Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Developing High-Quality TTS for Punjabi and Urdu: Benchmarking against MMS Models

Fatima Naseem , Maham Sajid , Farah Adeeba , Sahar Rauf , Asad Mustafa , Sarmad Hussain

Al-Khawarizmi Institute of Computer Science (KICS), University of Engineering and Technology, Lahore, Pakistan

fatimanaseem071@gmail.com, maham.sajid@kics.edu.pk, farah.adeeba@kics.edu.pk, sahar.rauf@kics.edu.pk, asad.mustafa@kics.edu.pk, sarmad.hussain@kics.edu.pk

Abstract

Existing Punjabi text-to-speech (TTS) solutions focus on Gurumukhi script, requiring transliteration from Shahmukhi. This leads to letter substitutions and omissions, resulting in pronunciation errors. In this study, speech corpus, phonetic lexicon, and text analysis module for Punjabi Shahmukhi were developed. Two model architectures: Tacotron 1 and Tacotron 2 with WaveGlow were used to build TTS models. In addition to Punjabi, Urdu TTS models were also developed. These models were benchmarked against Urdu and Punjabi Gurumukhi TTS models provided by Meta's Massively Multilingual Speech (MMS) which is a top profile multilingual speech project. Objective and subjective evaluations indicate that tacotron based Urdu and Punjabi models outperform MMS in intelligibility, naturalness, and phonetic accuracy, enhancing TTS quality for these languages.

Index Terms: speech synthesis, objective and subjective evaluations

1. Introduction

Text-to-speech (TTS) synthesis has witnessed remarkable advancements in recent years, driven by deep learning based models such as Tacotron 1 (Tac1) [1] and Tacotron 2 (Tac2) [2]. Tac2 combined with vocoders like WaveGlow (WG) [3], has significantly improved the naturalness and intelligibility of synthesized speech. However, for many low-resource languages, such as Urdu and Punjabi particularly spoken in Pakistan, the availability of high-quality TTS systems remains limited.

Existing Urdu TTS systems either use traditional synthesis approaches [4] or struggle with language-specific symbols, numbers, and special characters such as Massively Multilingual Speech (MMS) [5], leading to omissions in the synthesized speech which affects the intelligibility and naturalness. Similarly, while TTS systems for Punjabi Gurumukhi exist [6–9], they fail to adequately serve Punjabi Shahmukhi due to phonetic and structural differences between the two scripts.

In this study, deep learning based TTS systems for Urdu and Punjabi Shahmukhi are proposed which can do language-specific text analysis. These systems were built using two different TTS model architectures: Tac1 paired with Griffin-lim [10] and Tac2 paired with WG. These systems were benchmarked against MMS's Urdu and Punjabi Gurumukhi TTS models [5]. MMS provides the synthesis for Punjabi Gurumukhi, so it was evaluated by transliterating Shahmukhi script to Gurumukhi (via chatGPT) and then feeding it to the MMS model. To ensure a comprehensive evaluation, both objective and subjective metrics were employed. Objective measures such as Mel-Cepstral Distortion using Dynamic Time Wrapping (MCD-DTW) [11], SpeechBERTScore [12], and Word Error Rate/Character Error Rate (WER/CER) using whisper [13] were used to assess the naturalness and intelligibility of the synthesized speech. Subjective evaluations include the Diagnostic Rhyme Test (DRT) [14], Modified Rhyme Test (MRT) [15], comprehension test, Semantically Unstructured Sentence (SUS) Test, and Mean Opinion Score (MOS) to gauge the perceptual quality of the synthesized speech.

Our results demonstrate significant improvements in both Urdu and Punjabi TTS systems which highlight the effectiveness of our approach in addressing the phonetic and script-specific challenges posed by these languages.

2. Speech Corpus Development

2.1. Text Corpus

Text corpus is fundamental for TTS systems, as it provides the linguistic diversity necessary to capture a wide range of phonetic variations. For Urdu, the text corpus reported in [16], collected from Urdu news corpus, Urdu digest corpus, and 1M [1] word corpus was used ensuring linguistic coverage and phonetic balance. For Punjabi, text was sourced from publicly available news websites, articles, books, and journals. The dataset was carefully collected from multiple domains such as science, religion, education, engineering, health, entertainment, and agriculture to enhance phonetic and contextual diversity. In both the corpora, each sentence had a 16-word limit, keeping the corresponding audio within 10 seconds to meet TTS models requirement. Furthermore, each sentence was assigned a unique identifier to follow a structured naming convention to ensure alignment with the corresponding audio recordings during the speech corpus creation.

2.2. Phonetic Lexicon

To accurately represent the phonetic characteristics of a language, phonetic inventory is required which maps each letter to its corresponding International Phonetic Alphabet **(IPA)** and Case Insensitive Speech Assessment Method Phonetic Alphabet **(CISAMPA)** representation. The available Urdu phonetic inventory [2] was used for Urdu lexicon development. For Punjabi Shahmukhi, a dedicated phonetic inventory was developed by linguist to capture its unique phonetic features. The detailed Punjabi Shamukhi phonetic inventory can be accessed at https://zenodo.org/records/15532968. For Urdu and Punjabi phonetic lexicons, unique words were extracted from

[1] https://www.cle.org.pk/clestore/urdudigestcorpus1M.htm

[2] https://www.cle.org.pk/Downloads/ling_resources/phoneticinventory/UrduPhoneticInventory.pdf

10.21437/Interspeech.2025-469

the designed text corpora and then transcribed into their CISAMPA representations by expert linguists. This step ensured comprehensive coverage of phoneme combinations and syllable structures resulting in rich and diverse lexicons. Urdu phonetic lexicon contains 184k transcribed entries, while Punjabi lexicon contains 16k transcribed entries. These entries does not only include Urdu and Punjabi words but also numbers and English-transliterated words.

2.3. Audio Recordings

Professional speakers were recruited and recordings were conducted in a soundproof chamber using PRAAT software [3] to ensure acoustic clarity. Prior to recording, speakers were informed about the purpose of data collection, and written consents were signed. To maintain consistency, speakers were instructed to adhere to the same pitch range (f0) and intensity level across all sessions. Any mispronunciations, hesitations, or fumbling in the recorded audio were corrected in real time by re-recording the affected segments. The recordings were captured in mono (single channel) at a 48 kHz sampling rate to ensure high-quality audio creation. Each utterance was limited to 10 seconds. For the Punjabi corpus, we focused on the **Majhi accent** which is the standard dialect of Punjabi [17]. Table 1 shows the number of utterances, total duration of audio data in hours (hrs), average (avg) number of words per utterance, minimum (min), maximum (max), and avg audio durations in seconds (secs) in each dataset recorded.

Table 1: *Overview of Datasets used for TTS Training*

Dataset	# utterances	Total duration	Avg # words/utterance	Audio duration: (min, max, avg)
Punjabi male	14450	20 hrs	10	(2,18,5) secs
Urdu male	8081	10 hrs	9	(2,20,5) secs
Urdu female	8081	10 hrs	9	(2,20,5) secs

3. Experimental Setup

Experiments were conducted using three datasets: Male Urdu, Female Urdu, and Male Punjabi, as described in the Table 1. Tac1, Tac2, and WG models were trained on these datasets. Tac1 follows an encoder-decoder-based TTS approach [1] and uses Griffin-Lim as the default vocoder which uses phase reconstruction algorithm that estimates phase information iteratively from the spectrograms [10]. Tac2 improves upon Tac1 by using a more advanced attention-based architecture for spectrogram generation [2]. WG as vocoder is used with Tac2 which is a flow-based generative model that transforms mel spectrograms into high-quality speech waveforms [3].

The hyperparameters used for training each model are summarized in Table 2. The window length (1024) and hop length (256) were kept at their default values. The batch size was set to 32, considering the available GPU VRAM (12GB per GPU) for efficient training. To ensure smoother convergence and prevent instability due to the increased complexity of Tac2 and WG models, the learning rate was reduced from 0.001 (Tac1) to 0.0005 (Tac2 + WG). Each model was trained for 1500 epochs on **two NVIDIA GeForce RTX 3060 GPUs (12GB VRAM each)**.

It is important to note that the experiments conducted using Tacotron-based models utilized transcribed text in CISAMPA format, leveraging phoneme-based learning. In contrast, MMS operates on raw text, employing character-based text-to-sound mapping.

Table 2: *Values of Hyperparameters for Training*

Model	Window length	Hop length	Batch size	Learning rate	Epochs
Tac1	1024	256	32	0.001	1500
Tac2 + WG	1024	256	32	0.0005	1500

3.1. Urdu and Punjabi Text Processing Modules

At deployment, to handle language-specific numbers, special characters, and symbols in Urdu, the text-analysis approach described in [18] [19] was followed, and the same approach was then mapped to Punjabi which involved mapping numbers (including dates and figures), characters, and symbols into corresponding Punjabi words. Additionally, for handling out of vocabulary (OOV) words, the letter-to-sound (LTS) mapping proposed in [20] was used. For Punjabi, the same letter-to-sound mapping approach was applied to handle OOVs, after incorporating a few sound variations. Specifically, the mapping of certain consonant sounds in Punjabi depends on their position in a word. Sounds /b^{h}/, /dʒ^{h}/, /d̪h/, /g^{h}/, /d^{h}/ when occur at the **initial position** of a word undergo a transformation, mapping to /p/, /tʃ/, /t̪/, /k/, /t/ respectively in Majhi accent.

4. Evaluation and Results

4.1. Size and Real-Time-Factor of Models

The trained models were deployed on **Intel(R) Core(TM) i7-10700F CPU @ 2.90GHz CPU**. Table 4 shows the trained TTS models' sizes and their real-time factor (RTF).

Tac1, with 6.9 million parameters, demonstrates a significantly faster real-time factor (RTF) of 0.17 compared to Tac2+WG, which has a combined parameter size of 296 million and an RTF of 3.54. This suggests that while Tac2+WG are more computationally intensive, their complex architectures contribute to their better performance across languages and genders which will be visible in the evaluation and results section.

4.2. Objective Evaluation

The TTS systems were evaluated using objective metrics: WER and CER (after obtaining transcriptions from whisper [13]), MCD-DTW [11], and SpeechBERTScore [12] to analyze the pronunciation quality, spectral difference, and semantic similarity respectively between reference and synthesized audio samples. The results are summarized in Table 5.

Punjabi Male: For Punjabi male, Tac2 with WG outperformed other models with a WER of 16.1%, CER of 9.1%, SpeechBERTScore (semantic similarity) of 0.72, and an MCD-DTW of 6.6 dB, showing a significant improvement over Tac1. The MMS model, despite being trained on approximately 32 hours of male Punjabi data, showed the highest error rates, with a WER of 31.4% and CER of 16.1%.

Urdu Male: For Urdu male, Tac1 achieved the lowest WER (13.1%) and CER (4.0%), while Tac2+WG demonstrated a slight increase in WER (14.7%) and CER (5.3%) but showed an improvement in SpeechBERTScore (0.76) compared to Tac1 (0.68). This indicates that while Tac1 excels in transcription accuracy, Tac2+WG generates speech

[3] https://www.fon.hum.uva.nl/praat/

Table 3: *Phonetic Features Accuracy Across Models and Datasets*

Dataset	Model	Voicing (%)		Nasality (%)		Aspiration (%)		Sibilation (%)	
		WI	WF	WI	WF	WI	WF	WI	WF
Punjabi Male	Tac1	97.7	90.09	95.4	86.3	50	88.6	97.7	93.1
	Tac2+WG	97.7	86.3	95.4	97.7	65.9	88.6	95.4	95.4
	MMS	79.5	56.8	86.3	63.6	61.3	86.3	81.8	52.2
Urdu Male	Tac1	94.4	52.7	97.2	88.8	97.2	63.8	100	91.6
	Tac2+WG	94.4	61.11	94.4	94.4	94.4	55.5	97.2	97.2
	MMS	97.2	75	100	97.2	91.6	61.1	91.6	97.2
Urdu Female	Tac1	63.8	69.4	88.8	88.8	97.2	63.8	100	91.6
	Tac2+WG	88.8	63.8	100	97.2	91.6	44.4	100	97.2

Table 4: *Size and RTF of Models*

Model	RTF	# params
Tac1	0.17	6.9 M
Tac2+WG	3.54	28 M + 268 M

with higher perceptual quality. The MMS model, trained on 32 hours of male Urdu data, performed poorly in comparison, with a WER of 25.6% and CER of 10.3%.

Urdu Female: For Urdu female, Tac1 again provided the lowest WER (16.3%) and CER (4.0%), while Tac2+WG showed a higher WER (21.7%) and CER (9.4%). Interestingly, Tac2+WG achieved the lowest MCD-DTW (3.3 dB), suggesting better spectral fidelity, though it underperformed in SpeechBERTScore (0.66) compared to Tac1 (0.68). This discrepancy indicates that while the spectral quality improved in Tac2+WG, the overall perceptual quality might have been compromised. The MMS, on the other hand, doesn't come with female Urdu TTS.

In summary, for Punjabi, Tac2 with WG demonstrated superior performance in transcription accuracy (lower WER and CER), perceptual accuracy (higher precision), and spectral quality (lower MCD), establishing it as the leading model for Punjabi speech synthesis. This outcome suggests that the more complex architecture of Tac2+WG is better equipped to capture the intricate linguistic and phonetic features of the Punjabi language. For Urdu, the results were mixed: Tac1 performed comparable to Tac2+WG which indicates that simpler architectures can also effectively model the linguistic characteristics of Urdu. This finding highlights the adaptability of Tac1 for languages with specific phonetic and phonological structures. On contrary, the MMS framework consistently underperformed in both languages and exhibited lower transcription accuracy. Due to the unavailability of the exact training data used for MMS, MCD-DTW and SpeechBERTScore were not computed for this model, limiting the ability to fully assess its spectral and semantic accuracy.

To gain deeper insights into the intelligibility and naturalness of the synthesized speech across these models, a comprehensive series of subjective evaluations were also conducted which are detailed in the following section.

4.3. Subjective Evaluation

Subjective evaluation of the Urdu TTS systems was conducted using a comprehensive test design used in [4]. A similar test design was followed to evaluate Punjabi TTS systems available at https://zenodo.org/records/15532968.

Table 5: *Objective Evaluation Results*

Dataset	Model	WER (%)	CER (%)	MCD-DTW	Speech-BERTScore
Punjabi Male	Tac1	18.6	11.2	7.0	0.67
	Tac2+WG	16.1	9.1	6.6	0.72
	MMS	31.4	16.1	–	–
Urdu Male	Tac1	13.1	4.0	3.7	0.68
	Tac2+WG	14.7	5.3	3.8	0.76
	MMS	25.6	10.3	–	–
Urdu Female	Tac1	16.3	4.0	3.4	0.68
	Tac2+WG	21.7	9.4	3.3	0.66

The evaluation involved **20 participants: 10 male, 10 female aged between 18 to 35**, all of whom were experts in the native languages of Urdu and Punjabi.

4.3.1. DRT and MRT Test

These tests included the DRT and MRT that evaluated the Word initial (WI) and Word final (WF) sounds respectively. Unlike vowels, consonants are harder to recognize in synthetic speech due to abrupt spectral transitions and complex excitation signals [21] [22]. Furthermore, listeners perceive syllable initial and syllable final consonants differently [23]. Therefore, it is logical to evaluate the segmental quality of TTS systems by analyzing consonants in both initial and final positions within monosyllabic words. To achieve this, a test set was designed, consisting of multiple pairs of confusable rhyming words. The consonants were evenly distributed across four phonemic distinctive features such as voicing, nasality, aspiration, and sibilation. Table 3 shows the summarized results of all phonetic features in the evaluated systems.

Punjabi Male: The results indicate that for the Punjabi voice, Tac2+WG emerged as the most balanced system, consistently outperforming others across all phonetic features, particularly in nasality and sibilation. Tac1 demonstrated strength in word-initial (WI) and word-final (WF) voicing, as well as WI sibilation, but struggled with WI aspiration and WF nasality. In contrast, MMS showed significant weaknesses in WF features, making it the least effective system.

Urdu Male: Tac2 + WG remained better in Urdu Male synthesis which outperformed in WF nasality and sibilation with minimal WF aspiration loss. Tac1 was strong in WI features but weak in WF voicing and nasility. MMS performed better in WI voicing but remained limited in WF voicing and aspiration.

Urdu Female: In the the Urdu-Female voice, Tac2+WG led in nasality sibilation, but showed weakness in WF aspiration. Tac1 performed well in WI aspiration and sibilation but struggled with WF features. Tac2+WG is the better choice for Urdu Female, despite its limitations in WF aspiration, due to its overall superior performance. Unfortunately, female MMS Urdu TTS was not available for comparison.

4.3.2. SUS and Comprehension Test

SUSs are used to minimize contextual cues to ensure that intelligibility is measured purely on the basis of the clarity of the synthesized speech [24]. The comprehension test evaluates the system's ability to convey the underlying message of the synthesized speech [25]. Participants listened to the paragraphs sourced from Punjabi and Urdu content and responded to questionnaires designed to test their understanding. Results for SUS and Comprehension tests are shown in Table 6.

Punjabi Male: Tac2+WG achieved the highest intelligibility, as reflected in its SUS score, while Tac1 followed closely behind in a Punjabi voice. Both tacotron based models showed the same level of comprehension. MMS, however showed the lowest scores in both SUS and comprehension tests which highlights difficulties in producing clear and natural Punjabi speech by MMS.

Urdu Male: Tac1 demonstrated the highest SUS score, indicating better intelligibility. However, Tac2+WG showed the best comprehension score. MMS performed the weakest, with both its SUS and comprehension scores being the lowest.

Urdu Female: Both Tac1 and Tac2+WG demonstrated outstanding performance for Urdu Female speech, achieving more than 99% SUS scores and perfect comprehension scores. This indicates that both models generated highly natural and understandable speech.

4.3.3. MOS Test

For further detailed evaluation, intelligibility and naturalness were assessed using the MOS method [26], where participants rated the quality of speech on a 5-point scale, with higher scores indicating better quality. Figure 1 presents the MOS scores for intelligibility and naturalness across different languages. The results indicate the following trends:

Punjabi Male: Tac2+WG achieved the best intelligibility and naturalness score. Tac1 followed with lower scores, while MMS performed the worst, particularly in intelligibility, highlighting its limitations for Punjabi TTS.

Urdu Male: Tac2+WG achieved the highest score in both intelligibility and naturalness. Tac1 showed slightly lower performance in both metrics while MMS scored the lowest among all models with worst results in intelligibility.

Urdu Female: Tac2+WG again outperformed the Tac1 in evaluating the performance of Urdu female which reinforced the advantage of the Tac2+WG over Tac1.

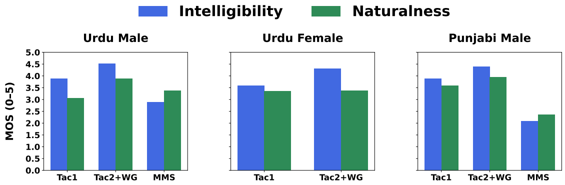

Figure 1: *MOS scores for intelligibility and naturalness*

Table 6: *SUS and Comprehension Test Results*

Dataset	Model	SUS (%)	Comp (%)
Punjabi Male	Tac1	97.8	93.9
	Tac2+WG	99.2	93.9
	MMS	90.8	87.9
Urdu Male	Tac1	97.6	96.3
	Tac2+WG	97.2	100
	MMS	93.6	88.9
Urdu Female	Tac1	99.6	100
	Tac2+WG	99.8	100

5. Conclusion

The objective evaluation results indicate that tacotron based Urdu and Punjabi TTS models performed better by achieving lower WER and CER values than MMS' Urdu and Punjabi models. Subjective evaluation results show that MMS struggled in MOS, SUS, and comprehension tests since it failed to synthesize speech against numbers and symbols in both Urdu and Punjabi. The transliteration layer in Punjabi resulted in letters omissions as well which further contributed to its poor performance. The tacotron based script-specific TTS models with text analysis modules gave better results in MOS, SUS, and comprehension tests. However, when it comes to specific sounds at WI and WF positions, MMS performed better in Urdu evaluations particularly in voicing and nasality. But in Punjabi, its performance again deteriorated particularly due to the phonetic mismatches between Shahmukhi and Gurumukhi scripts. Weak aspiration, voicing contrasts, and nasalization issues in MMS' speech degraded its naturalness. Also, the overall high MCD-DTW values in Punjabi suggest that the models struggle to learn the position-based phonetic variations (described in section 3.1) in Punjabi, Majhi accent.

As of model size and RTF, while Tac2 with WG demand more computational resources due to their larger parameter size and slower real-time factor (RTF), they deliver significantly improved synthesis quality. In contrast, MMS, though lightweight and efficient, sacrifices quality and lacks the language-specific and phoneme based optimizations required for high-quality synthesis. Tac1 emerges as a viable alternative for real-time and resource-constrained scenarios, maintaining a clear quality advantage over MMS.

Overall, the results demonstrated that tacotron based models, paired with Griffin-Lim and WG vocoders, significantly outperform the openly available MMS models for Urdu and Punjabi. Also, only a dedicated Shahmukhi based model can deliver intelligible speech since the transliteration layer does not serve the purpose. Another important point to note is that the poor performance of MMS models highlights that phoneme-based training results in more accurate sound generation. This shows that strong linguistic input is necessary to serve a language and that you cannot rely on character-to-sound mappings for highly intelligible speech synthesis. Additionally, the text analysis and LTS modules ensure phonetic accuracy of a TTS system which is visible in the subjective evaluation results of tacotron-based models. Hence, by integrating phoneme-based transcriptions, domain-specific dataset curation, and language-specific text handling, tacotron based trained models set a new benchmark for high-quality speech synthesis in underserved languages.

6. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[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 *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

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

[4] K. S. Shahid, T. Habib, B. Mumtaz, F. Adeeba, and E. ul Haq, "Subjective testing of urdu text-to-speech (tts) system," *LANGUAGE & TECHNOLOGY*, vol. 65, 2016.

[5] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi *et al.*, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[6] P. Singh and G. S. Lehal, "Text-to-speech synthesis system for punjabi language," in *Proceedings of international conference on multidisciplinary information sciences and technologies, Merida, Spain*, 2006.

[7] P. LEHAL, "Punjabi text-to-speech synthesis system," in *24th International Conference on Computational Linguistics*, 2012.

[8] M. Rashid, Priya, and H. Singh, "Text to speech conversion in punjabi language using nourish forwarding algorithm," *International Journal of Information Technology*, vol. 14, no. 1, pp. 559–568, 2022.

[9] R. Kaur, R. Sharma, P. Kumar, and N. Meghanathan, "Building atext-to-speech system for punjabi language," *ACSIT, SIPM, FCST, CoNeCo, CMIT*, 2017.

[10] N. Perraudin, P. Balazs, and P. L. Søndergaard, "A fast griffin-lim algorithm," in *2013 IEEE workshop on applications of signal processing to audio and acoustics*. IEEE, 2013, pp. 1–4.

[11] 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. IEEE, 1993, pp. 125–128.

[12] T. Saeki, S. Maiti, S. Takamichi, S. Watanabe, and H. Saruwatari, "Speechbertscore: Reference-aware automatic evaluation of speech generation leveraging nlp evaluation metrics," *arXiv preprint arXiv:2401.16812*, 2024.

[13] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[14] M. F. Cohen, J. Mickunas, J. Miller, and W. D. Voiers, "Diagnostic rhyme test for the evaluation of communications systems," *The Journal of the Acoustical Society of America*, vol. 37, no. 6_Supplement, pp. 1206–1206, 1965.

[15] A. S. House, C. Williams, M. H. Hecker, and K. D. Kryter, "Psychoacoustic speech tests: A modified rhyme test," *The Journal of the Acoustical Society of America*, vol. 35, no. 11_Supplement, pp. 1899–1899, 1963.

[16] W. Habib, R. H. Basit, S. Hussain, and F. Adeeba, "Design of speech corpus for open domain urdu text to speech system using greedy algorithm," in *Conference on Language and Technology (CLT14)*, 2014.

[17] M. Chuhan and M. Habib, "Phonemic comparison of majhi and shahpuri- dialects of punjabi," *Hamdard Islamicus: quarterly journal of the Hamdard National Foundation, Pakistan*, 05 2020.

[18] H. Kabir, S. R. Shahid, A. M. Saleem, and S. Hussain, "Natural language processing for urdu tts system," in *International Multi Topic Conference, 2002. Abstracts. INMIC 2002.* IEEE, 2002, pp. 58–58.

[19] K.-U. Lahore, "Text processing for urdu tts system."

[20] S. Hussain, "To-sound conversion for urdu text-to-speech system," in *Proceedings of the workshop on computational approaches to Arabic script-based languages*, 2004, pp. 74–79.

[21] V. Van Heuven and R. Bezooijen, "Quality evaluation of synthesized speech." *, - (1995)*, 01 1995.

[22] S. Lemmetty, "Review of speech synthesis technology," 1999.

[23] M. A. Redford and R. L. Diehl, "The relative perceptual distinctiveness of initial and final consonants in cvc syllables," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1555–1565, 1999.

[24] C. Benoît, M. Grice, and V. Hazan, "The sus test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences," *Speech communication*, vol. 18, no. 4, pp. 381–392, 1996.

[25] Y.-Y. Chang, "Evaluation of tts systems in intelligibility and comprehension tasks," in *Proceedings of the 23rd Conference on Computational Linguistics and Speech Processing (ROCLING 2011)*, 2011, pp. 64–78.

[26] Y. Yasuda and T. Toda, "Analysis of mean opinion scores in subjective evaluation of synthetic speech based on tail probabilities," in *Proc. INTERSPEECH*, vol. 2023, 2023, pp. 5491–5495.