Master Semester Project

Dana Kalaaji

Signal Processing for Communication Group Department

Pathological speech enhancement

Autumn 2023

Thesis Advisor
Prof. Daniel
Gatica-Perez
EPFL / IDIAP

Head of Department
and Supervisor
Dr. Ina Kodrasi
IDIAP

Acknowledgment

I express my gratitude to my thesis supervisor, Ina Kodrasi, whose continuous support, availability and constructive feedback greatly contributed to the development of my semester project. Additionally, I extend my thanks to Yacouba Kaloga for generously dedicating his time to assist me when I needed help.

Contents

1 Introduction

1.1 Idiap

Idiap is a research institute active in the field of artificial intelligence. It is dedicated to fundamental research, education, and technology transfer in the areas of artificial intelligence, machine learning, and signal processing. This report concludes my EPFL master's semester project done within Idiap's Signal Processing for Communication Group Department.

1.2 The project

Our human auditory gives us the remarkable ability to navigate conversations even when in the middle of chaotic environments with other conversations and ambient noises around. However, translating this natural skill into the realm of computers presents a significant challenge known as the "dinner party problem" within the speech processing field. While performant speech separation algorithm have been developed, they have been devised for neurotypical speakers, i.e., speakers without any speech impairments. However, this poses a problem as pathological conditions can disrupt the speech production mechanism, complicating the accurate processing by algorithm. This project aims to evaluate the performance of the state-of-the-art the speech separators models SepFormer and RE-SepFormer when handling such pathological signals in order to address these challenges.

The code and associated files for this project are accessible via the following website: `https://github.com/danaKalaaji/Master-Semester-Project_Dana`.

2 State of the art

Over the years, the field of speech separation has made remarkable progress, particularly since the incorporation of deep learning techniques like recurrent neural networks (RNNs). This improvement is illustrated in Figure 1 [11] showcasing a doubling of the SI-SDRI value in less than a decade when applying newer models to the WSJ0-2mix dataset [3]. This dataset consists of speech mixtures drawn from the WSJ0 dataset. The latter was designed in 1992 for automatic speech recognition, featuring read speech from the Wall Street Journal. These mixtures are generated by mixing pairs of recordings from different speakers at random signal-to-noise ratios uniformly drawn between 0 and 5 dB. In our project, we will evaluate the SepFormer and RE-SepFormer models created by SpeechBrain [8], an open-source initiative providing modular speech processing tool kits for diverse applications, including speech separation.

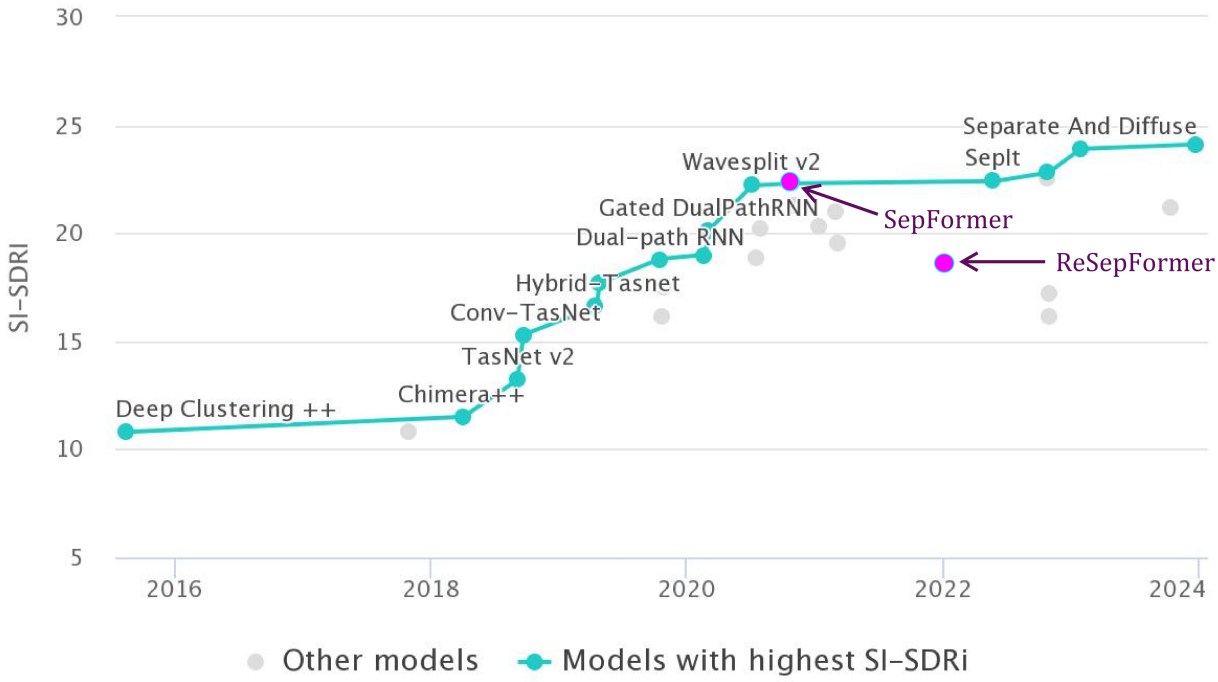

Figure 1: Performance of models on the WSJ0-2mix dataset through the years

2.1 SepFormer

Even though the incorporation of RNNs has allowed the speech separation field to evolve, their sequential nature poses challenges when handling larger datasets due to the inability to parallelize computations. SepFormer [9] represents a significant advancement in speech separation, as it addresses the limitations of RNNs by using a transformer-based architecture. This change allows for parallelization of computations over different time-steps. SepFormer is composed of three main components that work together to separate speech signals as depicted in Figure 2: the encoder, the decoder, and the masking network. The encoder processes input signals, the masking network split the input into temporal chunks and predicts optimal masks to separate the sources present in the mixtures. Finally, the decoder reconstructs the separated signals by using the masks predicted by the masking network. This architecture allowed the model to achieve state-of-the-art performance, with a SDRi of 22.4 dB and a SI-SDRi of 22.3 dB on the the WSJ0-2mix dataset.

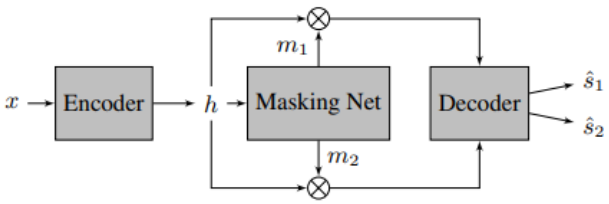

Figure 2: High-level description of SepFormer

2.2 Resource Efficient SepFormer

Though achieving state-of-the-art performance in speech separation, SepFormer is computationally demanding due to the large number of parameters it require. In response to this issue, a new model called Resource-Efficient Separation Transformer (RE-SepFormer) [10] was introduced by SpeechBrain. The RE-SepFormer optimizes efficiency by using non-overlapping chunks, reducing the workload by half when considering a default overlap rate of 50 %. Moreover, the inclusion of a Memory Transformer simplifies computations by working with summarized representations of entire chunks, removing the need to individually process each element. From Figure 3, we can see that these two modifications allows RE-SepFormer's to achieve efficiency in terms of computational resources and processing speed when compared to SepFormer, but at the cost of a performance decrease. Indeed, the number of parameters RE-SepFormer needs is reduced by 3 and we have a 9 times reduction in multiply-accumulate operations (MACs) when compared to the SepFormer. Nevertheless, despite having a performance drop compared to SepFormer (both the SDRi and the SI-SDRi on the WSJ0-2mix dataset were reduced to 18.8 and 18.6 respectively), the model still produces good quality speech separation.

Model	Overlap	Avg	SI-SNRi	SDRi	#Params	MACs/s
SepFormer	50%	×	22.3	22.4	25.7M	69.6
SepFormer	0%	×	15.9	16.2	25.7M	28.0
RE-SepFormer	0%	✓	18.6	18.8	8.0M	7.8

Figure 3: Comparison between SepFormer and RE-SepFormer (on WSJ0-2mix)

3 Data Generation

3.1 PC GITA Dataset

In our project, we used the PC-GITA [6] dataset to evaluate the performance of the SepFormer and ReSepFormer models, which were trained on the WSJO-2Mix dataset and are thus tailored to neurotypical speakers. The PC-GITA dataset was created in response to the challenges faced in analyzing pathological speech, particularly the lacking of databases with good technical conditions and diverse speech tasks.

This dataset contains speech recordings from 100 Spanish native speakers, including 50 patients suffering from Parkinson Disease (PD) and 50 age and gender-matched healthy controls. Each speaker articulated the same 10 spanish sentences, and these recordings were conducted in noise-controlled conditions within a soundproof booth. The recording protocol covers tasks related to phonation, articulation, and prosody, and was collaboratively designed with neurologists, linguists, and engineers to analyze different aspects of voice and speech in PD patients. Individuals with PD have distinct speech patterns, so the evaluation on this PC-GITA dataset will provide insights into how SepFormer and ReSepFormer handle pathological signals.

3.2 The scenarios

To conduct our analysis, we need to generate all possible mixtures of two audio signals from the PC-GITA dataset. Each mixture will then be processed through SepFormer and RE-SepFormer to evaluate the separation performance of these models across various factors. Our analysis will specifically consider health and gender factors, excluding age and sentences from consideration. Additionally, we will exclude mixtures signals where both audios come from the same speaker or from speakers saying the same sentence. Finally, it's worth noting that the order of signals within a pair $(x_1\, x_2)$ does matter since, as we will see later, the second signal x_2 acts like a noise in the resulting mixture signal.

We define three scenarios involving mixtures of audio signals from healthy speakers:

- Healthy male + Healthy male: 54 000 of audio signal pairs
- Healthy female + Healthy female: 54 000 of audio signal pairs
- Healthy male + Healthy female: 56 250 of audio signal pairs

Since SepFormer and ReSepFormer were trained on English audio signals and our dataset comprises Spanish audio signals, we anticipate a potential decrease in performance compared to literature values. This serves as our reference for evaluating model performance with pathological signals.

Similarly, we consider three scenarios involving mixtures of audio signals from speakers with parkinson:

- PD male + PD male: 54 000 of audio signal pairs
- PD male + PD female: 54 000 of audio signal pairs
- PD female + PD female: 56 250 of audio signal pairs

Finally, we examine four scenarios featuring mixtures of audio signals from speakers with different health conditions:

- Healthy male + PD male: 56 250 of audio signal pairs
- Healthy male + PD female: 56 250 of audio signal pairs
- Healthy female + PD male: 56 250 of audio signal pairs
- Healthy female + PD female: 56 250 of audio signal pairs

To determine the number of mixtures generated per scenario, we denote Y as the count of all pairs of distinct speakers, considering 25 speakers in each category when accounting for age and health conditions. This means that Y is all possible combination of pairs from speakers from the categories of a specific scenario, with the 25 pairs consisting of the same speaker removed when considering scenarios where we superpose audios from speakers belonging to the same category. We obtain the following mathematical expression for Y:

$$Y = \begin{cases} 25^2 & \text{if speakers have different (gender, health)} \\ (25^2 - 25) & \text{if speakers have the same (gender, health)} \end{cases}$$

For each pair of distinct speakers, we want all to create audio mixture from all possible pairs of distinct sentences they were uttering. Each speaker is recorded saying the same 10 sentences, this results in a total of 90 different pairs when excluding the same sentence for mixtures. Therefore, the overall number of pairs for the scenario is $Y \times 90$. Consequently, we have a total of 553 500 audio signal mixtures to be processed.

3.3 Generating the mixture of two audio signal

To conduct our analysis, we aim to generate mixtures of audio signals from the PC-GITA dataset. We will then feed the mixed signal y as input to SepFormer and RE-SepFormer, which will produce the two estimated signals $\hat{x}_1$ and $\hat{x}_2$. However, since both SepFormer and RE-SepFormer expect input recordings sampled at 8 kHz, we begin the process by resampling the entire dataset to match this frequency.

Given two audio signals x_1 and x_2 representing recordings of two different sentence from two different speaker, we want to create a mixed signal y using the formula $y = \alpha x_1 + \beta x_2$. We first ensure that both audio signals have the same duration by truncating the longer signal if needed.

$$\text{duration} = \text{len}(y) = \text{len}(\alpha x_1 + \beta x_2) = \min(\text{len}(x_1), \text{len}(x_2)).$$

For each mixed signal y, we generate a random Signal-to-Noise Ratio (SNR) uniformly drawn between 0 and 5 dB. This SNR represents the level of noise relative to the signal in y. When then adjust the levels of x_1 and x_2 by applying weight to them so that the signal of interest for estimation in y is x_1, while the noise to be separated from x_1 is x_2. The initial values of the weights for x_1 and x_2, denoted as α'' and β'' respectively, are set as:

$$\alpha'' = \frac{\text{SNR}}{2} dB$$

$$\beta'' = -\frac{\text{SNR}}{2} dB$$

We then convert them from decibels to a linear scale:

$$\alpha' = 10^{\left(\frac{\alpha''}{10}\right)}$$

$$\beta' = 10^{\left(\frac{\beta''}{10}\right)}$$

Next, we normalize both x_1 and x_2 by dividing them by their respective Root Mean Square (RMS). This measure provides a representation of the overall amplitude of the signal. The mathematical expression for the RMS of a signal x is given by:

$$\text{RMS}(x) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$$

Here, N represents the number of samples in the signal x, and x_i denotes each individual sample. This yields normalized signals x_1' and x_2'.

To ensure the amplitude of the resulting signal $y' = \alpha' x_1' + \beta' x_2'$ remains below 1, we multiply each signal by a scaling factor γ:

$$\gamma = \frac{0.9}{\max(\max(|y'|), \max(|x_1|), \max(|x_2|))}$$

The mathematical expressions for the final values of α and β reflecting the adjustment of the initial weights after normalization, and amplitude scaling are given by:

$$\alpha = \alpha' \cdot \left(\frac{\gamma}{\text{RMS}(x_1)}\right)$$

$$\beta = \beta' \cdot \left(\frac{\gamma}{\text{RMS}(x_2)}\right)$$

For each scenario, we generated a separate text file located in the subfolder `text_files` of the `PC-GITA_per_task_16000Hz` folder. Each line in these files represents an audio combination (x_1, x_2), providing the values for α, β, SNR, and duration allowing for consistent reconstruction of the corresponding mixture signal y.

4 Performance Evaluation

Although the Signal-to-Distortion Ratio (SDR) has been a commonly used metric for evaluating the quality of separation algorithms, it has limitations leading to potentially misleading results [4]. Indeed, the SDR is sensitive to issues like scaling and channel variations, impacting its reliability in evaluating the quality of separated signals. To overcome these limitations, we choose to employ the Scale-Invariant SDR (SI-SDR)[4] for evaluating the performance of SepFormer and RE-SepFormer on our signals. SI-SDR introduces scale invariance, making it more robust to changes in signal amplitude.

As previously mentioned, the audio mixture y was generated such that the signal of interest for estimation is x_1 and the associated noise is x_2. To assess how well the algorithms perform, we use SI-SDR$(\hat{x}, x_1)$, comparing the estimated source $\hat{x}$ to the original source x_1. However, when the SepFormer and RE-Sepformer algorithms return estimated signals $\hat{x}_1$ and $\hat{x}_2$, we

face uncertainty about which one actually corresponds to the estimation of x_1. To handle this, we take the maximum value between SI-SDR($\hat{x}_1, x_1$) and SI-SDR($\hat{x}_2, x_1$), as it represents the SI-SDR for the correct estimated source $\hat{x}$ and x_1. Additionally, to assess the algorithm's overall effectiveness compared to no separation, we compute SI-SDR(y, x_1), comparing the mixture y to the original source x_1.

5 Results

5.1 Overall view

For each scenario, we generated individual text files stored in the `processed_audio` subfolder of the `PC-GITA_per_task_16000Hz` directory. Each line within these files represents the same audio combination (x_1, x_2) as found in the corresponding scenario text file in the `text_files` subfolder. Each line includes the audio signals x_1 and x_2, along with SI-SDR(y, x_1), SI-SDR($\hat{x}_{Sepformer}, x_1$) and SI-SDR($\hat{x}_{RE-Sepformer}, x_1$). Figure 4 offers an overall view of performance by showing the average of the three SI-SDR values considered for each scenario. The scenarios are arranged based on their overall performance which is calculated based on the mean of those three different SI-SDR averages.

General patterns Upon examining the bar plot, a pattern emerges when comparing SISDR values in different scenarios. The values of 'Before' condition, representing SI-SDR(y, x_1), shows consistent performance across all scenarios. In contrast, the Sepformer and RE-Sepformer cases present notable variations, suggesting that the algorithm's performance depends on the specific scenario considered. Moreover, the average SISDR Before is consistently lower compared to those of Sepformer and RE-Sepformer, indicating that using these algorithms improves the separation performance. Another interesting observation is that Sepformer outperforms RE-Sepformer in all scenarios. This aligns with our expectations as the latter achieves efficiency in terms of computational resources and processing speed but at the cost of a performance decrease.

Influence of the gender factor The plot suggests that the algorithm's performance varies based on the gender composition of the audio mixtures and that they perform better for male voices. Indeed, the 4 top-performing scenarios are those involving a mix of genders, which aligns with expectations as distinguishing between male and female voices is easier. The next top three scenarios with the best overall performance involve only males, and the scenarios that have the lowest performance are those with only females. The scenario with the highest overall performance is the healthy male and female voices which aligns with our expectations since the models were trained on neurotypical data. A surprising results however is that the highest Sepformer average SISDR occurs in the scenario of healthy females and sick males implying that the model performs slightly better when separating audios mixture generated from sick males and healthy females than mixtures from healthy males and healthy females. Finally we observe that out of these 4 scenarios, the ones that have the lower performance for Sepformer are the ones involving sick females.

Influence of the health factor When looking at scenarios within the same gender composition (those with mixed gender, those involving only males and those with only females), we notice that the order the performance of both algorithms remains consistent:

- For Sepformer: healthy vs. healthy, healthy vs. sick, and sick vs. sick. (with the exception of the senarios with mixed genders)
- For RE-Sepformer: healthy vs. healthy, sick vs. sick, and healthy vs. sick.

This implies that the performance of both Sepformer and RE-Sepformer appears to be influenced by the health status of the speakers, with better results observed in scenarios involving healthy voices. Sepformer tends to perform better if one of the audios from the mixture involves a healthy person while RE-Sepformer does better in scenarios involving people with the same health condition

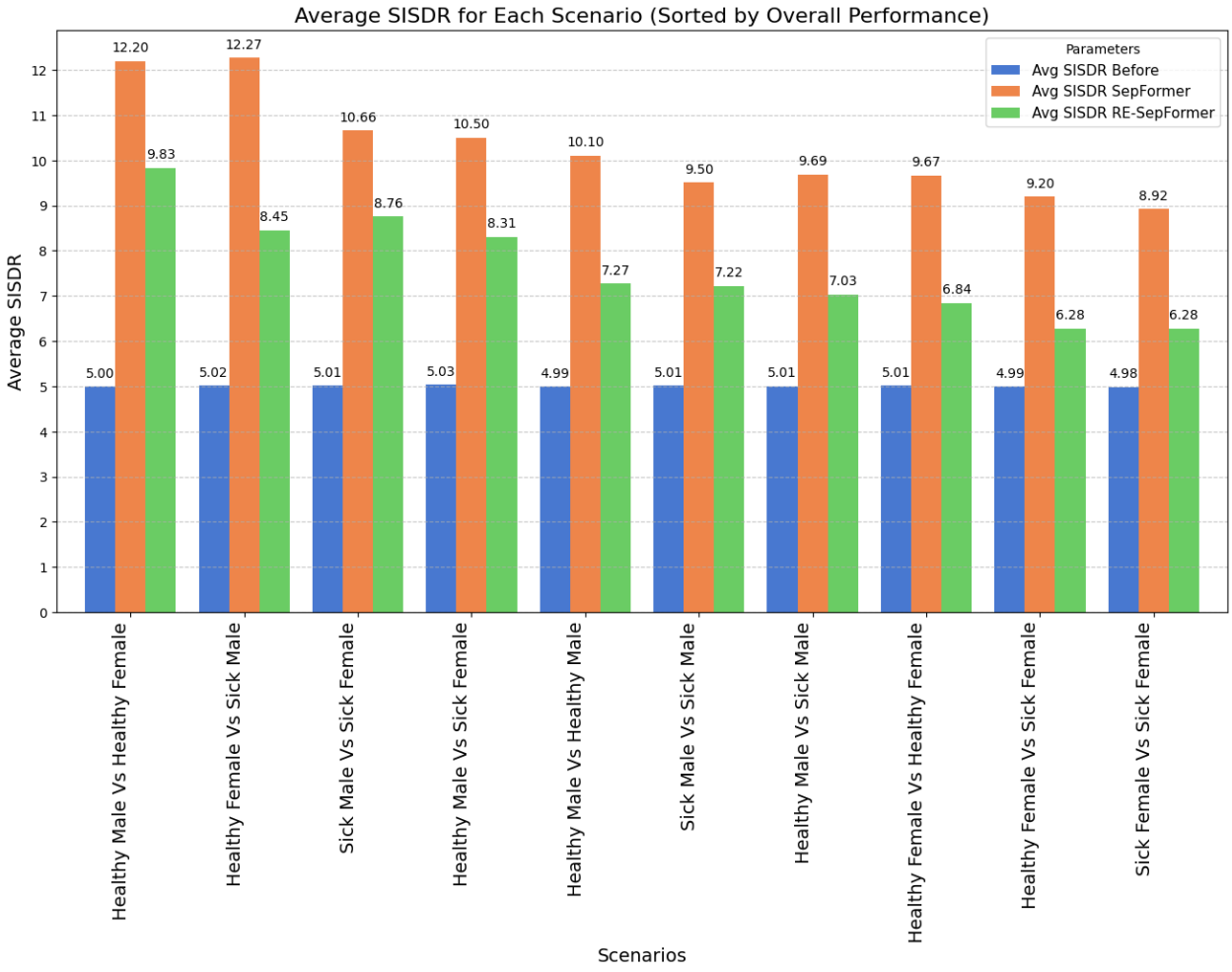

Figure 4: Comparison of separation performance of all scenarios for mixture signal y before and after going through Sepformer and RE-Sepformer

Influence of the language factor As mentioned earlier, SepFormer and RE-SepFormer were trained on English audio signals from neurotypical speakers, while our dataset consists of spanish audio signals, including both healthy and speakers with pathological speech. This language difference might impact the algorithms' performance, leading to variations from literature values. To gauge this impact, we specifically examine scenarios involving only healthy speakers (Healthy Male vs Healthy Female, Healthy Female vs Healthy Female, and Healthy Male vs Healthy Male). In literature, SepFormer and RE-SepFormer are reported to have SI-SDR values of 22.4 dB and 18.6 dB, respectively. Our analysis in a Spanish context focusing on neurotypical speakers, yields an average SI-SDR of 10.66 dB for SepFormer and 7.98 dB for RE-SepFormer. Comparing these results with literature values, we observe an almost halving of the values of the SI-SDR. This suggests that SepFormer and RE-SepFormer performance is indeed influenced by the language of the inputs and encounters challenges when processing Spanish audio signals.

5.2 Distribution

Examining the boxplots in Figure 5, we notice that the boxplots for SI-SDR without processing through SepFormer or RE-Sepformer show similar shapes. This suggests that the performance distributions are comparable across different scenarios. However, this pattern doesn't hold for the boxplots of signals after going through either SepFormer or RE-SepFormer which suggests that separation performance depends on the scenarios when using these algorithms. The differing median values tell us that the central tendencies of performance metrics vary between scenarios when using SepFormer or RE-SepFormer. Additionally, the presence of outliers indicates extreme performance values in specific scenarios, which may contribute to the differences in medians. Specifically, we observe a great number of outliers with low values in the Sepformer boxplot of the scenario Healthy Males vs Healthy Females which may explain the surprising observation made earlier that it is not the scenario achieving the highest for the Sepformer model. Another interesting result is that both model tend to have bigger variations for scenarios involving females only. Moreover the RE-Sepformer model tend to have larger variation overall compared to Sepformer, with the biggest ones observed for scenarios involving healthy speakers only.

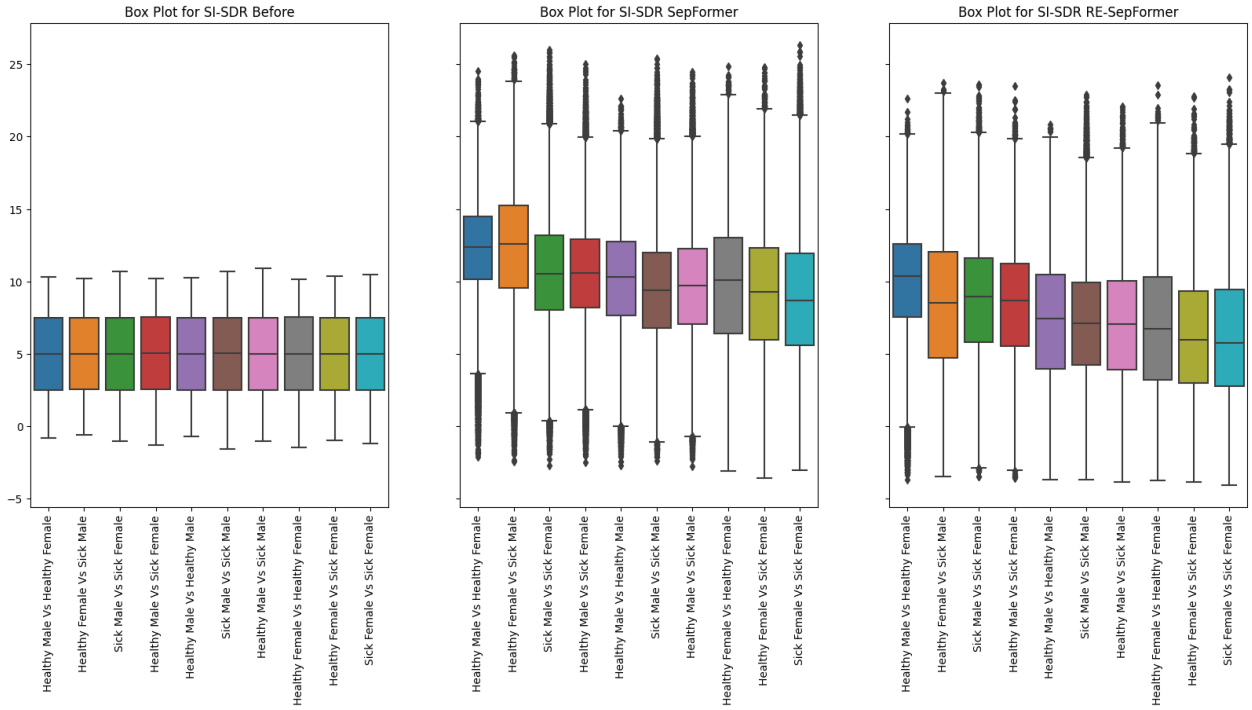

Figure 5: Box Plots of separation performance of all scenarios for mixture signal y before and after going through either Sepformer and RE-Sepformer

5.3 Statistical Testing

5.3.1 Kruskal-Wallis Test

We conducted statistical testing to determine if the observed patterns and variations are likely to be due to random chance or if they reflect genuine differences in the performance of SepFormer and RE-SepFormer across different scenarios. Given that the data did not satisfy the assumptions required for ANOVA testing due to the non-normality and lack of homogeneity of variances, we opted the Kruskal-Wallis test which is robust to these conditions.

- The Kruskal-Wallis test for SI-SDR Before data returned a p-value of 0.18, which is greater than the significance level of 0.05. Therefore, there is insufficient evidence to reject the null hypothesis that the medians of the groups are equal. This suggests that there is no significant difference in the SI-SDR values across the scenarios, supporting the idea that the performance of the separation of the mixture signal before going through either Sepformer or RE-Sepformer are consistent across scenarios.

- The Kruskal-Wallis test for both the SI-SDR Sepformer and SI-SDR RE-Sepformer data returned a very small p-value approaching zero. Therefore, the null hypothesis is rejected in both cases, suggesting a significant difference in SI-SDR values among the scenarios. This indicates that at least one scenario has a different median SI-SDR value compared to the others.

5.3.2 Dunn's Test

The statistical evidence supports the notion that the separation performance of both SepFormer and RE-Sepformer differs significantly among the scenarios. To pinpoints which scenarios specifically differ from one another, we performed Dunn's tests. The significance matrices obtained in Figures 6 and 7 show whether there are statistically significant differences between pairs of scenarios for Sepformer and RE-Sepformer. In the matrices, "True" indicates significance, while "False" indicates no significance. The index of the scenarios in the matrices are reported in Figure 8

SI-SDR SepFormer Statistically significant differences in SI-SDR values were observed across all scenarios, except for the comparison between Healthy Male vs. Sick Female (6) and Sick Male vs. Sick Female (9). This implies that SepFormer performance varies significantly between all scenarios, with only the mentioned pair showing no statistically significant difference.

SI-SDR RE-SepFormer Similar to SepFormer, almost all scenarios show statistically significant differences in SI-SDR values. However, the pairs of scenarios (Healthy Female vs. Sick Male (3), Healthy Male vs. Sick Female (6)) and (Healthy Male vs. Sick Male (7), and Sick Female vs. Sick Male (8)) show no statistically significant difference.

The Dunn's test results align with our findings from the Kruskal-Wallis tests, and emphasize that there are significant performance differences in SepFormer and RE-SepFormer across diverse scenarios. As these scenarios were crafted based on the gender and health status of speakers in the audio mixtures, the observed variations underscore the algorithms' sensitivity to these specific factors.

	1	2	3	4	5	6	7	8	9	10
1	False	True	True	True	True	True	True	True	True	True
2	True	False	True	True	True	True	True	True	True	True
3	True	True	False	True	True	True	True	True	True	True
4	True	True	True	False	True	True	True	True	True	True
5	True	True	True	True	False	True	True	True	True	True
6	True	True	True	True	True	False	True	True	False	True
7	True	True	True	True	True	True	False	True	True	True
8	True	True	True	True	True	True	True	False	True	True
9	True	True	True	True	True	False	True	True	False	True
10	True	True	True	True	True	True	True	True	True	False

Figure 6: Significance matrix for SI-SDR Sepformer

	1	2	3	4	5	6	7	8	9	10
1	False	True	True	True	True	True	True	True	True	True
2	True	False	True	True	True	True	True	False	True	True
3	True	True	False	True	True	False	True	True	True	True
4	True	True	True	False	True	True	True	True	True	True
5	True	True	True	True	False	True	True	True	True	True
6	True	True	False	True	True	False	True	True	True	True
7	True	True	True	True	True	True	False	True	True	True
8	True	False	True	True	True	True	True	False	True	True
9	True	True	True	True	True	True	True	True	False	True
10	True	True	True	True	True	True	True	True	True	False

Figure 7: Significance matrix for SI-SDR RE-Sepformer

Index	Scenario
0	Healthy Female Vs Healthy Female
1	Healthy Female Vs Sick Female
2	Healthy Female Vs Sick Male
3	Healthy Male Vs Healthy Female
4	Healthy Male Vs Healthy Male
5	Healthy Male Vs Sick Female
6	Healthy Male Vs Sick Male
7	Sick Female Vs Sick Female
8	Sick Male Vs Sick Female
9	Sick Male Vs Sick Male

Figure 8: Scenario Index Table

5.4 Limitation

Recent studies [2] have brought attention to significant performance drops when models trained on WSJ0-2Mix are tested on other datasets with similar characteristics. Moreover, the official documentation for SepFormer and RE-SepFormer explicitly mentions that the SpeechBrain team does not provide any warranty regarding the performance of these models when applied to alternative datasets. This limitation should be considered when quantifying the drop in performance when considering the language factor i.e. when transitioning from the neurotypical english dataset WSJO-2Mix to our neurotypical scenarios from the spanish PC-GITA dataset.

6 Conclusion

Our project aimed to evaluate the performance of SepFormer and RE-SepFormer, state-of-the-art speech separation models, in the context of handling pathological speech signals. Using the PC-GITA dataset, which consists of recordings from both healthy speakers and those with Parkinson's Disease, we conducted an analysis by considering all ten possible

scenarios based on gender and health factors. Our evaluation process involved generating audio mixtures and assessing the algorithms' performance through the SI-SDR metric. The results revealed variations in SepFormer and RE-SepFormer performance across the scenarios. Examining the influence of the gender factor, our analysis suggested better performance for male voices, and the scenarios with mixed genders exhibited higher overall performance. Surprisingly, SepFormer achieved its highest average SI-SDR in the scenario involving healthy females and sick males. Considering the health factor, both SepFormer and RE-SepFormer performed better in scenarios involving healthy speakers. SepFormer generally performed better when one of the speakers was healthy, while RE-SepFormer performed better in scenarios with speakers of the same health condition. Statistical testing confirmed the significance of observed variations, indicating that separation performance is significantly influenced by the gender and health status of the speakers present in the audio recordings. Notably, there is a decline in performance when dealing with pathological speech, especially if at least one of the signals in the mixture originates from a female speaker.

References

[1] Fahimeh Bahmaninezhad, Shi-Xiong Zhang, Yong Xu, Meng Yu, John H. L. Hansen, and Dong Yu. A unified framework for speech separation. *arXiv preprint arXiv:1912.07814*, 2019. Submitted on 17 Dec 2019. URL: `https://arxiv.org/abs/1912.07814`.

[2] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent. Librimix: An open-source dataset for generalizable speech separation. 2021. URL: `https://inria.hal.science/hal-03354695/document`.

[3] John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016. URL: `http://dx.doi.org/10.1109/ICASSP.2016.7471631`, `doi:10.1109/icassp.2016.7471631`.

[4] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. Sdr - half-baked or well done? *arXiv preprint arXiv:1811.02508*, 2018. URL: `https://arxiv.org/abs/1811.02508`.

[5] Mitsubishi Electric Research Laboratories (MERL). Deep Clustering, 2023. Accessed: February 1, 2024. URL: `https://www.merl.com/research/highlights/deep-clustering`.

[6] Juan Rafael Orozco-Arroyave, Julián David Arias-Londoño, Jesús Francisco Vargas-Bonilla, María Claudia González-Rátiva, and Elmar Nöth. New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 342–347, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL: `http://www.lrec-conf.org/proceedings/lrec2014/pdf/7_Paper.pdf`.

[7] Douglas B. Paul and Janet M. Baker. The design for the Wall Street Journal-based CSR corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992. URL: https://aclanthology.org/H92-1073.

[8] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624. arXiv:2106.04624.

[9] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *ICASSP 2021*, 2021.

[10] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Frédéric Lepoutre, and François Grondin. Resource-efficient separation transformer, 2022. arXiv:2206.09507.

[11] Papers with Code. Speech Separation on WSJ0-2Mix - State-of-the-Art, 2023. Accessed: February 1, 2024. URL: https://paperswithcode.com/sota/speech-separation-on-wsj0-2mix.