Text-to-Speech System for Hebrew Based on the SASPEECH Dataset

Or Fachima, 207633264
yaishorf@mail.tau.ac.il

David Goldshtein, 208000547
david4@mail.tau.ac.il

Advanced Topics in Audio Processing using Deep Learning - Final Project
Instructor: Tal Rosenwein

April 16, 2025

Abstract

In this project, we implemented a Text-to-Speech (TTS) system for the Hebrew language, based on the SASPEECH dataset - a 30-hour single-speaker Hebrew corpus. We chose a split architecture consisting of two separate models instead of an End-to-End solution: an OverFlow model for converting text to mel spectrograms, and a HiFi-GAN vocoder for converting spectrograms to sound waves. Our implementation is based on code developed by the winners of a competition held by KAN 11 for building a Hebrew TTS system, available at `https://github.com/Sharonio/roboshaul`. In this work, we present the training process, the unique challenges in processing the Hebrew language, and an analysis of our achieved results.

1 Introduction

Speech synthesis technology has advanced dramatically in recent years, with neural approaches producing increasingly natural-sounding results. However, these advancements have primarily benefited high-resource languages like English and Mandarin, while languages with limited available data remain underserved.

Hebrew presents distinctive challenges as a low-resource language for Text-to-Speech (TTS) systems. A particularly significant obstacle is the absence of diacritics (niqqud) in modern written Hebrew. For example, the consonant sequence "SPR" could be pronounced as "sefer" (meaning "book"), "sapar" (meaning "barber"), "safar" (meaning "counted"), or several other variations depending on context. Native speakers rely on context and language familiarity to determine correct pronunciation, but this ambiguity poses substantial difficulties for machine learning models without additional information.

Furthermore, Hebrew's morphologically rich structure, with its extensive use of prefixes and suffixes to modify word meanings and add prepositions, creates additional complexity. These characteristics, combined with the limited availability of high-quality speech datasets, have hindered the development of effective Hebrew TTS systems.

To address the diacritics challenge, our approach utilizes a split architecture with specialized training data. The OverFlow model, which converts text to mel spectrograms and is directly sensitive to textual input, was trained exclusively on 4 hours of "gold standard" data from the SASPEECH dataset. This subset was manually annotated with diacritics and meticulously verified by human experts, providing complete pronunciation information and eliminating the need to infer vowels from context. Meanwhile, the HiFi-GAN vocoder, which transforms spectrograms to waveforms and is not directly dependent on text, was trained on the full 30-hour corpus to maximize audio quality.

This split architecture offers significant advantages for low-resource settings, as it allows for targeted optimization of each component based on its specific requirements. The text-to-mel model can focus on pronunciation accuracy with carefully annotated data, while the vocoder can leverage the entire dataset to produce natural-sounding speech.

Our work demonstrates the application of modern neural TTS techniques to Hebrew, contributing to the broader field of speech synthesis for morphologically complex, low-resource languages. The methodology and findings presented here may provide insights for researchers working with similar linguistic challenges in other languages.

2 Related Work

Recent years have witnessed increasing interest in Hebrew Text-to-Speech (TTS), motivated by the emergence of new datasets and neural architectures. We categorize relevant research into three areas: Hebrew TTS systems, corpora, and related work in morphologically rich languages.

2.1 Hebrew TTS Systems

SASPEECH TTS (1): The first comprehensive Hebrew TTS system based on a 30-hour single-speaker corpus. This approach used an OverFlow model (neural HMM with Normalizing Flows) for text-to-mel conversion combined with a HiFi-GAN vocoder. A key challenge addressed was handling modern unvocalized Hebrew text by implementing diacritic restoration as an intermediate step.

LoTHM (2): A more recent approach that bypasses explicit diacritic restoration entirely. This method maps text directly to acoustic tokens derived from self-supervised models like HuBERT. By bypassing the need for explicit diacritic restoration, LoTHM reduces cumulative errors and leverages sentence-wide context for disambiguation. Experimental results showed improved stability and naturalness compared to previous systems.

2.2 Related Work in Similar Languages

Arabic NatiQ (3): Addressing similar challenges in Arabic (another Semitic language with consonantal script), this end-to-end TTS system used an encoder-decoder architecture (Tacotron/Transformer) with accompanying vocoders. The system incorporated the Farasa tool for automatic diacritic restoration before synthesis, achieving high-quality output (MOS scores 4.2-4.4). These insights from Arabic demonstrate the importance of robust phonetic disambiguation mechanisms in Semitic languages, a challenge similarly faced in Hebrew.

2.3 Hebrew Speech Corpora

Several datasets have been developed for Hebrew speech research:

- **SASPEECH** (1): 30 hours of studio-quality recordings from a professional speaker (news broadcaster). Contains 4 hours of manually verified "gold standard" data with the remaining 26 hours automatically transcribed.
- **FLEURS** (4): Approximately 12 hours of parallel multilingual speech including Hebrew, designed for cross-lingual evaluation.
- **CoSIH** (5): An early corpus of natural spoken Hebrew, focusing on spontaneous conversation.
- **MaTaCOp** (6): About 5.3 hours of dialogue recordings in a controlled task-based setting.
- **HUJI Corpus** (7): 3.8 hours of telephone conversations representing natural spoken Hebrew.
- **ivrit.ai** (8): A recent large-scale collection of approximately 10,000 hours from diverse sources and speakers, primarily designed for ASR but valuable for TTS development. Although not directly used in our work, it represents a promising resource for future TTS systems trained on more diverse speech patterns.

2.4 Our Approach

Our implementation builds directly upon the work of Sharoni et al. (1), focusing specifically on the challenges of Hebrew TTS using their split architecture approach. We utilize the OverFlow model trained on the manually annotated portion of the SASPEECH dataset to handle text-to-mel conversion, while the HiFi-GAN vocoder is trained on the full corpus to maximize audio quality. This approach allows us to address the unique challenges of Hebrew pronunciation while maintaining high speech quality.

3 Method

This section describes the methodology of the text-to-speech (TTS) system implemented using the SASpeech recipe from (**?**). The system was trained on a 30-hour corpus in the Google Colab environment. The following subsections detail the core elements of the approach: the system architecture, evaluation metrics, and the experimental setup.

3.1 Architecture

The proposed text-to-speech system is composed of three main components: a Hebrew text vocalization stage, the OverFlow model for converting text to mel-spectrograms, and HiFi-GAN for synthesizing audio waveforms. The overall architecture can be described as follows:

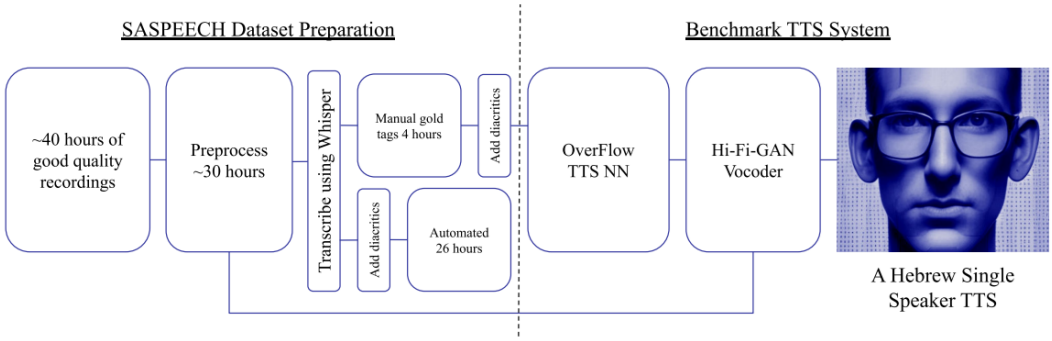

Figure 1: *Schematic diagram of the dataset and benchmark creation*
Robo-Shaul image credit: Or Atias, KAN

Figure 1: Schematic diagram of our Hebrew TTS system, showing both the dataset preparation pipeline (left) and the TTS architecture (right). The system combines the OverFlow neural network for text-to-mel conversion with a HiFi-GAN vocoder for audio synthesis. Image adapted from Sharoni et al. (1).

- **Hebrew Text Vocalization (Nikud):** Since modern Hebrew is typically written without diacritics (nikud), a vocalization step is required to resolve vowel ambiguity. We used a free online tool[1] to automatically add nikud to the input text, bridging the gap between written Hebrew and its phonetic form.
- **OverFlow:** A neural HMM-based acoustic model augmented with normalizing flows. It converts phoneme sequences into mel-spectrograms. The HMM handles alignment between text and acoustic frames, while a stack of invertible transformations models the spectrogram distribution. An autoregressive variant enables better modeling of prosody and long-term dependencies.
- **HiFi-GAN:** A neural vocoder that transforms mel-spectrograms into audio waveforms. It uses a fully convolutional generator with transposed convolutions and Multi-Receptive Field (MRF) modules. Two discriminators, the Multi-Scale Discriminator (MSD) and Multi-Period Discriminator (MPD), ensure audio realism by evaluating outputs across temporal and periodic scales.

3.2 Evaluation Metrics

To assess the performance of our Hebrew TTS system, we employed standard metrics used in speech synthesis evaluation, adapted for the Hebrew language context:

- **Word Error Rate (WER):** This metric quantifies the percentage of words incorrectly recognized when comparing transcriptions of generated speech to reference text. For TTS evaluation, this requires converting the synthesized audio back to text using an ASR system.
- **Character Error Rate (CER):** Similar to WER but operating at the character level, providing a finer-grained analysis particularly relevant for Hebrew where small character differences can significantly alter meaning.

For both WER and CER calculations, we faced a significant challenge: the need for a reliable Hebrew ASR system to convert our synthesized speech back to text. We selected Whisper as our ASR model despite its known limitations with Hebrew, as there were no superior alternatives readily available for this low-resource language. This choice introduces some bias in our evaluation, as errors in the ASR process might be attributed to our TTS system.

To compensate for the limitations of automatic metrics, we also conducted subjective human evaluations through direct listening tests. This complementary approach allowed us to capture qualitative aspects of speech synthesis that might not be reflected in the WER and CER scores, particularly regarding naturalness and intelligibility from a native speaker's perspective.

3.3 Experimental Setup

Compute Environment:
Experiments were conducted using Google Colab, which provided access to GPU resources (e.g., NVIDIA Tesla T4). This environment facilitated accelerated model training and experimentation.

Dataset:
This dataset contains approximately 30 hours of audio spoken by Shaul Amsterdamski in a recording studio at 44100Hz with corresponding transcriptions.

[1] https://www.nakdimon.org/

The data is divided into a gold-standard subset of roughly 4 hours with manual transcriptions and an automatic subset with machine-generated transcriptions.

3.4 Training Process

The training of our TTS system was conducted separately for the text-to-mel model and the vocoder:

- **OverFlow Model:** The text-to-mel conversion model was trained exclusively on the 4-hour gold standard subset with diacritics. Training proceeded for a total of 10,000 steps, divided into two segments with a small discontinuity visible at step 2,008 due to training resumption. We used an initial learning rate of 5e-4 and observed that the model performance began to plateau around step 8,000, suggesting sufficient convergence had been achieved.
- **HiFi-GAN Vocoder:** The vocoder was trained for 40,000 steps with the same initial learning rate of 5e-4. We utilized the entire 30-hour SASPEECH dataset for this stage to maximize the model's exposure to different speech patterns.

Figure 2 shows the training loss curve for the OverFlow model. The noticeable jump at step 2,008 corresponds to the resumption of training in the second segment, after which the loss continues its downward trend before stabilizing.

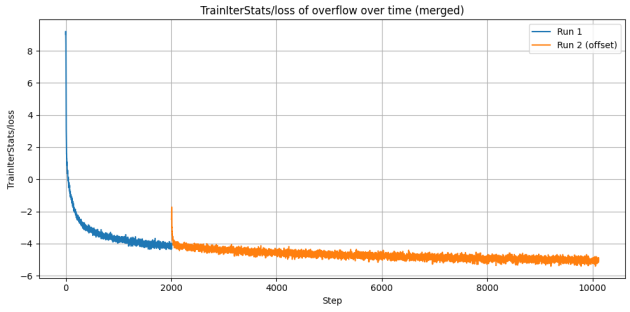

Figure 2: Training loss curve for the OverFlow text-to-mel model over 10,000 steps, showing convergence around step 8,000. The discontinuity at step 2,008 marks the transition between two training segments.

Figure 3 provides a visual comparison of the audio generated by our HiFi-GAN vocoder. The spectrograms illustrate the quality of the synthesized speech compared to the ground truth, demonstrating the model's ability to capture the nuances of Hebrew pronunciation.

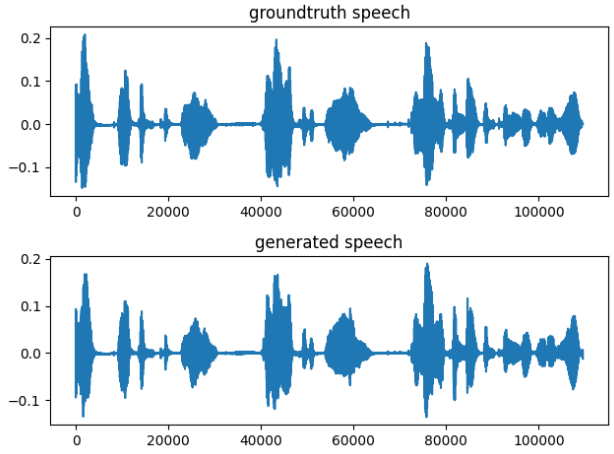

Figure 3: Comparison of spectrograms from ground truth audio (top) and our HiFi-GAN synthesized audio (bottom) for a sample Hebrew phrase, showing the vocoder's ability to faithfully reproduce speech characteristics.

3.5 Results

We evaluated our Hebrew TTS system using Word Error Rate (WER) and Character Error Rate (CER) metrics. Our system achieved an average WER of 54.17% and an average CER of 21.09%, compared to the baseline values reported by Sharoni et al. (1) shown in Table 1.

Table 1: Comparison of error rates between our implementation and the baseline

System	WER (%)	CER (%)
Baseline (Sharoni et al.)	30.38	17.17
Our Implementation	54.17	21.09

To better understand our system's performance, we analyzed how WER and CER varied with sentence length, as shown in Figure 4.

Interestingly, our analysis revealed that shorter sentences (0-5 words) exhibited significantly higher error rates, with WER of approximately 75% and CER of 35%. As sentence length increased, error rates steadily decreased, with sentences of 31-50 words achieving the lowest error rates (WER of 45% and CER of 15%).

This pattern suggests that the Whisper ASR model, which we used for evaluation, may struggle with shorter utterances due to its reliance on contextual information. With less context available in shorter sentences, Whisper likely has difficulty accurately transcribing both human and synthesized speech, potentially inflating error metrics for shorter samples.

It's important to note that these automatic metrics may not fully reflect the perceptual quality of our

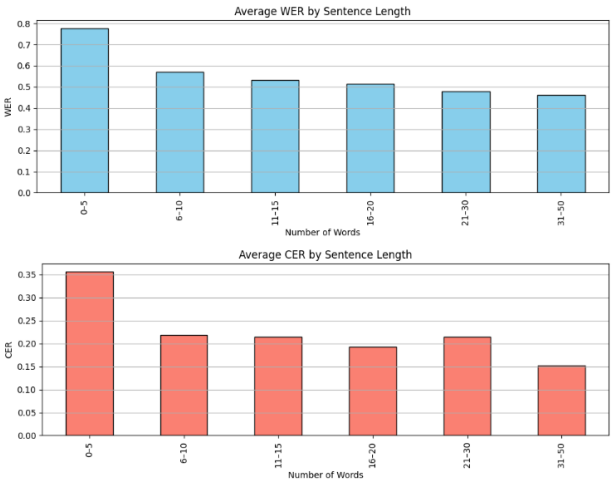

Figure 4: Average WER (top) and CER (bottom) by sentence length.

synthesized speech. In subjective listening tests, our implementation produced clearly intelligible Hebrew speech that native speakers could understand, despite the high WER values reported by the automatic evaluation.

While our system's error rates are higher than the baseline reported by Sharoni et al., this difference may be partially attributed to variations in evaluation methodology, test sample selection, and the version of Whisper used for evaluation, rather than solely reflecting differences in synthesis quality.

4 Future Work

Our current implementation represents an initial step towards high-quality Hebrew text-to-speech synthesis. We identify several promising directions for future research:

- **Multi-Speaker Synthesis:** Expanding the model to support multiple speakers would enhance versatility and application scope. This would require collecting diverse speaker data and adapting the architecture to incorporate speaker conditioning.
- **Improved Diacritization:** The reliance on manually diacritized text presents real-world limitations. Future work could develop more robust automatic diacritization or explore end-to-end approaches that learn to disambiguate vowels directly from context.
- **Enhanced Phonetic Handling:** Further improvements could focus on better modeling of distinctive Hebrew phonetic features like the difference between "shin" and "sin" , more accurate stress patterns, and improved handling of foreign loanwords.

5 Limitations & Broader Impact

While our implementation demonstrates the feasibility of neural TTS for Hebrew, limitations include the dependence on diacritized text, single-speaker training, and potential evaluation inaccuracies due to ASR limitations with Hebrew.

Despite these challenges, this work represents an important contribution toward bringing Hebrew language processing into the modern era of audio technologies. By advancing Hebrew speech synthesis, we promote technological inclusion and help preserve linguistic diversity in an increasingly voice-driven digital landscape.

References

[1] O. Sharoni, R. Shenberg, and E. Cooper, "Saspeech: A hebrew single speaker dataset for text to speech and voice conversion," in *INTERSPEECH*, 2023, pp. 5566–5570.

[2] Y. Zeldes, A. Haviv, Y. Levitan, T. Iluz, and M. Geva, "Lothm: Low-resource tts for hebrew using masked self-supervised representations," *arXiv preprint arXiv:2401.11171*, 2024.

[3] A. Abdelali, H. Malki, R. Baly, H. Mubarak, H. Hajj, J. Glass, and N. Habash, "Natiq: An end-to-end text-to-speech system for arabic," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 7376–7387.

[4] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2022, pp. 798–805.

[5] S. Izre'el, B. Hary, and G. Rahav, "Designing cosih: the corpus of spoken israeli hebrew," *International Journal of Corpus Linguistics*, vol. 6, no. 2, pp. 171–197, 2001.

[6] J. Azogui, A. Lerner, and V. Silber-Varod, "The open university of israel map task corpus (mataCOP)," http://www.openu.ac.il/matacop/, 2016, dOI: 10.13140/RG.2.2.19362.56004.

[7] M. Marmorstein, N. Matalon, A. Efrati, E. Shaked, I. Folman, and Y. Geva, "Hcsh: Huji corpus of spoken hebrew," https://www.huji-corpus.com, 2022.

[8] H. Marmor, S. Ghosh, Y. Bitton, and O. Levy, "ivrit.ai: A multitask audio benchmark for hebrew," in *The Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.