Beyond Adult Speech: Exploring SepFormer's Performance in Child Speech Separation

April 29, 2024

Wenjun meng

Abstract

This thesis explores the performance of SepFormer, a speech separation model, in processing child speech beyond its traditional application to adult speech. The research question investigates the effectiveness of SepFormer in separating speech in datasets consisting of child speech. The hypothesis posits a dip in performance when SepFormer confronts child speech, underscoring a need for recalibration or the inception of child-specific models. This investigation is important for the evolution of ASR systems, promising enhanced inclusivity and efficacy in contexts catering to children's educational and communicative needs.
Keywords: SepFormer, Child Speech Separation, Automatic Speech Recognition, Speech Processing, Speech Separation

Contents

1 Introduction

Speech separation plays a crucial role in enhancing the clarity and intelligibility of speech in environments where multiple sound sources overlap, often referred to as the cocktail party problem(Ephrat et al., 2018). While considerable advances have been made in adult speech separation, child speech separation has received less attention but holds significant potential, particularly for treating speech disorders in children. Wang et al. (2018) presents an LSTM design for speech separation between child and adult groups in a speaker-independent manner. Child speech separation is more challenging and distinct from adult speech separation due to many factors, for example, children's smaller vocal tracts which lead to higher fundamental and formant frequencies, slower speaking rate and more variability in speaking rate, vocal effort, and spontaneity of speech(Potamianos et al., 1997).

When it comes to child speech, few can match the performance of adult-targeted ASR. Still, several attempts have been made to classify children and adults speech. The researchers from Google(Liao et al., 2015) built a large vocabulary continuous speech recognition (LVCSR) system that works well for children, which is then used to recognize queries in the YouTube Kids app. Gray et al. (2014) presents a child specific LVCSR system that improves the accuracy for children speaking US English to interacting electronic devices.

In recent years, deep learning techniques have significantly advanced the challenge of extracting individual speech signals from a mix. The SepFormer model(Subakan et al., 2021), which leverages the Transformer architecture, introduces a transformative approach and sets a new benchmark for speech separation performance. Despite its impressive performance on benchmarks like WSJ0-2mix and WSJ0-3mix(Isik et al., 2016), which are from adult speeches, its performance on child speech remains unclear. This leads to a possible discrepancy between speech separation models trained primarily on adult speech and child speech.

This research aims to evaluate Sepformer's performance on child speech using a child phonology dataset PhonBank(Holliday et al., 2015). The objective is to assess whether SepFormer, initially trained on adult speech, experiences any degradation in performance when applied to child speech, evaluated by SI-SNR and SDR metrics(Vincent et al., 2006). The decision to focus on a specific age group is guided by the findings of Yeung and Alwan (2018), which demonstrated the significant influence of even a single year of age difference on child ASR performance. Their research suggests that ASR systems might achieve better accuracy with training data from slightly older children. Therefore, this study targets the 4 to 5-year-old age group, anticipating that this selection will provide the most effective training data and will help to clarify the role of age in ASR system performance. In order to maintain consistence with the baseline model and to enable a straightforward comparison, this study will concentrate on English-language recordings from the PhonBank database.
Additionally, this study seeks to highlight the importance of including child speech in the training datasets for speech separation models like SepFormer, to

ensure their effectiveness across a wider range of speakers.

This proposal is structured as follows: In part 2, I provide a brief literature review of existing research relevant to the topic. In part 3, I introduce my research question to articulate the central research question guiding the study, along with the hypothesis that the research will test. In part 4, I outline the methodology which describes the research design, data collection methods, and analytical techniques to be employed, as well as the timeline that presents a detailed schedule for the project. In part 5, I discuss risk mitigation to identify potential challenges and the strategies to address them. In part 6, ethical considerations are discussed.In part 7, I present analysis and outcomes which are more like methodology at this phase. Finally, in part 8, I describe the impact and relevance of my research to reflect on the anticipated outcomes.

2 Literature review

Studying child speech recognition is crucial for identifying and addressing speech disorders early on, which is essential for a child's cognitive and social development.(Sattorovich, 2022) Children can also benefit from ASR technology in everyday activities through the use of voice-driven educational resources, such as interactive gaming, reading assistance (Mostow, 2012; Yeung & Alwan, 2019)

Recent advancements in speech separation technology have predominantly focused on adult speech. Luo and Mesgarani (2018) introduced TasNet, leveraging an encoder-decoder framework to model audio directly in the time domain, a departure from traditional frequency-domain methods. This innovative approach significantly enhances real-time speech separation performance and computational efficiency. WaveSplit, introduced by Zeghidour and Grangier (2021), represents a breakthrough in end-to-end speech separation by utilizing speaker clustering to effectively address the permutation problem. Vaswani et al. (2017) introduced the Transformer model, using attention mechanisms to improve how machines understand and translate languages. This innovation led to better performance on language tasks, establishing new records for accuracy. As transformers developed over the years, in 2021, SepFormer was proposed(Subakan et al., 2021), a novel Transformer-based architecture for speech separation that leverages a multi-scale approach to learn short and long-term dependencies, and adopts a dual-path speech separation architecture with transformer blocks. The SepFormer model outperformed traditional RNN-based models while offering advantages in parallelization and reduced computational demands.

Recent research has consistently highlighted a significant performance gap when systems trained on adult speech are applied to children's speech. Bhardwaj et al. (2022) demonstrated that ASR systems based on adult speech patterns significantly underperform when applied to children's speech. Kennedy et al. (2017) provided empirical evidence of this challenge, revealing error rates ranging from 15% to 20% in child speech recognition within the dynamic context of real-world social human-robot interaction (HRI). These findings establish a critical baseline for investigating the SepFormer model's efficacy in child speech

separation.

Delving deeper into the root causes of these challenges, Bhardwaj, Kadyan, et al. (2020) identified the inherent acoustic variabilities in children's speech as a key factor contributing to lower recognition accuracy. Bhardwaj et al. (2022) and Shivakumar and Georgiou (2020) expanded on this by pointing to the continuously evolving nature of children's vocal mechanisms and language skills, which include developing mastery over prosodic elements such as pitch, volume, rhythm, and intonation. Kathania et al. (2021)further identified the acoustic differences between child and adult speech, noting that children's higher formant and fundamental frequencies are a result of their smaller vocal folds and shorter vocal tracts. These insights into the unique characteristics of children's speech patterns are crucial for informing the development and assessment of ASR systems for children.

While the existing literature provides compelling evidence of the inherent challenges in child speech recognition, it is crucial to assess whether the proposed technical solutions are sufficiently robust to address the variability and complexity of child speech. Grzybowska and Kacprzak (2016) extended age regression task by applying fusion of i-vectors and acoustic features regression to estimate the speaker age and for age classification. Zeng and Zhang (2007)proposed a novel speech classification system based on GMM to classify children speech and adults speech by applying the combined speech features of pitch, first three formants, 5-order RASTA-PLPC and Delta RASTA-PLPC aiming to model the behavior of children and adults speech respectively. Cristia et al. (2018) explored the complexities of talker diarization in real-world, child-centered audio recordings, highlighting the challenges posed by spontaneous conversations in diverse acoustic environments. A progressive learning approach was proposed to perform child-adult speech separation by leveraging a densely connected long short term memory (LSTM) architecture in a speaker independent manner(Wang et al., 2018). The authors then refined the model for enhanced robustness in realistic conditions by employing a new combined approach of speech enhancement and speech separation for child speech extraction(Wang et al., 2020). Although the aforementioned models have demonstrated the potential for success, further research is necessary to fully optimize their efficacy, particularly in environments that present significant challenges in terms of noise or reverberation.

The SepFormer model was initially trained using WSJ0-2mix and WSJ0-3mix datasets(Hershey et al., 2016), which consist of synthesized mixtures of adult speech in non-reverberant environments. While effective for adult speech, the model's performance on child speech remains untested. To address this question, this study proposes to test Sepformer on the PhonBank database(Li, 2012), which offers a rich repository of child language recordings in quiet rooms, including recordings from children between two to five years old and adults, covering several languages.

3 Research question and hypothesis

The LR has summarized the strides in speech separation technology, particularly models like SepFormer that have set benchmarks in adult speech separation. Despite these advancements, a gap persists in the application and evaluation of these technologies within the area of child speech. Importantly, child speech has unique acoustic and linguistic characteristics, distinguishing it markedly from adult speech. Nonetheless, speech separation in this domain remains little explored. The performance of advanced models like SepFormer, predominantly trained and validated on adult speech datasets, is yet to be evaluated against the distinct backdrop of child speech. This motivates a research avenue to assess whether these technologies can maintain their high performance standards when applied to child speech or if their efficacy wanes, necessitating tailored adaptations or the development of new child-centric models. Thus, stemming from this LR, the research question emerges: What is the difference in performance, measured through SI-SNR and SDR metrics, of the SepFormer model in separating child speech compared to adult speech using standardized datasets? The accompanying hypothesis, informed by the work of Bhardwaj et al. (2022), posits a statistically significant reduction in SepFormer's performance as measured by SI-SNR and SDR metrics compared to its performance on adult speech datasets when applied to child speech datasets. This will set the stage for an empirical investigation that could either validate this presumption or else reveal the versatility of SepFormer across diverse age demographics.

4 Execution

The primary objective of this study is to evaluate the performance of the SepFormer model on child speech datasets, with a comparison to its known performance levels on adult speech datasets.

4.1 Methodology

Data Collection: Speech samples will be collected from the PhonBank database, a comprehensive child phonology corpus. This study will specifically select English language samples from children aged four to five years.
Preprocessing: In alignment with the preprocessing protocols used for the baseline model, mixtures of speech have been generated, including female-female, male-male, and female-male combinations. These mixtures consist of both two-speaker and three-speaker configurations, created by randomly mixing utterances from the corpus to simulate a variety of interactive scenarios.
A significant preprocessing step involves the removal of long pauses present in the original audio recordings. To create a continuous flow of speech and to avoid the model learning from lengthy non-informative segments, silences longer than 1000 milliseconds have been removed. This parameter ensures that shorter pauses, which contribute to the natural rhythm of speech, are preserved. Addi-

tionally, a silence threshold has been set at -80 dBFS to remove audio segments quieter than this level.
Volume normalization is another essential process in the preparation of the dataset. RMS normalization has been applied to standardize audio levels across all samples. To mimic real-life variations in speaker volume and to maintain consistency with the SepFormer's original dataset preprocessing, the relative levels for the sources in each mixture have been uniformly varied between 0 dB and 5 dB.
To maintain uniformity with the SepFormer dataset, all audio tracks have been resampled to a sampling rate of 8 kHz. This consistency is crucial to prevent any potential performance issues that may arise from sampling rate discrepancies during model evaluation.For the purposes of this thesis, a total of 5 hours of audio data has been preprocessed. The choice of duration mirrors the test set size utilized in the original SepFormer study. This approach ensures that the evaluation of the model's performance on child speech is conducted under test conditions similar to those of its initial benchmarks. The dataset is essential in measuring the model's ability to separate speech in scenarios that include children's voices, thus providing valuable insights into its versatility and potential for wider real-world application.
Model Configuration: The SepFormer model will be implemented as described by its developers, with no modifications to the architecture. The training will be conducted with the same hyperparameters as the baseline model to maintain consistency.
Evaluation Metrics: This research is going to use the same two metrics that are used in Sepformer, SI-SNR and Signal-to-Distortion Ratio (SDR). SI-SNR measures the clarity of a speech signal relative to background noise, adjusting for signal scale to assess speech enhancement and separation regardless of volume. SDR quantifies the quality of a processed speech signal by comparing its strength to that of background noise and distortions, indicating how well speech is preserved or enhanced.
Experimental Procedure
Baseline Establishment: Utilize existing performance metrics of SepFormer on adult speech datasets as the baseline.
Child Speech Dataset Preparation: Select and preprocess speech samples from the PhonBank database to ensure compatibility with the evaluation criteria used for the baseline metrics.
Performance Comparison: Directly compare SepFormer's baseline performance metrics with those obtained from its application to the child speech dataset.
Analysis of Degradation: Analyze the results to determine if there is a performance degradation when SepFormer is applied to child speech, focusing on differences in SI-SNR and SDR.
Data Analysis: Apply statistical methods to compare the performance metrics of SepFormer across the two datasets.

4.2 Timeline

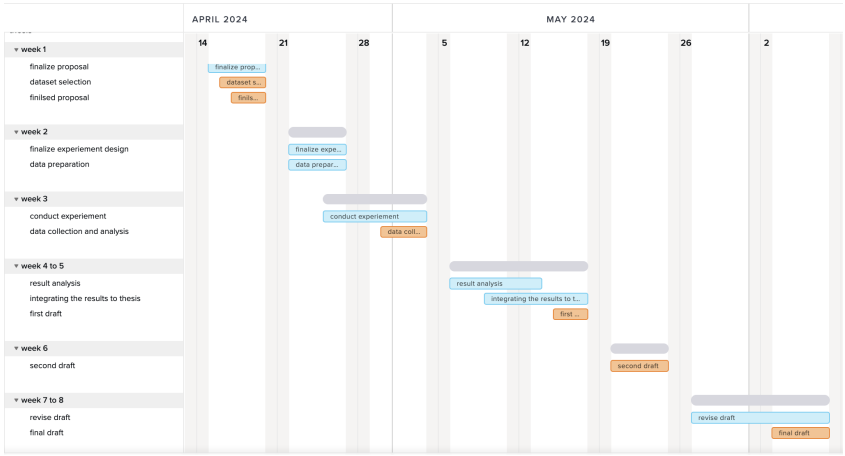

Figure 1: Timeline

5 Risk mitigation

The research question and methodology have been refined to mitigate anticipated risks; however, several challenges remain, as outlined below:

Risk	**Contingency**
Computational resource limitations	Proper usage of Habrok, or explore other cloud computing options
Unexpected model performance issues	Evaluation metrics backups across different scenarios
Unexpected delays	Buffer into timeline

Figure 2: Risk Mitigation

6 Ethical issues

As there's no direct human involvement in this study, the main ethical concerns center on the privacy and handling of the data. The dataset to be used is open-source; however, if there's a need to include additional datasets, it's important to ensure that these datasets can be used for the dissertation research. It's also important to confirm that any information that could identify individuals

has been removed or anonymised, particularly as this thesis focuses on children speech, a group that requires heightened privacy protection.

7 Analysis and outcomes

Data Preparation: To ensure consistency across all samples, the speech data sets are subjected to the necessary pre-processing to normalise the audio levels and remove any background noise.
Performance Evaluation: The model's performance metrics will be assessed using statistical methods such as SI-1 and SDR. SepFormer's performance will be benchmarked against existing standards for adult speech datasets and analyzed for any degradation when applied to child speech datasets.
Expected Outcomes: It is expected that SepFormer will show degradation in performance when on child speech compared to adult speech, which could indicate areas for model improvement or adaptation.
Implications of Findings: The findings will provide insights into the challenges of speech separation in child speech and may suggest directions for future research or practical modifications to existing models to better accommodate child speech characteristics.

8 Impact and relevance

Reflection on expected outcomes
After addressing the research question and validating or disproving the hypothesis that "SepFormer's performance may degrade when processing children's speech," several significant outcomes are expected. The validation of the hypothesis will highlight the need for specialized models or adaptations to existing models to effectively process children's speech. On the contrary, if the hypothesis fails and SepFormer performs well when processing children's speech, it may indicate that current speech separation techniques are applicable to a wider range of situations than previously understood.
Implications for future research
Model Adaptation and Development: Whether the hypothesis is tested or invalidated, the results of the study may prompt further research to improve speech separation models for children. For models with degraded performance, research could focus on understanding the specific challenges posed by children's speech and how models can be adapted to overcome these challenges.
Interdisciplinary applications: The findings may also facilitate interdisciplinary research, such as the application of speech separation technology in the treatment of speech disorder.
Results for hypothesis validation: Confirmation that SepFormer's performance degrades as it is applied to children's speech will emphasize the need to improve the model or customize the solution for children's speech processing. This has practical implications for designing more effective ASR systems

for educational tools, child-robot interactions, and assistive technology for children with speech disorders. This will emphasize the importance of customizing speech recognition technology to meet the unique characteristics of children's speech.

Consequences of hypothesis invalidation: If the hypothesis is invalidated, indicating that SepFormer can effectively handle children's speech, this will challenge current assumptions about the limitations of ASR technology in younger age groups. This would extend the range of applications of existing speech separation models and encourage their integration into more child-focused applications, potentially improving accessibility and quality of interaction for child users.

References

Bhardwaj, V., Ben Othman, M. T., Kukreja, V., Belkhier, Y., Bajaj, M., Goud, B. S., Rehman, A. U., Shafiq, M., & Hamam, H. (2022). Automatic speech recognition (asr) systems for children: A systematic literature review. *Applied Sciences*, *12*(9), 4419.

Bhardwaj, V., Kadyan, V., et al. (2020). Deep neural network trained punjabi children speech recognition system using kaldi toolkit. *2020 IEEE 5th international conference on computing communication and automation (ICCCA)*, 374–378.

Cristia, A., Ganesh, S., Casillas, M., & Ganapathy, S. (2018). Talker diarization in the wild: The case of child-centered daylong audio-recordings. *Interspeech 2018*, 2583–2587.

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W. T., & Rubinstein, M. (2018). Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*.

Gray, S. S., Willett, D., Lu, J., Pinto, J., Maergner, P., & Bodenstab, N. (2014). Child automatic speech recognition for us english: Child interaction with living-room-electronic-devices. *WOCCI*, 21–26.

Grzybowska, J., & Kacprzak, S. (2016). Speaker age classification and regression using i-vectors. *INTERSPEECH*, 1402–1406.

Hershey, J. R., Chen, Z., Le Roux, J., & Watanabe, S. (2016). Deep clustering: Discriminative embeddings for segmentation and separation. *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 31–35.

Holliday, J. J., Reidy, P. F., Beckman, M. E., & Edwards, J. (2015). Quantifying the robustness of the english sibilant fricative contrast in children. *Journal of Speech, Language, and Hearing Research*, *58*(3), 622–637.

Isik, Y., Roux, J. L., Chen, Z., Watanabe, S., & Hershey, J. R. (2016). Single-channel multi-speaker separation using deep clustering. *arXiv preprint arXiv:1607.02173*.

Kathania, H., Kadiri, S., Alku, P., & Kurimo, M. (2021). Spectral modification for recognition of children's speech under mismatched conditions. *Nordic Conference on Computational Linguistics*, 94–100.

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., Senft, E., & Belpaeme, T. (2017). Child speech recognition in human-robot interaction: Evaluations and recommendations. *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, 82–90.

Li, F. (2012). Language-specific developmental differences in speech production: A cross-language acoustic study. *Child development*, *83*(4), 1303–1315.

Liao, H., Pundak, G., Siohan, O., Carroll, M. K., Coccaro, N., Jiang, Q.-M., Sainath, T. N., Senior, A., Beaufays, F., & Bacchiani, M. (2015). Large vocabulary automatic speech recognition for children. *Sixteenth Annual Conference of the International Speech Communication Association*.

Luo, Y., & Mesgarani, N. (2018). Tasnet: Time-domain audio separation network for real-time, single-channel speech separation. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 696–700.

Mostow, J. (2012). Why and how our automated reading tutor listens. *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training (ISADEPT)*, 43–52.

Potamianos, A., Narayanan, S., & Lee, S. (1997). Automatic speech recognition for children. *Fifth European Conference on Speech Communication and Technology*.

Sattorovich, E. Z. (2022). Psychological influence of speech disorders and the causes that cause them on the child's psyche. *Academicia Globe*, *3*(01), 39–42.

Shivakumar, P. G., & Georgiou, P. (2020). Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. *Computer speech & language*, *63*, 101077.

Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., & Zhong, J. (2021). Attention is all you need in speech separation. *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 21–25.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vincent, E., Gribonval, R., & Févotte, C. (2006). Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, *14*(4), 1462–1469.

Wang, X., Du, J., Cristia, A., Sun, L., & Lee, C.-H. (2020). A study of child speech extraction using joint speech enhancement and separation in realistic conditions. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7304–7308.

Wang, X., Du, J., Sun, L., Wang, Q., & Lee, C.-H. (2018). A progressive deep learning approach to child speech separation. *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 76–80.

Yeung, G., & Alwan, A. (2018). On the difficulties of automatic speech recognition for kindergarten-aged children. *Interspeech 2018*.

Yeung, G., & Alwan, A. (2019). A frequency normalization technique for kindergarten speech recognition inspired by the role of f0 in vowel perception. *Interspeech 2019*.

Zeghidour, N., & Grangier, D. (2021). Wavesplit: End-to-end speech separation by speaker clustering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 2840–2849.

Zeng, Y., & Zhang, Y. (2007). Robust children and adults speech classification. *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, *4*, 721–725.

This document was compiled April 29, 2024.

References

Bhardwaj, V., Ben Othman, M. T., Kukreja, V., Belkhier, Y., Bajaj, M., Goud, B. S., Rehman, A. U., Shafiq, M., & Hamam, H. (2022). Automatic speech recognition (asr) systems for children: A systematic literature review. *Applied Sciences*, *12*(9), 4419.

Bhardwaj, V., Kadyan, V., et al. (2020). Deep neural network trained punjabi children speech recognition system using kaldi toolkit. *2020 IEEE 5th international conference on computing communication and automation (ICCCA)*, 374–378.

Cristia, A., Ganesh, S., Casillas, M., & Ganapathy, S. (2018). Talker diarization in the wild: The case of child-centered daylong audio-recordings. *Interspeech 2018*, 2583–2587.

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W. T., & Rubinstein, M. (2018). Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*.

Gray, S. S., Willett, D., Lu, J., Pinto, J., Maergner, P., & Bodenstab, N. (2014). Child automatic speech recognition for us english: Child interaction with living-room-electronic-devices. *WOCCI*, 21–26.

Grzybowska, J., & Kacprzak, S. (2016). Speaker age classification and regression using i-vectors. *INTERSPEECH*, 1402–1406.

Hershey, J. R., Chen, Z., Le Roux, J., & Watanabe, S. (2016). Deep clustering: Discriminative embeddings for segmentation and separation. *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 31–35.