IKM_Lab at BioLaySumm Task 1: Longformer-based Prompt Tuning for Biomedical Lay Summary Generation

Yu-Hsuan Wu, Ying-Jia Lin, Hung-Yu Kao
Intelligent Knowledge Management Lab
Department of Computer Science and Information Engineering
National Cheng Kung University
Tainan, Taiwan
p76104655@gs.ncku.edu.tw, yingjia.lin.public@gmail.com
hykao@mail.ncku.edu.tw

Abstract

This paper describes the entry by the Intelligent Knowledge Management (IKM) Laboratory in the BioLaySumm 2023 task1. We aim to transform lengthy biomedical articles into concise, reader-friendly summaries that can be easily comprehended by the general public. We utilized a long-text abstractive summarization longformer model and experimented with several prompt methods for this task. Our entry placed 10th overall, but we were particularly proud to achieve a 3rd place score in the readability evaluation metric. Our code is available at https://github.com/IKMLab/BioLaySumm.

1 Introduction

The BioLaySumm 2023 Task 1: Lay Summarization of Biomedical Articles (Goldsack et al., 2023) is to generate lay summaries for two biomedical datasets (Goldsack et al., 2022), PLOS and eLife, respectively. The aim of this task is to generate simplified and easy-to-understand summaries of the original articles that can be readily comprehended by the general public. Figure 1 provides an information of this task dataset, showcasing examples that highlight the importance of not only generating abstracts for the text but also simplifying the content. In the examples, the bold vocabulary in the article text represents biomedical terminology that may be difficult for the general public to understand. Therefore, it is crucial to transform such terminology into language that is more accessible and easily understood by a lay audience, as depicted by the bold vocabulary in the lay summary. This transformation plays a vital role in ensuring that the generated lay summaries are both comprehensible and effectively convey the key information to non-expert readers.

Article:
The virus SARS-CoV-2 can exploit biological vulnerabilities (e.g. host proteins) in susceptible hosts that predispose to the development of severe COVID-19. To identify host proteins that may contribute to the risk of severe COVID-19, we undertook proteome-wide genetic colocalisation tests, ...

Lay Summary:
Individuals who become infected with **the virus that causes COVID-19** can experience a wide variety of symptoms. These can range from no symptoms or minor symptoms to severe illness and death. Key demographic factors, such as age, gender and race, ...

Figure 1: The initial sentences of the article and lay summary of an eLife demonstrate distinctions in language usage and emphasis on background information.

The average length of articles in the eLife dataset is 8,442 words, while for the PLOS dataset, it is 5,864 words. (Table 1.). Using a state-of-the-art summary generation model with good performance on shorter texts would not be sufficient to capture all the important information in these lengthy articles, as the model would not be able to process such a large number of tokens. In order to solve the problem of dealing with long texts, we utilized models like Longformer (Beltagy et al., 2020) as the main framework in our competition.

In our implementation, we go beyond using just the article field and also incorporate the information from the keyword and heading fields. By extracting and utilizing these additional fields during the training process, our goal is to generate lay summaries that are more informative, accurate, and easily comprehensible to a general audience. This approach allows us to leverage the rich information contained in the keyword and heading fields, thereby enhancing the quality and effectiveness of the generated summaries.

Dataset	#Docs	Article	Sum.	Comp. %
		#words	#words	
eLife	4,346	8,441.7	348	4.1
PLOS	24,773	5,864	176.6	3.0

Table 1: The statistics of the eLife and PLOS training datasets, including the number of datasets, the average word count for articles and summaries, and the word compression ratio (%) from article to summary.

2 System Overview

Our system employs a pre-trained longformer (Beltagy et al., 2020) as the primary architecture and implements diverse prompt formats for the keyword and heading fields in the dataset during training. In the following sections, each component of the system will be described in detail.

2.1 Longformer

In our implementation, we used the pre-trained LED model (Beltagy et al., 2020) available on HuggingFace[1] (Wolf et al., 2020) to fine-tune each of the two datasets because LED accepts a maximum sequence length of 8,192 tokens, which shows potential for fitting the two datasets. As we observe the data statistics in Table 1, there are variations in the average length of the lay summaries between the eLife and PLOS datasets. Considering that the average length of eLife's lay summaries is 348 words, we set a maximum output sequence length of 512 during the decoding process. This adjustment also accommodates some lay summaries in the PLOS dataset that exceed the limit of 256 words.

We trained the model for 2 epochs on each dataset. Compared with our own fine-tuning experiments with the pre-trained BART model (Lewis et al., 2020), we found that the LED model performed better during the fine-tuning stage, which was in line with our expectations.

2.2 Prompt

We drew inspiration from existing research on optimized summarization techniques. Controlled tokens (Luo et al., 2022) to regulate the form and style of summary generation, using explicit tokens (Martin et al., 2020) to improve the effect of simplification, and enhancing the learning effect of the model by adding prompts (Zhang et al., 2022).

Based on the aforementioned method, we examined the various information provided in both datasets and discovered the existence of a “keywords” column, which describes the topic of the article, as well as a “headings” column. We utilized the information contained in these two fields to guide our prompt form design.

Keywords: To leverage the information provided in the keywords column, we devised two types of prompts. The first prompt form involved concatenating the keywords column with the article “Keywords $(k_1, k_2, ..., k_m)$ summarize: ", The second prompt form is to replace the [KEYWORDS] and [LAY_SUM] placeholders in the prompt with special tokens “[KEYWORDS] $(k_1, k_2, ..., k_m)$ [LAY_SUM]: ", this was done to highlight the importance of these specific tokens and make it easier for the model to differentiate them during training.

Headings: The title of each section in the text is represented by a heading, such as “Abstract” ,“Methods” etc., and each section is separated by “\n" in the article. To help the model better understand the correlation between each section, we added heading prompts. Compared to the keyword prompts, the form we designed using headings is h_1 Sections1, h_2 Section2,... and $h_i \in H(n = 1, 2, ..., n)$ where H is a list of headings names.

2.3 Prompt Tuning

During the prompt tuning stage, we implemented the prompt forms described earlier by converting them into string format. These prompts were then concatenated with the article text, creating a connected text that incorporated the relevant prompts. This connected text was subsequently used as input for the Longformer model during the prompt tuning process. By integrating the prompts directly into the text, we aimed to guide the model towards better understanding and generating higher quality lay summaries that aligned with our desired objectives.

3 Experiments

3.1 Evaluation Metrics

In this competition, three evaluation aspects have been defined to assess the effectiveness of sum-

[1] https://huggingface.co/models

Model	Relevance ↑				Readability ↓		Factuality ↑
	R-1	R-2	R-L	BERTScore	FKGL	DCRS	BARTScore
BART	0.4786	0.1525	0.4452	0.8486	12.3617	9.9345	-2.7569
T5	0.4358	0.1214	0.4095	0.8398	**10.1728**	**9.1107**	-3.7528
LED	**0.4858**	**0.1552**	**0.4502**	**0.8571**	11.8577	9.8441	**-2.0367**

Table 2: Results of the pre-trained model fine-tuned on both datasets.

marization generation and simplification, and each aspect comprises multiple automatic metrics:

- Relevance: ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L) (Lin, 2004), and BERTScore (Zhang et al., 2019).
- Readability: Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975) and Dale-Chall Readability Score (DCRS) (Chall and Dale, 1995).
- Factuality: BARTScore (Yuan et al., 2021). The evaluation model is provided by the official.

Of the three evaluation aspects, excluding readability, a lower index value is indicative of better performance, while higher scores in relevance and factuality are desirable.

3.2 Model Analysis

Considering the substantial size of the article, prioritizing the implementation of summarization over simplification is of utmost importance. Hence, the initial implementation involves fine-tuning both datasets within a range of prominent abstract summarization models, including the LED model of interest.

The results of various metrics are presented in Table 2, where the average of the two validation datasets evaluated separately is reported. The results indicate that for pure fine-tuning, the LED model performs as expected. Although it exhibits slightly lower readability compared to the T5 model (Raffel et al., 2020), other evaluation metrics show significant improvements. Hence, a minor decrease in readability metrics is deemed acceptable. Therefore, we chose the LED model as our primary architecture for further experimentation.

3.3 Prompt Analysis

This section focuses on the utilization of various prompts to fine-tune the pretrained LED model for handling long texts, as discussed in Section 2.2. Our main objective was to improve the quality of summaries and simplify the text through prompt tuning. Additionally, we conducted a comparative analysis with the BART model, employing similar prompts simultaneously, to comprehensively evaluate the effectiveness of the LED model after prompt integration.

Model	**Readability ↓**	
	FKGL	DCRS
LED	15.1239	11.6738
LED-keywords w/o ST	14.4514	**11.3581**
LED-keywords w/ ST	14.4429	11.3690
LED-headings	**14.3973**	11.3598

Table 3: Readability metric results on PLOS validation dataset for each prompt method. (“ST” means “Special Tokens”)

Upon analyzing the prompt tuning results of the LED model, we observed that when using the PLOS dataset, the overall scores of the relevance and factuality metrics remained relatively stable across different prompt methods compared to the absence of prompt tuning. Moreover, we found that both readability metrics (FKGL and DCRS) exhibited simultaneous improvements, resulting in the simplification of the text, the results show in Table 3 (with “ST” representing “Special Tokens”).

The final evaluation results of the two validation datasets, as presented in Table 4, demonstrate that the performance of the prompt utilizing special tokens to represent `KEYWORDS` and `LAY_SUM` is comparable to that of the prompt without special tokens. The scores only exhibited slight fluctuations, and the performance of the seven evaluation metrics was superior to that of using headings as a prompt. In terms of model selection based on the prompts, the LED model consistently outperformed the BART model in overall performance, aligning with our initial expectations.

Model	Relevance ↑				Readability ↓		Factuality ↑
	R-1	R-2	R-L	BERTScore	FKGL	DCRS	BARTScore
BART-keywords	0.4776	0.1512	0.444	0.8558	12.4748	10.0333	-2.93
LED-keywords w/o ST	**0.4849**	0.1536	**0.4534**	0.855	11.8637	**9.8021**	-2.3375
LED-keywords w/ ST	**0.4849**	**0.1537**	**0.4534**	**0.8554**	11.9689	9.8715	-2.3358
LED-headings	0.4839	0.1525	0.4515	0.8552	**11.8577**	9.8441	**-2.3237**

Table 4: Results of the pre-trained BART and LED model through prompt tuning with keywords and headings.

3.4 Keyword Analysis

In this section, we analyze the keywords field in the two datasets to discuss the potential negative impact of keywords on our method.

1. **Coarse keywords in the datasets.** Through observations on the datasets, we found that the granularity of the keyword field is relatively coarse. The words used in this field typically represent the overall theme or topic of the entire article, such as “biochemistry” and “cell biology”. These keywords often fail to capture the intricate details of an article and may lead to lower scores on ROUGE-1 and ROUGE-2 in Table 4, compared with the pre-trained LED in Table 2.

2. **Missing keywords in PLOS.** The number of empty keywords field over the data points in the PLOS dataset is much higher than that in the eLife dataset (Table 5), which may cause inconsistent keyword prompts for the instances of the two datasets.

3. **Low keyword diversity in eLife.** The number of unique keywords and the average number of keywords per example in the PLOS dataset are much higher than in the eLife dataset, as listed in Table 5. In addition, eLife provides keywords with an even coarser granularity for each data point than PLOS. These may further hinder our model from capturing the article details in eLife.

	PLOS	eLife
Num. of empty keywords field	3471	2
Avg. keywords per example	16.71	2.28
Num. of non-repeat keywords	7235	31

Table 5: Keywords field statistics. The first row indicates the number of examples with an empty list in the keywords field, the second shows the average number of keywords per example, and the last row is the counts of overall non-repeated keywords in each dataset.

3.5 Headings Analysis

Analysis of headings indicates that prompt tuning has a relatively weaker impact compared to keywords. The evaluation metrics show a slight increase only in the FKGL readability metric and BARTscore, while other indicators decline to some extent. Our observation reveals that headings are more complex than anticipated. Although common headings like Abstract, Introduction, Method, and Discussion are present, there are also unique and specific headings tailored to individual articles, such as “Bacterial Films” or “The Roles of VSG N-linked Oligosaccharides”. These headings provide less generalizable information compared to keywords. Additionally, variations in paragraph information across articles may hinder the model’s understanding of paragraph importance. Consequently, headings offer less comprehensive information compared to keywords.

4 Submission Options

4.1 Submission 1

To maintain consistency, we used the pure fine-tuned LED model as the initial submission (the LED model in Table 2) and compared its results with the evaluation results of the local validation dataset. We also conducted an analysis to determine whether the validation data and testing data had similar properties.

4.2 Submission 2

For the second submission, we selected the model with the best prompt tuning results. Although the performance of special tokens was better in terms of relevance indicators, the improvement rate was insignificant, and the readability metric showed a significant drop. Therefore, we decided to use the keywords prompt without special tokens as our second submission (that is, LED keywords w/o ST model in Table 4).

Model	Relevance ↑				Readability ↓		Factuality ↑
	R-1	R-2	R-L	BERTScore	FKGL	DCRS	BARTScore
Submission 1	**0.4744**	**0.1497**	**0.4431**	0.8558	**11.8394**	**9.8404**	-2.3399
Submission 2	0.4715	0.1476	0.4399	**0.8562**	11.8782	9.8989	**-2.3335**

Table 6: The evaluation results of the testing data for the two submissions.

4.3 Results

We present the evaluation scores for each metric in Table 6 for both of the submissions and show the generation comparison in Appendix A. Based on the overall judgment, Submission 1 outperformed Submission 2 and was chosen as our final submission for comparison with other teams. Our system achieved an overall 10th ranking and placed third in the readability metrics.

5 Conclusion

Overall, fine-tuning the pre-trained LED model for lay summary generation tasks on the PLOS and eLife datasets produced satisfactory results. Surprisingly, the model demonstrated excellent scores on readability metrics, indicating its effectiveness in text simplification. However, the use of information design prompts such as keywords and headings did not improve the LED model's ability to learn and generate better lay summaries as expected, and most of the evaluation metrics declined.

The final results of our experiments have revealed the need for further investigation and research into the prompt strategy that was initially intended to enhance the generation of lay summaries. We aim to determine the underlying causes of the decline observed in the evaluation metrics. Possible factors contributing to this decline include inadequate quality or insufficient differentiation of keywords from the content or abstract, excessive links or correlations within the keywords, or inconsistencies in the formatting of headings across articles, which may hinder the model's ability to grasp the importance of each paragraph's content. To address these issues and improve the effectiveness of the prompts, we will explore and develop new methods in future research.

Limitations

The performance of the model heavily relies on the quality and relevance of the keywords and headings provided in the dataset. If the dataset lacks rigor during the column labeling stage or if the information in these fields is inadequate, it can lead to a decrease in the overall effectiveness of the model. Moreover, incorporating additional prompt tokens may introduce length constraints, potentially resulting in article truncation and the loss of crucial information.

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023. Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yubo Zhang, Xingxing Zhang, Xun Wang, Si-qing Chen, and Furu Wei. 2022. Latent prompt tuning for text summarization. *arXiv preprint arXiv:2211.01837*.

A Appendix

We provide an example of the eLife validation set for Submission 1 in Table 7 and Submission 2 in Table 8, with the following highlighting guidelines:

1. Factually consistent to **Gold** and fluent.
2. Factually consistent to **Gold** but not fluent.
3. Facutally inconsistent to **Gold** but fluent.
4. Facutally inconsistent to **Gold** and not fluent

The sentences in the gold article corresponding to the relevant generated sentences are highlighted in orange. All the highlights are initially provided by ChatGPT (GPT-4) and then manually reviewed by the authors. Note that the second category (purple) is not found in this example.

Table 7: Example of the eLife dataset (Gold and Submission1).

(Gold) Our ears were not designed for the society our brains created . The World Health Organization estimates that a billion young adults are at risk for hearing problems due to prolonged exposure to high levels of noise . For many people , the first symptoms of hearing loss consist in an inability to follow a single speaker in crowded places such as restaurants . However , when Parthasarathy et al . examined over 100 , 000 records from the Massachusetts Eye and Ear audiology database , they found that around 10% of patients who complained about hearing difficulties were sent home with a clean bill of hearing health . This is because existing tests do not detect common problems related to understanding speech in complex , real-world environments: new tests are needed to spot these hidden hearing disorders . Parthasarathy et al . therefore focused on identifying biological measures that would reflect these issues . Normally , the brain can 'unmix' different speakers and focus on one person , but even in the context of normal hearing , some people are better at this than others . Parthasarathy et al pinpointed several behavioral and biological markers which , when combined , could predict most of this variability . This involved , for example , measuring the diameter of the pupil while people are listening to speech in the presence of several distracting voices (which mirrors how intensively they have to focus on the task) or measuring the participants' ability to detect subtle changes in frequency (which reflects how fast-changing sound elements are encoded early on in the hearing system) . The findings show that an over-reliance on high-level cognitive processes , such as increased listening effort , coupled with problems in the early processing of certain sound traits , was associated with problems in following a speaker in a busy environment . The biological and behavioral markers highlighted by Parthasarathy et al do not require specialized equipment or marathon sessions to be recorded . In theory , these tests could be implemented into most hospital hearing clinics to give patients and health providers objective data to understand , treat and monitor these hearing difficulties.
(Submission 1) When we hear someone speak, our ears adjust their volume to match the volume of the speaker's voice. This process, known as speech processing, is particularly important in noisy environments, such as those in the workplace or social settings. However, it is not clear why some people struggle to process speech in quiet environments. To address this question, Koelewijn et al. recruited 23 young or middle-aged adults to a hearing clinic in Massachusetts, where they underwent a series of tests to determine how well they could process speech. The results showed that people with reduced hearing were less likely to be able to distinguish speech from background speakers. In addition, people with impaired hearing were more likely to struggle to distinguish between speech and background noise. These findings suggest that people who struggle to hear background speakers may have difficulties with speech processing. In the future, it may be possible to develop new ways to help people with hearing difficulties. For example, using a technique called electroencephalography (EEG) to measure the activity of the brain in response to speech could help to identify people who may have difficulty with speech. This approach could also be used to help identify people with hidden hearing disorders. In particular, it could help identify individuals who may struggle to understand speech. To test this idea, the volunteers were asked to perform a task in which they had to listen to a male speaker producing a string of four digits simultaneously. The volunteers were then asked to answer questions about their brain activity and their ability to process the four digits. This revealed that the volunteers' ability to distinguish sounds from background noise depended on a number of factors, including how much time they spent listening to the speaker, and how sensitive they were to changes in the frequency of the speech. These factors were also found to affect how well the volunteers could distinguish speech between different speakers. These results suggest that the individuals with a hidden hearing disorder may be better able to detect speech than people with a normal hearing disorder. The next step following on from this work will be to develop a better way to detect hidden hearing conditions.

Table 8: Example of the eLife dataset (Submission2), corresponding to Table 7.

(Gold) Our ears were not designed for the society our brains created . The World Health Organization estimates that a billion young adults are at risk for hearing problems due to prolonged exposure to high levels of noise . For many people , the first symptoms of hearing loss consist in an inability to follow a single speaker in crowded places such as restaurants . However , when Parthasarathy et al . examined over 100 , 000 records from the Massachusetts Eye and Ear audiology database , they found that around 10% of patients who complained about hearing difficulties were sent home with a clean bill of hearing health . This is because existing tests do not detect common problems related to understanding speech in complex , real-world environments: new tests are needed to spot these hidden hearing disorders . Parthasarathy et al . therefore focused on identifying biological measures that would reflect these issues . Normally , the brain can 'unmix' different speakers and focus on one person , but even in the context of normal hearing , some people are better at this than others . Parthasarathy et al pinpointed several behavioral and biological markers which , when combined , could predict most of this variability . This involved , for example , measuring the diameter of the pupil while people are listening to speech in the presence of several distracting voices (which mirrors how intensively they have to focus on the task) or measuring the participants' ability to detect subtle changes in frequency (which reflects how fast-changing sound elements are encoded early on in the hearing system) . The findings show that an over-reliance on high-level cognitive processes , such as increased listening effort , coupled with problems in the early processing of certain sound traits , was associated with problems in following a speaker in a busy environment . The biological and behavioral markers highlighted by Parthasarathy et al do not require specialized equipment or marathon sessions to be recorded . In theory , these tests could be implemented into most hospital hearing clinics to give patients and health providers objective data to understand , treat and monitor these hearing difficulties.
(Submission 2) People with hidden hearing disorders often have difficulty hearing. This is because they are unable to distinguish the speech of other people in the room. This difficulty is caused by the loss of the ability to distinguish between speech waves from background speakers. This loss of hearing may be due to the loss in the ability of the brain to process speech in a noisy environment. However, it is not clear why people with hidden sound disorders struggle to distinguish speech waves. To address this question, Johnson et al. examined over 100, 000 patients who had been referred to a hearing clinic because of their hearing loss. The analysis revealed that the individuals with hidden sounds disorders were not only less able to distinguish different speech waves than those with normal hearing. Instead, they were more likely to struggle to process the speech in noisy environments. This suggests that the brain may be using a combination of two mechanisms to distinguish distinct speech waves: high frequency sound waves and low frequency soundwaves. Johnson et Al. also found that the patients with hidden speech disorders were more sensitive to high frequency sounds than those without. This was particularly true for those individuals whose hearing was impaired. The experiments show that the brains of people with hearing disorders are more sensitive than those of people without. These findings suggest that the ability for the brain's ability to process different speech sounds is due to a combination: high-frequency sound waves are processed by the brain, and low-frequency sounds are processed more by the ear. The findings of Johnson et et al' suggest that people with a hidden hearing disorder may struggle to discriminate between speech waveforms. The next step is to develop new tests that can detect hidden hearing impairments in people with these conditions.