Deepfake Defense: Constructing and Evaluating a Specialized Urdu Deepfake Audio Dataset

Sheza Munir, Wassay Sajjad, Mukeet Raza, Emaan Mujahid Abbas,
Abdul Hameed Azeemi, Ihsan Ayyub Qazi, Agha Ali Raza
Lahore University of Management Sciences
Pakistan
shezamnr@umich.edu, {24100226, 23100313, 23100276,
abdul.azeemi, ihsan.qazi, agha.ali.raza}@lums.edu.pk

Abstract

Deepfakes, particularly in the auditory domain, have become a significant threat, necessitating the development of robust countermeasures. This paper addresses the escalating challenges posed by deepfake attacks on Automatic Speaker Verification (ASV) systems. We present a novel Urdu deepfake audio dataset for deepfake detection, focusing on two spoofing attacks – Tacotron and VITS TTS. The dataset construction involves careful consideration of phonemic cover and balance and comparison with existing corpora like PRUS and PronouncUR. Evaluation with AASIST-L model shows EERs of 0.495 and 0.524 for VITS TTS and Tacotron-generated audios, respectively, with variability across speakers. Further, this research implements a detailed human evaluation, incorporating a user study to gauge whether people are able to discern deepfake audios from real (bonafide) audios. The ROC curve analysis shows an area under the curve (AUC) of 0.63, indicating that individuals demonstrate a limited ability to detect deepfakes (approximately 1 in 3 fake audio samples are regarded as real). Our work contributes a valuable resource for training deepfake detection models in low-resource languages like Urdu, addressing the critical gap in existing datasets. The dataset is publicly available at: https://github.com/CSALT-LUMS/urdu-deepfake-dataset.

1 Introduction

Automatic Speaker Verification, a method for biometric person recognition, has gained popularity in recent years. However, this surge in popularity has also given rise to new challenges in the form of spoofing or deepfake attacks. Initially coined on Reddit in 2017, the term 'deepfake' (Bitesize, 2019) denotes the application of deep learning techniques for face swapping in videos. Presently, the term has evolved to broadly encompass any audio or video manipulation where key attributes are digitally altered or swapped using artificial intelligence (AI) technologies. The ASVspoof community classifies these attacks into two main categories: logical access, involving deepfake-generated audios, speech synthesis, and voice conversion, and physical access, which includes replay attacks and impersonation (Wang et al., 2020b).

Deepfakes, a complex way of manipulating media, make fake content easier to generate and harder to detect. Speech synthesis models now allow the creation of deepfakes that are undetectable by the human ear or even verification systems (Mirsky and Lee, 2021). In 2019, impostors leveraged AI-driven software to replicate the voice of a corporate executive, orchestrating a fraudulent transfer of USD 243,000 (Stupp, 2019). This incident underscores the imperative of developing robust methods to accurately identify deepfake audio in order to counteract such fraudulent activities. In a behavioral study, Kobis et al. (2021) revealed that people cannot easily detect deepfakes, yet they perceive that they can. Thus, these fake audios have the potential to spread misinformation, create mass panic and havoc, malign personalities, and change narratives. Moreover, beyond this social impact, deepfakes have the power to break through systems protected by voice recognition through the spoofing attacks listed above. Considering the adverse effects of deepfake audios, it is crucial to develop systems capable of discerning between real and deepfake audio. The ASVspoof challenge, a community-led initiative, promotes the development of such countermeasures against deepfakes and audio spoofing (Wu et al., 2015; Kinnunen et al., 2017; Todisco et al., 2019; Yamagishi et al., 2021).

Countermeasures against deepfakes include detection algorithms designed to identify features in deepfake audios. The physical attributes of sound, encompassing pitch, texture, loudness, and duration, can now be accurately replicated in artificially generated deepfake audios. To detect the features that differentiate bonafide (actual utter-

Findings of the Association for Computational Linguistics ACL 2024, pages 14470–14480
August 11-16, 2024 ©2024 Association for Computational Linguistics

ances of the people) and fake audios, the model needs to train on a large amount of data (Azeemi et al., 2022). These differentiations are based on spectral and temporal differences and micro features (Delgado et al., 2021; Dhamyal et al., 2021; Tak et al., 2020). Widely used datasets created for this purpose include WaveFake (Frank and Schönherr, 2021), FakeAVCeleb (Khalid et al., 2021), and the ASVspoof dataset (Wang et al., 2020b) itself. These datasets, from high-resource languages, exemplify the large amount of data required to train deepfake detection models. Unfortunately, in low-resource languages, this large amount of data is unavailable. To cater to this lack of data in Urdu, we create and evaluate a dataset that can be used to train against spoofing attacks.

1.1 Contributions

The presented research offers the following contributions:

- We present an audio deepfake dataset, containing 20,451 utterances of bonafide and 16,830 utterances of deepfake audio, to train detection models in Urdu, a low-resource language. The dataset is hosted on a publicly accessible repository[1].

- We assess the dataset through human evaluation and discover that about one out of every three audio samples goes undetected by individuals as being fake. This finding carries implications for the potential spread of misinformation.

- We evaluate the dataset qualitatively and qualitatively. Qualitative measures include examining the variations in the relative distribution of deepfake-generated and real audios using t-SNE plotting and comparing L2 norms between bonafide audios and each set of deepfake audios. For quantitative analysis, we calculate the Equal Error Rate (EER) across various speakers and spoofing attacks.

2 Related Works

2.1 Deepfake Detection Models

The field of audio deepfake detection has seen remarkable growth recently, focusing on using machine learning to differentiate real speech from synthetic audio (Wu et al., 2020; Wang et al., 2020a; Chen et al., 2020). This research typically follows either a conventional pipeline method, combining feature extraction with classification, or newer end-to-end methodologies that process raw audio data directly for both tasks.

A key hurdle in this domain is the extensive data required for training advanced deep learning Text-to-Speech (TTS) models (Ping et al., 2017; Shen et al., 2017; Sotelo et al., 2017; Tachibana et al., 2017; Wang et al., 2017). Research has shown high efficacy for multi-speaker TTS models, especially when data for a specific speaker is limited (Latorre et al., 2018; Luong et al., 2019). The study by Luong et al. (2019) emphasized the superiority of multi-speaker models using oversampling techniques in scenarios with sparse data. While undersampling generally showed negative impacts, ensemble methods were noted for their ability to improve speech naturalness, albeit at the cost of higher computational resources (de Korte et al., 2020).

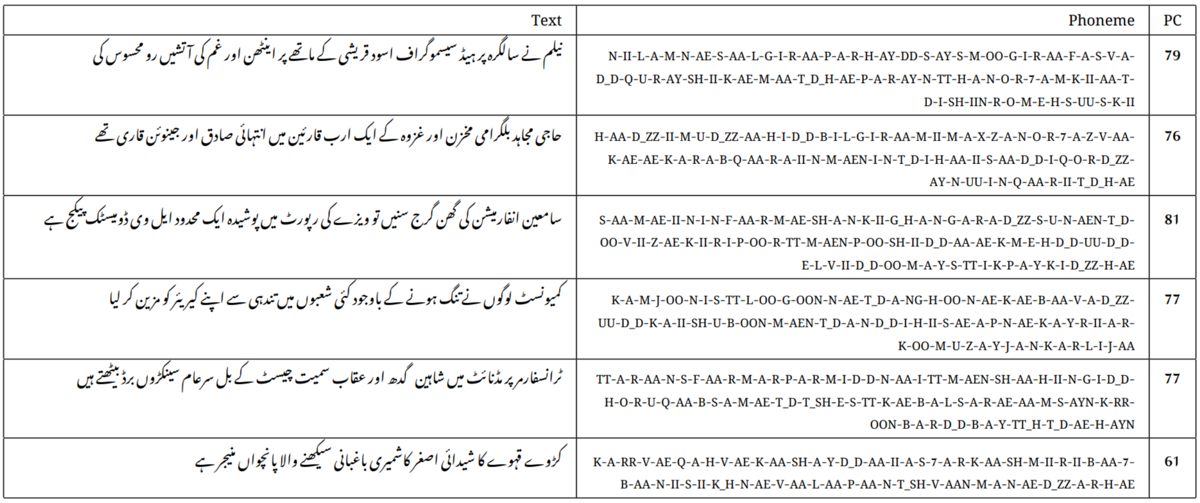
Furthermore, the majority of research and competitions in audio deepfake detection, such as ASVspoof and ADD, are focused on English and Chinese, reflecting a language bias due to easier data collection (Wang et al., 2020b; Yi et al., 2022).

2.2 Deepfake Detection Datasets

The creation of robust TTS datasets is vital for the development of effective detection models. These datasets should be of high quality, featuring diverse speakers, accurate transcripts, and ample audio content per speaker (Bakhturina et al., 2021). Best practices for TTS dataset creation underscore the necessity for error-free, clear recordings, uniformity in tone and pitch, comprehensive phoneme representation, and overall naturalness. Rigorous quality assessments, such as examining the length of clips and transcripts and inspecting spectrograms for noise, are also advised to maintain dataset integrity (coq, 2023).

Recent trends in audio deepfake research include using alternative data sources to address the lack of target data. Efforts to build TTS datasets through community-driven or automated collection and transcription processes have been observed (Gutkin et al., 2016; Xu et al., 2020; Wibawa et al., 2018). However, these methods might result in datasets with lower recording quality and naturalness, which could impact the effectiveness of TTS models when compared to traditional datasets (Guo

[1] Public dataset repository: https://github.com/CSALT-LUMS/urdu-deepfake-dataset



Text	Phoneme	PC
نیلم نے سالگرہ پر ہیڈ سیسموگراف اسود قریشی کے ماتھے پر اینٹھن اور غم کی آتشیں رو محسوس کی	N-II-L-A-M-N-AE-S-AA-L-G-I-R-AA-P-A-R-H-AY-DD-S-AY-S-M-OO-G-I-R-AA-F-A-S-V-A-D_D-Q-U-R-AY-SH-II-K-AE-M-AA-T_D_H-AE-P-A-R-AY-N-TT-H-A-N-O-R-7-A-M-K-II-AA-T-D-I-SH-IIN-R-O-M-E-H-S-UU-S-K-II	79
حاجی مجاہد بلگرامی مخزن اور غزوہ کے ایک ارب قارئین میں انتہائی صادق اور جینوئن قاری تھے	H-AA-D_ZZ-II-M-U-D_ZZ-AA-H-I-D_D-B-I-L-G-I-R-AA-M-II-M-A-X-Z-A-N-O-R-7-A-Z-V-AA-K-AE-AE-K-A-R-A-B-Q-AA-R-A-II-N-M-AEN-I-N-T_D-I-H-AA-II-S-AA-D_D-I-Q-O-R-D_ZZ-AY-N-UU-I-N-Q-AA-R-II-T_D_H-AE	76
سامعین انفارمیشن کی گھن گرج سنیں تو ویزے کی رپورٹ میں پوشیدہ ایک محدود ایل وی ڈومیسٹک پیکج ہے	S-AA-M-AE-II-N-I-N-F-AA-R-M-AE-SH-A-N-K-II-G_H-A-N-G-A-R-A-D_ZZ-S-U-N-AEN-T_D-OO-V-II-Z-AE-K-II-R-I-P-OO-R-TT-M-AEN-P-OO-SH-II-D_D-AA-AE-K-M-E-H-D_D-UU-D_D-E-L-V-II-D_D-OO-M-A-Y-S-TT-I-K-P-A-Y-K-I-D_ZZ-H-AE	81
کمیونسٹ لوگوں نے تنگ ہونے کے باوجود کئی شعبوں میں تندہی سے اپنے کیریئر کو مزین کر لیا	K-A-M-J-OO-N-I-S-TT-L-OO-G-OON-N-AE-T_D-A-NG-H-OO-N-AE-K-AE-B-AA-V-A-D_ZZ-UU-D_D-K-A-II-SH-U-B-OON-M-AEN-T_D-A-N-D_D-I-H-II-S-AE-A-P-N-AE-K-A-Y-R-II-A-R-K-OO-M-U-Z-A-Y-J-A-N-K-A-R-L-I-J-AA	77
ٹرانسفارمر پر مڈنائٹ میں شاہین گدھ اور عقاب سمیت چیسٹ کے بل سرعام سینکڑوں برڈ بیٹھتے ہیں	TT-A-R-AA-N-S-F-AA-R-M-A-R-P-A-R-M-I-D-D-N-AA-I-TT-M-AEN-SH-AA-H-II-N-G-I-D_D-H-O-R-U-Q-AA-B-S-A-M-AE-T_D-T_SH-E-S-TT-K-AE-B-A-L-S-A-R-AE-AA-M-S-AYN-K-RR-OON-B-A-R-D_D-B-A-Y-TT_H-T_D-AE-H-AYN	77
کڑوے قہوے کا شیدائی اصغر کاشمیری باغبانی سیکھنے والا پانچواں منیجر ہے	K-A-RR-V-AE-Q-A-H-V-AE-K-AA-SH-A-Y-D_D-AA-II-A-S-7-A-R-K-AA-SH-M-II-R-II-B-AA-7-B-AA-N-II-S-II-K_H-N-AE-V-AA-L-AA-P-AA-N-T_SH-V-AAN-M-A-N-AE-D_ZZ-A-R-H-AE	61

Figure 1: PRUS Corpus

et al., 2022).

Recently, the focus on enhancing TTS systems for under-resourced languages has gained traction. Researchers are exploring how well-structured datasets in various languages can improve TTS for languages with scarce resources. Techniques like cross-lingual transfer learning and multilingual TTS are being investigated for this purpose (Azizah et al., 2020; Tu et al., 2019; He et al., 2021), aiming to democratize TTS technology and extend its reach to a wider range of languages and dialects.

2.3 Benchmark Dataset

The Phonetically Rich Urdu Speech Corpus (PRUS) and the PronouncUR lexicon are crucial resources for developing and benchmarking Urdu Text-to-Speech (TTS) systems, particularly in the context of audio deepfakes in Urdu, a low-resource language.

PRUS, consisting of 70 minutes of transcribed read speech, with its comprehensive phonetic coverage, including 62 out of 67 total phonemes in Urdu and a wide array of tri-phonemes, offers a detailed representation of Urdu's phonetic diversity. This corpus, balancing high-frequency word focus with practical dataset size, serves as an ideal benchmark for phonetic diversity and quality assessment in TTS systems. Figure 1 shows a snippet of PRUS corpus and its phoneme counts (PC).

PronouncUR's lexicon, encompassing approximately 46,000 words and covering 64 out of 67 phonemes, provides a broad spectrum of Urdu sounds. Its phoneme frequency distribution and expert tagging make it invaluable for evaluating TTS system comprehensiveness and phonetic accuracy.

The availability of PRUS and PronouncUR open up the opportunity to develop benchmark datasets for audio deepfakes in languages like Urdu. These resources are not only vital for TTS system development but also offer a framework for detecting and authenticating audio deepfakes, addressing a significant challenge in digital communication in low-resource languages.

3 Methodology

To create the text corpus for the dataset, we randomly select sentences from reputable Urdu news sources. We then analyze the phonemic structure of the text corpus, ensuring its alignment with natural language patterns. Statistical measures confirm the dataset's phonemic cover and balance. For the spoofing attacks, advanced text-to-speech models Tacotron and VITS TTS are utilized to generate deepfake audios. Figure 2 highlights the steps taken in dataset construction.

3.1 Phonemic Analysis of the Datasets

The text corpus (referred to as the news corpus here onwards) for our dataset has been curated by randomly selecting 495 sentences from reputable Urdu news sources, with permission. Given the rich phonemic inventory inherent in the Urdu language (Raza et al., 2009), it is imperative to ensure that our dataset possesses a comprehensive phonemic cover and balance. To achieve this, we conduct a careful analysis to ascertain the presence of all possible phonemes within the text and to verify

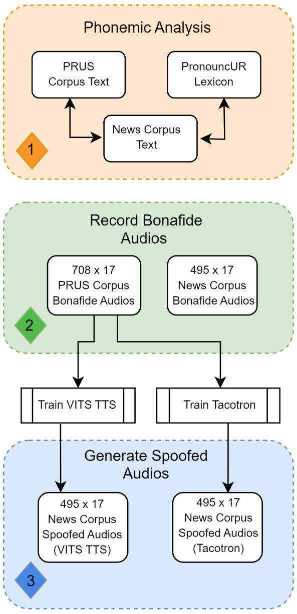

Phonemic Analysis
PRUS Corpus Text
PronouncUR Lexicon
News Corpus Text
1
Record Bonafide Audios
708 x 17 PRUS Corpus Bonafide Audios
495 x 17 News Corpus Bonafide Audios
2
Train VITS TTS
Train Tacotron
Generate Spoofed Audios
495 x 17 News Corpus Spoofed Audios (VITS TTS)
495 x 17 News Corpus Spoofed Audios (Tacotron)
3

Figure 2: Step-by-step summary of dataset construction.

whether their frequencies aligned with those observed in natural language (Zia et al., 2018).

To establish the phonemic fidelity of our dataset, we conduct a comparative analysis with established Urdu corpora known for their adherence to Urdu's phonemic distribution patterns. Notably, we employ the Phonetically Rich Urdu Corpus (PRUS) (Raza et al., 2009) and PronouncUR (Zia et al., 2018) as references.

In our linguistic research, we conducted a comparative analysis of phoneme ranks across two different corpora: the PRUS Corpus and the PronouncUR Corpus, each compared against the News Corpus. We formulate the null hypothesis stating no significant correlation between the phoneme distributions of the two datasets. The visual data from the line graphs illustrate a striking similarity in phoneme distribution in both comparisons. Figure 4 shows the phoneme rank comparison between PRUS Corpus and the News Corpus, while Figure 5 shows the phoneme rank comparison between PronouncUR training lexicon and the News Corpus. This visual correlation is statistically substantiated by Spearman's Rank Correlation Coefficient. It can be understood as ranging from no association (coefficient = 0) to a perfectly monotonic relationship (coefficient = −1 or +1). We observe values of 0.977 for the PRUS Corpus comparison and 0.958 for the PronouncUR comparison, both suggesting exceptionally strong positive monotonic correlations. These high coefficients are coupled with near-zero p-values, confirming that these correlations are statistically significant and not products of chance. Spearman's metric was particularly apt for these analyses as it adeptly captures monotonic relationships without the need for data normality, and it remains robust in the presence of outliers.

Metric	PRUS vs PronounceUR	P-Value
Spearman's Rank Correlation	0.956	< 2.2e-16
Kendall's Tau Coefficient	0.845	5.67e-40
Average Rank Difference	3.34	-

Table 1: Phoneme Rank Evaluation Metrics for PRUS vs PronounceUR

Metric	PRUS vs News Corpus	P-Value
Spearman's Rank Correlation	0.977	< 2.2e-16
Kendall's Tau Coefficient	0.888	5.67e-40
Average Rank Difference	2.66	-

Table 2: Phoneme Rank Evaluation Metrics for PRUS vs News Corpus

Metric	PronouncUR vs News Corpus	P-Value
Spearman's Rank Correlation	0.958	< 2.2e-16
Kendall's Tau Coefficient	0.841	1.60e-22
Average Rank Difference	4.04	-

Table 3: Phoneme Rank Evaluation Metrics for PronouncUR vs News Corpus

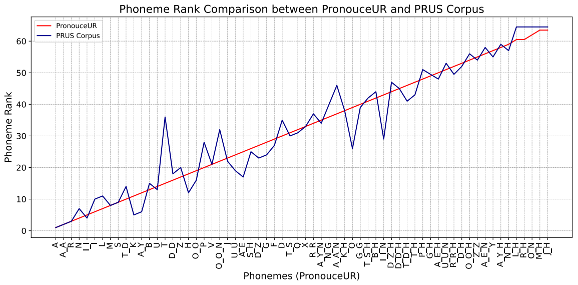


Figure 3: Phoneme Rank Comparison between PRUS Corpus and PronounceUR Corpus.

In addition to comparing the News Corpus with established Urdu corpora, we conducted a detailed phonemic analysis comparing the PRUS Corpus and the PronounceUR Corpus. The results of this comparison are visualized in Figure 3 showing the rank correlation of phonemes between the two corpora.

The Spearman's Rank Correlation coefficient of 0.956 and Kendall's Tau coefficient of 0.845 both indicate a strong positive correlation between the phoneme ranks in the PRUS and PronounceUR corpora. The Average Rank Difference of 3.34

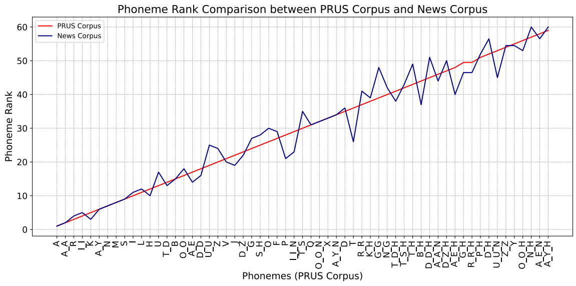


Figure 4: Phoneme Rank Comparison between PRUS Corpus and News Corpus.

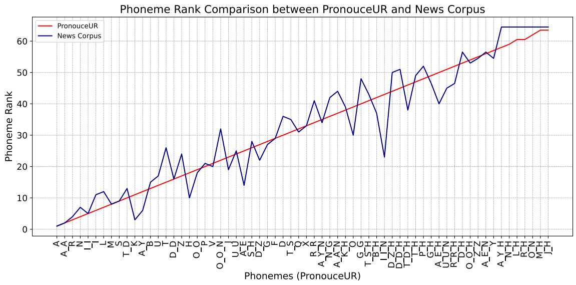


Figure 5: Phoneme Rank Comparison between PronouncUR and News Corpus.

suggests a close similarity in the rank order of phonemes between the two datasets. These results further confirm the consistency and reliability of phoneme usage patterns across different linguistic resources.

Our investigation extended to lexical distribution via Zipf's Law, which posits an inverse relationship between word frequency and its rank in a corpus. Analyzing our dataset against this law, we observed a distribution pattern closely aligning with Zipfian expectations. The linear regression analysis of the log-log plot, as illustrated in Figure 6, yielded a slope of -0.8676, close to the ideal Zipfian slope of -1, and an R-squared value of 0.9595. These results underscore a strong adherence to Zipf's Law, indicating a natural linguistic patterning within the Urdu news corpus. This adherence not only highlights the corpus's linguistic representativeness but also validates its utility for computational linguistics research. The close alignment with Zipfian expectations reinforces the dataset's suitability for exploring language models and comprehension studies, affirming its value in linguistic and phonemic research endeavors.

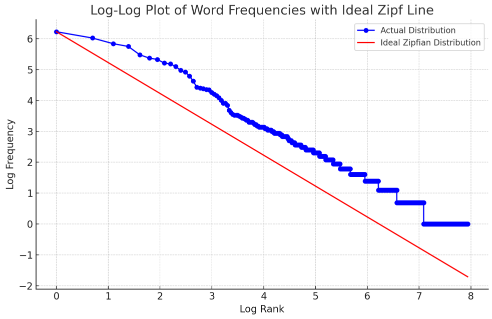
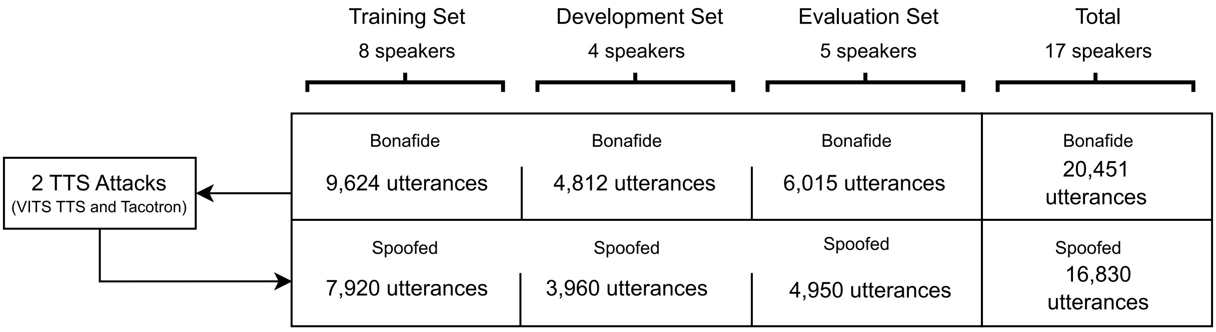


Figure 6: Log-Log plot of word frequencies in Urdu news corpus exhibiting a Zipfian distribution

Furthermore, the strength of these relationships is reinforced by Kendall's Tau Coefficient. It can again be understood as ranging from no association (coefficient = 0) to a perfectly monotonic relationship (coefficient = −1 or +1). We observe values of 0.888 for the PRUS comparison and 0.841 for the PronouncUR comparison. These coefficients mirror the strong positive correlations indicated by Spearman's, and their very low p-values support the notion of a significant, non-random association between the phoneme ranks in the respective corpora. The conservative nature of Kendall's Tau makes it a suitable choice for the datasets, especially considering that it is less influenced by small sample sizes and the non-parametric nature of the data.

Additionally, the Average Rank Difference metric complements these findings, showing minimal discrepancies in phoneme rankings between the PRUS Corpus and the News Corpus at approximately 2.66, and a slightly larger yet modest variation of approximately 4.04 when comparing the PronouncUR Corpus to the News Corpus. Despite the slight differences indicated by this metric, the strong Spearman's and Kendall's correlations confirm a general consistency in phoneme rank order across the examined linguistic resources. The coefficients and p-values from both hypothesis tests indicate a significant correlation, thereby rejecting the null hypothesis.

The integration of Spearman's Rank Correlation, Kendall's Tau, and Average Rank Difference in these analyses provides a robust, multifaceted validation of the initial graphical observations. It collectively supports the conclusion that there is a substantial overlap in phoneme usage patterns within the compared linguistic resources. While the PronouncUR Corpus exhibits a slightly greater



	Training Set 8 speakers	Development Set 4 speakers	Evaluation Set 5 speakers	Total 17 speakers
Bonafide	9,624 utterances	4,812 utterances	6,015 utterances	20,451 utterances
Spoofed	7,920 utterances	3,960 utterances	4,950 utterances	16,830 utterances

Figure 7: Distribution and splits of the dataset

variability in phoneme rank compared to the PRUS Corpus, both corpora maintain a significant parallelism with the News Corpus, underscoring the reliability of phoneme usage patterns across different linguistic datasets. Table 1 and 2 summarize the results of the phonemic analysis.

3.2 Spoofing Attacks

We create a dataset consisting of a combination of bonafide and deepfake audios. In order to achieve this, we choose two advanced text-to-speech (TTS) models, Tacotron (Wang et al., 2017) and VITS TTS (Kim et al., 2021), to generate the deepfake audio. This selection is based on their demonstrated effectiveness in processing the Urdu language, essential due to its complex phonetic structure, and the popularity of these models in deepfake generation. Additionally, these models represent the cutting edge in TTS technology, providing high-quality, realistic audio outputs. The choice of two distinct models, one based on a sequence-to-sequence model with attention (Tacotron) and the other on a Conditional Variational Autoencoder with Adversarial Learning (VITS TTS), allowed for a comprehensive exploration of audio deepfake generation methodologies. The models have been fine tuned to work on Urdu datasets.

3.2.1 Spoofing Attack 1: Tacotron

Tacotron serves as an end-to-end text-to-speech (TTS) model based on the sequence-to-sequence (seq2seq) paradigm with an attention mechanism. In our study, we train and utilize a Tacotron model to generate deepfake audios. This model incorporates PronouncUR (Zia et al., 2018) as a pronunciation lexicon, functioning as a grapheme-to-phoneme (G2P) converter. During the training process, sentences from the PRUS corpus (Raza et al., 2009) are initially passed to PronouncUR to convert them into a string of phonemes, which are then fed into the pre-trained Tacotron model.

3.2.2 Spoofing Attack 2: VITS TTS

VITS (Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech) stands as an end-to-end text-to-speech model that combines an encoder and vocoder. In our study, VITS TTS serves as the second attack method. This attack analyzes input text using natural language processing (NLP) techniques to extract linguistic features, including phonemes, stress patterns, and intonation. To train the VITS TTS model, we use the list of sentences from the PRUS Corpus (Raza et al., 2009), along with their corresponding audios.

We train the Tacotron and VITS TTS models on the voice of 17 individuals separately. We then generate the deepfake audios through the trained models. These audios were then compared with the bonafide audios.

3.3 Training Data Collection

We train Tacotron and VITS TTS on the PRUS corpus audios. To achieve this, we select a sample of 20 student volunteers who record the 708 sentences from the PRUS corpus. Each speaker receives a set of pre-recorded audios, articulating every sentence of the PRUS corpus. Participants attentively listen to each audio before reproducing the sentence in their own voice. We also document the laptop make, model, and headphones used by each speaker during recording, and they are instructed to record in a quiet, closed environment. Upon completing the recording stage, we carefully choose a sample of 17 speakers (7 female, 10 male) with high-quality complete audio recordings to advance to the next phase of the experiment, and get written consent for the public sharing of their recordings (and derivatives) for research.

	Bonafide Part 1	**Bonafide Part 2**	**Tacotron**	**VITS TTS**
Total Duration (mins)	1,302.66	1,271.65	1,061.96	1,340.79
Maximum Sample Length (mins)	112.42	120.75	80.34	111.01
Minimum Sample Length (mins)	61.73	56.45	44.64	65.53
Average Sample Length (mins)	76.63	74.80	62.47	78.87
Audio files for each speaker	708	495	495	495

Table 4: Summary details of the audios in each dataset split.

3.4 Generation of Deepfake Audios

We assign a unique speaker ID to each speaker based on their training order. This ensures distinct identification while preserving anonymity for the public dataset release. We generate deepfake audios using the final checkpoint of each model, using the 495 sentences of the News Corpus for both attacks. The speakers also record the bonafide audios of the News Corpus. This process yields PRUS and News Corpus recordings as bonafide audios and two sets of deepfake audios (one for each attack) for each speaker. In Figure 7, the distribution of bonafide and deepfake utterances in the final dataset is depicted. The duration and lengths of audios for each split are shown in Table 4. The dataset is segmented across 8, 4, and 5 speakers for training, development, and evaluation, respectively.

3.5 Evaluation of the Dataset

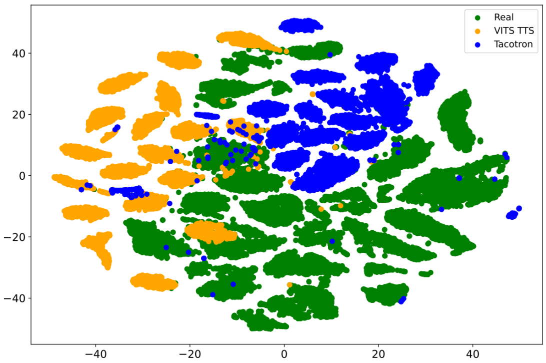


Figure 8: Visualization of Audio Sample Distribution using t-SNE. The graph illustrates the separation of bonafide and deepfake audio samples in a two-dimensional space. Real audio samples are represented by green dots. Yellow dots indicate audio samples generated by VITS TTS model and blue dots represent audio samples synthesized by the Tacotron model.

To understand the differences in the bonafide and deepfake audios in the dataset, it is important to analyze the spectral composition of these subsets. We visualize these subsets by obtaining the Mel Frequency Cepstral Coefficients (MFCCs) of each audio. MFCCs are a representation of the short-term power spectrum of a sound signal. They are commonly used in audio processing and speech recognition. We reduce the dimensions of MFFC features through the tree-based t-SNE algorithm — with a perplexity value of 40 as suggested in (Wang et al., 2020b) and plotting the reduced dimensions. Figure 8 shows the scatter plot of the processed features for each subset. The colors represent different subsets of the dataset, i.e. bonafide audio (green), VITS TTS deepfake audios (yellow), and Tacotron deepfake audios (blue). The smaller clusters within each subset represent individual speakers. We notice differences in the position and distribution of each attack as compared to the bonafide audios. Both deepfake subsets exhibit considerable overlap with the bonafide audios, especially those generated using the Tacotron model, highlighting the spectral similarity between these subsets.

In addition to computing t-SNE of the Mel-frequency cepstral coefficients (MFCCs) from the audio samples, we also calculate the L2 norm of the MFCCs to compare bonafide recordings with those generated by the Tacotron and VITS TTS models. Figure 9 illustrates a notable trend: Tacotron-generated audios exhibit a smaller disparity from bonafide audios compared to VITS TTS-generated audios.

We further evaluate the quality of the generated audios by running it on AASIST-L and RawNet2. AASIST-L (Jung et al., 2022) is a lightweight end-to-end audio anti-spoofing model that can efficiently model spoofing artefacts in temporal and spectral domains. RawNet2 (Tak et al., 2021) is an end-to-end convolutional neural network for audio anti-spoofing. We obtain an overall equal error rate of 0.495 and 0.524 through AASIST-L for audios generated through TTS and Tacotron respectively. The EER breakdown for each speaker through AASIST-L and RawNet2 is presented in Table 5. The EER score for AASIST-L varies from 0.44 to 0.58 depending upon the quality of the generated audios for each speaker. This range for the

EER score indicates that the real and fake audios cannot be distinguished reliably.

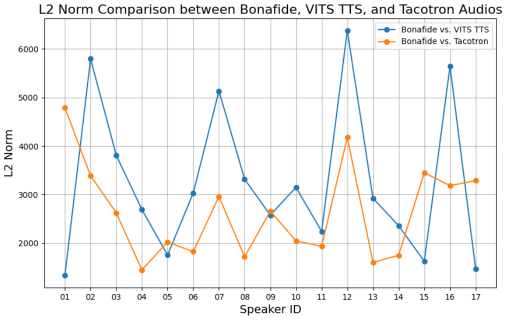


Figure 9: L2 norm comparison between Tacotron and bonafide audios, and VITS TTS and bonafide audios

Speaker	**AASIST-L**		**RawNet 2**	
	TTS	**Tacotron**	**TTS**	**Tacotron**
Speaker 01	0.48	0.47	0.60	0.55
Speaker 02	0.50	0.47	0.61	0.63
Speaker 03	0.50	0.44	0.50	0.48
Speaker 04	0.52	0.46	0.51	0.43
Speaker 05	0.44	0.57	0.43	0.50
Speaker 06	0.44	0.48	0.49	0.49
Speaker 07	0.52	0.50	0.57	0.55
Speaker 08	0.51	0.51	0.52	0.54
Speaker 09	0.47	0.58	0.49	0.43
Speaker 10	0.54	0.47	0.49	0.54
Speaker 11	0.56	0.52	0.50	0.45
Speaker 12	0.53	0.48	0.57	0.43
Speaker 13	0.47	0.47	0.52	0.55
Speaker 14	0.48	0.50	0.57	0.48
Speaker 15	0.49	0.53	0.56	0.48
Speaker 16	0.49	0.53	0.51	0.47
Speaker 17	0.50	0.48	0.51	0.45

Table 5: EER breakdown by speaker ID for VITS TTS and Tacotron audios evaluated through AASIST-L and Raw Net

4 Human Evaluation

4.1 User Study

To assess the quality of our dataset, we employ a human evaluation-based approach. Participants in our study listen to a set of 30 random audios in a controlled environment and classify each as either Fake (deepfake) or Real (bonafide). We employ a convenience sample of 100 participants between the ages of 10 to 48, with a male-to-female ratio of 70-30, with varying tech literacy. The participants are paid PKR 500 per evaluation (approximately 10 minutes) Each random sample of 30 audios includes 10 random bonafide audios, 10 Tacotron-generated, and 10 VITS TTS-generated audios.

We conduct the evaluation in a controlled environment to eliminate biases stemming from variations in speaker quality. During the assessment, we ask each participant to listen to each audio and give the following instructions: "The audio sample that you will listen to is audio produced by humans or produced artificially by artificial intelligence. Please listen to the audio sample and determine whether the voice is artificially generated or is uttered by a person, judging only on the basis of the voice you hear. You can listen to it as many times as you like. And then share your reasons for the classification." Each participant categorizes each audio in the assigned group of recordings into two distinct groups, real or fake. We document their reasons for classifying the audios as fake or real. We observe that most participants base their judgment on factors such as audio distortion and length. Audios containing longer sentences with minimal pauses for breath are often categorized as deepfake generated.

4.2 Analyzing User Study Results

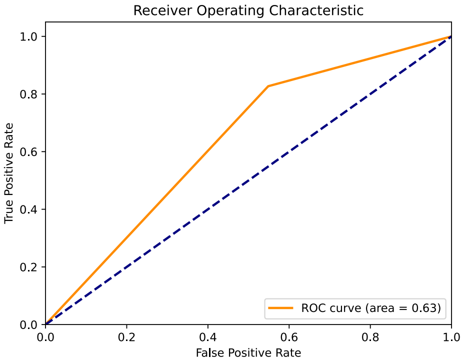


Figure 10: ROC Curve for human evaluation results

The evaluation results, illustrated by the ROC curve in Figure 10, shed light on how well human participants performed in distinguishing between genuine and deepfake audio samples at various classification thresholds. The ROC curve, plotting True Positive Rate against False Positive Rate, indicated a moderate level of discriminative performance with an Area Under the Curve (AUC) value of 0.63.

This AUC suggests that individuals demonstrated a limited ability to detect deepfakes, with approximately 1 in 3 fake audio samples being misidentified as real. When considering the consequences of such limitations in distinguishing between genuine and manipulated content, especially in contexts like political situations or audio leaks in Pakistan, there is a heightened risk of misinformation spreading. This misinformation could contribute to a climate of mistrust, political polarization, and potentially erode public confidence in state institutions.

The societal impact of these findings on democracy underscores the need for more robust detection mechanisms to mitigate the potential threats posed by deepfakes. Developing reliable methods to differentiate between genuine and manipulated content becomes crucial for safeguarding public trust, political discourse, and the integrity of democratic processes.

5 Limitations and Conclusion

In presenting our Urdu deepfake detection dataset, we recognize limitations and suggest areas for future improvement. The dataset currently emphasizes two text-to-speech (TTS) synthesis methods—Tacotron and VITS TTS. Expanding to a broader range of TTS techniques in future iterations will enhance deepfake detection. The dataset's reliance on a convenience sample leads to a gender imbalance in the speakers, highlighting the need for a more diverse dataset in future work. Additionally, our dataset primarily covers logical access scenarios; future research could include physical access scenarios for added detection challenges. In conclusion, our dataset lays a solid foundation for deepfake detection research in the Urdu language. Addressing the outlined limitations and pursuing future research directions will further enhance the dataset's value and contribute to the advancement of deepfake detection technologies in low-resource languages.

6 Ethical Impact

Deepfakes pose risks of spreading misinformation, causing panic, damaging reputations, and manipulating narratives. While improving detection models is a key solution, it inadvertently fosters the development of more sophisticated deepfake generation models that can evade detection. The creation of extensive deepfake audio datasets raises ethical concerns as it may inadvertently contribute to refining audio deepfake generation techniques. Responsible management of such datasets is crucial to address potential ethical challenges in their deployment.

References

2023. What makes a good tts dataset - coqui tts documentation. Accessed: 2023-12-14.

Abdul Hameed Azeemi, Ihsan Ayyub Qazi, and Agha Ali Raza. 2022. Dataset pruning for resource-constrained spoofed audio detection. In *INTERSPEECH*, pages 416–420.

Kurniawati Azizah, Mirna Adriani, and Wisnu Jatmiko. 2020. Hierarchical transfer learning for multilingual, multi-speaker, and style transfer dnn-based tts on low-resource languages. *IEEE Access*, 8:179798–179812.

Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, and Yang Zhang. 2021. Hi-fi multi-speaker english tts dataset. In *Interspeech*.

B. Bitesize. 2019. Deepfakes: What are they and why would i make one? Available: https://www.bbc.co.uk/bitesize/articles/zfkwcqt.

Tianxiang Chen, Avrosh Kumar, Parav Nagarsheth, Ganesh Sivaraman, and Elie el Khoury. 2020. Generalization of audio deepfake detection. In *The Speaker and Language Recognition Workshop*.

Marcel de Korte, Jaebok Kim, and Esther Klabbers. 2020. Efficient neural speech synthesis for low-resource languages through multilingual modeling. In *Interspeech*.

Héctor Delgado, Nicholas W. D. Evans, Tomi H. Kinnunen, Kong-Aik Lee, Xuechen Liu, Andreas Nautsch, Jose Patino, Md. Sahidullah, Massimiliano Todisco, Xin Wang, and Junichi Yamagishi. 2021. Asvspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *ArXiv*, abs/2109.00535.

Hira Dhamyal, Ayesha Haider Ali, Ihsan Ayyub Qazi, and Agha Ali Raza. 2021. Fake audio detection in resource-constrained settings using microfeatures. In *Interspeech*.

Joel Cameron Frank and Lea Schönherr. 2021. Wavefake: A data set to facilitate audio deepfake detection. *ArXiv*, abs/2111.02813.

Haohan Guo, Fenglong Xie, Xixin Wu, Hui Lu, and Helen M. Meng. 2022. Towards high-quality neural tts for low-resource languages by learning compact speech representations. *ArXiv*, abs/2210.15131.

Alexander Gutkin, Linne Ha, Martin Jansche, Knot Pipatsrisawat, and Richard Sproat. 2016. Tts for low resource languages: A bangla synthesizer. In *International Conference on Language Resources and Evaluation*.

Mutian He, Jingzhou Yang, Lei He, and Frank K. Soong. 2021. Multilingual byte2speech models for scalable low-resource speech synthesis. *ArXiv*, abs/2103.03541.

Jee-Weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-Jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. 2022. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. pages 6367–6371.

Hasam Khalid, Shahroz Tariq, and Simon S. Woo. 2021. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *ArXiv*, abs/2108.05080.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *ArXiv*, abs/2106.06103.

Tomi H. Kinnunen, Md. Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas W. D. Evans, Junichi Yamagishi, and Kong-Aik Lee. 2017. The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In *Interspeech*.

Nils Kobis, Barbora Doležalová, and Ivan Soraperra. 2021. Fooled twice - people cannot detect deepfakes but think they can. *iScience*, 24:103364.

Javier Latorre, Jakub Lachowicz, Jaime Lorenzo-Trueba, Thomas Merritt, Thomas Drugman, S. Ronanki, and Klimkov Viacheslav. 2018. Effect of data reduction on sequence-to-sequence neural tts. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7075–7079.

Hieu-Thi Luong, Xin Wang, Junichi Yamagishi, and Nobuyuki Nishizawa. 2019. Training multi-speaker neural text-to-speech systems using speaker-imbalanced speech corpora. In *Interspeech*.

Yisroel Mirsky and Wenke Lee. 2021. The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54:1–41.

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Ö. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2017. Deep voice 3: 2000-speaker neural text-to-speech. *ArXiv*, abs/1710.07654.

Agha Ali Raza, Sarmad Hussain, Huda Sarfraz, Inam Ullah, and Zahid Sarfraz. 2009. Design and development of phonetically rich urdu speech corpus. *2009 Oriental COCOSDA International Conference on Speech Database and Assessments*, pages 38–43.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Z. Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2017. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.

Jose M. R. Sotelo, Soroush Mehri, Kundan Kumar, João Felipe Santos, Kyle Kastner, Aaron C. Courville, and Yoshua Bengio. 2017. Char2wav: End-to-end speech synthesis. In *International Conference on Learning Representations*.

C. Stupp. 2019. Fraudsters used ai to mimic ceo's voice in unusual cybercrime case. In *The Wall Street Journal*. Available online: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara. 2017. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4784–4788.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. 2021. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6369–6373. IEEE.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas W. D. Evans, and Anthony Larcher. 2020. End-to-end anti-spoofing with rawnet2. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6369–6373.

Massimiliano Todisco, Xin Wang, Ville Vestman, Md. Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas W. D. Evans, Tomi H. Kinnunen, and Kong-Aik Lee. 2019. Asvspoof 2019: Future horizons in spoofed and fake audio detection. In *Interspeech*.

Tao Tu, Yuan-Jui Chen, Cheng chieh Yeh, and Hung yi Lee. 2019. End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. *ArXiv*, abs/1904.06508.

Run Wang, Felix Juefei-Xu, Yihao Huang, Qing Guo, Xiaofei Xie, L. Ma, and Yang Liu. 2020a. Deepsonar: Towards effective and robust detection of ai-synthesized fake voices. *Proceedings of the 28th ACM International Conference on Multimedia*.

X. Wang, J. Yamagishi, M. Todisco, et al. 2020b. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Journal of Computer Speech and Language*, 64.

Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Z. Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Robert A. J. Clark, and Rif A. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. In *Interspeech*.

Jaka Aris Eko Wibawa, Supheakmungkol Sarin, Chenfang Li, Knot Pipatsrisawat, Keshan Sanjaya Sodimana, Oddur Kjartansson, Alexander Gutkin, Martin Jansche, and Linne Ha. 2018. Building open javanese and sundanese corpora for multilingual text-to-speech. In *International Conference on Language Resources and Evaluation*.

Zhenjie Wu, Rohan Kumar Das, Jichen Yang, and Haizhou Li. 2020. Light convolutional neural network with feature genuinization for detection of synthetic speech attacks. In *Interspeech*.

Zhizheng Wu, Tomi H. Kinnunen, Nicholas W. D. Evans, Junichi Yamagishi, Cemal Hanilçi, Md. Sahidullah, and Aleksandr Sizov. 2015. Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Interspeech*.

Jin Xu, Xu Tan, Yi Ren, Tao Qin, Jian Li, Sheng Zhao, and Tie-Yan Liu. 2020. Lrspeech: Extremely low-resource speech synthesis and recognition. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md. Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong-Aik Lee, Tomi H. Kinnunen, Nicholas W. D. Evans, and Héctor Delgado. 2021. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. *ArXiv*, abs/2109.00537.

Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, Shan Liang, Shiming Wang, Shuai Zhang, Xin Yan, Le Xu, Zhengqi Wen, and Haizhou Li. 2022. Add 2022: the first audio deep synthesis detection challenge. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9216–9220.

Haris Bin Zia, Agha Ali Raza, and Awais Athar. 2018. Pronouncur: An urdu pronunciation lexicon generator. *ArXiv*, abs/1801.00409.

A Reproducibility and Hyperparameters

The hyperparameters used for training and evaluation of the TTS models are added below. Table 6 contains the hyperparameters for Tacotron and Table 7 shows the hyperparameters for VITS TTS.

Table 6: Training and Evaluation Parameters for Tacotron.

Parameter	Value
Training	
batch_size	32
adam_beta1	0.9
adam_beta2	0.999
initial_learning_rate	0.002
decay_learning_rate	True
use_cmudict	False
Evaluation	
max_iters	450
griffin_lim_iters	60
power	1.5

Table 7: Training and Evaluation Parameters for VITS TTS

Parameter	Value
Training	
batch_size	32
use_speaker_embedding	True
epochs	1000
do_trim_silence	False
learning_rate	0.0002
num_mels	80
sample_rate	16000
Evaluation	
eval_batch_size	16

B Datasets and Evaluation Model

We use the PRUS Corpus available under the Creative Commons license, which allows distribution, remixing, tweaking, and building upon the work, as long as we credit the creators for the original creation. We use PronouncUR and AASIST-L available under the MIT License.