

FASA: a generic Flexible & Automatic Speech Aligner for the ASR community

Anonymous submission to Interspeech 2024

Abstract

Automatic Speech Recognition (ASR) with DL models has significantly benefited adults, but applying ASR technology to children introduces unique challenges due to distinct characteristics of the children speech. The pre-trained deep neural networks on adult speech data often struggle in generalizing to children’s speech, which calls for diverse training dataset for model finetuning. However, human annotation are not scalable to large scale datasets, and existing force alignment toolkits make impractical assumptions on the provided transcriptions, hindering their usages. To bridge the performance gap between children and adult ASR models, we introduce the Flexible & Automatic Speech Aligner (FASA), a novel force alignment toolkit. FASA force-aligns children’s speech datasets, providing accurate, well-segmented audio segments with transcriptions under flexible conditions. We also present the first young children ASR dataset under clinical environments.

Index Terms: automatic speech recognition, force alignment, children ASR, dataset, deep learning

1. Introduction

In recent years, the field of Automatic Speech Recognition (ASR) has witnessed remarkable advancements in the development of deep learning (DL), transforming our interactions with technology and enhancing various applications, from virtual assistants [1, 2, 3] to transcription services [4]. While these developments have significantly benefited adults, the application of ASR technology to children’s speech presents a unique set of challenges. Children’s speech exhibits distinct characteristics, including rapid changes in pitch, articulation patterns, and vocabulary development [5], making it not been generalized from deep neural networks (DNNs) pre-trained with adult speech corpus.

Previous studies [6, 7, 8, 9, 10, 11] have demonstrated that fine-tuning pre-trained ASR models on large and diverse children ASR corpora can alleviate the performance gap between children and adult ASR DL models. Thus, one crucial aspect of advancing ASR for children is the creation and refinement of datasets tailored to their speech patterns. However, while high quality data - which is defined as the transcription and the audio matches exactly - are heavily needed for children’s speech patterns, manual annotation faces two significant challenges. First, it is a labor-intensive process that requires domain expertise and considerable time investment. From our experience, manual annotations with pre-segmented audio files take at least 3x times of the audio’s duration, and segmenting the audio files takes another 1-1.5x times of the duration. Work from [12] supports our experience and claims that it takes 7-8 minutes to transcribe a minute of audio to SALT format. Second, even if there are tran-

scriptions, the qualities of different datasets vary a lot due to differences in human annotators, as well as different focuses of those datasets (for example, stuttering detection, speech sound disorder, dialects, etc.). Thus, human annotations are clearly not scalable on large-scale high quality datasets.

A natural question that arises from the two challenges is that: how could we produce high quality datasets from available low quality datasets? To address this and facilitate further research in children’s ASR, the development of a toolkit for force-aligning children’s speech emerges as a pivotal solution. In this work, we notice the deficiencies among existing alignment toolkits, and we bring forth these contributions to the ASR community. First, we present a novel open-sourced force alignment toolkit, Flexible & Automatic Speech Aligner (FASA), for children’s ASR that provides accurate, aligned, well-segmented audio segments with its transcriptions under flexible conditions. Force alignment, the process of aligning audio segments with their transcriptions, is crucial for generating accurate ASR datasets. However, previous methods like MFA [13] rely on the correctness of provided transcriptions, which is often impractical to obtain for many datasets. On the other hand, FASA leverages state-of-the-art DL model as the backbone to automatically align children’s speech datasets under minimal requirements of provided transcriptions. Second, we use this toolkit to compile a new dataset from CHILDES [14]. To the best of our knowledge, this is the first at-scale ASR dataset from clinical data on young children. Third, we show the superiority and consistency of FASA compared to existing force-alignment toolkits and even human annotators.

The rest of the paper will be organized as follows. In Section 2, we will formally define the force alignment problem, as well as introduce the related works in the area of ASR and alignment tools. In Section 3, we will describe FASA in details. In Section 4, we will showcase the superior performance of this new toolkit under noisy and unorganized datasets using CHILDES [14] as an example, and finally we will conclude the paper in Section 5. The code and instructions are available at ¹

2. Related Works

2.1. Definition of the Task

Force alignment with provided transcription refers to the task of aligning an audio with accurate timestamps given its transcription. Modern ASR systems expect the input audio to be segmented into smaller pieces during their training process. For example, the Whisper model [15] pads or trims the audio input to 30 seconds. Thus, when a transcription without timestamp is associated with a long audio, a force-alignment toolkit

¹anonymous under review

is essential for the creation of the dataset. Formally, the force alignment task is defined over an audio sample composed of n utterances, $A := \{A_1, A_2, \dots, A_n\}$, and a transcription of m “words”, $T := \{T_1, T_2, \dots, T_m\}$. It is noteworthy that a “word” in T refers to the basic element of the transcription. Depending on the use case, it could represent an utterance, a word, or even a phonetic symbol. The force alignment task is to associate each A_i with its corresponding words in T , which are from T_{si} to T_{ei} , or report that A_i is not transcribed in T . For the ease of notation, we will define the association between A_i and its transcription as $A_i = (T_{si}, T_{ei})$.

A robust auto-alignment system will need to have two important features. First, it shall not assume that if $A_i = (T_{si}, T_{ei})$, $A_j = (T_{sj}, T_{ej})$, and $i < j$, then $ei < sj$. That is, the transcription does not have time-dependency with the audio. An utterance appears early in the audio does not necessarily mean it appears early in the transcription. Second, A_i does not necessarily have a corresponding (T_{si}, T_{ei}) , and it is possible that $A_i = \emptyset$. That is, not all audio information will be transcribed into the transcription, and some audio will be left as un-transcribed.

Such two features are important because they do not require the completeness and orderliness of the provided transcription, which are usually infeasible to achieve in real-world scenarios. A popular direction of obtaining large amount of audio and transcriptions is from scraping the Internet, but the vast majority of transcriptions from the Internet contains missing and wrongly-ordered materials.

2.2. Existing Alignment Tools

Traditionally yet still prevalently, alignment between audio and its transcription are done via human annotators on various software [16, 17]. However, as discussed earlier, such practice is not scalable for large dataset. Work from [18] contains some parts of a complete force-alignment pipeline, but it does not address the fundamental problem of aligning audio with its transcription. Work from [19] uses generative-adversarial networks (GAN) to perform data augmentation on children ASR dataset, but their work does not introduce diverse *new* data to the field. Recently, Talkbank project announces its data processing pipeline that converts raw audio into CLAN-annotated transcriptions [20]. While their work uses similar backbone structure as ours, they rely on transcription generated by ASR models, whereas we faithfully adhere to the provided transcription as the ground truth. Thus, on downstream tasks such as fine-tuning ASR models, our dataset will be more usable because datasets generated by ASR models might cause severe degradation according to [15]. On the other hand, there have been several works on force-alignment ASR datasets with the assistance of human-labelled transcriptions [13, 21, 22], with Montreal-Force-Aligner (MFA) being the most popular toolkit [13]. MFA incorporates Kaldi [23] as the backbone, which uses Gaussian Mixture Model (GMM) for its transcription generation process. However, while MFA [13] works well with carefully annotated transcriptions, it requires the transcription to have a perfect matching with the audio. That is, $A_1 \rightarrow \{t_1, t_2, \dots, t_i\}$, $A_2 \rightarrow \{t_{i+1}, t_{i+2}, \dots, t_j\}$, and so forth. Thus, under the two practical assumptions defined in Section 2.1, MFA [13] does not generate satisfactory results. Moreover, while recent multi-modal large language models (MLLM) might have the potential of automating the alignment process [22], they are much more resource-intensive compared to specific ASR models.

2.3. Dataset

Recent adoptions towards children ASR usually consider general datasets such as PF-STAR [24], My Science Tutor (MyST) [25], CMU Kids Corpus [26], OGI Kids Corpus [27], or smaller-scale clinical datasets for children with older ages such as ETLT, a German-English ASR dataset for older children [28]. Among them, MyST is the largest dataset, containing 393 hours of speech data between children and a virtual science tutor, which is still significantly smaller than the adult datasets. We’ve observed that the majority of children ASR models are typically fine-tuned using general datasets, leading to a lack of task-specific abilities in crucial areas, such as clinician settings for young children. This limitation arises from the unavailability of high-quality datasets tailored for these specific task requirements. However, we are optimistic that our toolkit will address and resolve this issue effectively.

3. FASA Toolkit Design

3.1. Features of FASA

Similar to existing auto-alignment toolkits, FASA requires an audio file and a transcription associated with the audio file. However, due to the high uncertainty in raw dataset, FASA assumes only minimum format from the input. In particular, FASA does not assume the correctness of the transcription. The ground truth (GT) of an utterance A_k is defined by Equation 1 in the pipeline of FASA. Compared to previous force alignment toolkits, FASA will choose to ignore the utterances without proper transcriptions, and thus reach significantly better dataset quality from this practice.

$$GT_k = \begin{cases} T_{A_k} = \{T_i, T_{i+1}, \dots, T_j\}, & \text{if } T_{A_k} \in T \\ \emptyset, & \text{otherwise} \end{cases} \quad (1)$$

Besides the improvement on flexibility of datasets, FASA incorporates beneficial design elements from established toolkits. In a bid to enhance user convenience, FASA follows the same design principles as MFA for its usage. Users simply have to compile the audio file and its transcriptions into a designated folder, then execute a program. The subsequent processes are all automatic, streamlining the user experience. FASA further integrates a crucial feature present in both MFA [13] and PonSS [21]. This feature empowers users to select and to manually input transcriptions for utterances in instances where the provided transcriptions raise suspicions of inaccuracy. This flexibility ensures precision and user control over the transcription process. At the same time, FASA incorporates an optional post-generation checking schema that enables automatic exclusion to incorrect alignments in the generated dataset, minimizing the possibility of incorrectness from the underlying model.

3.2. Workflow

FASA follows a five-module pipeline to automatically segment, label, and align a long audio file with its transcription, as illustrated in Figure 1. Among the five modules, the second and third are mandatory, whereas the other three are optional for enhancing the quality and quantity of the dataset. These five modules together maximize the correctness of force alignment under flexible conditions. The first module applies regular expression to clean the provided transcriptions and to exclude any non-alphanumeric characters. For the second module, modern ASR

Algorithm 1 sliding window to find best matching

Input: $A, T = \{T_1, \dots, T_m\}, \bar{T}$, alignment threshold σ_a , inclusion threshold σ_i .

Step 1: Initialize holder for dataset of aligned segments: $\text{DATA}_{\text{align}} = []$

Initialize holder for questionable segments: $\text{DATA}_{\text{verify}} = []$

Step 2: **for** $A_k \in A$ **do**

 Get A_k 's transcription: $T_{A_k} = \{\bar{T}_i \dots \bar{T}_j\} \in \bar{T}$

 Initialize minimum distance $D_{\min} = \infty$, best starting index BEST_i , best length BEST_l

for $a = 1, 2, \dots, m$ **do**

for $b = 1, 2, \dots, (j - i)$ **do**

if $\text{DIS}(T_{A_k}, T[a : a + b + 1]) < D_{\min}$ **then**

$D_{\min} = \text{DIS}(T_{A_k}, T[a : a + b + 1])$

$\text{BEST}_i = a$

$\text{BEST}_l = b + 1$

end if

end for

end for

Step 3: let $GT_k = T[\text{BEST}_i : \text{BEST}_i + \text{BEST}_l]$

if $\text{WER}(GT_k, T_{A_k}) < \sigma_i$ **then**

if $\text{WER}(GT_k, T_{A_k}) < \sigma_a$ **then**

 append (A_k, GT_k) to $\text{DATA}_{\text{align}}$

else

 append (A_k, GT_k, T_{A_k}) to $\text{DATA}_{\text{verify}}$

end if

end if

end for

Output: $\text{DATA}_{\text{align}}, \text{DATA}_{\text{verify}}$

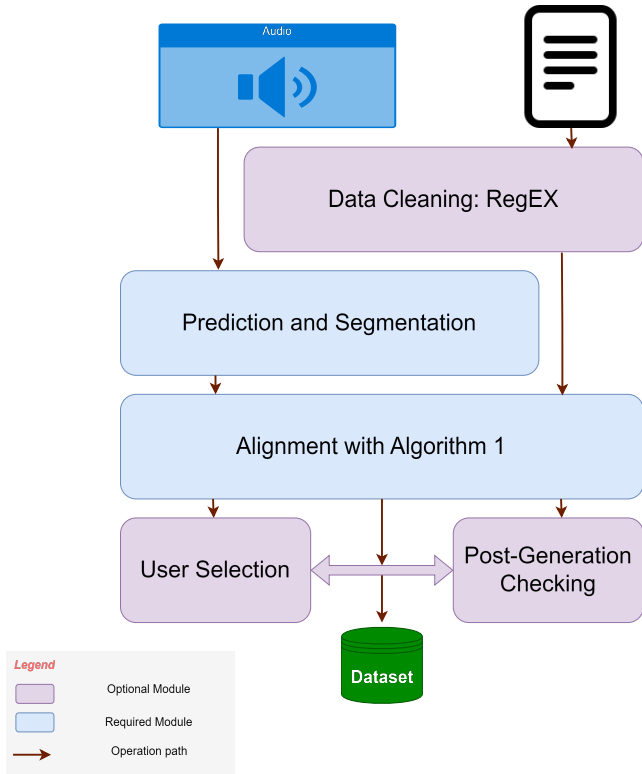


Figure 1: This figure illustrates the pipeline of FASA. The input is an audio file and a transcription. The blue modules are required, whereas the purple modules are optional that could be enabled by user. The entire system besides user selection is automatic.

models will be used obtain word-level timestamps of the transcriptions. Currently, sentence-level separations from the provided model is used as the segmentation marks for long audio. After the second module, a folder consisted of audio segments and their corresponding predictions will be generated. The set of predictions for sentence-level utterances will be denoted as $\bar{T} = \{\bar{T}_1, \bar{T}_2, \dots, \bar{T}_n\}$. For each utterance A_k , its predicted transcription will be $A_k = \bar{T}_{A_k}$. The third module will apply a sliding-window Algorithm 1 to find the best matching from the provided transcription (T) for each utterance (A_k). In the algorithm, DIS is the Levenshtein distance between two sentences. After this module, two datasets will be generated. The first dataset $\text{DATA}_{\text{align}}$ is when the algorithm finds close alignment between the prediction and provided transcription that is within a threshold, and the second $\text{DATA}_{\text{verify}}$ is when the algorithm finds slight mismatches between the prediction and the transcription. For $\text{DATA}_{\text{align}}$, the provided transcription from T will be used as ground truth of the utterance. The user selection module is an optional module that launches a graphical-user-interface (GUI) that allows user to listen to, select, or input correct transcription for each utterance in $\text{DATA}_{\text{verify}}$ so that they could be added to the dataset. The post-generation checking (PGC) is an optional module that iterates through $\text{DATA}_{\text{align}}$ to find if there are significant mismatches between a second-round prediction and the aligned transcription on sentence length. The implemented metric for PGC is based on the difference in sentence length between results of a second-round prediction and the aligned transcription. If the difference is greater than a threshold, the utterance and its transcription will be removed from $\text{DATA}_{\text{align}}$. After the two optional modules, FASA assumes the validity of $\text{DATA}_{\text{align}}$, which will be used as the final output dataset.

Currently, speaker identification is not supported by FASA. If the user has specific needs for diarization, they need to modify stage 2 and 3 to incorporate diarization features.

Table 1: *Manual inspections from the authors on the generation quality of FASA on two randomly selected audio files and their transcriptions. AU is the number of aligned utterances in $Data_{align}$, VU is the number of utterances in $Data_{verify}$. AU Error is the number of utterances that have any incorrect transcription. AW is the number of aligned words, and AW Error is the number of words that are incorrect in the aligned words. The percentage by the AW Error is the WER percentage of the aligned words. PGCU is the number of utterances that are removed from $Data_{align}$ in post-generation checking, and PGCU FP is the number of false positives in PGCU.*

Backbone Model	AU	VU	AU Error	AW	AW Error (%)	PGCU	PGCU FP
Stable whisper [29]	77	32	2	814	3 (0.37%)	18	11
Whisperx [30]	81	33	1	903	2 (0.22%)	5	3

4. CHILES as an Example

4.1. CHILDES [14] dataset

The Child Language Data Exchange System (CHILDES) [14] is a component of the TALKBANK project². CHILDES is a collection of many existing research works that contains massive amount of audio and transcriptions of children speech under various conditions. Among all of the audio files, we select four datasets that contain audio spoken by children in English and accompanied with English transcriptions, and we use FASA to convert them into smaller audio/transcription segments that are compatible with the training requirements of the current DL models. We report the specifications of each dataset in Table 2. Among the four datasets, Narrative records 352 children from age 4 to 9. The children are performing the Edmonton Narrative Norms Instrument (ENNI) test [31]. To the best of our knowledge, this is the first at-scale dataset for young children ASR from clinical recordings. Clinical-Eng includes additional ENNI test utterances, with some of them being duplicates from Narrative. Clinical-Other contains conversations between children and the clinicians, and the utterances are much shorter, often at word-level. English-NA contains recordings of 5332 conversations in household when a child is involved. Compared to the other datasets, Eng-NA dataset has varying audio qualities and contains not only children speech but also adult speech as the audios are recorded in home setting. During dataset generation, we set $\sigma_a = 0.1$, $\sigma_i = 0.3$; we also enable post-generation checking with maximum length tolerance as 1 word. For the backbone model, we uses WhisperX [30] because of its superior performance over other ASR models, but we will also show the performance of another variant [29] in the evaluations.

Table 2: *Specifications for the four datasets extracted from CHILDES [14]. Original are the number of participants (or the number of audio files in the original dataset). AU is the number of aligned utterances, whereas VU is the number of utterances to be verified if user chooses the manual selection. Time is the total duration of all the utterances in the aligned dataset, the format of Hour:Minute:Second.*

Dataset	Original	AU	VU	Time (H:M:S)
Narrative	352	14654	4402	15:16:51
Clinical-Eng	1540	59539	14610	30:11:40
Clinical-Other	292	21982	677	4:18:6
English-NA	5332	778978	180146	285:18:22

The processed datasets are organized in a similar way as LibriSpeech [32]. Each (anonymized) participant’s audio file is segmented, and the segmented audio utterances are put under

a folder named by the hashed participant. A transcription is paired with each utterance’s audio file.³

4.2. Evaluations

Due to the size of the dataset, we are not able to performance manual quality verification on the entire generated dataset. Thus, we randomly selected two audio files and transcriptions from the 352 recordings in Narrative dataset, and we report the manual inspection results for data generated by FASA with these two audio files in Table 1. Several results need to be emphasized here. First, since the two transcriptions are noisy, MFA [13] completely fails to properly align the audio segments with the correct transcription. To be specific, one of the document has missing transcriptions for the beginning of audio, which results in 100% AU Error with MFA; the other document was not successfully processed by MFA due to their internal software errors. Second, WhisperX [30] shows better performance than Stable-whisper [29], and given its faster speed, we recommend user to use WhisperX [30] wherever possible. Third, FASA using WhisperX [30] as the backbone incorrectly aligns one utterance with its transcription. For that utterance, it misses “so the” sound at the end of the utterance, and the two words are not recorded into the aligned transcription. Manual inspection found out that the speaker stuttered and repeated “so the”, which might be the issue of model not picking up that sound. At last, WhisperX [30] has a WER of **0.22%** for the aligned words. This result is much better than human annotators. [33] reported that 5 out of 393 hours speech in MyST dataset [25] are potentially incorrect with WER > 50%, resulting in **3%** increase in WER for the entire training dataset. Compared to human annotators that were used to annotate MyST, FASA achieves one magnitude lower WER without requiring any human labor.

5. Conclusion

In this paper, we present the Flexible & Automatic Speech Aligner (FASA) toolkit, which leverages DL models as its backbone, to automate the alignment of children’s speech datasets with minimal human intervention. At the same time, we introduce the first children ASR dataset sourced from clinical data, filling a significant gap in available resources. Fine-tuning state-of-the-art DL models on this new dataset demonstrates a tangible reduction in the performance gap between children and adult ASR, which aligns with existing works in this area. The availability of our toolkit, code, and dataset aims to propel further research and collaboration in advancing ASR for children, ultimately enhancing technology’s ability to understand and interact with the unique characteristics of children’s speech.

³Due to restrictions on commercial usage from raw data, we will not publish processed dataset. Contact the authors for research usage, or use the software to generate by yourself.

²<https://talkbank.org/>

6. References

- [1] D. Bernard and A. Arnold, "Cognitive interaction with virtual assistants: From philosophical foundations to illustrative examples in aeronautics," *Computers in Industry*, vol. 107, pp. 33–49, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0166361518304445>
- [2] L. M. Thomas M. Brill and R. J. Miller, "Siri, alexa, and other digital assistants: a study of customer satisfaction with artificial intelligence applications," *Journal of Marketing Management*, vol. 35, no. 15-16, pp. 1401–1436, 2019. [Online]. Available: <https://doi.org/10.1080/0267257X.2019.1687571>
- [3] C. Van Gysel, "Modeling spoken information queries for virtual assistants: Open problems, challenges and opportunities," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 3335–3338. [Online]. Available: <https://doi.org/10.1145/3539618.3591849>
- [4] J. Y. Kim, C. Liu, R. A. Calvo, K. McCabe, S. C. R. Taylor, B. W. Schuller, and K. Wu, "A comparison of online automatic speech recognition systems and the nonverbal responses to unintelligible speech," *CoRR*, vol. abs/1904.12403, 2019. [Online]. Available: <http://arxiv.org/abs/1904.12403>
- [5] V. Bhardwaj, M. T. Ben Othman, V. Kukreja, Y. Belkhier, M. Bajaj, B. S. Goud, A. U. Rehman, M. Shafiq, and H. Hamam, "Automatic speech recognition (asr) systems for children: A systematic literature review," *Applied Sciences*, vol. 12, no. 9, 2022. [Online]. Available: <https://www.mdpi.com/2076-3417/12/9/4419>
- [6] R. Jain, A. Barcovski, M. Yiwere, P. Corcoran, and H. Cucu, "Adaptation of whisper models to child speech recognition," 2023.
- [7] S. Wills, Y. Bai, C. Tejedor-García, C. Cucchiari, and H. Strik, "Automatic speech recognition of non-native child speech for language learning applications (short paper)." Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023. [Online]. Available: <https://drops.dagstuhl.de/opus/volltexte/2023/18521/>
- [8] P. Gurunath Shivakumar and S. Narayanan, "End-to-end neural systems for automatic children speech recognition: An empirical study," *Computer Speech & Language*, vol. 72, p. 101289, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0885230821000905>
- [9] S. Shraddha, J. L. G. and S. K. S., "Child speech recognition on end-to-end neural asr models," in *2022 2nd International Conference on Intelligent Technologies (CONIT)*, 2022, pp. 1–6.
- [10] E. Booth, J. Carns, C. Kennington, and N. Rafla, "Evaluating and improving child-directed automatic speech recognition," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, May 2020, pp. 6340–6345. [Online]. Available: <https://aclanthology.org/2020.lrec-1.778>
- [11] T. Rolland, A. Abad, C. Cucchiari, and H. Strik, "Multilingual transfer learning for children automatic speech recognition," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, Jun. 2022, pp. 7314–7320. [Online]. Available: <https://aclanthology.org/2022.lrec-1.795>
- [12] J. F. Miller, K. Andriacchi, and A. Nockerts, "Using language sample analysis to assess spoken language production in adolescents," *Language, Speech, and Hearing Services in Schools*, vol. 47, no. 2, p. 99–112, Apr 2016.
- [13] 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.
- [14] B. MacWhinney, "The childes project: Tools for analyzing talk, third edition," 2000. [Online]. Available: <https://childes.talkbank.org/>
- [15] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.
- [16] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (version 5.3.51)," 01 2007.
- [17] M. S. Grover, P. Bamdev, Y. Kumar, M. Hama, and R. R. Shah, "audino: A modern annotation tool for audio and speech," 2020.
- [18] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0885230816302418>
- [19] P. Sheng, Z. Yang, and Y. Qian, "Gans for children: A generative data augmentation strategy for children speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 129–135.
- [20] H. Liu, B. MacWhinney, D. Fromm, and A. Lanzi, "Automation of language sample analysis," *Journal of Speech, Language, and Hearing Research*, vol. 66, no. 7, p. 2421–2433, Jul 2023.
- [21] J. Rodd, C. Decuyper, H. R. Bosker, and L. ten Bosch, "A tool for efficient and accurate segmentation of speech data: announcing POnSS," in *Behavior Research Methods* 53, 2021, pp. 744–756.
- [22] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities," 2023.
- [23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.
- [24] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, "The PF.STAR children's speech corpus," in *Proc. Interspeech 2005*, 2005, pp. 2761–2764.
- [25] S. S. Pradhan, R. A. Cole, and W. H. Ward, "My science tutor (myst) – a large corpus of children's conversational speech," 2023.
- [26] M. Eskenazi, J. Mostow, and D. Graff, "The CMU Kids Corpus," in *LDC97S63*, 1997.
- [27] K. Shobaki, J.-P. Hosom, and R. A. Cole, "The OGI kids² speech corpus and recognizers," in *Proc. 6th International Conference on Spoken Language Processing (ICSLP 2000)*, 2000, pp. vol. 4, 258–261.
- [28] R. Gretter, M. Matassoni, D. Falavigna, A. Misra, C. Leong, K. Knill, and L. Wang, "ETLT 2021: Shared Task on Automatic Speech Recognition for Non-Native Children's Speech," in *Proc. Interspeech 2021*, 2021, pp. 3845–3849.
- [29] jianfch, "stable-ts," <https://github.com/jianfch/stable-ts/tree/main>, 2023.
- [30] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," *INTER-SPEECH 2023*, 2023.
- [31] P. Schneider, R. V. Dubé, and D. Hayward, "The edmonton narrative norms instrument," 2005.
- [32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [33] A. A. Attia, J. Liu, W. Ai, D. Demszky, and C. Espy-Wilson, "Kid-whisper: Towards bridging the performance gap in automatic speech recognition for children vs. adults," 2023.