.19**	**0.95**	**0.90**	**-0.004**
BAANI-V1	**3.12**	**6.65**	**0.18**	**0.96**	**0.93**	**-0.003**

TABLE VI: Comparison of BAANI with different SOTA architectures using Quantitative measures.

Model	WER (↓)	CER (↓)
UnivNET [16]	0.681	0.168
P.WaveGAN [8]	0.694	0.174
HiFi-GAN [9]	1.000	0.990
BigVGAN [11]	0.670	0.163
VocGAN [12]	0.682	0.170
BAANI-V2 (Proposed)	**0.689**	**0.243**
BAANI-V1 (Proposed)	**0.681**	**0.238**

detail within limited training steps, mainly due to data scarcity and high computational demands.

REFERENCES

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

[2] 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.

[3] 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.

[4] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[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] 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}.

[7] R. Huang, M. W. Lam, J. Wang, D. Su, D. Yu, Y. Ren, and Z. Zhao, "Fastdiff: A fast conditional diffusion model for high-quality speech synthesis," *arXiv preprint arXiv:2204.09934*, 2022, {Last Accessed: March 16^{th}, 2025}.

[8] 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 *Proc. ICASSP*. IEEE, 2020, pp. 6199–6203.

[9] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial network for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc., 2014.

[11] 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}.

[12] J. Yang, J. Lee, Y. Kim, H. Cho, and I. Kim, "Vocgan: A high-fidelity real-time vocoder with a hierarchically-nested adversarial network," *arXiv preprint arXiv:2007.15256*, 2020.

[13] 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}.

[14] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates, Inc., 2016.

[15] Speech Technology Consortium, Hema A. Murthy, and S. Umesh, "Indic TTS: A text-to-speech database for indian languages," 2023. [Online]. Available: https://www.iitm.ac.in/donlab/indictts/

[16] 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}.

[17] 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.

[18] K. Lee, S. Kim, J. Kim, J. Kong, and S.-H. Kim, "Bigvgan: A universal neural vocoder with large-scale training," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[19] TechSpot, "Nvidia geforce gtx 1080 review," https://www.techspot.com/review/1174-nvidia-geforce-gtx-1080/, 2016.

[20] "ITU-T Recommendation P.800: Methods for subjective determination of transmission quality," ITU-T, Tech. Rep., 1996 {Last Accessed: August 18^{th}, 2024}.

[21] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," *arXiv preprint arXiv:2104.09494*, 2021.

[22] 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," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 749–752, 2001.

[23] 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.

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

[25] N. Komatsu, K. Takeda, and K. Tanaka, "A statistical analysis of modulation spectra for reverberant speech recognition," in *Interspeech*, 2016.

[26] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 1988.

[27] A. Singhal, "Modern information retrieval: A brief overview," *IEEE Data Engineering Bulletin, Vol. 24(4), pp. 35–43*, 2001.

[28] A. Radford, J. Gao, J. W. Kim, T. Xu, G. Brockman, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," https://github.com/openai/whisper, 2023, openAI Technical Report.

[29] S. Goldwater and M. Johnson, "Words Worth: How robust automatic speech recognition is to speech variations," in *IEEE Workshop on Spoken Language Technology (SLT), pp. 248–253*, 2010, Berkeley, California, USA.

[30] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4960–4964*, 2016, Lujiazui, Shanghai, China.