



Building an Accurate Open-Source Hebrew ASR System through Crowdsourcing

Yanir Marmor^{1,3}, Yair Lifshitz³, Yoad Snopir³, Kinneret Misgav^{2,3}

¹Mathematics and Computer Science Faculty, Weizmann Institute of Science, Israel

²Hadassah Medical Organization, Israel

³ivrit.ai, Israel

yanir.marmor@weizmann.ac.il, yair@ivrit.ai, yoad@ivrit.ai, mkinneret@hadassah.org.il

Abstract

Automatic Speech Recognition (ASR) for Hebrew faces significant challenges due to limited resources and rich morphology. While recent advances have improved high-resource languages ASR, Hebrew still lacks robust open-source solutions. Through crowdsourcing efforts, we created a dataset of 314 hours of transcribed speech, which we used to train a new Hebrew ASR model based on the Whisper architecture. Our model demonstrates up to 29% reduction in error rates compared to existing Whisper solutions, particularly excelling in producing verbatim transcriptions. Additionally, we introduce a new evaluation dataset designed specifically for Hebrew ASR assessment. By making both the model and methodology freely available, we provide a framework that can be adapted for developing ASR systems in other under-resourced languages. This work represents a step toward making speech technology more accessible in different languages.

Index Terms: speech recognition, Hebrew ASR, crowdsourcing, under-resourced languages, open-source models, Whisper architecture

1. Introduction

Automatic Speech Recognition (ASR) has become an important technology in our digital world, enabling essential applications, such as virtual assistants and real-time transcription, allowing accessibility tools for the hearing impaired and automated customer service systems. Traditional ASR systems used algorithmic approaches to tackle speech recognition challenges but often struggled with limited generalization in different languages, which significantly affected their real-world effectiveness [1, 2, 3, 4]. In Hebrew, early ASR systems faced unique challenges due to the language’s distinctive characteristics, including rich morphology and the absence of written vowels in most texts. Examples of traditional tools in Hebrew include the DARPA-funded BBN Hebrew recognizer from the 1990s [5], and the subsequent HMM-based systems [6]. Recent advances in deep learning and neural architectures have made advancements in Hebrew ASR in the last decade [7]. The emergence of foundation model-based ASR systems, such as Whisper [8] and VALL-E [9], has revolutionized the field of speech recognition by offering open-source multilingual capabilities. The OpenAI Whisper system [8] leverages an extensive corpus of subtitle data for its training process. The results on applications in high-resource languages, such as English, Spanish, and Mandarin, are accurate and thus useful: the system demonstrates a nuanced understanding of written language conventions. Whisper is powerful in its ability to produce high-quality transcriptions that incorporate proper capitalization and punctuation, effectively bridging the gap between spoken and written language

forms.

However, Whisper’s effectiveness encounters certain constraints when dealing with other languages [10, 11], including Hebrew. In the next section, we will elaborate on the specific challenges in using Whisper for Hebrew speech.

2. Current Challenges in Hebrew ASR

The first challenge in creating efficient ASR in Hebrew is the limited amount of available speech data and linguistic resources (e.g., pronunciation dictionaries and annotated datasets), since Hebrew is spoken daily by less than 10 million people [12]. The other challenge is linguistic. Hebrew has rich morphology, making the words highly variable due to its root-and-pattern system and extensive affixation, where a single root can generate numerous word forms through the addition of prefixes, suffixes, and internal vowel patterns, which significantly increases the vocabulary size and makes it more challenging for ASR systems to handle out-of-vocabulary words and recognize all possible word forms accurately. For example, consider the base word *mšmrt* (*mishmeret*, “shift”), which can generate numerous variations through the systematic addition of prefixes, suffixes, and internal vowel patterns. Simple inflectional morphology produces forms such as *mšmrwt* (*mishmarot*, “shifts”) for plural and possessive forms like *mšmrty* (*mishmarti*, “my shift”), while the addition of prefixes creates variations such as *hmšmrt* (*hamishmeret*, “the shift”) and *lmšmrt* (*lemishmeret*, “to/for a shift”). The complexity compounds when multiple affixes combine, generating forms like *wlmšmrwtyhm* (*ulemishmeroteihem*, “and to their shifts”), *šbmšmrk* (*shebeshmartecha*, “that is in your shift”) which incorporates multiple grammatical elements into a single word. These differences from Latin languages are not problematic, but place challenges when trying to use the same tools and measurements in languages like Hebrew. We suggest enhancing existing benchmarking datasets to include comprehensive coverage of morphological variations, ensuring that both train and test sets reflect the full range of prefix-suffix combinations and vowel patterns that occur in natural Hebrew speech.

Recent work has made advancements in Hebrew ASR development due to the challenges mentioned above. Turetzky et al. [13] introduced HebDB, a substantial weakly supervised dataset comprising approximately 2,500 hours of natural Hebrew speech across diverse speakers and topics. Their work included both self-supervised and fully supervised baseline ASR systems. While HebDB represents a step in Hebrew ASR research by providing a large-scale training dataset and baseline models, we identify two critical areas needed to further their work. First, we introduce the ivrit.ai crowd-transcribe dataset of manually corrected transcriptions comprising hundreds of hours

of speech, addressing the need for high-quality evaluation data that can serve as a reliable benchmark for testing ASR systems. We adopted a crowdsourcing-based approach, both for collecting a large-scale dataset and for engaging the target community in the annotation process. This strategy follows former trends in language and speech technologies, where contributions from non-expert users have enabled the creation of valuable resources in a cost-effective and scalable manner [14, 15]. Second, we present an enhanced ASR architecture that leverages state-of-the-art transformer-based models (Whisper), specifically optimized for Hebrew. While our work focuses on developing Hebrew ASR, the methodology can serve as a blueprint for developing ASR systems in other under-resourced languages. Our approach of combining community efforts with modern models demonstrates a practical path forward for languages that currently lack robust ASR solutions. By making our methodology and tools openly available, we aim to make speech technology more accessible across different languages and communities.

To sum up, the current work presents three contributions to Hebrew ASR. First, we introduce the largest manually verified and corrected, openly licensed Hebrew speech dataset to date (ivrit.ai crowd-transcribe), providing a robust foundation for ASR development. Second, we demonstrate a new open-source Hebrew-Whisper model that achieves state-of-the-art performance while remaining freely available to the research and industrial community. Finally, addressing the limitations observed in existing multilingual corpora, we propose a new test set specifically designed for evaluating Hebrew ASR models, better reflecting modern spoken Hebrew.

3. Datasets

3.1. The Full ivrit.ai Corpus

The ivrit.ai corpus is presented in [16]. Since its original publication, the corpus has expanded and currently encompasses approximately 15,000 hours of audio content that spans diverse domains, topics, and speakers. The corpus marks a significant step in the available Hebrew speech corpora, as can be seen in Table 1 in the original paper, and has recently been used for new Hebrew speech developments [17, 18]. For labeling and training use, the ivrit.ai corpus was segmented as follows: First, the long audio segments were split into short segments of 2-29 seconds. This was done using Silero VAD [19] to identify the different segments, then utilizing `ffmpeg` to generate a matching `mp3` file for each segment. Silero VAD was chosen due to its efficiency: a lightweight process that can operate on a single CPU for the entire dataset of 15K hours. The choice of segment durations was made to ensure that the segment is long enough to be legible (at least 2 seconds), and fits into Whisper’s maximum segment duration (30 seconds) [8]. The ivrit.ai corpus is not manually transcribed. In the current work, we manually transcribed the speech segments from the ivrit.ai corpus, creating the sub-dataset ivrit.ai crowd-transcribe.

3.2. The Sub Dataset: ivrit.ai crowd-transcribe

The ivrit.ai crowd-transcribe¹ was created in a crowd-sourcing process, by asking the Hebrew-speaking community to help manually transcribe the ivrit.ai corpus. 2,100 volunteers participated in the effort through an interface (see Figure 1),² con-

¹<https://huggingface.co/datasets/ivrit-ai/crowd-transcribe-v4>

²The interface available at: <http://serve.ivrit.ai> and its source code available at: <https://github.com/ivrit-ai/>

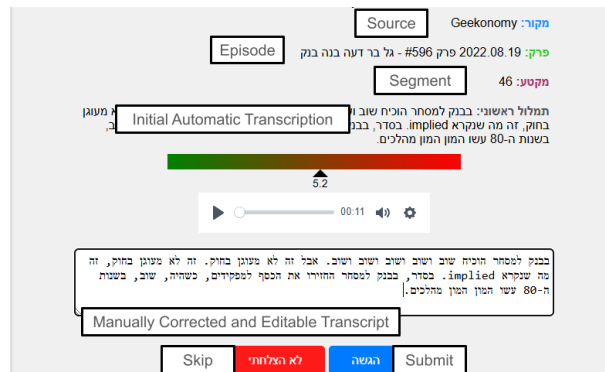


Figure 1: Screenshot of the User Interface for Correcting Automatic Transcriptions.

taining a random speech segment from the ivrit.ai corpus, accompanied by its automated transcription, created by Whisper large-v2. In addition, the interface presents the transcription in an editable format, asking the volunteer to correct the transcription if needed. The volunteer then can submit the corrected transcription or ask to skip to another segment if they had difficulties understanding it (and report the reason for skipping).

The segments were retrieved from the ivrit.ai corpus into the crowd-transcribe interface. The randomization occurs at the individual segment level. Selection probability for each segment is proportional to its source’s total duration in hours: normalizing the number of segments multiplied by the average length shows that we select episodes from each podcast proportionally to the podcast’s share of total hours. About 10% of segments undergo re-transcription. This is intended to enable in-depth verification of transcription by comparing between two or more transcribers. Additionally, we store and release information for each transcription, such as browser type, number of times the volunteer listened to the segment, number of playback pauses, and a unique identifier for the volunteer so we can check the quality of different segments transcribed by the same volunteer. All manual transcriptions were stored in a PostgreSQL database for further usage.

The final ivrit.ai crowd-transcribed sub-dataset contains a total of 246 hours of Hebrew speech, composed of 169,376 segments. Segments length ranged between 0.10-28.98 seconds (mean: 5.22s, SD: 4.54), taken from 47 ivrit.ai sources across diverse domains: science, tech, medicine, history, economics, interviews, politics, Jewish studies, psychology, family relations, self-development, lifestyle, sports, news, politics and entrepreneurship. Of the 5,310 speakers across all segments, we obtained demographic data about 3,561 (67%) through Wikidata. Of the identified speakers, 1,194 (33%) were women. Speakers were aged 20-90 years old. 72% of the identified speakers were native Hebrew speakers, and the others speak other languages as their first language: Arabic, English, Russian, Amharic, Spanish, German, and French. To ensure rigorous evaluation, we implemented a strict train-test split (90%-10%), completely excluding all segments from one randomly selected podcast, lengthened 25.45 hours from the training data, thus the training set of ivrit.ai crowd-transcribe includes 220.55 hours (151,849 segments). In addition to the crowd-transcribe dataset, we included in the training dataset 93.7 hours (14,508 segments, mean: 23.27, SD: 7.8) of Hebrew audio contributed

crowd-transcribe.

Corpus	Hours	Domain	Source	Description
ivrit.ai Evaluation Set (eval-d1)	2 hours	Informal conversation between two persons	Part of ivrit.ai test-set (not included in the train-set)	Audio is segmented into approximately 5-minute chunks. Verbatim transcription.
SASpeech [20]	4 hours	Economics and Politics podcast	Narrative podcast	Audio segments are several seconds in length. Verbatim transcription.
FLEURS [21]	2 hours	Read speech	-	The Hebrew subset of FLEURS
Mozilla Common Voice (version 17.0) [22]	2 hours	Read speech	-	The Hebrew test-set of the corpus
Kan subtitles [23]	1.7 hours	Broadcast content subtitles	"Kan" Youtube channel (news)	Subtitles transcription.

Table 1: Overview of Hebrew Speech Test Datasets

by commercial news media along with its corresponding subtitles, which was not part of the original ivrit.ai dataset. In total, 314.2 hours of fully annotated Hebrew speech were used in the current training.

4. Model Architecture and Training Pipeline

We used the basic Whisper architecture to fine-tune the model. Specifically, Whisper’s Large v2 model was used, as it showed a better starting point for Hebrew STT compared to Whisper’s Large v3 model in our initial results testing. Training was performed on a single H100 GPU with 80GB of VRAM, via vast.ai, for 2 epochs. The batch size was set to 8.

4.1. Experimental Setup

We evaluated the models across multiple datasets, as shown in Table 1. In addition to the FLEURS [21] multilingual dataset, widely used for ASR evaluation, the other test sets are used for the first time here for evaluation purposes, to the best of our knowledge. To enhance the evaluation options, we created a verbatim oriented test set, eval-d1, based on two hours from the ivrit.ai crowd-transcribe test set. Both reference transcription and the transcription output from each model were normalized by the `BasicNormalizer` from the original OpenAI source code (for removing symbols, punctuation, etc.), such that the final results compared only the text. We used the common measure word error rate (WER) to evaluate the results.

5. Results

As can be seen in Table 2, the ivrit.ai Whisper model demonstrates performance improvements across datasets, marking a significant advancement in Hebrew ASR technology. Specifically, the model achieves the lowest WER across all benchmark datasets in comparison to other Whisper-based models, with a particularly strong performance on spontaneous conversational speech, as in the eval-d1 and the SASpeech datasets. However, one commercial model (AWS Transcribe Batch) showed the best performance on the FLEURS, Common Voice and KAN datasets. We discuss the differences and possible reasons in the next section.

6. Discussion

We present ivrit.ai Whisper: a new ASR model based on the Whisper architecture, designed for better performance in transcribing Hebrew speech into text. Our model training is based on 220.5 audio hours, collected from recorded content, and an additional 93.7 audio hours with subtitles retrieved from media content. We aimed to address issues related to the limited Hebrew resources by collecting a large corpus for free use. We did not address the issues derived from Hebrew’s special linguistics directly, but enhanced the available data for training and the options to explore it by testing the various Hebrew-capable models on diverse datasets, including a new transcribed evaluation dataset (eval-d1). The new model demonstrates 8-29% reduction in error rates compared to other Whisper versions, marking an advance given the challenging nature of spontaneous Hebrew speech recognition. Importantly, WER is a metric that’s sensitive to different types of errors, which have varying significance in Hebrew [24, 25]. For example, defective and full spelling (for example, *šlḥm/šwlḥm*: ‘table’ written with and without the vowel letter *w*), isn’t necessarily considered an error but rather a preference, yet it counts as an error in the WER metric just like complete word substitutions do. Additionally, writing numbers in digits versus words (for example, 5 or “five”) and the nature of the transcription used as ground truth has an effect: manual transcriptions that emphasized disfluencies (i.e. when speakers repeated word fragments or said “um”) received lower scores when the model tended to omit such segments and only represent complete words (similar to subtitles).

The improvement was not uniform across all tested datasets, with specifically lower performance on the Kan dataset. The Kan dataset includes speech from traditional media (mostly news), along with their subtitles as a transcription. The superior performance of the new model on verbatim transcriptions (eval-d1 and SASpeech), compared to edited content, suggests that our model has developed a capability to capture natural speech patterns, including common disfluencies and linguistic variations specific to spoken Hebrew. However, since our training dataset also included 93.7 hours of news media and subtitles, other variables may influence the results. Furthermore, ASR applications may also differentiate between producing exact transcription versus what might be considered a “clean” transcription, often prioritizing readability and com-

Model	ivrit-ai/eval-d1	SASpeech	FLEURS	common_voice	KAN
Open Source Models					
ivrit.ai Whisper	6.2	8	24.1	20.7	11.3
OpenAI Whisper Large v2	8	9.8	26.6	23.3	16.4
OpenAI Whisper Large v3	9.6	9.4	26.2	23.1	13.4
OpenAI Whisper Large v3 Turbo	8.5	10.4	28.9	28	15.6
Commercial (Not Open Source) Models					
AWS Transcribe Batch (Dec 2024)*	6.9	8.6	23*	14.1*	9*
AWS Transcribe Stream (Dec 2024)*	8.1	9	28.7	20	13.1
Google Speech (Jan 2025)*	21.2	18.9	38.5	38	29.2

Table 2: Word Error Rate (WER) Comparison Across Speech Models and Datasets. The best model per dataset is bolded. Models marked with * are proprietary, non-reproducible, and thus limited for direct comparison.

prehension over verbal accuracy, frequently omitting disfluencies, hesitations, and making grammatical corrections. Future work may produce different sub-models for different ASR use cases, according to their purpose [26]. Our work joins a growing body of research in natural language processing and speech recognition that has leveraged crowdsourcing. We did so on two levels: first, by collecting the ivrit.ai corpus through voluntary contributions from content creators, in a manner similar to efforts such as Common Voice [27] and UncommonVoice [28]; and second, by inviting the community expected to benefit from the effort to assist in annotating the data. Our approach is comparable to other annotation initiatives that have used non-expert contributors for labeling tasks [14, 15], and we similarly found that non-expert work can be highly effective. However, we differ from those works, as our contributors were not paid, but volunteered; and although we turned to the general public, our participants were *domain-experts* in the Hebrew language. This is different from paid platforms like AMT, where workers are often unfamiliar with low-resource languages—a limitation noted in previous research [15]. Our study is thus closer to UncommonVoice [28], which also engaged volunteers from a specific community of interest, but the dataset we compiled and the model we refined are substantially larger in scale. The data collection strategies, model adaptation techniques, and evaluation frameworks developed in this study could potentially extend beyond Hebrew to benefit other languages with limited resources. This work demonstrates how targeted fine-tuning with carefully curated data can significantly improve speech recognition for languages traditionally underserved by mainstream ASR technologies.

The AWS Transcribe Batch commercial model showed the best performance on three datasets. However, we cannot analyze or explain their advantages as they do not detail their training data, the model architecture, or the training process. Furthermore, they provide their services at a cost, while we aimed to provide free access to a state-of-the-art ASR model.

6.1. Limitations and Future Directions

Despite the improvements and contributions the new model showed, our work has several limitations. First, although the ivrit.ai dataset represents an advancement in Hebrew speech datasets, containing more than 15,000 hours of Hebrew speech,

the subset used for developing the model was smaller, containing 314 hours of manually transcribed audio, due to limited budget and reliance on volunteers. Importantly, even this dataset is larger than former datasets used for similar tasks. Another limitation is that we did not use a validation set or conduct hyperparameter optimization in our approach, though our evaluation on multiple unseen datasets suggests good generalization capability. Future work should incorporate a formal validation set and systematic hyperparameter optimization to potentially achieve better performance. Future work may also enhance the dataset by collecting more manually transcribed data. In addition, we may implement data augmentation strategies, including the addition of background noise types, audio stretching, and other perturbations to improve model robustness. Future research could also investigate whether synthetic data generated through text-to-speech systems can supplement or replace portions of manually transcribed data while maintaining comparable performance. This approach could potentially provide a cost-effective way to expand the training dataset. Better performance may come from enhancing its quality: Our current method of segmenting continuous speech into shorter chunks, while practical for training, can lead to loss of important contextual information and may affect the model’s ability to maintain coherence across longer utterances. Future work may also develop improved methods for handling continuous audio, maintaining time-code alignment, and preserving contextual information across longer speech segments. More broadly, our work would benefit from a wider demographic representation of diverse accents, ages, and genders as speakers.

6.2. Conclusion

We aimed to develop an improved, freely accessible ASR system for Hebrew speech recognition, based on the Whisper architecture. Our approach demonstrates that specialized open-source models can effectively handle languages with rich morphology despite limited resources. The methodology we developed - combining community efforts, data curation, and architectural adaptation - provides a practical framework for developing ASR systems in other under-resourced languages. This work thus contributes not only to Hebrew ASR but also to making speech technology more accessible across different languages and communities.

7. Acknowledgment

We would like to thank all the content contributors who made the creation of the dataset possible, and all the volunteers who worked on its labeling. Thanks to Adv. Eli Greenbaum who created the license and provides us with legal guidance. Thanks also to AWS for contributing the credits that enabled the computing power, and to Dr. Shimon Shahar and Yam Peleg who assisted us along the way.

8. References

- [1] S. Khare, A. R. Mittal, A. Diwan, S. Sarawagi, P. Jyothi, and S. Bharadwaj, "Low resource asr: The surprising effectiveness of high resource transliteration." in *Interspeech*, 2021, pp. 1529–1533.
- [2] K. Bhogale, A. Raman, T. Javed, S. Doddapaneni, A. Kunchukuttan, P. Kumar, and M. M. Khapra, "Effectiveness of mining audio and text pairs from public data for improving asr systems for low-resource languages," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [3] 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.
- [4] B. Thai, R. Jimerson, R. Ptucha, and E. Prud'hommeaux, "Fully convolutional ASR for less-resourced endangered languages," in *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, D. Beermann, L. Besacier, S. Sakti, and C. Soria, Eds. Marseille, France: European Language Resources association, May 2020, pp. 126–130. [Online]. Available: <https://aclanthology.org/2020.sltu-1.17>
- [5] F. Kubala, S. Austin, C. Barry, J. Makhoul, P. Placeway, and R. Schwartz, "Byblos speech recognition benchmark results," in *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*, 1991.
- [6] T. Vaich and A. Cohen, "HMM phoneme recognition with supervised training and viterbi algorithm," in *Eighteenth Convention of Electrical and Electronics Engineers in Israel*. IEEE, 1995, pp. 3–2.
- [7] V. Silber-Varod, I. Siegert, O. Jokish, Y. Sinha, and N. Geri, "A cross-language study of speech recognition systems for english, german, and hebrew," *Online Journal of Applied Knowledge Management*, vol. 9, no. 1, pp. 1–15, 2021.
- [8] 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.
- [9] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.
- [10] G. Paraskevopoulos, C. Tsoukala, A. Katsamanis, and V. Katsouros, "The greek podcast corpus: Competitive speech models for low-resourced languages with weakly supervised data," *Proc. Interspeech 2024*, 2024.
- [11] P. E. Kummervold, J. de la Rosa, F. Wetjen, R.-A. Braaten, and P. E. Solberg, "Whispering in norwegian: Navigating orthographic and dialectic challenges." *Proc. Interspeech 2024*, 2024.
- [12] P. Zhakevich and B. Kantor, "Modern hebrew," in *The Semitic Languages*. Routledge, 2019, pp. 571–610.
- [13] A. Turetzky, O. Tal, Y. Segal-Feldman, Y. Dissen, E. Zeldes, A. Roth, E. Cohen, Y. Shrem, B. R. Chernyak, O. Seleznova *et al.*, "Hebdb: a weakly supervised dataset for hebrew speech processing," *arXiv preprint arXiv:2407.07566*, 2024.
- [14] R. Snow, B. O'connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks," in *Proceedings of the 2008 conference on empirical methods in natural language processing*, 2008, pp. 254–263.
- [15] S. Novotney and C. Callison-Burch, "Cheap, fast and good enough: Automatic speech recognition with non-expert transcription," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 207–215.
- [16] Y. Marmor, K. Misgav, and Y. Lifshitz, "ivrit. ai: A comprehensive dataset of hebrew speech for ai research and development," *arXiv preprint arXiv:2307.08720*, 2023.
- [17] E. Zeldes, O. Tal, and Y. Adi, "Enhancing tts stability in hebrew using discrete semantic units," in *The 50th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2025.
- [18] A. Roth, A. Turetzky, and Y. Adi, "A language modeling approach to diacritic-free hebrew tts," in *Proc. Interspeech 2024*, 2024, pp. 2775–2779.
- [19] S. Team, "Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier," <https://github.com/snakers4/silero-vad>, 2024.
- [20] O. Sharoni, R. Shenberg, and E. Cooper, "Saspeech: A hebrew single speaker dataset for text to speech and voice conversion," in *Proc. Interspeech 2023*, 2023, p. To Appear.
- [21] 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)*. IEEE, 2023, pp. 798–805.
- [22] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.
- [23] V. Gurevich, "Hebrew speech recognition dataset: Kan," hugging-face, 2022.
- [24] S. Kim, A. Arora, D. Le, C. feng Yeh, C. Fuegen, O. Kalinli, and M. L. Seltzer, "Semantic distance: A new metric for asr performance analysis towards spoken language understanding," in *Interspeech*, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:233033738>
- [25] Y.-Y. Wang, A. Acero, and C. Chelba, "Is word error rate a good indicator for spoken language understanding accuracy," in *2003 IEEE workshop on automatic speech recognition and understanding (IEEE Cat. No. 03EX721)*. IEEE, 2003, pp. 577–582.
- [26] J. D. Cintas and A. Remael, *Subtitling: Concepts and practices*. Routledge, 2020.
- [27] W. Phatthiyaphaibun, C. Chaksangchaichot, T. Rakthammanon, E. Chuangsuwanich, and S. Nutanong, "Crowdsourced data validation for asr training," in *Proc. Interspeech 2023*, 2023, pp. 551–555.
- [28] M. Moore, P. Papreja, M. Saxon, V. Berisha, and S. Panchanathan, "Uncommonvoice: A crowdsourced dataset of dysphonic speech." in *Interspeech*, 2020, pp. 2532–2536.