university of groningen

campus fryslân

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

Wenjun Meng

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

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

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Associate Prof. Dr. Matt Coler (Voice Technology, University of Groningen)

Wenjun Meng (S5613329)

June 9, 2024

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Associate Prof. Dr. Matt Coler, for his invaluable insights and patient guidance throughout my thesis. His expertise and thoughtful mentoring have been crucial in shaping both the direction and success of my work. I am also grateful to Dr. Phat Do, Assistant Prof. Dr. Vass Verkhodanova, Assistant Prof. Dr. Shekhar Nayak, and Assistant Prof.Dr. J.K. (Joshua) Schäuble for their supportive teaching and guidance. Additionally, I am thankful for the friends in class for their help and companionship, which have made this academic journey both enjoyable and enriching.

Special thanks to our study advisor, H. (Hieke) Hoekstra, for the assistance and encouragement she provided throughout this process.

I appreciate the support from the Center for Information Technology at the University of Groningen for providing access to the Hábrók high-performance computing cluster, which was essential for my research.

On a personal note, I am deeply grateful to my parents and my sister for their unwavering love and support. My thanks also go to my little niece who has been such a wonderful source of joy. Lastly, I would like to extend my thanks to my boyfriend, who has been a constant source of strength, encouragement, and love. Their support has been invaluable in helping me pursue and complete this journey.

Abstract

This thesis investigates the performance of SepFormer, a state-of-the-art speech separation model, in processing child speech, which has been less explored compared to adult speech. The study aims to evaluate the effectiveness of SepFormer in separating speech in datasets comprising child speech, with the hypothesis that SepFormer's performance will significantly decline due to the unique acoustic properties of child speech. The research utilizes the PhonBank database and employs evaluation metrics such as Scale-Invariant Signal-to-Noise Ratio and Signal-to-Distortion Ratio to assess performance. The findings are expected to highlight the need for recalibrating existing models or developing child-specific speech separation models. This investigation is crucial for advancing automatic speech recognition systems, ensuring they are inclusive and effective in educational and communicative contexts for children.

Keywords: Sepformer, Child Speech Separation, Automatic Speech Recognition, Speech Processing, Speech Separation

Contents

1 Introduction **6**
1.1 Research Question and Hypothesis 7

2 Literature Review **8**
2.1 Research on the Acoustic Differences Between Child and Adult Speech 8
2.2 Research on Adult Speech Separation Models 9
2.3 Research on the Sepformer Model 9
2.4 Research on Child Speech Separation 10
2.5 Research Synthesis 10

3 Methodology **12**
3.1 Data Collection 12
3.2 Preprocessing 12
3.3 Model Configuration 13
3.4 Evaluation Metrics 13
3.5 Ethical considerations 14

4 Experimental Setup **15**
4.1 Description of the Sepformer Model 15
4.2 Dataset Preparation 15
4.3 Experiment: 20 hours of dataset 17

5 Results **19**
5.1 Performance metrics 19
5.2 Statistical Analysis 19
5.3 Comparative Analysis 20

6 Discussion **24**
6.1 Summary of Main Findings 24
6.2 Interpretation of Results 24
6.3 Comparison with Previous Research 25
6.4 Practical Implications 26
6.5 Limitations and Future Research 26
6.6 Final Conclusion 27

References **29**

references.bib **31**

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). This is of particular importance in the context of automatic speech recognition (ASR) and voice-activated systems.

In recent years, there have been notable advancements in the field of speech separation. One of the promising techniques involves the use of Recurrent Neural Networks (RNNs) due to their ability to model temporal dependencies in audio signals (Luo, Z. Chen, and Yoshioka 2020; Xu et al. 2021). However, RNNs are less effective at recognizing patterns over extended periods, which is a crucial aspect of speech processing. The transformer model, introduced by Vaswani et al. (2017), addresses this shortcoming with its attention mechanism that can manage long-range dependencies. Building on the strengths of transformers, Subakan et al. (2021) introduced the SepFormer, a new transformer-based architecture designed for speech separation. This model has shown improved performance over traditional RNN approaches and offers the added benefits of better parallel processing and reduced computational requirements.

While significant advances have been made in adult speech separation, child speech separation has received comparatively less attention despite its potential benefits, particularly in the early identification and treatment of speech disorders in children(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 (Yeung and Alwan 2019; Mostow 2012)

The unique characteristics of child speech, such as higher fundamental and formant frequencies due to smaller vocal tracts, slower and more variable speaking rates, and differences in vocal effort and spontaneity (Potamianos, Narayanan, and Lee 1997), present distinct challenges compared to adult speech separation.

This study aims to evaluate the performance of SepFormer, a state-of-the-art speech separation model leveraging the Transformer architecture (Subakan et al., 2021), on child speech datasets. Despite its impressive performance on benchmarks like WSJ0-2mix and WSJ0-3mix (Isik et al., 2016), which consist of adult speech, SepFormer's effectiveness in separating child speech remains unexplored, potentially revealing a performance gap when confronted with the unique acoustic and linguistic properties of child speech.

The current research utilizes the child phonology dataset, PhonBank (Holliday et al. 2015; Edwards and Beckman 2008), as a testbed for the SepFormer model. The performance of SepFormer on this dataset is evaluated using SI-SNR and SDR metrics (Vincent, Gribonval, and Févotte 2006). Both SI-SNR and SDR are common metrics used in audio signal processing to assess the quality of a processed signal, while SI-SNR compares the level of the desired signal to the level of the noise or interference present in an audio signal after processing, and SDR compares the level of the desired signal to the level of all other unwanted components distortions that are present, which includes noise, interference, and distortions introduced. Both SI-SNR and SDR are expressed in decibels (dB), and higher values indicate better performance of the audio processing system. By utilizing the same evaluation metrics as the vanilla model, this research seeks to provide insights into its capability to handle the variability and complexity of child speech.

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 thesis is structured as follows: In section 2, I provide a brief literature review of existing research relevant to the topic. In section 3, the methodology is introduced, including data collection, preprocessing, model configuration, evaluation metrics, and ethical considerations. In section 4, the experiment setup is explained, detailing data preparation and the experiment procedure. In section 5, the results of the experiment are presented, including statistical and comparative analysis. Finally, in section 6, I summarize the main findings, interpret the results, compare them with previous research, describe the impact of my research, discuss the limitations of the research, and suggest future research. I then conclude with a final conclusion.

1.1 Research Question and Hypothesis

The introduction has highlighted significant progress 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 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 effectiveness decrease, necessitating tailored adaptations or the development of new child-centric models.

In light of the preceding discussion, the research question at the core of this study can be formulated as follows:

> **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 provide the foundation for an empirical investigation that could either validate this assumption or else demonstrate the adaptability of SepFormer across diverse age demographics.

2 Literature Review

This section provides a review of research relevant to child speech separation using the SepFormer model. By carefully reviewing what has been published in this area, the goal is to understand better and share knowledge about separating speech in children.

The literature review is organized to guide the reader through various key aspects systematically. First, Section 2.1 explores the acoustic differences between child and adult speech, highlighting the unique challenges these differences pose for speech separation models. Next, Section 2.2 reviews the development and performance of various speech separation models, focusing on traditional and SOTA approaches with an emphasis on adult speech. Section 2.3 delves into the specifics of the SepFormer model, detailing its architecture, advantages, and performance on standard datasets. Following this, Section 2.4 examines previous research efforts specifically addressing child speech separation, discussing the methodologies and findings of these studies. Finally, Section 2.5 synthesizes the reviewed literature to identify gaps and limitations, thereby justifying the need for the current study on SepFormer's performance with child speech. This structured approach ensures a clear and logical progression, culminating in a strong connection to the research question and hypothesis.

2.1 Research on the Acoustic Differences Between Child and Adult Speech

Kathania et al. (2021) have pinpointed the inherent acoustic variabilities in children's speech as a principal factor leading to decreased accuracy in child speech recognition systems. These variabilities are particularly evident in the spectral characteristics of children's voices, which exhibit higher fundamental and formant frequencies, as well as increased spectral variability. Notably, children under the age of 10 demonstrate greater within-vowel spectral variability than adults (Lee, Potamianos, and Narayanan 1999).

Temporal features and speech segment durations also reflect developmental differences. A spectrogram analysis of speech segment durations across three age groups—4, 6, and 12 years by Kent and Forner (1980) revealed that 4-year-olds generally had longer segment durations and greater variability in these durations compared to adults and older children. Additionally, children's speech is marked by a higher variability in speaking rate, vocal effort, and spontaneity (Gerosa et al. 2009).

The observed discrepancies can be attributed to both physiological and linguistic factors. Kathania et al. (2021) identified the smaller size of vocal folds and the shorter vocal tracts in children as the primary reasons for their higher formant and fundamental frequencies. Lee, Potamianos, and Narayanan (1999) suggested that children younger than 10 years have not yet fully developed stable articulatory targets for vowels. Moreover, Shivakumar and Georgiou (2020) and Bhardwaj et al. (2022) highlighted the ongoing evolution of children's vocal mechanisms and language skills, which encompasses a developing command of prosodic elements such as pitch, volume, rhythm, and intonation. This is further evidenced by the presence of additional words and premature phonations in children's speech at significantly higher levels than in adults (Strommen and Frome 1993).

Interestingly, the research findings indicate that there is a greater degree of intra-speaker variability in young children, particularly those under 10 years (Gerosa et al. 2009). This suggests that speech separation tasks may be more straightforward for children's speech than for adult speech, providing a potential counterpoint to the hypotheses proposed in this research.

Nonetheless, the divergent characteristics between child and adult speech inherently suggest potential negative implications for the performance of ASR models that are primarily trained on adult

speech data. 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.

In conclusion, children's speech has distinct acoustic characteristics and greater variability due to physiological and linguistic factors, which poses a challenge for ASR models trained primarily on adult speech. In this study, we will investigate whether the SepFormer model, which performs well in adult speech segregation, shows significant performance degradation in child speech segregation.

2.2 Research on Adult Speech Separation Models

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.

While models such as TasNet, WaveSplit, and SepFormer have made significant progress in adult speech separation, it remains critical to study how these models perform on child speech.

2.3 Research on the Sepformer Model

While the SepFormer has demonstrated exceptional performance on the wsj0-2mix and wsj0-3mix datasets consisting of speech in quite conditions, its effectiveness degrades in the presence of noise and reverberation (Ho, J.-w. Hung, and B. Chen 2022). To improve the robustness of the SepFormer under different acoustic conditions, the authors introduced a dual-encoder system with different time resolutions, coupled with a bi-projection fusion (BPF) module to merge information from both the time and frequency domains. Despite this innovation, the BPF module can sometimes complicate the learning process of the mask estimator, especially in simpler acoustic environments.

Later, Ho, J.-W. Hung, and B. Chen (2023) developed ConSep, a framework that conditions the magnitude spectrogram to avoid domain mismatch or confusion. This approach aims to ensure consistent performance even when the audio environment changes.

In addition, research by Yip et al. (2023) has shown that SepFormer's performance can be significantly affected by emotional content in speech, with degradation of up to 5.1 dB in SI-SDRi for

mixtures containing strong emotions.

In summary, while the SepFormer model performs well in clean speech conditions, performance degrades in noisy and emotional speech environments. This study will investigate whether the effectiveness of SepFormer is significantly reduced when applied to children's speech than when applied to adult speech, thereby assessing its robustness in different age populations.

2.4 Research on Child Speech Separation

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. One of the primary challenges has been the scarcity of children's speech data. To address this, some researchers have turned to data augmentation using text-to-speech technology, which has shown promise in improving the performance of children's ASR systems (W. Wang et al. 2021). Cristia et al. (2018) delved into the complexities of talker diarization within child-centric audio recordings, emphasizing the difficulties presented by spontaneous dialogues occurring in a range of acoustic settings.

In an effort to distinguish between child and adult speech, Zeng and Zhang (2007) developed an innovative speech classification system utilizing Gaussian Mixture Models (GMM). This system combines speech features such as pitch and the first three formants to model the distinct characteristics of children's and adults' speech. X. Wang, Du, Sun, et al. (2018) introduced a progressive learning method that utilizes a densely connected Long Short-Term Memory (LSTM) network for child-adult speech separation, operating independently of the speaker's identity. This model was later refined to increase its robustness under realistic conditions, integrating speech enhancement and separation techniques specifically tailored for extracting child speech (X. Wang, Du, Cristia, et al. 2020).

Efforts have also been directed toward integrating ASR for children into interactive educational devices. Gray et al. (2014) presents a child specific LVCSR system that improves the accuracy for children speaking US English to interacting electronic devices. 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.

Recently there has been a growing inclination towards transfer learning, a technique commonly used in adult speech recognition. Shivakumar et al. (2020) tackled the principal challenges using transfer learning from adult models to child models within a Deep Neural Network (DNN) framework for children's ASR. Furthermore, Rolland et al. (2022) demonstrated the effectiveness of a two-step training strategy for children's speech, which begins with multilingual learning and is followed by language-specific transfer learning, which has been shown to surpass the performance of conventional single language/task training methods.

Research on child speech separation highlights the limitations of current ASR systems with child speech, underscoring the need for exploring SepFormer's performance in this area.

2.5 Research Synthesis

The literature review has highlighted distinct acoustic differences between child and adult speech, which are of critical importance in the context of speech recognition and separation systems. Studies such as Kathania et al. (2021) and Lee, Potamianos, and Narayanan (1999) have detailed the higher fundamental and formant frequencies and increased spectral variability present in children's speech.

These differences are not only acoustic but also temporal, with children exhibiting longer segment durations and greater variability in these durations, as well as in speaking rate and vocal effort (Gerosa et al. 2009; Kent and Forner 1980). Such variability, due to physiological and linguistic developmental factors, presents unique challenges for ASR systems, which are often trained on adult speech data and therefore underperform when applied to child speech (Bhardwaj et al. 2022; Kennedy et al. 2017).

While recent advancements in adult speech separation models have been significant, with innovations such as TasNet and WaveSplit (Luo and Mesgarani 2018; Zeghidour and Grangier 2021), and the transformer model's attention mechanisms (Vaswani et al. 2017), these have predominantly been focused on adult speech. The SepFormer model, a Transformer-based architecture, has shown superior performance in adult speech separation tasks (Subakan et al. 2021). However, its effectiveness is challenged in noisy and reverberant conditions, and its performance is influenced by the emotional content of speech (Ho, J.-w. Hung, and B. Chen 2022; Ho, J.-W. Hung, and B. Chen 2023; Yip et al. 2023).

Research on child speech separation is less mature, with the scarcity of children's speech data being a significant hurdle (W. Wang et al. 2021; Cristia et al. 2018). Innovations like the use of Gaussian Mixture Models for speech classification (Zeng and Zhang 2007) and progressive learning methods for speech separation (X. Wang, Du, Sun, et al. 2018) have been developed, but the field lacks a robust, child-specific model that can handle the acoustic complexities of child speech. Efforts to integrate ASR in educational devices for children (Gray et al. 2014; Liao et al. 2015) and the application of transfer learning techniques (Shivakumar and Georgiou 2020; Rolland et al. 2022) indicate a move towards more child-focused approaches.

Despite significant advancements in speech separation technologies and the demonstrated efficacy of models like SepFormer on adult speech datasets, the performance of these models on child speech remains underexplored. The unique acoustic and linguistic characteristics of child speech pose distinct challenges that current models, primarily trained on adult speech, may not adequately address. The reviewed literature underscores a critical gap: the lack of studies evaluating SOTA speech separation models on child speech datasets.

Furthermore, while some studies have begun to explore adaptive techniques and TL to improve ASR systems for children, there is a need for focused research on speech separation models specifically tailored for child speech.

This study aims to fill this gap by evaluating the performance of the SepFormer model on child speech datasets. By leveraging the PhonBank database and employing rigorous evaluation metrics such as SI-SNR and SDR, this research will provide insights into the adaptability of SepFormer to the unique challenges of child speech. The findings will inform future developments in speech separation models by revealing performance degradation, or highlight the model's robustness and versatility across different age groups and speech characteristics.

3 Methodology

In this section, I will outline the methodology used to address the research question and validate the hypothesis on a high-level. First, subsection 3.1 details the dataset selected for model evaluation. This is followed by subsection 3.2, which describes the techniques applied in data preprocessing. Subsequently, subsection 3.3 provides an overview of the model configuration and its implementation. The evaluation approach and metrics used are then explained in subsection 3.4. Lastly, subsection 3.5 considers the ethical implications associated with this study.

3.1 Data Collection

Speech samples will be collected from the PhonBank database(Edwards and Beckman 2008), a comprehensive child phonology corpus. This study will specifically select English language samples from children aged four to five years.

3.2 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, silences longer than 1000 milliseconds have been removed. This parameter ensures that shorter pauses, which contribute to the natural rhythm of speech, are preserved. Additionally, a silence threshold has been set at -80 dBFS to remove audio segments quieter than this level.

```
min_silence_len = 1000
silence_thresh = -80
for subdir, dirs, files in os.walk(source_root_dir):
    for filename in files:
        if filename.endswith(".mp3"):
            # Construct the full file path
            audio_path = os.path.join(subdir, filename)
            audio = AudioSegment.from_file(audio_path)
            chunks = silence.split_on_silence(audio,
                                              min_silence_len=min_silence_len,
                                              silence_thresh=silence_thresh)

            processed_audio = AudioSegment.empty()
            for chunk in chunks:
                processed_audio += chunk
```

Volume normalization is another essential process in the preparation of the dataset. To mimic real-life variations in speaker volume and to maintain consistency with the SepFormer's original

dataset preprocessing (Subakan et al. 2021), 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 20 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.

3.3 Model Configuration

The SepFormer model will be implemented as described by its developers, with no modifications to the architecture. The testing will be conducted with the same hyperparameters as the baseline model to maintain consistency.

```
model = Separator.from_hparams(source="speechbrain/sepformer-wsj02mix",
                               savedir='pretrained_models/sepformer-wsj02mix')

audio_files = glob.glob(os.path.join(mixture_base_dir, '*.wav'))

for file_path in audio_files:
    # Perform separation
    est_sources = model.separate_file(path=file_path)

    base_filename = os.path.basename(file_path).replace('.wav', '')
    output_filename_1 = f"{base_filename}_sep1.wav"
    output_filename_2 = f"{base_filename}_sep2.wav"
```

3.4 Evaluation Metrics

$$\begin{cases} \mathbf{s}_{\text{target}} = \dfrac{\langle \hat{\mathbf{s}}, \mathbf{s} \rangle \mathbf{s}}{||\mathbf{s}||^2} \\ \mathbf{e}_{\text{noise}} = \hat{\mathbf{s}} - \mathbf{s}_{\text{target}} \\ \text{SISNR} = 10\log_{10} \dfrac{||\mathbf{s}_{\text{target}}||^2}{||\mathbf{e}_{\text{noise}}||^2} \end{cases}$$

Figure 1: SI-SNR Formula

$$\text{SDR} := 10\log_{10}\left(\frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}\|^2}\right)$$

Figure 2: SI-SNR Formula

This research is going to use the same two metrics that are used on Sepformer, SI-SNR and Signal-to-Distortion Ratio (SDR)(Vincent, Gribonval, and Févotte 2006). 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.

3.5 Ethical considerations

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 employed is Phonbank, a component of the TalkBank system (Edwards and Beckman 2008), which investigates phonological development in children and its effects on language learning across different languages. All participants in the dataset have given their consent, and data collection has been conducted in accordance with the guidelines of the Institutional Review Board (IRB). The resulting corpus is shared under a Creative Commons license that allows for non-commercial use.

In order to deidentify the audio data, specific measures were employed, including the removal of names and addresses from the audio tracks. This was guided by the occurrence of the terms "Lastname" and "Address" in the transcripts, ensuring that personal identifiers were not included in the study's data. Additionally, the dataset complies with the General Data Protection Regulation (GDPR), which provides a robust framework for data privacy and security.

The preprocessing steps taken to align with the baseline model training dataset include the removal of long silent gaps, the splitting of audio files into 5-second segments, the normalization of volume, and the conversion of the sampling rate to 8 kHz. These steps not only prepare the data for the experiment but also enhance data privacy. The processed data is securely stored in a folder as part of the Research Package, which will then be uploaded and electronically archived in the University of Groningen Thesis Repository.

Objective metrics were employed for evaluation without the need for subjective methods involving human participants, thus there are no concerns regarding the ethics of involving human participants or any other issues that do not align with the ethics of the faculty.

This concludes the methodology section, outlining the methods used in this study; the next section will provide more details into the experimental setup.

4 Experimental Setup

To address the research question ”What is the difference in the performance, measured by SI-SNR and SDR metrics, of the SepFormer model in separating child speech from adult speech using standardized data sets?”, the underlying hypothesis is that there will be a noticeable decrease in the performance of the SepFormer model, as measured by SI-SNR and SDR metrics, when separating child speech compared to adult speech from standardized datasets.

The following sections are organized as follows: A brief overview of the SepFormer model, which is the focus of this study, is presented in section 4.1. This is followed by a detailed description of the preprocessing steps applied to the dataset in preparation for the experiments, which are explained in section 4.2. Finally, the experimental procedures and evaluation methods are described in section 4.3.

4.1 Description of the Sepformer Model

The SepFormer is a novel RNN-free transformer-based neural network for speech separation. Designed to capture both short and long-term dependencies within speech signals, the SepFormer employs a multi-scale strategy that uses the power of transformers to effectively separate overlapping speech.

The SepFormer has achieved state-of-the-art (SOTA) performance on the widely recognized WSJ0-2mix and WSJ0-3mix datasets, which are benchmarks in the field of speech separation, as a testament to its innovative design. It has demonstrated remarkable performance, achieving a scale-invariant signal-to-noise ratio (SI-SNR) improvement of 22.3 dB on the WSJ0-2mix dataset and a SI-SNRi of 19.5 dB on the WSJ0-3mix dataset.

One of the key advantages of the SepFormer is its ability to take advantage of the inherent parallelization benefits of the Transformer architecture. This allows for more efficient computation, making the SepFormer not only faster, but also less demanding in terms of memory usage compared to other contemporary systems that deliver similar levels of performance. As a result, SepFormer stands out as a highly efficient and effective solution to the task of speech separation.

The SepFormer model employs 256 convolutional filters with a kernel size of 16 samples and a stride of 8 samples. The model processes data in chunks of size 250 with 50% overlap, employs eight Transformer layers in both intra- and inter-chunk processing, and repeats this dual-path pipeline twice. The model comprises eight parallel attention heads, 1024-dimensional feed-forward networks, and contains 26 million parameters. The model employs dynamic mixing and speed perturbation for data augmentation, and is trained with the Adam optimizer, gradient clipping, and automatic mixed-precision over 200 epochs.

The SepFormer is accessible within the SpeechBrain toolkit[1].

4.2 Dataset Preparation

The dataset used in this study comes from the English subset of the Paidologos project, a cross-linguistic study of elicited phonological forms. This dataset features an equal distribution of male

[1] speechbrain.github.io/

and female speakers across four age groups: 2, 3, 4, and 5 years old, with each group consisting of 10 individuals of each gender.

Given the time constraints and the findings of Yeung and Alwan (2018), which suggest that ASR systems may achieve higher accuracy with training data derived from older children, this research has chosen to focus on the 4- to 5-year-old group. This decision is based on the expectation that data from this age group will serve as the most effective training material. In addition, this focus aims to demonstrate the influence of age on the performance of ASR systems.

A series of preprocessing steps were applied to the child speech dataset in order to ensure compatibility with the Sepformer model and to optimize the quality of the input data for speech separation tasks. The following subsections provide a detailed account of each step in the preprocessing pipeline.

Change in Sampling Rate

The Sepformer model requires input audio at a sampling rate of 8 kHz. However, the original recordings in the dataset were sampled at 16 kHz. To reconcile this discrepancy, the first preprocessing step involved downsampling the audio files to the required 8 kHz rate, thereby aligning them with the model's specifications.

Removal of Silent Gaps

Silent intervals within the original audio can have a detrimental impact on the performance of the speech separation model. To address this, silent gaps exceeding 1000 milliseconds were excised from the recordings. Furthermore, a silence threshold of -80 dBFS was set to eliminate audio segments quieter than this level. This threshold was chosen to effectively reduce background noise while preserving the subtle qualities of child speech, which often exhibits lower volume and greater variability than adult speech.

Manual Inspection

A manual review was conducted to identify and exclude any anomalous recordings. This included recordings with additional speakers, such as a teacher's voice, and those with exceedingly low volume levels. Such audios were removed to maintain the integrity of the dataset and to ensure that only the target child speech was present.

Segmentation into Five-Second Clips

In order to ensure the reliability of the speech separation performance evaluation, it was necessary to maintain consistency with the Sepformer's test set samples, which range from 2 to 6 seconds in length. This was achieved by segmenting the dataset's audio files into 5-second clips.

Metadata Generation

A metadata file was created to facilitate the subsequent audio mixing process. This file serves as a reference for pairing audio clips to generate mixtures. To prevent redundancy, a duplicate check was incorporated, ensuring the uniqueness of each audio pair.

A snippet for generating a 2-speaker mixed metadata file:

```
with open('5_mix_2_spk_tr.txt', 'w') as metadata_file:
    while line_count < max_lines and len(audio_files) >= 2:
        selected_files = random.sample(audio_files, 2)
        positive_level = generate_positive_level()
        levels = [positive_level, -positive_level]
        random.shuffle(levels)  # Shuffle the levels to assign them randomly
```

```
        for file, level in zip(selected_files, levels):
            metadata_file.write(f"{file} {level} ")
        metadata_file.write("\n")

        line_count += 1
```

Audio Normalization and Uniform Sampling
The audios were first normalized to -1 dBFS to prevent clipping during processing or playback. This also preserves the dynamic range of the audio by preventing peaks from causing digital distortion. In accordance with the Sepformer's training and testing protocols, the volume levels of the sources were uniformly adjusted to fall between 0 dB and 5 dB prior to mixing. This normalization step was applied to each audio path listed in the metadata to maintain consistent relative levels.

The snippet for uniform sampling:

```
intended_loudness_1 = loudness_1 + level1
intended_loudness_2 = loudness_2 + level2

# the intended difference in dB after adjustments
intended_difference_in_dB = abs(intended_loudness_1 - intended_loudness_2)
max_difference = 5.0

# Check if the intended difference is more than the maximum allowed
if intended_difference_in_dB > max_difference:
    scale_factor = max_difference / intended_difference_in_dB
    # Apply the scaled adjustments
    adjusted_audio_1 = audio_1.apply_gain(level1 * scale_factor)
    adjusted_audio_2 = audio_2.apply_gain(level2 * scale_factor)
else:
    # Apply the intended levels from the metadata if within the max_difference
    adjusted_audio_1 = audio_1.apply_gain(level1)
    adjusted_audio_2 = audio_2.apply_gain(level2)
```

Audio Mixing
The prepared audios were then mixed to simulate multi-speaker environments, creating 2-speaker and 3-speaker mixtures. These mixtures formed the test set for the Sepformer model, designed to test the model's ability to separate individual child speakers from the composite audio.

4.3 Experiment: 20 hours of dataset

Sepformer Inference
For each age critiera and models for 2 speakers and 3 speakers, there are 5 hours of audios respectively, so in total there are 20 hours of recordings. With the test set prepared, inference was conducted by loading pre-trained models that had been previously trained on the wsj0-2mix and wsj0-3mix datasets to separate the mixed speech into individual audio streams corresponding to each child speaker.

Performance Metrics Calculation
The performance of the Sepformer was quantified by calculating the Scale-Invariant Signal-to-Noise Ratio (SI-SNR) and Signal-to-Distortion Ratio (SDR)[2] improvement metrics using the torchmetrics package. Due to the possibility that the separated audio may not match its original sources, the Permutation Invariant Training (PIT) metrics[3] from the same package was used. PIT allows for a permutation invariant evaluation of the model's performance in multi-speaker speech separation tasks.

SI-SNRi calculation for 2-speaker scenario:

```
mixture_audio = load_audio_tensor(mixture_path)

# Load the separated audios and stack them to match the shape (batch, spk, time)
separated_audios = torch.stack([load_audio_tensor(p) for p in separated_paths])
clean_audios = torch.stack(
    [load_audio_tensor(os.path.join(clean_dir, speaker_id[4:6],
    f"{speaker_id}.wav")) for speaker_id in speaker_ids])

separated_audios = separated_audios.unsqueeze(0)  # Add a batch dimension
clean_audios = clean_audios.unsqueeze(0)
mixture_audio = mixture_audio.unsqueeze(0).unsqueeze(0)

# Calculate PIT-si_snr for the separated audio
separated_pit_si_snr = pit(separated_audios, clean_audios)
mixture_pit_si_snr= pit(mixture_audio.repeat(1,
                        len(speaker_ids), 1), clean_audios)

# Calculate si_snr improvement (si_snr)
si_snr_improvement = separated_pit_si_snr.item() - mixture_pit_si_snr.item()
si_snr_values.append(si_snr_improvement)
```

Post-Inference Checks
Following the separation process, a manual inspection was performed to identify any outliers in the results. Specifically, instances where all two or three separated audios returned negative values or were significantly lower than expected were flagged for further investigation.

The experimental setup outlines the process for evaluating SepFormer's performance on child speech datasets. By leveraging the PhonBank database and implementing preprocessing steps, this study ensures that the data reflects the unique characteristics of child speech. The detailed configuration of the SepFormer model and the use of robust evaluation metrics such as SI-SNR and SDR provide a foundation for assessing the model's effectiveness. These preparations are crucial for answering the RQ and testing the H regarding the model's performance. The subsequent section will present the results of these experiments, offering insights into the adaptability and efficacy of SepFormer in handling child speech separation.

[2]https://lightning.ai/docs/torchmetrics/stable/audio/signal$_d istortion_r atio.html$
[3]https://lightning.ai/docs/torchmetrics/stable/audio/permutation$_i nvariant_t raining.html$

5 Results

The objective of this study is to assess the performance of the SepFormer model in separating child speech compared to adult speech, measured through SI-SNR and SDR metrics. The hypothesis is that the SepFormer's performance will demonstrate a significant reduction when applied to child speech datasets in comparison to its performance on adult speech datasets.

The performance of the SepFormer on speech from 4 to 5-year-old children is compared with its baseline performance on the WSJ0-2mix and WSJ0-3mix datasets. The WSJ0-2mix consists of mixtures of two speakers, and the WSJ0-3mix includes mixtures of three speakers. The comparative results are presented in Table 1 and Table 2.

5.1 Performance metrics

Dataset	SI-SNRi	SDRi	Test Set Size
WSJ0-2mix	**20.4**	**20.5**	5 hours
4-year-old	4.83	5.41	5 hours
5-year-old	2.39	3.08	5 hours

Figure 3: Baseline results on WSJ0-2mix compared to SepFormer's performance on the 4 to 5-year-old children's speech dataset.

Dataset	SI-SNRi	SDRi	Test Set Size
WSJ0-3mix	**17.6**	**17.9**	5 hours
4-year-old	3.06	4.07	5 hours
5-year-old	2.59	2.78	5 hours

Figure 4: Baseline results on WSJ0-3mix compared to SepFormer's performance on the 4 to 5-year-old children's speech dataset.

The tables provided offer a summary of the mean SI-SNRi and SDRi values for the child and adult datasets. The following section will provide further statistical information, including medians, standard deviations, and ranges to provide a comprehensive overview of the central tendencies and dispersions within the data.

5.2 Statistical Analysis

The box plot clearly reveals the distribution of the data, offering insights into the performance of the SepFormer model under different conditions. In the two-speaker condition among 4-year-olds, the median SI-SNRi (Scale-Invariant Signal-to-Noise Ratio improvement) was 5 dB, with an interquartile range (IQR) from -0.5 dB to 10 dB. This wide distribution indicates a significant variability in

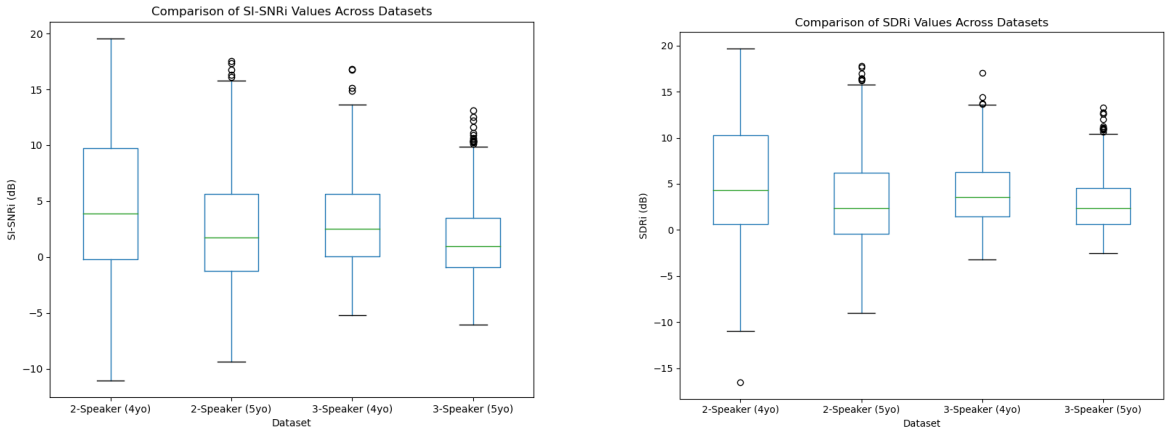

Figure 5: Comparison of SI-SNRi and SDRi Values Across Datasets

how well the model can separate speech. Some results show improvement up to 20 dB, while others decline to as low as -10 dB. Such variability suggests that while the model can be effective, its performance is inconsistent, particularly with younger children's speech, which can vary greatly in pitch and pronunciation.

In contrast, the 5-year-old group in the same two-speaker condition exhibited a slightly lower median SI-SNRi of 2 dB and a narrower IQR from 0 dB to 5 dB, indicating more consistent but overall less effective speech separation. Fewer outliers were observed, which suggests that the speech characteristics of 5-year-olds may be slightly easier for the model to handle compared to 4-year-olds.

In the three-speaker condition, the median SI-SNRi for 4-year-olds decreased to 3 dB, with an IQR of 0 dB to 5.5 dB. The presence of outliers above 14 dB and below -5 dB further highlights the challenges faced by the SepFormer model in more complex acoustic environments. For five-year-olds, the median SI-SNRi decreased further to 1 dB, with an IQR between -1 dB and 4 dB. This indicates a decline in performance as the number of speakers increased.

Similarly, the signal-to-distortion ratio improvement (SDRi) values followed a comparable pattern. For example, the median SDRi for the two-speaker group of four-year-olds was 5 dB, indicating that the average improvement in speech quality after separation was moderate but could vary considerably, as evidenced by outliers reaching 20 dB and dropping to -10 dB. The variability in SDRi, as well as that in SI-SNRi, highlights the necessity for the model to be further refined in order to more effectively handle the acoustic variability present in children's speech.

5.3 Comparative Analysis

SI-SNRi values in the two-speaker condition showed higher medians and wider IQRs for 4-year-olds compared with 5-year-olds. This shows that the model performs better when processing the speech of younger children. The 4-year-old group has a wider distribution of values and the presence of outliers, indicating greater variability in performance, sometimes with extremely high or very low SI-SNRi. In contrast, the 5-year-old group has a lower median and fewer outliers, indicating more stable but generally lower performance.

The median SI-SNRi decreased for both age groups when the number of speakers increased

to three, highlighting the increased difficulty in isolating more speakers. Medians were higher in the 4-year-old group than in the 5-year-old group, consistent with the trend observed in the two-speaker condition. However, the IQR narrowed and the number of outliers decreased, indicating that although the task was more challenging, the performance was more stable over a smaller range of improvements.

The changing trend of SDRi values is similar to that of SI-SNRi values. In the two-speaker condition, the 4-year-old group had a higher median SDRi and a wider IQR than the 5-year-old group. This again suggests better speech performance and greater variability in younger children. High outliers indicate that the model occasionally performs exceptionally well.

In the three-speaker condition, the median SDRi decreased for both age groups, reflecting the increased difficulty in isolating the three speakers. The 4-year-old group continued to outperform the 5-year-old group, with a higher median and slightly wider IQR. The consistency of the SI-SNRi and SDRi metrics further supports the conclusion that the model handles the speech of younger children more effectively, while the challenge of additional speakers has a consistent impact on performance across age groups.

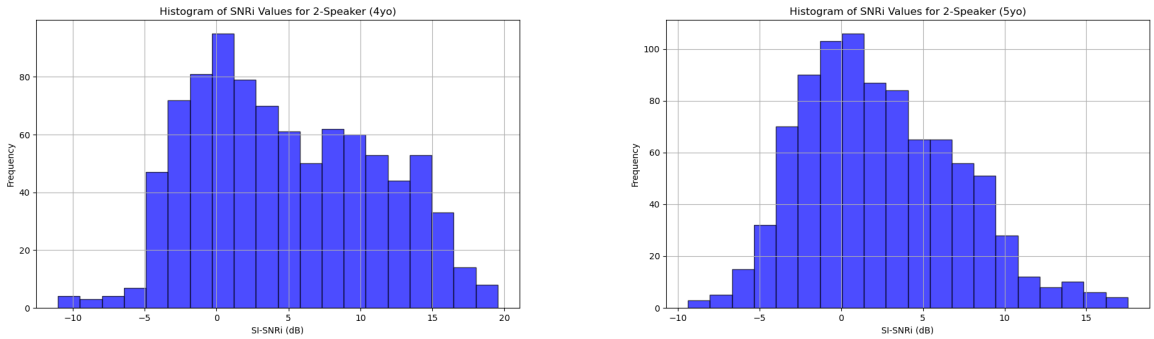

Figure 6: Histogram of SI-SNRi values for 2 Speakers

The histograms above illustrate the distribution of SI-SNRi values for the two-speaker cases. These distributions provide insight into the performance of the speech separation model in different age groups. By comparing these histograms, we can understand the model's performance with regard to age-related speech features.

The distribution of SI-SNRi values for the 2-speaker mixtures reveals a range from -10 dB to 20 dB for 4-year-olds, with a notable peak at 2.5 dB, indicating a higher variability in performance. This indicates that while there are instances of excellent performance (above 15 dB), the overall consistency varies. In contrast, the SI-SNRi values for five-year-olds are more tightly clustered around 0 to 5 dB, with a distinct peak at 2.5 dB, reflecting a more uniform performance with less spread.

This indicates that while the central tendencies are comparable for both age groups, with similar median performance levels, the 4-year-old group exhibits a broader spread, suggesting a greater diversity in response to the separation algorithm. This could be attributed to the fact that younger children exhibit a greater variety of speech patterns. The presence of negative SI-SNRi values in both groups, with a greater frequency in the 4-year-olds, indicates instances where the separation algorithm may have failed to enhance the signal quality or may have even degraded it.

In conclusion, the analysis indicates that the speech separation model tends to perform with greater consistency for five-year-olds than for four-year-olds in a two-speaker setting. The reduced variability and fewer extreme values in the older children's group indicate that the model's performance becomes somewhat more predictable and robust with increasing age within this young cohort.

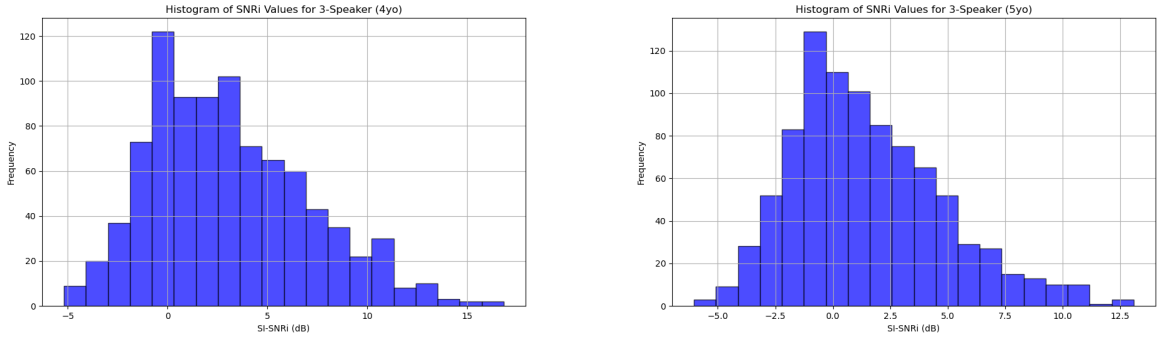

Figure 7: Histogram of SI-SNRi Values for 3 Speakers

In the 3-speaker scenario, the histograms for the 4- and 5-year-olds show a right-skewed distribution of SI-SNRi values, with the majority ranging from about 0 to 2.5 dB for the 4-year-olds and a slightly lower number, from about -1 to 2 dB, for the 5-year-olds. The range of SI-SNRi values for the 4-year-old group is from -5 dB to 15 dB, and the peak frequency is from 0 to 2.5 dB, suggesting a moderate expansion into the higher performance range. In contrast, the values for the 5-year-old group were more concentrated, ranging from -5 dB to 12.5 dB, with peak frequencies below 0 dB, suggesting a lower tendency toward concentration and fewer instances of high performance.

The smaller range of SI-SNRi values for both age groups compared to the 2-speaker case suggests less variability and a greater challenge. Notably, the histograms for both groups had fewer outliers in the higher performance range, emphasizing the difficulty of achieving superior separation as the number of speakers increases.

Although the concentration trend is slightly better in the 4-year-olds, the presence of negative SI-SNRi values in both age groups reveals that the separation process may degrade the signal. This problem was more pronounced in the 5-year-old group.

In summary, for both age groups, the performance of the model in the 3-speaker condition appeared to be reduced compared to the 2-speaker condition, with the 4-year-olds performing slightly better than the 5-year-olds. The decrease in variability and the reduction in high performance outliers highlight the serious challenges of the 3-speaker separation task. The SDRi histograms for both age groups and speaker scenarios show similar patterns to the SI-SNRi histograms, with some minor differences in spread and skewness.

In the 2-speaker scenario, the SDRi values for both the 4-year-old and 5-year-old age groups show that the mean and median are concentrated around the 0 to 5 dB range, indicating that the average performance of the speech separation model is not significantly different between the two age groups. However, the variance of the SDRi values is slightly higher for the 4-year-olds, indicating increased variability and consequently less consistency. In addition, the distributions for both age groups are right-skewed, indicating a tail of higher SDRi values. This skew is more pronounced for the 4-year-olds, suggesting occasional instances where the model achieved exceptional separation

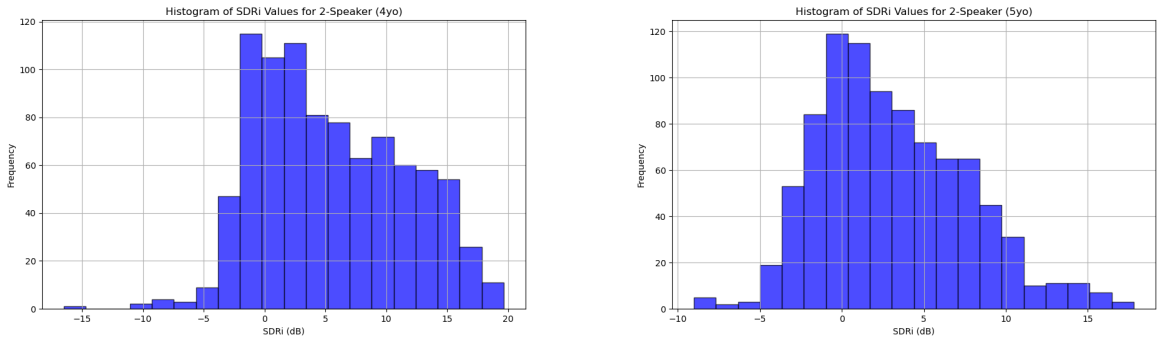

Figure 8: Histogram of SDRi Values for 2 Speakers

performance. This reflects the presence of outliers and underscores the potential for the model to perform well above average, although less frequently.

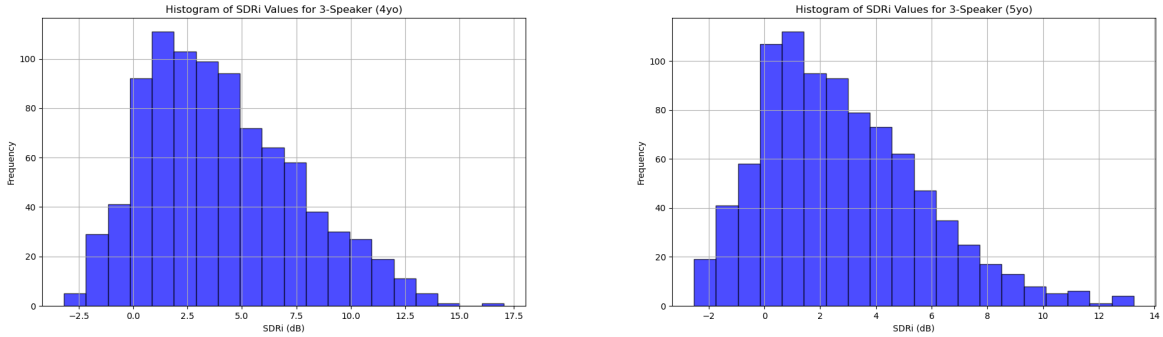

Figure 9: Histogram of SDRi Values for 3 Speakers

In the three-speaker situation, the SDRi values for the 4-year-old group showed a wider range from -2.5 dB to 17.5 dB, while the SDRi values for the 5-year-old group ranged from -2 dB to 14 dB. Both histograms peaked at around 2.5 dB. However, the right-skewed distribution observed in the histograms suggests that lower SDRi values are more common than higher SDRi values and higher SDRi values are less common in both groups. Notably, some outliers as high as 17.5 dB were seen in the 4-year-old group, suggesting occasional instances of superior model performance. In contrast, the 5-year-old group had fewer outliers, with a maximum value of about 14 dB, suggesting a more limited range of extreme performance.

Overall, both the SDRi and SI-SNRi histograms show that the performance of the model is generally consistent, while the 4-year group shows slightly better performance, it also shows more variability compared to the 5-year group.

6 Discussion

6.1 Summary of Main Findings

The focus of this study is to measure the difference in performance of the SepFormer model in separating children's speech from adult speech using a standardized dataset through the SI-SNR and SDR metrics. The hypothesis suggests that there will be a statistically significant decrease in SepFormer's performance on the children's speech dataset compared to the adult speech dataset. The experimental results clearly validate the above hypothesis, confirming a general decrease in the model's effectiveness.

These findings underscore the challenges that child speech presents to current speech separation models, in alignment with the anticipated negative result. The inherent complexity and variability of child speech, which differ significantly from adult speech, highlight the need for future speech processing models to adapt more effectively. This study not only confirms the hypothesis but also sets the stage for a comprehensive discussion on how these results answer the research question and hypothesis. It also explores the specific performance limitations and potential enhancements needed for broader applicability in real-world scenarios.

6.2 Interpretation of Results

The results of the SI-SNRi and SDRi analysis provide a detailed insight into the performance of the speech separation model on children's speech. In the two-speaker condition, the median performance of the model was better for 4-year-olds than for 5-year-olds, which may reflect the model's greater sensitivity to the acoustic characteristics of younger children's speech. However, in terms of wider IQRs and outliers, there was greater variability and a wider range of performance for this group of children, suggesting that the model was less successful in predicting younger children. This may be due to the inherent variability in the speech patterns of 4-year-olds, which may present both opportunities and challenges for the separation algorithm.

The median SI-SNRi was reduced for both age groups in all three speaker conditions, which highlights the model's difficulties in dealing with more complex auditory scenarios. In this case, the smaller the IQR, the fewer the outliers, suggesting that when faced with the task of separating more speakers, the model's performance is limited to a smaller effective range. Interestingly, the median value for 4-year-olds remains high, implying that the model maintains some relative advantage in processing the speech of younger children even as the task complexity increases.

The results of SDRi largely confirm those of SI-SNRi, with the group of 4-year-olds showing a trend towards better performance in the two-speaker condition. This suggests that the model occasionally achieves excellent separation quality, although this is uncommon. In the three-speaker condition, the median SDRi decreased for both groups, again highlighting the additional difficulties posed by the extra speakers.

The histogram analysis further confirmed these observations, with the distribution of SI-SNRi values for the two-speaker mixtures showing more pronounced changes for the 4-year-olds. For 5-year-olds, the clustering around the median was tighter, suggesting more stable model performance despite the generally lower level of separation quality.

These findings prompted a discussion about the balance between the ability of models to handle phonological variation and the consistency of model performance. In less complex situations, the

greater phonological variability in younger children may provide richer cues for separation, which the model can utilize to obtain higher SI-SNRi values. However, this advantage diminishes as the acoustic environment becomes more complex, as seen in the three-speaker condition.

In addition, both SI-SNRi and SDRi showed negative values across age groups and across conditions, which calls into question the robustness of the model. This suggests that in some cases the model may introduce distortion rather than clarifying the speech signal. This is of particular concern for younger age groups, where such negative effects are more prevalent.

It is commonly assumed that the performance of ASR improves with the age of the child speaker. However, it is notable that there is a decline in both the SI-SNRi and SDRi performance of 5-year-olds compared to 4-year-olds. One potential explanation for these differences in performance is the influence of phonological features and variability. For 4-year-olds, children's speech may still be in the early stages of development, with greater variability in pitch, rhythm, and articulation. While this variability can be challenging, it can also indicate a reduction in speech patterns and an increased ability of distinguishing between speakers. For 5-year-olds, their speech may become more consistent and closer to adult speech patterns. However, this consistency may incidentally lead to the emergence of more similar speech features, making it more difficult for the model to distinguish between speakers in separation tasks.

In conclusion, the findings validate the hypothesis to a significant degree, confirming a general decrease in the performance of the SepFormer model on child speech datasets, especially as the number of speakers increases. This supports the hypothesis that child speech presents additional challenges to current speech separation models, a finding consistent with the expected negative result. The complexity and variability inherent in child speech are factors that future speech processing models will need to address more effectively in order to improve performance across different age groups.

6.3 Comparison with Previous Research

Previous studies underscores the unique acoustic and temporal characteristics of child speech, which present significant challenges for speech separation systems traditionally designed for adult speech. Previous studies, such as those by Kathania et al. (2021) and Lee, Potamianos, and Narayanan (1999), have detailed how the higher fundamental frequencies, increased spectral variability, and greater temporal variability in children's speech impact the performance of ASR systems. These systems often underperform with child speech due to their primary training on adult speech datasets (Bhardwaj et al. 2022; Kennedy et al. 2017). The observed variability in performance in this study is consistent with the increased spectral and temporal variability highlighted in the literature and the degradation of models trained on adult speech.

The SepFormer model, which has shown superior performance in adult speech separation tasks (Subakan et al. 2021), presents a novel approach to addressing the challenges in speech separation. However, as indicated by the findings of this study, while SepFormer exhibits potential, its performance varies significantly when applied to child speech.

This study makes a significant contribution to the field by specifically evaluating the SepFormer model on a child speech dataset, a gap that has been rarely addressed in previous research. The results align with the established challenges highlighted by Gerosa et al. (2009) and Kent and Forner (1980), showing that the increased variability in pitch and speaking patterns in children's speech affects the performance of speech separation models designed primarily for adults. This research

demonstrates a clear decline in both SI-SNR and SDR metrics as the number of speakers increases, which is consistent with the findings for adult models.

6.4 Practical Implications

This research has a profound impact on the development of speech separation technology, with particular relevance to the field of child speech separation. It establishes an important foundation for subsequent research in this area and advances the development of inclusive and effective speech technologies. The necessity for child-specific speech separation models is highlighted, which will greatly improve the utility and accuracy of speech-driven applications targeting young populations. The potential benefits of improved speech separation are considerable, particularly in the field of educational software and therapy, which can support children with speech impairments and help create more effective learning environments. Furthermore, insights from this study can guide future research to improve and adapt advanced speech separation models to ensure their effectiveness in different age groups and acoustic scenarios.

6.5 Limitations and Future Research

One of the limitations of this study is the focus on a specific age group (children aged 4-5 years) and the limited size of the data set. While this focus allows for a detailed analysis of the model's performance in this particular group, it may not capture the full variation in children's speech across ages and contexts. This limitation is particularly notable given the variability in speech characteristics that can occur across different ages. The dataset used does not adequately capture this variability, as it lacks a broad representation of different developmental stages.

This insufficiency is a critical issue that, unfortunately, cannot be fully addressed within the timeframe of the current research. However, it is essential for future studies to consider these aspects. Expanding the dataset to include a more diverse range of age groups and increasing the sample size would allow for a more comprehensive understanding of the SepFormer model's performance across various developmental stages.

Another notable limitation of this study is the lack of a clear comparative benchmark. This study did not conduct a comprehensive comparative analysis of the SepFormer model's performance on child speech with other models or baselines designed specifically for this age group. Without such a comparison, it is difficult to contextualize and assess the effectiveness of the SepFormer model. The lack of a priori comparative data makes it challenging to idenitfy whether the observed performance in this study is significant in comparison to existing models tailored for child speech.

It will be critical to address this shortcoming in future research. A comparative study that includes other transformer-based methods tuned for child speech could provide valuable insights into the strengths and weaknesses of the SepFormer model. Such a benchmark would not only clarify the relative performance of the models, but also provide a more nuanced understanding of how different models handle the complexity of child speech. Such an approach would greatly increase the depth and applicability of the study and provide clearer guidance for future developments in child speech processing technology.

6.6 Final Conclusion

This study highlights the challenges and potential of the SepFormer model in child speech separation. The findings validate the need for specialized approaches to improve speech separation quality for children.

First, a critical gap in the application of advanced speech separation models, such as SepFormer, to the domain of children's speech was identified. This highlighted the need for research. A dataset from the PhonBank corpus was carefully selected and pre-processed to match the original SepFormer benchmarks. The data set consists of 20 hours of English speech samples from 4- to 5-year-old children. Preprocessing included creating speech mixtures with different gender combinations in both two- and three-speaker configurations, removing silence longer than 1000 milliseconds while preserving natural pauses, normalizing the volume of the audio source, and resampling all audio to a consistent 8 kHz frequency.

The SepFormer model was evaluated on this child speech dataset using the established hyperparameters and evaluation metrics SI-SNR and SDR to allow a direct comparison with its performance on adult speech.

The analysis of the performance of the SepFormer model on children's speech based on the SI-SNRi and SDRi metrics provides insights relevant to the research questions and hypothesis. The main findings are as follows:

In the two-speaker condition, the median SI-SNRi for 4-year-olds was 5 dB, with a large interquartile range (IQR) and outliers ranging from -10 dB to 20 dB. This shows that there is considerable variation in the performance of the model in this age group, with some being very successful and others being clear failures. The 5-year-old group had a lower median SI-SNRi of 2 dB and a smaller IQR, indicating that their performance was more stable but also declined.

In the 3-speaker condition, the median SI-SNRi for the 4-year-olds was 3 dB, and the distribution of outliers was similar but not as extreme as in the 2-speaker condition. The 5-year-olds' median SI-SNRi dropped to 1 dB, with smaller IQRs and fewer outliers, suggesting that their performance degraded as the number of speakers increased.

SDRi values mirror the trend in SI-SNRi across age groups and conditions, reinforcing the observed patterns.

These results support the hypothesis that the performance of the SepFormer model decreases when applied to children's speech data. The degradation suggests that the model, while effective for adult speech, may need to be adapted to better handle the acoustic characteristics of children's speech

This study significantly contributes to the field of speech separation technology by focusing on child speech, an area that has been less explored compared to adult speech. It demonstrates the challenges and limitations of applying existing models, specifically the SepFormer, to child speech, which has inherently different acoustic and temporal characteristics. The research underscores the necessity for developing specialized models that cater specifically to the unique needs of children's speech. This is crucial for enhancing the performance of speech-driven applications in educational and therapeutic settings, ultimately supporting better learning and communication aids for young children. The findings lay a critical groundwork for future research aimed at refining speech separation technologies to be more inclusive and effective for all age groups.

The study's focus on a narrow age range and a small dataset limits its generalizability. Future research should expand the dataset to cover a broader age range to better understand model perfor-

mance across different developmental stages. Additionally, incorporating comparative benchmarks with other models designed for child speech is essential to evaluate the SepFormer model's relative effectiveness and guide further advancements in the field.

References

Bhardwaj, Vivek et al. (2022). "Automatic speech recognition (asr) systems for children: A systematic literature review". In: *Applied Sciences* 12.9, p. 4419.

Cristia, Alejandrina et al. (2018). "Talker diarization in the wild: The case of child-centered daylong audio-recordings". In: *Interspeech 2018*, pp. 2583–2587.

Edwards, Jan and Mary E Beckman (2008). "Methodological questions in studying consonant acquisition". In: *Clinical linguistics & phonetics* 22.12, pp. 937–956.

Ephrat, Ariel et al. (2018). "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation". In: *arXiv preprint arXiv:1804.03619*.

Gerosa, Matteo et al. (2009). "A review of ASR technologies for children's speech". In: *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, pp. 1–8.

Gray, Sharmistha S et al. (2014). "Child automatic speech recognition for US English: child interaction with living-room-electronic-devices." In: *WOCCI*, pp. 21–26.

Ho, Kuan-Hsun, Jeih-Weih Hung, and Berlin Chen (2023). "ConSep: a Noise- and Reverberation-Robust Speech Separation Framework by Magnitude Conditioning". In: *2023 24th International Conference on Digital Signal Processing (DSP)*, pp. 1–5. DOI: 10.1109/DSP58604.2023.10167992.

Ho, Kuan-Hsun, Jeih-weih Hung, and Berlin Chen (2022). "Bi-Sep: A Multi-Resolution Cross-Domain Monaural Speech Separation Framework". In: *2022 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pp. 72–77. DOI: 10.1109/TAAI57707.2022.00022.

Holliday, Jeffrey J et al. (2015). "Quantifying the robustness of the English sibilant fricative contrast in children". In: *Journal of Speech, Language, and Hearing Research* 58.3, pp. 622–637.

Kathania, Hemant et al. (2021). "Spectral modification for recognition of children's speech under mismatched conditions". In: *Nordic Conference on Computational Linguistics*. Linköping University Electronic Press, pp. 94–100.

Kennedy, James et al. (2017). "Child speech recognition in human-robot interaction: evaluations and recommendations". In: *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, pp. 82–90.

Kent, Raymond D and Linda L Forner (1980). "Speech segment durations in sentence recitations by children and adults". In: *Journal of phonetics* 8.2, pp. 157–168.

Lee, Sungbok, Alexandros Potamianos, and Shrikanth Narayanan (1999). "Acoustics of children's speech: Developmental changes of temporal and spectral parameters". In: *The Journal of the Acoustical Society of America* 105.3, pp. 1455–1468.

Liao, Hank et al. (2015). "Large vocabulary automatic speech recognition for children". In: *Sixteenth Annual Conference of the International Speech Communication Association*.

Luo, Yi, Zhuo Chen, and Takuya Yoshioka (2020). "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 46–50.

Luo, Yi and Nima Mesgarani (2018). "Tasnet: time-domain audio separation network for real-time, single-channel speech separation". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 696–700.

Mostow, Jack (2012). “Why and how our automated reading tutor listens”. In: *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training (ISADEPT)*. KTH Stockholm, Sweden, pp. 43–52.

Potamianos, Alexandros, Shrikanth Narayanan, and Sungbok Lee (1997). “Automatic speech recognition for children”. In: *Fifth European Conference on Speech Communication and Technology*.

Rolland, Thomas et al. (2022). “Multilingual transfer learning for children automatic speech recognition”. In.

Sattorovich, Elov Ziyodullo (2022). “Psychological influence of speech disorders and the causes that cause them on the child’s psyche”. In: *Academicia Globe* 3.01, pp. 39–42.

Shivakumar, Prashanth Gurunath and Panayiotis Georgiou (2020). “Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations”. In: *Computer speech & language* 63, p. 101077.

Strommen, Erik F and Francine S Frome (1993). “Talking back to big bird: Preschool users and a simple speech recognition system”. In: *Educational Technology Research and Development* 41.1, pp. 5–16.

Subakan, Cem et al. (2021). “Attention is all you need in speech separation”. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 21–25.

Vaswani, Ashish et al. (2017). “Attention is all you need”. In: *Advances in neural information processing systems* 30.

Vincent, Emmanuel, Rémi Gribonval, and Cédric Févotte (2006). “Performance measurement in blind audio source separation”. In: *IEEE transactions on audio, speech, and language processing* 14.4, pp. 1462–1469.

Wang, Wei et al. (2021). “Towards data selection on tts data for children’s speech recognition”. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6888–6892.

Wang, Xin, Jun Du, Alejandrina Cristia, et al. (2020). “A study of child speech extraction using joint speech enhancement and separation in realistic conditions”. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7304–7308.

Wang, Xin, Jun Du, Lei Sun, et al. (2018). “A progressive deep learning approach to child speech separation”. In: *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, pp. 76–80.

Xu, Yong et al. (2021). “Generalized spatio-temporal rnn beamformer for target speech separation”. In: *arXiv preprint arXiv:2101.01280*.

Yeung, Gary and Abeer Alwan (2018). “On the difficulties of automatic speech recognition for kindergarten-aged children”. In: *Interspeech 2018*.

— (2019). “A frequency normalization technique for kindergarten speech recognition inspired by the role of f0 in vowel perception”. In: *Interspeech 2019*.

Yip, Jia Qi et al. (2023). *Analysis of Speech Separation Performance Degradation on Emotional Speech Mixtures*. arXiv: 2309.07458 [cs.SD].

Zeghidour, Neil and David Grangier (2021). “Wavesplit: End-to-end speech separation by speaker clustering”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29, pp. 2840–2849.

Zeng, Yumin and Yi Zhang (2007). “Robust children and adults speech classification”. In: *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*. Vol. 4. IEEE, pp. 721–725.