INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

SASPEECH: A Hebrew Single Speaker Dataset for Text To Speech and Voice Conversion

Orian Sharoni[1*], *Roee Shenberg*[1*], *Erica Cooper*[2]

[*] Co-first authors
[1]Up·AI, Israel
[2]National Institute of Informatics, Japan
orian.sharoni@upai.dev, roee.shenberg@upai.dev, ecooper@nii.ac.jp

Abstract

We present SASPEECH, a 30-hour single speaker Hebrew corpus accompanied by a text-to-speech (TTS) benchmark. Our TTS benchmark was developed with other low resource languages in mind, allowing it to be adapted and potentially generalized. For the proposed method to work, one must have several hours of recordings and transcripts or have their language included in the Whisper model. SASPEECH is the first large-scale high-quality open dataset of its kind. Thus, it allows a discussion of challenges Hebrew presents when incorporated into generative models. For instance: bridging the gap between modern Hebrew lettering which lacks diacritics and correct pronunciation. We also tackle prominent issues shared by low resource languages and examine how to evaluate output quality without a benchmark. We believe our work will facilitate future generative Hebrew tools and low resource language research. The corpus is publicly accessible at https://www.openslr.org/134.

Index Terms: Hebrew, speech corpus, low resource languages, TTS, text to speech, single speaker, voice conversion.

1. Introduction

We built a single speaker dataset in Hebrew, and added a TTS benchmark model to it (Figure 1). We did it because existing Hebrew open source datasets were small and did not allow the creation of a Hebrew TTS system, nor did they allow testing single speaker voice synthesis.

The corpus was designed to follow the scale and structure of the widely used English corpus LJSpeech [1] and it was developed in collaboration with KAN, the Israeli Public Broadcasting Corporation. In a notable contribution to Hebrew speakers, journalist Shaul Amsterdamski generously offered his voice to help advance scientific knowledge, while also documenting the corpus creation process on public radio [2]. The corpus contains the following data:

- Speech recordings: 30 hours of Amsterdamski's voice recordings for the podcast "Hayot Kis". The recordings are of journalistic spoken Hebrew, read in a semi-casual intonation from written text.
- 26 hours of automated sentence/utterance level transcriptions and alignments created using Whisper [3].
- 4 hours of "gold tags": a randomly selected subset of recording sessions was manually corrected for accuracy of transcriptions and alignments.
- Automated diacritics, using Nakdimon [4]. Note that the diacritics were not manually verified and contain mistakes.

The dataset is hosted online and is available for non-commercial use. Audio examples are available at https://shenberg.github.io/saspeech_site/

1.1. The need for a Hebrew single-speaker corpus

Hebrew, a low-resource language spoken by 9 million people worldwide [5], presents unique challenges that constrain research and product development in speech technology. The lack of open-source resources for spoken Hebrew necessitates generating a corpus for each machine learning model or adapting models trained on other languages to suit Hebrew. These challenges include:

- Hebrew uses non-Latin letters, which sets it apart from many languages.
- While ancient Hebrew has diacritics, modern Hebrew rarely uses them. Readers are expected to assume most phonemes based on familiarity with the language, making it challenging for automatic speech recognition (ASR) and text-to-speech (TTS) systems to accurately learn the connection between audio and text (i.e. the word "dog" *kelev* would be written as "dg" *klv*). Additionally, as the norm in modern Hebrew is to write without diacritics, Whisper and other ASR systems do not output diacritics in their transcripts.
- Hebrew is a morphologically rich language, with common use of prefixes and suffixes to modify words' meanings and to add prepositions. This characteristic makes it challenging to understand the meaning of word error rate results in terms of correlation with generated speech intelligibility.
- Often, different Hebrew words sound phonetically the same but have different spellings, which can pose challenges when evaluating TTS models.

Consequently, there is a paucity of research demonstrating new machine learning capabilities on Hebrew alongside a scarcity of speech-related ML products that support Hebrew. This situation emphasizes the need for a Hebrew single-speaker corpus, which would be a valuable resource for further research and product development.

1.1.1. Speech Datasets in Hebrew

Hebrew speech datasets are rare, and only a few offer high-quality recordings in quiet conditions. One such dataset is the FLEURS corpus, which contains data from 102 languages, including approximately 12 hours of Hebrew sentences read from Wikipedia by different speakers [6]. Another dataset is the MaTaCOp, which consists of 32 several minutes dialogues between pairs of speakers [7]. Other noteworthy datasets with less than optimal recording conditions include the HUJI Corpus of Spoken Hebrew, with approximately 4 hours of phone-call recordings [8], and the CoSIH Corpus of Spoken Israeli Hebrew, which comprises of recordings gathered from 53 vol-

10.21437/Interspeech.2023-430

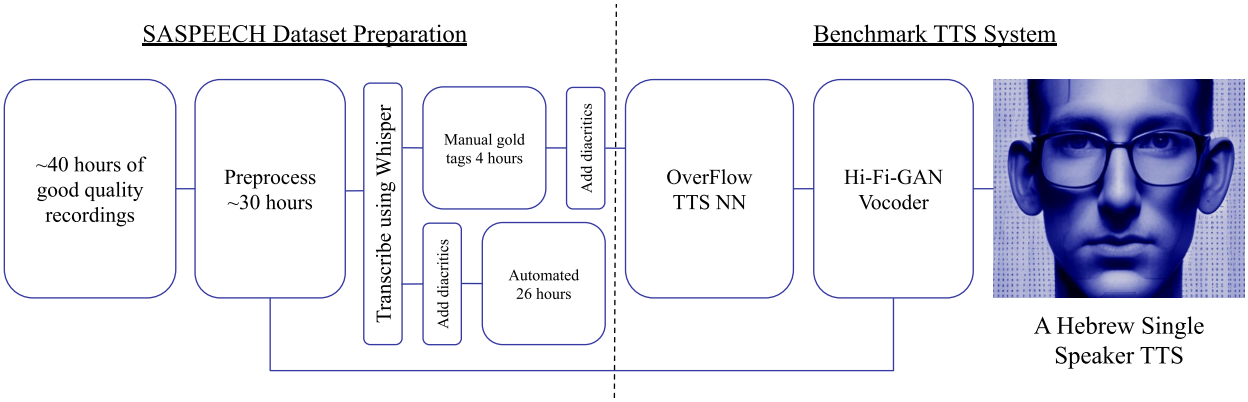


Figure 1: *Schematic diagram of the dataset and benchmark creation*
Robo-Shaul image credit: Or Atias, KAN

unteers of different backgrounds as they went about their daily lives [9].

1.2. Studies on TTS for other low resource languages

Text-to-Speech systems seek to convert sequences of characters to waveforms of the spoken text. Data-driven approaches based on learning the text to audio mapping using neural networks have reached human-level quality [10]; however, common approaches expect dozens of hours of transcribed speech, spoken by professionals, with consistent environments, speaking rates, etc, and carefully edited for quality.

Previous work on low-resource TTS has mostly focused on pre-training neural networks on resource-rich languages, such as Mandarin or English, and using the trained weights to initialize networks for the target language. LRSpeech [11] epitomizes this approach, jointly training ASR and TTS systems this way, while using each sub-system to generate synthetic data for the other system. The resulting system functions well in objective and subjective measures, but it has a three-stage training procedure totalling 9 days of training on 4x NVIDIA V100 GPUs, limiting the applicability of the system for smaller institutions.

Pine et al. [12] recently investigated using small amounts of parallel English text and audio in order to directly train FastSpeech 2 [13] (modified following Luo et al. [14]), using Montreal Forced Aligner [15] trained on the same dataset for phoneme alignments. This approach is then adapted to three indigenous languages in Canada, with training data ranging from 25 minutes to three hours. The limitation of this work is the necessity of a grapheme-to-phoneme lexicon.

SPEAR-TTS [16] trained a model to translate text tokens into the discrete semantic audio tokens learned by a self-supervised semantic audio representation network. These semantic tokens are translated then to discrete acoustic tokens learned by a neural audio codec. Only the conversion from text to semantic tokens requires parallel data, and 15 minutes suffice, but audio token learning requires many hours of data (60k hours in this case). While collection of unlabelled audio is cheap, the size of the dataset and compute required to train it keep it out of reach for smaller institutions.

2. Corpus Curation

The initial release of SASPEECH contains roughly 30 hours of modern Hebrew speech from 147 different recording sessions from a single speaker, Shaul Amsterdamski.

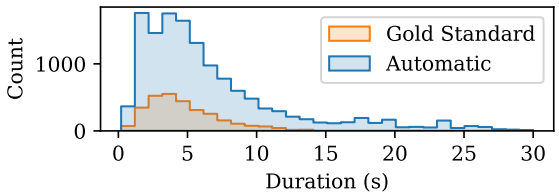


Figure 2: *Histogram of utterance durations*

The speech is from the unedited recordings of the podcast "Hayot Kis". The material is colloquial journalistic spoken Hebrew, mostly read from a script, with less than 10 recording sessions of spontaneous conversations. Sometimes there are lapses of the spoken Hebrew, remarks from the speaker to himself and his editor, and several takes with different intonations for a given sentence (and the occasional slang curse word when he struggles to achieve the desired narration in one take).

Detailed statistics of the corpus are listed in Table 1, and the distribution of utterance lengths is seen in Figure 2.

2.1. Corpus Recording Details

The original recordings were made in the years 2017-2023. The speaker's age ranged between 36 and 42.

Recordings were made in two quiet rooms, the speaker's home-studio and the IPBC studios.

In the IPBC studio, the speech was recorded on an Electro-Voice RE320 dynamic microphone and captured using Audacity. In the home-studio, the speech was recorded on a Sennheiser MKE 600 shotgun microphone, captured by a TASCAM DR-40X audio recorder and transferred to a computer using the REAPER DAW.

All sessions except two were captured at 44.1KHz, with two recordings captured at 96KHz and resampled to 44.1KHz using SoX[1].

Twelve sessions were encoded as MP3 files, ten at a bit-rate of 320Kbps, one at 256Kbps and one at 128Kbps.

[1]SoX 14.4.2, VHQ resampling algorithm with linear phase and a -1dB gain applied to counteract clipping, the command used was `sox input.wav output.wav gain -1 rate -v 44100`

2.2. Corpus Annotations

The corpus provides utterance-level machine-generated transcriptions of the speech with additional machine-generated diacritics, and a subset of the dataset was manually corrected by human annotators, as detailed below.

2.2.1. *Whisper Transcriptions*

Machine-generated transcriptions were created using the Whisper[2] pre-trained ASR model, which also provides segmentation of speech into utterances, as is required during TTS training.

The preprocessing performed consisted of dividing the recording sessions into separate files by silences (a second where the RMS energy is below -55dB is considered silence). This was done in order to combat "hallucinations" where Whisper generates non-existent transcriptions during long silences.

A common issue with Whisper transcriptions is errors in its segmentation into utterances - a single utterance is sometimes split into two or more segments, start and end times of segments often cut off the first and last words spoken or contain some of the preceding and following words.

Additionally, Whisper attempts to correct speech lapses, fillers and repetitions in its transcription, which is a desirable property for a speech-to-text system, but not ideal for a dataset intended for text-to-speech purposes.

2.2.2. *Manual Annotation*

A subset of 22 recordings containing ≈ 4 hours of speech was manually corrected by human annotators in the following ways:

- Transcriptions were corrected for mistakes
- Lapses, fillers and repetitions were transcribed
- Numbers below ten were spelled out
- Acronyms were marked: in Hebrew, the marking consists of a double-quote between the last two characters of an acronym
- Occasional English words were transliterated to Hebrew
- Audio segment bounds were corrected
- Adjacent segments from the same utterance were joined
- Utterances longer than 14 seconds were split based on speaker's natural pauses to shorten input length for TTS system

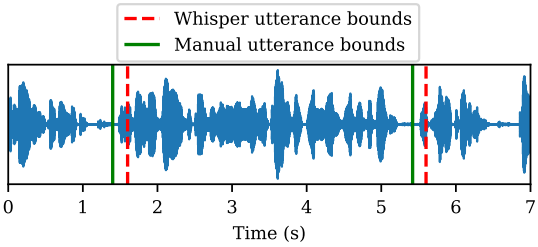
The software used in order to perform the corrections was the open-source Aegisub subtitle typesetting software[3].

The annotators were Authors 1 & 2, and the speaker, all native Hebrew speakers with editing and transcribing experience. Figure 3 shows an example utterance and manual corrections.

2.2.3. *Diacritic Restoration*

Diacritics were restored using Nakdimon, a character-level LSTM predicting missing diacritics. This system is competitive in performance with proprietary solutions such as Nakdan [17] and has its code, data and trained models available online[4].

Our initial TTS system received as input undiacritized Hebrew text, and produced output with incorrect vowels, as expected given that this information is missing from the text. One avenue of improvement would be to use an existing grapheme to phoneme converter for undotted Hebrew as a front-end to the TTS system, however, while solutions exist, such as eSpeak NG's[5] Hebrew front-end and Hasegawa-Johnson et al. [18], they all suffer from high error rates ($> 10\%$ Phoneme error rate) due to the complexities of Hebrew and lack of semantic context used by readers to disambiguate between heteronyms, while Nakdimon reports 96.37% per-character accuracy.



Whisper transcript: בשבוע שעבר ביקר בישראל מזגל ה-OECD החדש,
bashavua she'avar biker be'israel mazgal ha oh ee see dee haxadash

Corrected transcript: בשבוע שעבר ביקר בישראל מזכ"ל האו-אי-סי-די החדש,
bashavua she'avar biker be'israel mazkal ha oh ee see dee haxadash

An example utterance introducing the secretary general of the OECD. Manual corrections are correcting the timing of the utterance, spelling out OECD in hebrew, correcting a mistake due to partial assimilation, "mazgal"→"mazkal" ("mazkal" is valid hebrew, "mazgal" is not, but is closer to what is pronounced), and adding an acronym signifier.

Figure 3: *Example of Manual Correction*

Table 1: *Statistics for the entire dataset and its subsets*

	Entire Dataset	Gold Standard	Automated
#Recordings	147	22	125
#Utterances	16,522	2,986	13,536
Total Duration	29:39:18	3:55:50	25:43:27
Min Duration (s)	0.2	0.32	0.2
Mean Duration (s)	6.5 ± 5.2	4.7 ± 2.6	6.8 ± 5.5
Max Duration (s)	30.0	14.0	30.0
0.5% F0 (Hz)	59.2	73.8	58.2
Mean F0 (Hz)	160 ± 47	160 ± 44	160 ± 48
99.5% F0 (Hz)	333.2	323.7	335.1

3. A Hebrew Single-Speaker Text To Speech Benchmark

3.1. Method

3.1.1. *Benchmark TTS System*

Our benchmark TTS system is based on OverFlow [19], using the HiFi-GAN [20] vocoder, implemented using Coqui TTS[6].

OverFlow follows Tacotron 2 [21], in which a sequence-to-sequence network maps character inputs to mel spectograms, followed by a neural vocoder predicting waveforms from spectograms. Following Mehta et al. [22], OverFlow uses a variant of a neural transducer [23] - an autoregressive neural Hidden Markov Model (NHMM). In NHMM, speech is modelled probabilistically and NHMM can be optimized by maximizing output likelihood, while enforcing monotonic input-output alignment. NHMM requires less data and fewer training steps than Tacotron 2 to produce intelligible speech, and does not require

[2]OpenAI Whisper large-v2 model

[3]Precise version at `https://github.com/wangqr/Aegisub/releases/tag/v3.3.3`, has the option for right-to-left text support, enabled by de-selecting "Use styled edit box" in the software's interface preferences

[4]`https://github.com/elazarg/nakdimon/`

[5]eSpeak NG text-to-speech software - `https://github.com/espeak-ng/espeak-ng`

[6]`https://github.com/coqui-ai/TTS`

external input-output alignments as in FastSpeech 2 [13].

OverFlow introduces a normalizing flow between the NHMM and the output spectrograms, allowing the system to model a greater range of output distributions. This change improves measures of speech quality (ASR, WER, MOS), leading OverFlow to produce speech of comparable quality to Tacotron 2, while retaining the advantage in low-compute scenarios.

While the original OverFlow system has a text to phoneme converter as a front-end, no reliable front-end exists for Hebrew. Therefore the system was trained to directly map between character sequences and mel spectrograms.

3.1.2. Training Procedure

The OverFlow model was trained on the SASPEECH corpus gold-standard subset, resampled to 22,050Hz, for 7750 steps with learning rate of 10^{-4} and batch size of 10, on a single NVIDIA TITAN RTX GPU.

HiFi-GAN was fine-tuned from a model trained on LJSpeech for 250,000 steps, by training for an additional 250,000 steps using an initial learning rate of 10^{-5}, batch size 32, on the entire SASPEECH dataset resampled to 22,050Hz.

3.2. Evaluation Methodology

Evaluating generative acoustic models typically involves comparing them to existing works to gauge their performance. However, in the absence of prior works, we present objective measures in this paper and provide audio examples on our website[7] to showcase the model. Although subjective measures are preferable for evaluating speech synthesis systems, the limited number of Hebrew speakers and inadequate infrastructure hindered us from conducting subjective experiments in this study. Therefore, we opted for objective measures to assess the quality of our proposed model.

3.2.1. Word error rate using Whisper as ground truth

We assessed the model's quality using Whisper with a method adapted to Hebrew from the OverFlow paper. First, we tested Whisper's consistency in transcribing Hebrew on our recordings, followed by calculating word error rate between manual tags and Whisper's outputs, and finally, compared a Whisper transcriptions of our TTS system to manual transcriptions.

To test Whisper's reliability, 162 utterances were randomly sampled from the gold standard subset and transcribed 50 times using Whisper. Only one utterance failed to produce identical transcripts across all 50 repetitions. We concluded that Whisper consistently produces Hebrew transcriptions with minimal variation in quality.

We then continued to calculate word and character error rate between 100 utterances sampled randomly and transcribed manually from the automatically generated transcripts. After discarding 3 utterances that turned out to be nonverbal short clicks, we ended up with 97 utterances. Results show an error rate of 12% between the manual transcripts and Whisper, and 30% error rate between our benchmark model and the manual transcripts (Table 2). It is reasonable to assume that the difference between Whisper and the manual transcripts can account for some of the error rate of the model output.

It is also worth noting that the morphological richness of Hebrew and other challenges discussed previously might have contributed to the word error rate. Additionally, as Hebrew has many homophones, our character error rate (CER) values also reflect Whisper's struggles with context-based transcription.

Table 2: *Word and Character Error Rate*

Transcripts	WER	CER
Manual vs. Whisper	12.1%	8.03%
TTS output vs. Manual	30.38%	17.17%

3.2.2. MCD score

The mean mel cepstral distortion value of the 97 generated model outputs and original wav recordings was 8.42 with a standard deviation of 1.33.

4. Discussion and Conclusion

In this paper, we have presented SASPEECH, a 30-hour single speaker Hebrew corpus, and a TTS system based on it, which are the best of their open source kind. Our study illustrates the utility of SASPEECH for TTS, as well as the challenges of integrating Hebrew into generative models.

We are releasing SASPEECH along with a key baseline approach and objective quality measures that will allow future model comparisons.

In creating our corpus, we followed the footprint of LJSpeech, the academic single speaker standard, so that progress in English neural models would more likely translate into Hebrew neural models. Despite the resemblance, SASPEECH was not planned for the TTS mission: it comprises of organic unedited raw materials of podcast journalism. Which means it is not as clean and accurate as LJSpeech who is based on edited materials. Nevertheless it allowed us to use it for building a unique pioneering open source model.

Our work also highlighted the ability to use OverFlow with a relatively small amount of data, emphasizing it as an exceptional architecture for low resource languages. However, we note that the gap between state of the art English TTS and our Hebrew benchmark is prominent. This may be alleviated by improving Hebrew ASR models, as their quality serves as an upper bound to generative models.

Further research into Hebrew-specific model pitfalls and the development of phonetically balanced test sets will improve model quality testing methods in Hebrew. In the same vein, subjective listening test methods optimized for Hebrew can also help establish more reliable metrics for evaluating TTS systems.

Despite the challenges, we believe that the work presented in this paper opens a new chapter in Hebrew machine learning research, providing an invaluable resource for future studies.

5. Acknowledgements

We extend our gratitude to Shaul Amsterdamski for generously providing his voice recordings. We are also thankful to KAN and the Hayot Kis podcast team for their unwavering support.

Uri Eliabayev, consultant and founder of MDLI community, facilitated the connection between all parties involved.

We are also grateful to Marc Brockschmidt for his fantastic ideas and proofreading and to Andre Loose of SmartML, LLC, and HarmonAI for the donated compute power, which contributed to the success of this project.

EC was supported by MEXT KAKENHI grant 21K11951.

[7]Model audio examples can be fount at: `https://shenberg.github.io/saspeech_site/#benchmark-tts-system-examples`

6. References

[1] K. Ito, "The LJ Speech Dataset," https://keithito.com/LJ-Speech-Dataset/, 2016.

[2] S. Amsterdamski, "Shaul turns into a robot – part 1," 2023. [Online]. Available: https://www.roboshaul.com/

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[4] E. Gershuni and Y. Pinter, "Restoring Hebrew Diacritics Without a Dictionary," in *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1010–1018. [Online]. Available: https://aclanthology.org/2022.findings-naacl.75

[5] (2022) Ethnologue: Languages of the world. SIL International. Dallas, Texas. [Online]. Available: http://www.ethnologue.com

[6] 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.

[7] J. Azogui, A. Lerner, and V. Silber-Varod, "The Open University of Israel Map Task Corpus (MaTaCOp)," http://www.openu.ac.il/matacop/, 2015, dOI: 10.13140/RG.2.2.19362.56004.

[8] 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.

[9] 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.

[10] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He, F. Soong, T. Qin, S. Zhao, and T.-Y. Liu, "NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality," 2022. [Online]. Available: https://arxiv.org/abs/2205.04421

[11] J. Xu, X. Tan, Y. Ren, T. Qin, J. Li, S. Zhao, and T.-Y. Liu, "LRSpeech: Extremely Low-Resource Speech Synthesis and Recognition," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 2802–2812. [Online]. Available: https://doi.org/10.1145/3394486.3403331

[12] A. Pine, D. Wells, N. Brinklow, P. Littell, and K. Richmond, "Requirements and motivations of low-resource speech synthesis for language revitalization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 7346–7359. [Online]. Available: https://aclanthology.org/2022.acl-long.507

[13] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=piLPYqxtWuA

[14] R. Luo, X. Tan, R. Wang, T. Qin, J. Li, S. Zhao, E. Chen, and T.-Y. Liu, "Lightspeech: Lightweight and Fast Text to Speech with Neural Architecture Search," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5699–5703.

[15] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[16] E. Kharitonov, D. Vincent, Z. Borsos, R. Marinier, S. Girgin, O. Pietquin, M. Sharifi, M. Tagliasacchi, and N. Zeghidour, "Speak, Read and Prompt: High-Fidelity Text-to-Speech with Minimal Supervision," 2023. [Online]. Available: https://arxiv.org/abs/2302.03540

[17] A. Shmidman, S. Shmidman, M. Koppel, and Y. Goldberg, "Nakdan: Professional Hebrew diacritizer," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, Jul. 2020, pp. 197–203. [Online]. Available: https://aclanthology.org/2020.acl-demos.23

[18] M. Hasegawa-Johnson, L. Rolston, C. Goudeseune, G. Levow, and K. Kirchhoff, "Grapheme-to-phoneme transduction for cross-language asr," in *Statistical Language and Speech Processing - 8th International Conference, SLSP 2020, Proceedings*, ser. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), L. Espinosa-Anke, I. Spasic, and C. Martín-Vide, Eds. Germany: Springer, 2020, pp. 3–19.

[19] S. Mehta, A. Kirkland, H. Lameris, J. Beskow, E. Székely, and G. E. Henter, "Overflow: Putting flows on top of neural transducers for better tts," 2022. [Online]. Available: https://arxiv.org/abs/2211.06892

[20] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 17 022–17 033. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/c5d736809766d46260d816d8dbc9eb44-Paper.pdf

[21] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[22] S. Mehta, E. Székely, J. Beskow, and G. E. Henter, "Neural HMMs are all you need (for high-quality attention-free TTS)," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7457–7461.

[23] A. Graves, "Sequence transduction with recurrent neural networks," in *International Conference of Machine Learning (ICML) Workshop on Representation Learning*, 2012. [Online]. Available: https://arxiv.org/pdf/1211.3711.pdf