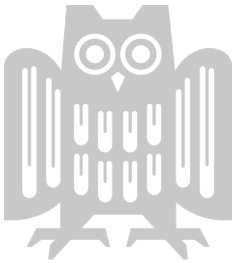
Exploring Paraphrasing for Enhancing Speech Perception in Noisy Environments



A dissertation submitted towards the degree
Doctor of Engineering (Dr.-Ing.)
of the Faculty of Mathematics and Computer Science
of Saarland University

Anupama Chingacham

Saarbrücken
2024

DAY OF COLLOQUIUM:

09.12.2024

DEAN OF THE FACULTY:

Prof. Dr. Roland Speicher

EXAMINATION BOARD:

Chair	Prof. Dr. Jürgen Steimle
Reviewer	Prof. Dr. Dietrich Klakow
Reviewer	Prof. Dr. Preethi Jyothi
Academic assistant	Dr. Badr M. Abdullah

To my *Amma, Acha & Chechi*, who showed me,

Life is what you make it.

ABSTRACT

In the event of speech distortions caused by echoes, reverberations, or background noise like in a busy cafeteria, listening can be challenging even for individuals with normal hearing thresholds. When noise hinders listening, the meaning of a message perceived by a listener can be different from the intended meaning and it may lead to misunderstandings or even communication breakdowns in extreme cases. Unlike human speakers who can adapt speech to accommodate their interlocutor's listening difficulties, current spoken dialogue systems are limited in their ability to produce noise-robust speech. Most algorithmic solutions to synthesize noise-robust speech are based on acoustic modification, which is not beneficial in all cases as it may lead to signal distortions, deteriorating the naturalness or the quality of the synthesized speech. This thesis proposes to utilize an alternative strategy of *utilizing paraphrases to improve speech perception in noise*, which involves no signal distortions.

Noise impacts lexical units differently – some are more noise-robust while others are more prone to misperception. Hence, paraphrasing does not guarantee better speech perception. If the lexical units used for paraphrasing are perceived in a listening situation in a similar way to the original formulation, they do not represent an improvement. Hence, the first study in this thesis aims to gain a better understanding of whether and to what extent a simple yet common paraphrasing strategy – lexical replacement with synonyms – could reduce word misperceptions in noise. Human listening experiments were conducted to capture the perception differences among synonyms in noise. Analyzing the newly created dataset – Synonyms-in-Noise – it was found that replacing a lexical unit with its synonym that is less risky to be misheard can improve word recognition by up to 37% in a highly noisy environment like babble noise at SNR-5 dB. Furthermore, a modeling experiment was performed to explain the observed gain in intelligibility. The results show that the intelligibility gain is attributed to the linguistic cues of synonyms, in low and medium noisy conditions; while the gains are mainly driven by acoustic cues in highly noisy conditions.

In order to consider more generic types of paraphrases, the second study of the thesis focuses on sentential paraphrases and their impacts on the whole utterance intelligibility. By collecting human speech perception data of sentential paraphrases,

a new dataset called Paraphrases-in-Noise was created. It was found that the intelligibility scores of sentential paraphrases are also significantly different in a highly noisy condition and choosing the right paraphrase within a pair can introduce an overall gain in intelligibility as high as 33%. Additionally, the study proposed an intelligibility-aware paraphrase ranking model to correctly identify the more intelligible paraphrases using their linguistic and acoustic features. The proposed model outperformed both baseline models (random and majority), achieving the highest performance of 67% at a high noise condition.

The final study of the thesis aims to *generate acoustically better intelligible paraphrases*, which is potentially useful to build noise-adaptive spoken dialogue systems. The investigation begins with an evaluation of the extent to which modern text generation models such as Large Language Models (LLMs) can produce texts that fulfill both textual (such as semantic equivalence) and non-textual attributes (such as acoustic intelligibility). Modeling results showed that LLMs in standard prompting setups struggle to improve acoustic intelligibility, while effectively maintaining semantic equivalence. Additionally, it was found that the proposed post-processing approach – *prompt-and-select* – performs better than fine-tuned models at generating paraphrases that are acoustically more intelligible.

In summary, this thesis explored the potential of paraphrasing to improve speech perception in noise. As a result, we created two new datasets and proposed a new framework to synthesize noise-robust speech, which introduces no signal distortions.

ZUSAMMENFASSUNG

Im Falle von Sprachsignalverzerrungen durch Echos, Nachhall oder Hintergrundgeräusche, wie zum Beispiel in einem belebten Café, kann das Zuhören selbst für Personen mit normalem Hörvermögen eine Herausforderung darstellen. Wenn Lärm das Zuhören behindert, kann die Bedeutung einer Nachricht, die der Hörer wahrnimmt, von der vom Sprecher beabsichtigten Bedeutung abweichen. Das kann zu Missverständnissen oder in extremen Fällen sogar zu Kommunikationsabbrüchen führen. Im Gegensatz zu menschlichen Sprechern, die ihre Sprechweise an die Hörschwierigkeiten ihres Gesprächspartners anpassen können, sind die derzeitigen Sprachdialogsysteme nur begrenzt in der Lage, geräuschrobuste Sprache zu produzieren. Die meisten algorithmischen Lösungen zur Synthese geräuschrobuster Sprache basieren auf akustischen Modifikationen, die nicht in allen Fällen von Vorteil sind, da sie zu Signalverzerrungen führen können, die die Natürlichkeit oder Qualität der synthetisierten Sprache beeinträchtigen. In dieser Arbeit wird eine alternative Strategie zur Verbesserung der Sprachwahrnehmung bei Störgeräuschen vorgeschlagen, die keine Signalverzerrungen mit sich bringt: die Verwendung von Paraphrasen.

Geräusche wirken sich unterschiedlich auf verschiedene lexikalische Einheiten aus - einige sind geräuschresistenter, während andere anfälliger für Fehlwahrnehmungen sind. Daher ist die Verwendung von Paraphrasen keine Garantie für eine bessere Sprachwahrnehmung. Wenn die lexikalischen Einheiten, die zur Umschreibung verwendet werden, in einer Hörsituation ähnlich wahrgenommen werden wie die ursprüngliche Formulierung, stellen sie keine Verbesserung dar. Daher zielt die erste Studie in dieser Arbeit darauf ab, ein besseres Verständnis dafür zu erlangen, ob und inwieweit eine einfache, aber weit verbreitete Paraphrasierungsstrategie - die lexikalische Ersetzung durch Synonyme - Wortfehlwahrnehmungen im Lärm reduzieren kann. Es wurden Hörexperimente durchgeführt, um die Wahrnehmungsunterschiede zwischen Synonymen im Lärm zu erfassen. Die Analyse des zu diesem Zweck neu erstellten Datensatzes – *Synonyms-in-Noise (SiN)* – ergab, dass das Ersetzen einer lexikalischen Einheit durch ein Synonym mit geringerem Risiko, falsch verstanden zu werden, die Worterkennung in einer stark verrauschten Umgebung (z.B. bei Babble Noise mit einem Signal-Rausch-Verhältnis von -5 dB) um bis

zu 37% verbessern kann. Außerdem wurde ein Modellierungsexperiment durchgeführt, um den beobachteten Gewinn an Verständlichkeit zu erklären. Die Ergebnisse zeigen, dass der Verständlichkeitsgewinn bei geringem und mittlerem Lärm auf die linguistischen Merkmale von Synonymen zurückzuführen ist, während der Gewinn bei starkem Lärm hauptsächlich von akustischen Merkmale bestimmt wird.

Um allgemeinere Arten von Umschreibungen zu berücksichtigen, konzentriert sich die zweite Studie der Arbeit auf Satzumschreibungen und ihre Auswirkungen auf die Verständlichkeit der gesamten Äußerung. In weiteren Hörexperimenten wurden Satzparaphrasen verglichen und ein neuer Datensatz namens *Paraphrases-in-Noise (PiN)* erstellt. Es wurde festgestellt, dass sich die Verständlichkeitswerte von verschiedenen Satzparaphrasen auch unter stark verrauschten Bedingungen signifikant unterscheiden und die Wahl der richtigen Paraphrase innerhalb eines Paares einen Gesamtgewinn an Verständlichkeit von bis zu 33% bewirken kann. Darüber hinaus wurde in der Studie ein verständlichkeitsorientiertes Paraphrasen-Ranking-Modell vorgeschlagen, um die verständlichsten Paraphrasen anhand ihrer linguistischen und akustischen Merkmale korrekt zu identifizieren. Das vorgeschlagene Modell übertraf beide Basismodelle (Zufalls- und Mehrheitsmodell) und erreichte mit 67% die höchste Verständlichkeit bei starkem Rauschen.

Die abschließende Studie dieser Arbeit zielt darauf ab, akustisch besser verständliche Paraphrasen zu generieren, die potenziell nützlich sein könnten, um geräuschadaptive Sprachdialogsysteme zu entwickeln. Wir evaluieren, inwieweit moderne Textgenerierungsmodelle wie Large Language Models (LLMs) Texte produzieren können, die sowohl textuelle Anforderungen (z.B. semantische Äquivalenz) als auch nicht-textuelle Anforderungen (z.B. akustische Verständlichkeit) erfüllen. Die Ergebnisse der Studie zeigen, dass LLMs in Standard-Prompting-Setups Schwierigkeiten haben, die akustische Verständlichkeit zu verbessern und gleichzeitig die semantische Äquivalenz effektiv zu erhalten. Außerdem wurde festgestellt, dass der vorgeschlagene Nachbearbeitungsansatz - *prompt-and-select* – besser abschneidet als fein abgestimmte Modelle, wenn es darum geht, Paraphrasen zu erzeugen, die akustisch besser verständlich sind.

Zusammenfassend lässt sich sagen, dass in dieser Arbeit das Potenzial der Paraphrasierung zur Verbesserung der Sprachwahrnehmung im Lärm untersucht wurde. Als Ergebnis haben wir zwei neue Datensätze erstellt und einen neuen Rahmen für die Synthese von geräuschrobuster Sprache vorgeschlagen, der keine Signalverzerrungen verursacht.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the unwavering support, guidance, and encouragement of numerous individuals.

Foremost, I extend my heartfelt gratitude to my esteemed supervisors, Dietrich Klakow and Vera Demberg, for graciously welcoming me into the realm of research and providing an excellent opportunity to learn under their tutelage. I am indebted to Dietrich for fostering an environment of intellectual freedom, empowering me to explore new avenues, and guiding me back on track whenever I felt disoriented. Likewise, I express my profound appreciation to Vera for her insightful critiques and constructive feedback. Her dedication to critically scrutinizing my research has greatly contributed to empowering me as a researcher. Most importantly, I am thankful to both of them, for cultivating environments in their respective laboratories that celebrate diversity and inclusivity, providing a vibrant space to learn and grow.

My deepest gratitude also goes to my colleagues who made my entire PhD journey not only academically enriching but also immensely enjoyable and inspiring. Special thanks to Marjolein and Pratik for their numerous insightful discussions, particularly during the initial stages of my PhD, when I was struggling to gain clarity. I am immensely thankful to Badr, Marius, Michael, Dana, Dave, David, and Iona, whose constructive feedback and comments at various stages of my work propelled me forward and encouraged me to explore new avenues in my research. I am particularly grateful to my collaborator, Miaoran, for sharing her expertise and insightful comments to improve the quality and rigor of this work. Also, thanks to Vagrant, Marian, and Merel, for their generous sharing of feedback that significantly enhanced the presentation of this work. I extend my heartfelt appreciation to Aravind, Alex, Volha, Dawei, Xudong, Margarita, Iza, Omnia, Jacek, Julia, and SFB folks, for the delightful moments that we shared during board game evenings, LSV breakfasts, dinner parties, and group hikes, which were much-needed respite and camaraderie amidst the rigors of academic pursuit. A special mention goes to our lab admin, Nico, for his promptness and diligence in ensuring a robust lab infrastructure to run our experiments on time. Also, I am indebted to the office secretaries – Patricia, Gabi, Angelika, Khanda, and Claudia – for their willingness to go above and beyond in helping me navigate the intricacies of German administration, both within and

beyond the university. I am also thankful to my academic mentors Aghilas Sini, Heiner Drenhaus, Helena Moniz, and Dennis Paperno for their invaluable guidance at various stages of my academic journey.

Finally, but most importantly, I am deeply grateful to my family and friends, for being there for me, keeping a tap of my well-being, and reassuring me that everything will be alright in the end (and if it is not, that is not the end :)). Interestingly, my greatest inspiration to embark on doctoral studies is none other than my husband, Sunit, who was a doctoral student at the time. I am immensely thankful to him, for rekindling my academic interest and being my rock support throughout this journey. I am also infinitely grateful to my dad, Venugopal, my sister, Archana, and her husband, Ranjith, for providing me the best opportunities to learn and thrive, from a very young age. Their unwavering expression of trust in my abilities greatly bolstered me and helped me overcome my self-doubt, on multiple occasions. Likewise, I am profoundly appreciative of the support from Sunit's parents, his fantastic family, as well as, my extended family of wonderful cousins, uncles, and aunts. Their support significantly helped me to focus on my studies, while staying sane. I won't be able to thank my friends Maya, Blessy, Paru, Aparna, Akhi, Adi, Harish, Pavan, and Niharika, enough for generously expressing their love and care for me, which has greatly strengthened me throughout my studies.

Lastly, I am profoundly grateful to my mom, Hemalatha, who exemplified empathy and perseverance throughout her life. Her values, words, and memories have been my guiding lights in this journey, enabling me to overcome challenges and move forward every day.

CONTENTS

1 INTRODUCTION

Conversations are critical for human life. However, not all conversational environments are ideal for message exchanges. For instance, in real-world conversational setups, there is often background noise like competing speech, appliance noise, automobile noise, loud music, or public announcements. As noise distorts clean speech, listening can be challenging in noisy environments, even for individuals with normal hearing threshold. When a listener incorrectly recognizes the spoken sounds, the meaning perceived by the listener can be different from the actual meaning intended by the speaker, and it may lead to misunderstanding or even conversation breakdowns in extreme cases (Grimshaw, 1980; Dua, 1990). This highlights the need to enhance the intelligibility of speech produced in noisy environments.

With the growing presence of voice assistants in the modern world, human mishearing is not just limited to human-human interactions; it can also occur in human-machine interactions. However, mishearing mitigation efforts vary between these types of interactions. In an extensive review on speech production, Cooke *et al.* (2014) have showcased a wide range of speech modification strategies employed by human speakers to reduce their interlocutors' listening difficulties. They refer to human speakers as *listening talkers*, indicating their potential to adapt speech, in such a way that their interlocutors can hear and understand more easily. Compared to human speakers, spoken dialogue systems (SDSs) are still less adaptive to the listening challenges of their users. This could be a factor in limiting the scope of SDS to only users in ideal listening conditions, ignoring a large set of individuals who live or work in adverse listening conditions. In the last two decades, research in speech technology has therefore focused on developing adaptive SDS to synthesize noise-robust speech, thereby improving the human-machine interactions, especially in adverse listening environments (Patel *et al.*, 2006; Bonardo and Zovato, 2007; Anumanchipalli *et al.*, 2010; Cooke *et al.*, 2013a; Rennies *et al.*, 2020; Ohashi and Higashinaka, 2022).

1.1 MOTIVATION

Interestingly, research on algorithmic solutions to enhance synthetic speech intelligibility is predominantly driven by acoustic modification. In other words, text-to-speech (TTS) systems are trained to modify a range of acoustic features like pitch, intensity, and duration of synthesized speech to improve its intelligibility in noisy environments (Mayo *et al.*, 2012; Bollepalli *et al.*, 2019). On one hand, such speech modification strategies like imitating Lombard speech (Huang *et al.*, 2010) have resulted in improved intelligibility in noise. On the other hand, the benefits were not large enough to bridge the gap in intelligibility introduced by the lack of linguistic features like predictability – even with acoustic modification, words in less predictable contexts remain less intelligible than words in highly predictable contexts (Valentini-Botinhao and Wester, 2014). More critically, acoustic modification is not beneficial in all cases as it may lead to signal distortions, which could negatively influence the naturalness and the quality of the synthesized speech (Cooke and Lecumberri, 2012; López-Peláez and Clark, 2014; Valentini-Botinhao and Wester, 2014). Given the ubiquity of voice assistants in today's world (like in traffic navigation, customer care services, and medical assistance), it has become critical to additionally *explore alternate strategies to improve the synthetic speech intelligibility in noise*.

Considering the fact that the ultimate objective of a dialogue exchange is to convey the meaning of a message, enhancing the intelligibility of synthetic speech can be achieved by performing not only acoustic modification but also linguistic modification. In fact, human speakers are known for their ability to simplify the linguistic structure and word complexity in speech, especially when they converse with non-natives or kids in the early stage of language acquisition (Cooke *et al.*, 2014). This motivates us to investigate a speech modification strategy based on linguistic modification to improve speech perception in noise. Notably, a linguistic-oriented strategy could lead to *an approach of enhancing speech intelligibility without introducing any signal distortions*.

To this end, we identified the three main objectives of this work and they are described in the following section.

1.2 RESEARCH OBJECTIVES

1. **Verify the potential of paraphrases:** It is well-established that noise impacts words differently as linguistic features like word familiarity and predictability can influence word intelligibility in noise (Luce and Pisoni, 1998; Kalikow *et al.*, 1977). Similarly, the underlying sounds of a word (*ie.,* consonants and vowels) are differently misrecognized in different listening conditions (Weber and Smits, 2003), leading to higher misperceptions for some words than others. It is however not understood ***whether the noise impact is different among semantically equivalent phrases*** (ie., synonyms or paraphrases) and its interaction with different noise levels. A significant intelligibility difference among paraphrases is critically important, as paraphrasing may not be impactful for improving perception if paraphrases exhibit similar intelligibility scores. Therefore, we identify the need to study the intelligibility differences among (spoken) paraphrases, which are synthesized by a TTS and listened to by native listeners in different adverse listening environments.

2. **Explain the potential of paraphrases:** A rephrasing-based approach to enhance synthetic speech intelligibility was initially proposed in Zhang *et al.* (2013). They focused on developing an objective metric to distinguish the paraphrases' intelligibility, with the assumption that all paraphrases are different in intelligibility under all listening conditions. Their study focused little on ***explaining why and to what extent paraphrasing supports better intelligibility in noisy conditions*** and therefore it is essential to fill this gap in understanding the potential of paraphrases. Since paraphrases consist of a wide range of syntactic and lexical variations (Bhagat and Hovy, 2013), an elaborate analysis to verify whether a simple yet common paraphrasing approach like lexical replacement contributes to better utterance intelligibility in noise would be beneficial for further research.

3. **Utilize the potential of paraphrases:** More recently, there has been an increasing interest in building adaptive natural language generation (NLG) systems that could adapt to users in different listening conditions (Ohashi and Higashinaka, 2022). They employed a word confusion model to simulate the noise impact on the synthesized utterance. Then, the noisy utterance is given as input to a natural language understanding (NLU) module to determine whether

the generated utterance is understood as intended. Their study employed an automatic speech recognition (ASR) system as a simulated listener to facilitate large-scale analysis. However, actual listening is more complex than replacing words based on a confusion model, especially in noisy environments, wherein human listeners leverage their world knowledge and language proficiency to compensate for their listening difficulties. Thus, the question of ***how to employ paraphrasing for better perception of synthetic speech in noise*** remains a challenge. For better understanding, it is also critical to evaluate proposed approaches with human listeners in a given noisy environment.

1.3 CONTRIBUTIONS

This section discusses the main contributions of this thesis which address the aforementioned research objectives.

- **A method to improve the intelligibility of synthetic speech in noise *without* signal distortions.**
  Instead of modifying the acoustic characteristics like pitch, intensity, and duration of an utterance, we employed paraphrasing, which introduces no signal distortions, to improve the intelligibility of an utterance in noise. We demonstrated that replacing a lexical item with an ideal synonym can significantly improve word intelligibility in noisy conditions, even among native listeners with no hearing impairments. Additionally, the outcome of our results provides evidence that not just the individual word intelligibility, but even the intelligibility of whole sentences can be improved by rephrasing sentences with syntactic and lexical variations, especially in highly noisy conditions.

- **An extensive investigation into *why* paraphrases improves hearing in noise.**
  We built models to explain the features that drive the intelligibility gain introduced by paraphrases. Modeling experiment results show that the intelligibility gain is primarily attributed to the linguistic characteristics of paraphrases, only in listening setups with low/medium noise levels – words with higher linguistic predictability were better recognized because the supporting context was available to listeners, and that helped them in compensating the noisy audio of target utterance. However, in a highly noisy condition, paraphrases are differently perceivable because of their noise-robust acoustic cues – the

sentence with more acoustic cues that survived the energetic masking was significantly more audible than its paraphrase. Overall, the study explains the factors that drive the intelligibility gain through paraphrases.

- **Creation of two *new datasets* for paraphrases perceived in noise.**
  Evaluating the paraphrase intelligibility in noise involves speech perception experiments with human subjects, which is expensive and time-consuming. We created two large perception experiments on a crowd-sourcing platform with about 200 native English listeners. Human annotations of paraphrases perceived in different listening environments were collected and the resultant datasets (Synonyms in Noise and Paraphrases in Noise) were released for further research in this direction.

- **Proposal of a new framework for *noise-adaptive* spoken dialogue systems.**
  Modern SDSs are less adaptive to the listening difficulties of human users, especially in noisy environments. The system utterances are typically synthesized by a TTS module, which takes the input from an NLG module that generates the text to convey a message. Considering the potential of paraphrases in enhancing human hearing in noise, we proposed to employ paraphrasing in an SDS framework to reduce misperceptions by users in adverse listening conditions. We developed a paraphrase-pair ranking model that can be used with any paraphrasing model, to identify the sentential paraphrases that are more noise-robust for a given noise condition.

- **Evaluation of Large Language Models for *generating intelligible paraphrases*.**
  The task of generating intelligible paraphrases involves both textual and non-textual attributes: the generated text needs to be *semantically equivalent* to as well as *acoustically more intelligible* than the given input sentence. Modeling results showed that in zero-shot and in-context learning setups, Large Language Models (LLMs) fail to learn the non-textual attributes that are hard to describe in text. Fine-tuning models with datasets of downstream tasks is another approach for controllable text generation. However, it demands large-scale datasets. We developed a data augmentation pipeline with an automatic speech intelligibility metric to develop a large parallel dataset of paraphrases with different acoustic intelligibility. Finally, we proposed a new prompting approach called prompt-and-select, which avoids the compute-intense fine-tuning step, to generate text that satisfies both the desired textual and non-textual attributes.

The contributions discussed in this dissertation center around the following publications:

1. Anupama Chingacham, Vera Demberg and Dietrich Klakow (2021). *Exploring the Potential of Lexical Paraphrases for Mitigating Noise-Induced Comprehension Errors.* In Proceedings of the International Speech Communication Association (**INTERSPEECH**). URL. **Best student paper award**.

2. Anupama Chingacham, Vera Demberg and Dietrich Klakow (2022). *A Data-Driven Investigation of Noise-Adaptive Utterance Generation with Linguistic Modification.* In Proceedings of the IEEE Spoken Language Technology Workshop (**IEEE SLT**). URL.

3. Anupama Chingacham, Miaoran Zhang, Vera Demberg and Dietrich Klakow (2024). *Human Speech Perception in Noise: Can Large Language Models Paraphrase to Improve It?* To appear in Proceedings of the Human-Centered Large Language Modeling Workshop (**HuCLLM @ ACL 2024**).

1.4 THESIS STRUCTURE

This section provides a brief introduction to all chapters of this thesis.

Background. Conversations are prone to mishearing and misunderstanding, especially in adverse listening environments. This is true not only for human-human interactions but also for human-machine interactions, wherein the human listeners are located in a noisy environment. With the increasing presence of speech technologies in human life, it becomes critical to understand the different factors and existing algorithms to improve speech perception in noisy conditions. In **Chapter 2**, we present a survey of different factors that influence human misperceptions and some of the mishearing mitigation strategies that are adopted by human speakers, to reduce the listening challenges of interlocutors in less ideal environments. We also describe the existing methods to define, measure, and represent misperception. We then define paraphrases and introduce existing methods to identify and generate paraphrases, which we will adopt in later chapters of this thesis. Overall, this chapter sets the context and defines the scope of our investigation.

Algorithmic solutions to improve speech perception in noise. Speech perception is the task of recognizing spoken tokens (like phonemes, words, or sentences) in an utterance. On the other hand, speech comprehension involves understanding the meaning, in addition to recognizing spoken tokens. As our research focuses on improving speech perception with paraphrases (*ie.,* different linguistic forms of similar meaning), it is then important to formally define speech perception and speech comprehension and we present it in **Chapter 3**. We describe mishearing and how paraphrasing is a potential solution to reduce comprehension errors introduced by mishearing. In the same chapter, we discuss the two possible frameworks to mitigate mishearing, highlighting the critical distinction between linguistic modification and acoustic modification for enhancing the intelligibility of synthetic speech. Finally, we present a strong motivation to use paraphrasing to improve speech perception, showcasing how word confusions are influenced by masker types.

Lexical paraphrases to mitigate word misperception. Many studies have demonstrated the significant influence of lexical characteristics on word intelligibility in noise. More precisely, they showcase that noise impacts lexical units differently (Luce and Pisoni, 1998; Vitevitch, 2002). We hypothesize that replacing a lexical unit with its synonym which is more noise-robust, can improve the overall word intelligibility in noise. But first, we need to study whether synonyms differ in intelligibility under noisy listening conditions. This is a critical aspect of this investigation – if the masking effect of noise on a word and its synonyms are equally the same, then choosing one over the other is less likely to improve intelligibility. In **Chapter 4**, we present the listening experiment, which was conducted to collect human perception data of synonyms in different noise conditions. We then describe the newly created dataset called Synonyms in Noise (SiN), which is annotated by native listeners with normal hearing thresholds. We found that synonyms' intelligibility scores are significantly different in noise, especially at a high noise level (SNR -5 dB). Interestingly, we observed a similar pattern of intelligibility difference among synonyms, even when they were presented with linguistic contexts. With further investigation using linear regression models, we observed that the intelligibility gain introduced by synonyms is mainly driven by their linguistic characteristics under clean/low noise conditions and acoustic characteristics under high noise conditions.

Sentential paraphrases to improve intelligibility. At this point, we established evidence that replacing a lexical unit with its synonym, which is more noise-robust,

can significantly improve speech perception, leading to improved spoken language comprehension. While this is an interesting finding, an intelligibility improvement strategy solely based on lexical replacements is constrained by the availability of synonyms that fit the context of a given sentence. Also, the words that undergo the lexical replacement might not be crucial for the perception of the whole sentence, as certain misperceptions can be corrected by high-level signals like linguistic contexts or situational cues. It is then important to extend our investigations to sentential paraphrases, which introduce both syntactic and lexical variations. We present in **Chapter 5**, a new perception experiment that we conducted with 300 sentential paraphrases and the resultant dataset called Paraphrases in Noise (PiN). To measure the intelligibility difference between sentential paraphrases, we then define a metric called Sentence-level Intelligibility (Sent-Int), which captures how well an utterance is perceived by a group of listeners in a listening condition. Analyzing the perception data, we observed that even at the sentence level, the noise impact is significantly different among sentential paraphrases, and choosing the more noise-robust paraphrase can improve the overall perception with a relative gain of 33%, under high noise conditions (babble noise at SNR -5 dB). Once again, we found that the intelligibility difference among paraphrases is driven by their acoustic cues, highlighting the benefits of paraphrasing to represent a message with a less energetic masking effect. We also demonstrate the potential of building a noise-adaptive spoken dialogue system, with a prototype of an intelligibility-aware paraphrase ranking model to select the more noise-robust sentence within a pair of paraphrases.

Generate acoustically intelligible paraphrases. An intelligibility-aware paraphrase ranking method is useful. However, the benefits of paraphrasing will be limited by the diversity of paraphrases available for selection. In other words, when the paraphrase candidates involve only trivial paraphrases (*i.e.,* phrases that differ in tense/voice) could lead to limited benefits in using one phrase over the other. Hence, in **Chapter 6**, we shift our focus to specifically generate paraphrases that are acoustically more intelligible than an input sentence, in a given listening environment (*ie.,* babble noise at SNR -5 dB). The generation of paraphrases with better acoustic cues involves both the *textual attribute* (semantic equivalence) and the *non-textual attribute* (acoustic intelligibility in a listening condition). In other words, the objective of this study is to perform a paraphrase generation task controlled by a non-textual attribute. To do so, we use LLMs, which have shown an incredible capability for several text generation tasks including paraphrase generation and controllable text

generation. We observe that fine-tuning LLMs with a small dataset does not help improve acoustic intelligibility. However, increasing the size of the fine-tuning dataset helps the model generate paraphrases with improved acoustic intelligibility. We also show that a pre-trained language model without any fine-tuning (ie., in zero-shot and in-context learning setups), fails to generate text with the desired non-textual attribute, while efficiently handling the desired textual attribute. We propose a new approach called *prompt-and-select*, which involves generating multiple paraphrases that are semantically equivalent to the input text and employing a paraphrase selection method based on acoustic intelligibility, thereby decoupling the desired textual and non-textual attributes in the text generation pipeline.

In summary, this thesis establishes the evidence that paraphrasing is a useful strategy to improve the intelligibility of synthetic speech in noise and describes some of the frameworks to use this potential of paraphrases to build better solutions for noise-adaptive SDSs. An overview of this thesis is represented in Figure 1.1.

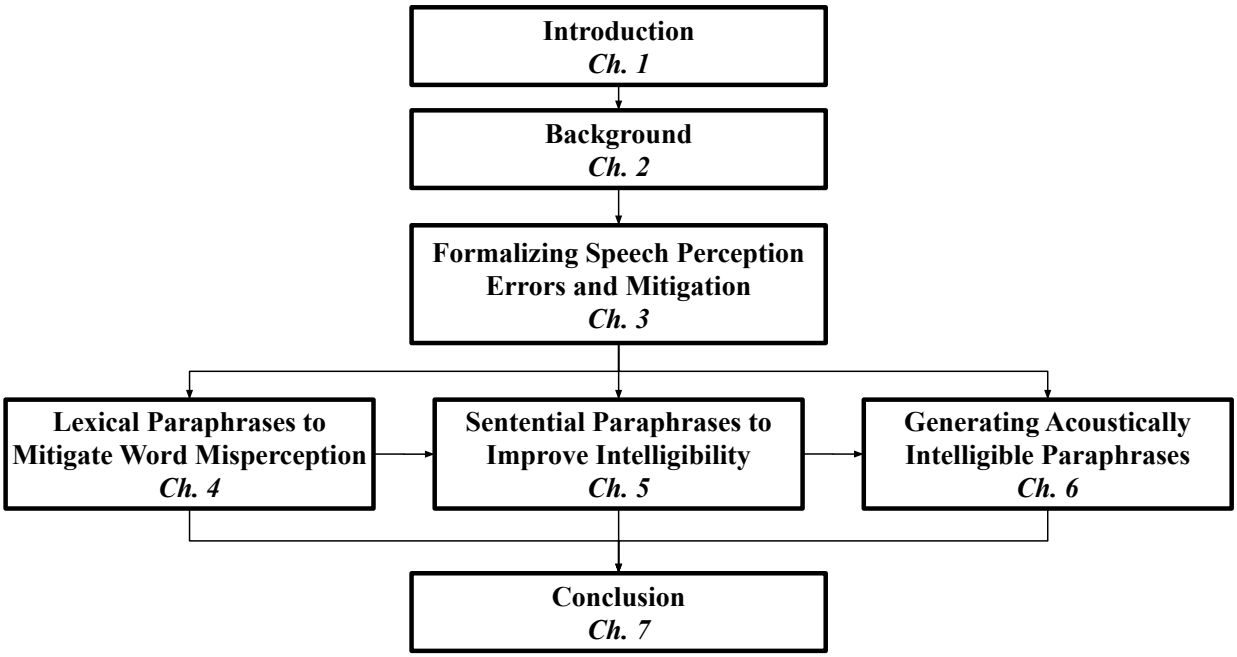


Figure 1.1: A pictorial representation of the thesis overview.

2 BACKGROUND

For humans, speech is one of the primitive modes of communication. Individuals with no speaking and hearing disorders primarily use *conversational speech* to interact with the external world. Typically, in a conversational setup, a speaker produces speech in such a way that an intended message is realized with a sequence of sounds, which listeners interpret by decoding a meaning from the perceived sounds. To achieve conversational success, listeners must interpret the meaning as intended by the speaker. Any deviation between the actual *intended meaning* and the *perceived meaning* can lead to confusion and misinterpretations in the dialogue exchange. The task of interpreting the meaning of an utterance – *Spoken Language Comprehension* (SLC) – is not trivial; rather, it is a combination of several complex processes like the auditory perception of sounds, the identification and the retrieval of meaningful sounds from the human lexicon, decryption of the underlying meaning, integration of contextual cues, etc. For successful retrieval of the intended meaning, listeners must *perceive the spoken tokens correctly* in the first place.

In the last 50 years, several models and theories were proposed to explain the underlying process of human speech perception (McClelland and Elman, 1986; Norris, 1994; Gaskell and Marslen-Wilson, 1997); for review see (Weber and Scharenborg, 2012). The two well-known frameworks – top-down and bottom-up processing – have been frequently revisited in speech science to understand how humans process speech to retrieve a message. Early proposed spoken word recognition models demonstrated that humans utilize both bottom-up, data-driven systems *as well as* top-down, knowledge-driven systems to map the low-level acoustic signals to high-level meaning representations (Marslen-Wilson and Welsh, 1978). Existing models of human speech recognition showed that high-level signals like linguistic and situational contexts are particularly useful to compensate for ambiguous or weak low-level acoustic information, such as in noisy listening environments.

Speech perception in noise. Speech perception in noisy listening environments is particularly interesting for two reasons: (1) the real-world conversational setups are usually noisy and less ideal, compared to the quiet lab environments, and (2) acoustic

noise distorts the clean speech signal, which may lead to challenges in bottom-up processing, and studying this scenario is particularly useful to gain a better understanding on the role of top-down and bottom-up processing pipelines. Additionally, prior work has shown that acoustic noise in the background can introduce listening difficulties and such difficulties may lead to mishearing, misunderstandings, and communication breakdowns (Grimshaw, 1980). Hence, to achieve communication success in real-world noisy environments, human listeners should be able to perceive speech with limited perception errors, even when they are interacting with another human or a machine.

Mishearing occurs when a listener *incorrectly recognizes* the actual spoken tokens. Mishearing is different from misunderstanding, which occurs when a listener incorrectly comprehends the meaning of the actual spoken message (see Section 3.1 for a formal distinction). In speech science, mishearing is also referred to as *Slips-of-the-Ear* (Bond, 1999) and *misperception* (Marxer *et al.*, 2016a; Albert Felty *et al.*, 2013; Cooke *et al.*, 2019), which indicates the *deviation in the actual and perceived spoken tokens*. Errors in speech recognition are also used to compare the acoustic intelligibility of speech produced by different speech synthesis systems or a group of speakers. More precisely, speech that results in a lesser number of recognition errors is identified as *acoustically more intelligible*.

Measuring misperception. Measuring misperception involves human listening experiments to collect the perception data and compare the actual and perceived speech. The recognition errors can be measured with different speech tokens like phonemes, triphones, or words. *Word Error Rate* (WER) and *Phoneme Error Rate* (PhER) refer to the proportion of insertion (I), deletion (D), and substitution (S) of tokens in the perceived speech to be correctly recognized as the actual spoken utterance, measured in terms of words and phonemes, respectively. To count the errors, *minimal edit distance* is calculated between two sequences of spoken tokens and perceived tokens. More generally, the error rate (ER) is defined as below:

$$ER = \frac{\#deletions + \#insertions + \#substitutions}{\#Total\ tokens\ in\ the\ spoken\ utterance} \tag{2.1}$$

The minimum value for ER is zero when the perceived utterance is the same as the spoken utterance. However, its ER value ranges from 0.0 to ∞, as the perceived speech can (theoretically) be infinitely long. In a classical setting, all three error types (*ie.*, I, D, and S) have equal weight of 1.0. However, different weights can be

assigned to error types to weigh them differently in the ER calculation.

Speech Reception Threshold (SRT) is another metric to measure the human ability to perceive speech in noisy environments. The SRT calculation requires an adaptive listening experiment wherein the noise level of stimuli audio is systematically increased/decreased after each trial. SRT_{50} and SRT_{80} denote the Signal-to-Noise Ratio (SNR) levels of audio, indicating the points where 50% and 80% of spoken words are accurately recognized, respectively.

For simple perception experiments like word/phoneme recognition, the proportion of correct recognitions of a token among all its listening instances is used to represent its intelligibility. This score is referred to as *recognition rate*.

Representing misperception. Perception data is not only useful for identifying the intelligibility of an utterance but also provides insights into the type of confusion/misperception. Several research studies in the past have analyzed the patterns of misperceptions under different listening setups, which are defined by the listener groups and the noise in the listening environment. Their findings are summarized in confusion matrices (Miller and Nicely, 1955; Weber and Smits, 2003; Cutler *et al.*, 2004; Phatak *et al.*, 2008). For instance, Weber and Smits (2003) demonstrated the confusion of vowels and consonants, in listening environments with babble noise, among native English speakers with normal-hearing abilities. Figure 2.1 showcases the summary of their perception experiment with babble noise at SNR 0 dB, illustrating how often a stimulus vowel is recognized correctly or misrecognized as another vowel – all diagonal entries represent the correct recognitions of a vowel and all non-diagonal values show their misperception.

In comparison to the intelligibility scores discussed above, a confusion matrix provides a more detailed view of speech misperceptions in noise. For example, the confusion matrix shows that the vowel in the lexical item *hot* is one of the most misrecognized vowels and it is mostly misrecognized as the vowel in the lexical item *caught*. Prior studies have also explored the potential of phoneme confusion matrices to build a spoken word recognition model (Luce and Pisoni, 1998), phrasal confusion models in a closed vocabulary domain (Cox and Vinagre, 2004), etc. However, generating confusion matrices involves large-scale human listening experiments, which demand both time and resources.

		response															
		beat	bit	wait	bet	bat	hot	cut	caught	boat	cook	boot	buy	boy	shout	bird	
		i	ɪ	eɪ	ɛ	æ	ɑ	ʌ	ɔ	oʊ	ʊ	u	aɪ	ɔɪ	aʊ	ɝ	miss
stimulus	i	**86.3**	4.2	.1	3.2				.3		.1	1.6	.6		.1	1.9	1.5
	ɪ	1.2	**83.1**	.6	9.3	.6		1.6	.1	.1	.1	.7	.4		.1	.9	1.0
	eɪ	3.1	3.6	**83.1**	3.2	4.2				.1	.1		.6		.1	.6	1.2
	ɛ	.6	5.2	2.3	**78.8**	5.7		1.7	1.2		.1		.3		.1	1.6	2.3
	æ		.9	3.8	8.1	**80.5**	.3		1.2			.1	.3	.1	2.3	1.2	1.2
	ɑ		.3	1.2	.6	9.0	**37.8**	18.6	26.9	.7	.1		.4		1.0	.9	2.5
	ʌ			.6	1.2	3.6	11.9	**65.1**	9.7	1.0	1.0	.1	.6	.1	2.2	1.5	1.3
	ɔ		.1	.6		1.6	29.9	4.1	**56.5**	2.3	.9		.3	1.0	.9		1.7
	oʊ	.1	.3		.1		3.2	.6	.7	**80.4**	4.5	3.6		2.3	2.2	.1	1.7
	ʊ	.1			.3		2.0	17.9	1.3	.7	**66.0**	4.7	.3	1.6	1.0	.4	3.6
	u	3.6	.7	.1	.4		.4	1.9	1.0	1.0	12.9	**72.4**	.1	.7	2.2	.6	1.7
	aɪ		7.1	1.5	.1		.1	.1	.1	.1			**89.4**	.3	.1	.3	.6
	ɔɪ	.1		.3	.1		.3	.3	.7	1.3	.7	.3	.1	**92.4**	2.6		.6
	aʊ	.3		.1	1.5	.6	2.6		4.4	5.4	.4	.1		1.3	**82.0**	.4	.9
	ɝ	.4	.3		1.3			.9	.1							**96.7**	.3

Figure 2.1: The confusion matrix of vowels, recorded as part of a listening experiment with native American English listeners in a listening condition with babble noise at SNR 0 dB (Weber and Smits, 2003).

2.1 FACTORS THAT INFLUENCE MISPERCEPTION IN NOISE

Even though misperception occurs at the listener's end, every component of a conversational setup – *listener, speaker, communication channel, and the speech itself* – can influence the perception errors. In this section, the existing literature on each of these factors is discussed in detail.

2.1.1 Listener

Listener characteristics like age, hearing ability, and language proficiency have shown a significant influence on speech misperception, especially in noisy environments (Rogers *et al.*, 2012; Taitelbaum-Swead and Fostick, 2016; van Os *et al.*, 2021). Their findings demonstrate that the human ability to perceive speech in noise deteriorates with aging, resulting in more perception errors among older adults compared to younger adults. Similarly, listeners with hearing loss suffered more recognition errors than individuals with normal hearing abilities in adverse listening setups (Jürgens *et al.*, 2010; Brons *et al.*, 2014; van Knijff *et al.*, 2018). Language proficiency of listeners

is another feature that has shown a significant impact on perception errors in noise (Warzybok *et al.*, 2015); non-native listeners showed a native-like performance in speech perception only for linguistically easy sentences. Also, compared to non-natives, native listeners produced significantly fewer recognition errors, as they compensate for their listening difficulties with their linguistic knowledge.

2.1.2 Speaker

Like listeners, the attributes of speakers such as age, gender, and language proficiency have been studied in the context of speech perception in noise. For instance, the intelligibility and listening efforts of *accented speech* and *native speech* have been compared and it was found that they are significantly more different in noisy environments (Munro, 1998; Van Engen and Peelle, 2014). More specifically, native listeners in adverse listening environments produced far fewer perception errors while listening to native speakers, compared to non-native speakers. Interestingly, it was also demonstrated that the perceivable attributes of a speaker like age, gender, and ethnicity influence listeners' ability to perceive speech in noise (Drager, 2011; Kutlu, 2023).

In addition to those common attributes, earlier research has investigated the impact of rare speaker attributes such as speech impediments (like lisp speech) and pathological disorders (like dysarthria) on intelligibility, in noisy environments. The study by Eadie *et al.* (2021) demonstrated that speakers who are completely intelligible in quiet, but exhibit mild speech impairments, are considerably more vulnerable to the effects of background noise, compared to those speakers with intact speech. Interestingly, non-linguistic factors like familiarity with the speaker/voice have also shown that they can significantly improve spoken language processing in noisy environments (Nygaard and Pisoni, 1998).

2.1.3 Communication Channel

The communication channel refers to the environment in which a conversation takes place. Depending on the type of communication, speakers and listeners in conversation may or may not share the same environment. For example, conversations over a telephone network could result in a listening setup with the speaker in quiet and the listener in a noisy environment or vice-versa. Despite such differences, both

speech production and *speech perception* are considerably influenced by the adverse communication channel, leading to increased efforts of participants in conversations. In this section, we discussed some of the existing findings on recognition errors induced by a noisy listening environment.

Acoustic noise is prevalent throughout the real world; hence, a majority of human conversations occur in the presence of noises like road traffic, appliances/machinery noise, or competing speech of other people in the background. Noise interferes with spoken language comprehension, as it may distort the clean signal resulting in a degraded signal that is difficult to perceive and understand. In early studies on acoustic noise, Pollack (1975) have identified the two forms of noise masking – *energetic masking and informational masking* – which could lead to listening difficulties and perception errors. More specifically, energetic masking is defined as the signal distortions introduced by masking the spectral features of the clean signal. However, informational masking is not based on low-level acoustic characteristics; instead, it pertains to high-level interference such as challenges in employing cognitive resources like attention and working memory.

Energetic masking varies with masker type. In Cooke (2006), a speech perception model was proposed for noisy environments by measuring the presence of *glimpses* – spectrotemporal regions that survived the energetic masking. Noisy utterances with better glimpse proportion are more intelligible. They also demonstrated that the availability of glimpses varied with masker types – maskers with high spectral and temporal energy modulations result in high glimpse proportion, leading to a better perception. In other words, stationary maskers like speech-shaped noise resulted in a limited amount of glimpses, compared to a masker type like single talker competing speech, leading to more misperceptions in the presence of speech-shaped noise.

Acoustic noise can be categorized in several ways and one of the most commonly used criteria is based on its source. On a high level, the source of noise can be of two types: synthetic and naturalistic. Synthetic noise is *generated* by modifying the acoustic properties of a signal. White noise, pink noise, speech shaped noise (SSN) are some examples of artificially generated acoustic noise. Naturalistic noise, on the other hand, *occurs* in everyday conversational environments such as babble noise, traffic noise, etc. Noisex-92 is one of the early noise corpora that consists of eight different types of naturalistic noises that were recorded at different physical locations (Varga and Steeneken, 1993).

In addition to the source, acoustic noise is categorized based on a signal charac-

teristic - stationarity. A signal is said to be stationary when its frequency or spectral contents remain constant over some time. In other words, a signal with equal energy distributed in all its time-frequency bands is stationary. Speech perception errors are comparatively less in non-stationary noise, compared to those in stationary noises. In prior literature, a common stationary signal - white noise - is also referred to as *non-fluctuating* noise. On the other hand, signals with changing frequency/spectral contents over time are non-stationary. Most of the naturalistic noise is non-stationary.

2.1.4 Spoken Tokens

Early investigations on misperceptions in noise were primarily concentrated on sub-lexical units like vowels, consonants, and syllables (Miller and Nicely, 1955; Pickett, 1957; Luce and Pisoni, 1998). The choice of sub-lexical sounds was motivated by the need to study consistent confusions of sounds that occur in the presence of a masker (Miller and Nicely, 1955; Pickett, 1957; Weber and Smits, 2003). Pickett (1957) conducted an elaborate analysis of vowel perception errors in the presence of white noise. They found that spoken vowels with higher natural intensity are perceived better only in low-frequency noise conditions. In the case of high-frequency noise, they observed that the second formant is masked by noise, leading to lower intelligibility for high-intensity vowels.

Kalikow *et al.* (1977) demonstrated that word recognition errors are significantly less when the linguistic context is highly predictable. In their work, they proposed a hearing test called Speech Perception in Noise (SPIN) which is currently being used in clinical analysis of human hearing ability in noise.

In addition to phonetic and linguistic features, lexical features like the length of the target word is particularly important for word recognition. Several studies have shown that longer words (which have fewer neighbors (Pisoni *et al.*, 1985)) are easier to recognize (Vitevitch, 2002; Vitevitch and Rodríguez, 2005). On the contrary, we can also find evidence from the literature that familiar words (which are usually shorter in length as per Zipf's law) are easier to recognize. However, whether the word familiarity favors its recognition in noise is not well documented in prior studies.

Since familiarity is a subjective measure, studies have also utilized objective metrics like word frequency as its alternative. In addition, the Neighborhood Activation Model (NAM) proposed by Luce and Pisoni (1998) demonstrated that word recognitions are also influenced by additional factors such as the neighborhood

density (*i.e.*, the number of words that are easily confusable with the target word) and the frequency of neighboring words.

2.2 MITIGATION OF MISPERCEPTION

Because mishearing can negatively influence message comprehension, mitigating mishearing is essential to achieving communication goals, especially in noisy environments. In the case of human-human interactions, both listeners and speakers in a conversational setup have shown greater effort to reduce the instances of mishearing as much as possible. However, in comparison to listeners, human speakers have demonstrated a larger involvement in reducing mishearing, probably because of their higher level of control over the rendered speech that could be misheard. In the following sections, we will discuss some of the well-known strategies adopted by humans and a handful of those implemented by algorithms.

2.2.1 Speaker Strategy

A large share of human-human interactions in the real world, occurs in noisy environments with sounds like people talking in the background, vehicles/machines being operated, or loud music/announcements being played. Yet, to a great extent, humans manage to converse in such non-ideal listening setups. One reasonable explanation for not breaking down in every conversation in the real world is that human speakers can adapt their speech to accommodate the listening difficulties of their interlocutors in a particular listening environment. In the review paper, Cooke *et al.* (2014) demonstrated that humans employ almost 32 different speech modification strategies in different listening setups to improve their interlocutor's hearing. They refer to human speakers as *listening talkers*. We identified that those strategies can be classified broadly into two groups based on the underlying feature that is being modified: (1) acoustic features and (2) linguistic features.

2.2.1.1 *Modulating acoustics*

Prior studies on speech production in noise have shown that human speakers attempt to improve the audibility of the speech, even if they are present in a clean listening environment but their interlocutors are in a noisy environment. It is the instinct of

humans to modify speech by increasing acoustic features like pitch, intensity, and duration. This modified speech is referred to as Lombard Speech. The benefits of using Lombard speech have been extensively studied for several noise environments and hence, it inspired the synthesis of Lombard speech in some of the recent Text-To-Speech (TTS) systems. However, Lombard speech is less intelligible than normal speech in quiet. Also, Lombard benefits are more likely to be influenced by the linguistic features and thus it is critical to determine when to use Lombard speech, if implemented as algorithms to produce speech in noise.

Human speech production in noise has also shown a phenomenon of enhanced speech for information-bearing words and for those words which are easy to be misheard. Slowing down the speech rate, using additional pauses, or increasing the loudness are some of the commonly used speaker strategies to reduce the hearing difficulties of their interlocutors. However, when implemented in the speech synthesis pipeline, not all strategies were found to be beneficial. On the contrary, some of the techniques were detrimental as they reduced the user's attention and engagement in the conversation, leading to poor comprehension.

2.2.1.2 *Modulating linguistics*

Earlier work in speech perception in noise has investigated how different speech tokens such as vowels (Pickett, 1957; Cutler *et al.*, 2004) or consonants (Weber and Smits, 2003; Jürgens and Brand, 2009) are affected by noise, and have considered word intelligibility in isolation (Luce and Pisoni, 1998; Clopper *et al.*, 2010; Wilson and Cates, 2008) as well as in context (Kalikow *et al.*, 1977). Although earlier studies on word misperception in noise (Albert Felty *et al.*, 2013; Cooke *et al.*, 2019; Marxer *et al.*, 2016a; Cooke, 2009) have shown that the noise impact is dependent on the lexical items, very few studies have explored the potential of linguistic modification to enhance speech perception in noise.

However, human speakers employ several modifications at the linguistic level to reduce listening difficulties like the use of less complex words and phrases in foreigner-directed speech (FDS) or hearing-impaired directed speech (HIDS). Similarly, human speakers are known to use small sentences and common words in infant-directed speech (IDS). More specifically, human speakers in noisy environments, use repetitions and rephrasing to produce noise-robust speech.

However, implementing repetitions as a template-based approach has shown no significant improvement in speech perception (Cooke *et al.*, 2014). The use of less

confusable words was implemented earlier to improve the correct word recognition in the ATC domain (Cox and Vinagre, 2004). Although it was shown as a useful strategy, scaling it to large-vocabulary domains is still an open problem. Similarly, even though paraphrasing is recognized as a useful human strategy to improve noise-robustness, a model to generate noise-robust paraphrase is still an unexplored area.

2.2.2 Listener Strategy

In addition to speakers, listeners in a conversational environment also engage in mishearing mitigation strategies, which are also called as dialogue repair mechanisms. When listeners realize a certain message turn was too hard for them to hear or it sounds in-coherent with the dialogue history, they usually probe the speaker for clarification with statements like '*huh*', '*pardon*', '*say it again*', '*sorry, what was that?*', etc. Such dialogues usually indicate an instance that suggests the speaker repeat or rephrase the dialogue (Skantze, 2005).

In addition to those explicit actions, listeners also employ a few implicit actions like integrating high-level signals such as linguistic cues (Van Os *et al.*, 2022) and situational cues (Ward *et al.*, 2017), to complement the distorted audio in challenging listening environments. Although such predictability has shown benefits in reducing misperceptions in noise (Kalikow *et al.*, 1977), a few studies have also demonstrated the risk of *false hearing*, which occurs when an utterance is misheard as something else with strong, but incorrect certainty (Rogers *et al.*, 2012).

2.3 MEASURES, METHODS, AND MODELS

In this section, we discuss some of the existing literature on different measures, methods, and models that contributed to a better understanding of human misperception and mishearing mitigation strategies in noisy environments.

2.3.1 Measures

This section outlines the different acoustic and linguistic measures that are employed in this thesis. The acoustic measures are used for capturing the characteristics of noisy utterances. Linguistic measures are employed to define the underlying text of

utterances, using features like predictability and the semantic equivalence between two texts. These measures are later used in the thesis to define and evaluate our proposed approach of using paraphrases to improve speech perception in noise.

Noise level. To measure the noise level in a processed/distorted signal, a commonly used metric is the signal-to-noise ratio (SNR). SNR represents the ratio of the power of a clean (undistorted) signal and a noise signal, which combine to form the distorted signal. Simply put, it is a fraction of powers as defined in Equation (2.2). It is commonly measured on a logarithmic scale and referred to in units of decibels (dB), as defined in Equation (2.3). The power of a signal is the sum of the absolute squares of signal magnitudes averaged across the time domain. In other words, it is the square of the root mean square (RMS) of the signal. Hence, SNR can also be defined as the ratio of RMS values of clean and noise signals as shown in Equation (2.4).

$$SNR = \frac{P_{signal}}{P_{noise}} \tag{2.2}$$

$$SNR_{dB} = 10\log_{10}(SNR) = 10\log_{10}(\frac{P_{signal}}{P_{noise}}) \tag{2.3}$$

$$SNR_{dB} = 20\log_{10}(\frac{RMS_{signal}}{RMS_{noise}}) \tag{2.4}$$

When a clean speech signal is mixed with a noise signal with equal power, the SNR of the resultant distorted speech is 0 dB. Similarly, when the power of the clean signal is higher than that of the noise, the SNR of the resultant signal is positive (> 0 dB). Higher SNR scores indicate better audibility. On the other hand, when the noise power is more in the processed signal, the SNR value is negative (< 0 dB).

Speech Intelligibility measures. Automatic metrics support scalable and cheaper methods to identify the intelligibility of audio. There are two types of Speech Intelligibility (SI) metrics exist: (a) intrusive and non-intrusive. The underlying theory of intrusive metrics is that the intelligibility of audio corresponds to the amount of signal that survived the energetic masking effect. Hence, they require both the distorted signal as well as its corresponding clean signal. On the other hand, non-intrusive metrics do not require a clean signal and they require only the distorted speech, which is used to estimate its clean signal and then its intelligibility.

STOI. Short-Time Objective Intelligibility (STOI) is an intrusive measure (Taal *et al.*, 2010) that captures the *similarity* between a distorted/processed signal, $x(n)$ and the corresponding clean signal, $s(n)$ of equal duration n. The similarity is calculated by measuring the distance between the temporal envelopes of two (time-aligned) signals in frequency sub-bands. This distance is also referred to as the *intermediate intelligibility index* (d_m) for a time segment m. More precisely, linear correlation coefficients are calculated for the clean and distorted short-time envelope spectrograms, and the overall intelligibility of the distorted signal is captured by averaging the correlation coefficients (as shown in Equation 2.5) across all M time segments.

$$d = \frac{1}{M} \sum_{m=1}^{M} d_m \tag{2.5}$$

Before calculating the correlation coefficients, the audio signals must undergo a clipping procedure including normalization to avoid clipping on all Time-Frequency units. The higher the STOI score, the better the intelligibility of the signal as it indicates its better correlation with the clean audible signal. Because STOI represents the correlation, the STOI value ranges between -1 (least intelligible) and $+1$ (most intelligible).

Language Model score. The objective of a language model is to represent a natural language (*like* Malayalam) such that the model is capable of identifying/generating a plausible sequence of tokens, just like humans. In other words, an English language model learns to assign a higher probability for the sentence $s1$ than $s2$ in the following example, as $s2$ is grammatically incorrect and implausible in the language.

$s1$: *They want to learn .*

$s2$: *They learn want to .*

The probability of a sequence (e.g. $s1$) is then computed by combining the conditional probabilities of all words in the sequence using the chain rule of probability. The probability of each word in a sequence is conditioned on its *context* (*ie.,* all words until then) as shown below:

$$\begin{aligned} P(s1) = {} & P(They)\ P(want|They)\ P(to|They\ want)\ P(learn|They\ want\ to) \\ & P(.|They\ want\ to\ learn) \end{aligned} \tag{2.6}$$

Traditionally, language models were built using *n-grams* such that the context of a word is approximated by a set of few previous n-grams (or words). For instance, the conditional probabilities (in Equation 2.6) are further simplified by considering only the previous word as the context of a word (*ie.,* bigrams), forming the following equation:

$$P(s1) = P(They)\ P(want|They)\ P(to|want)\ P(learn|to)\ P(.|learn) \tag{2.7}$$

More recently, large language models have been built using an architecture of more than a billion parameters and trained on an extremely large dataset (of size in Terabytes). Despite its humongous size and the complexity of the model architecture, the core of the language model remains the same - *learn to predict the next token for a given context*. Such large language models are incredibly more powerful than small n-gram models, as they consider a large context (of $\approx$ 1024 tokens) at each instance of token prediction.

Hence, a trained language model (with parameters θ) is useful to estimate the perplexity (PPL) of a sentence, $x_{1:n}$ (tokenized as $x_1, x_2, \ldots x_n$), as shown in Equation 2.8. More specifically, perplexity is the exponentiation of the average negative log likelihood of a sentence. Thus, the sequence of tokens that has high perplexity indicates a less likelihood in the linguistic domain of the trained language model. We use a pre-trained language model to estimate the PPL of sentences, thereby capturing the linguistic predictability of sentences. The perplexity value ranges from zero to $+\infty$.

$$\text{PPL} = \exp -\frac{1}{n}\sum_{i}^{n} \log p_\theta(x_i|x_{<i}) \tag{2.8}$$

Semantic Textual Similarity. Comparing the semantic equivalence between a pair of texts is particularly important in the NLP domain. On one hand, natural languages permit their users to represent a meaning/message in different linguistic forms by choosing different words, phrases, or syntactic structures. On the other hand, this degree of freedom also introduces ambiguity and may risk a significant deviation in the semantics of different linguistic realizations. Hence, the verification of semantic equivalence has been a critical step for multiple language generation tasks, including but not limited to, machine translation, question answering, and story generation.

In NLP, the task of verifying whether two texts are semantically equivalent is referred to as **paraphrase identification**. In early studies, paraphrase identification models were evaluated with human-annotated paraphrase corpus, in which the pairs of texts are labeled as paraphrases or not. However, verifying whether a generated text is semantically equivalent to the ground truth is more challenging, with the latest text generation models that produce linguistically variant texts. Even though, human annotations are the gold standard for evaluating text generation models, several automatic metrics were proposed in recent years to address the cost and scalability issues of human annotations.

Earlier, metrics like BLEU and ROUGE were used to validate the closeness in the actual and expected text by considering the n-gram overlaps. More recently, multi-dimensional representations of tokens were used to avoid the limitations of n-gram-based metrics. Sentence-BERT, BERTScore, and BLEURT are some of the recent STS metrics, which utilize the pre-trained models and their representations of language tokens. For instance, BERTScore aggregates the individual representations of tokens for each sequence in a pair and then the aggregated representation is combined to get a single score.

2.3.2 Methods

This section briefly introduces some of the methods that we used in the human perception experiments.

2.3.2.1 *Additive noise mixing*

The stimuli for the perception experiment are created by mixing a noise signal with a clean speech signal. We used an open-source tool, audio-SNR (Sato, 2018: accessed July 6, 2022) for noise-mixing. At first, a random snippet of the noise signal is trimmed, for the same length as the clean signal. Both signals are then normalized and depending on the SNR required energy of the clean signal is modified before combining with the noise signal amplitudes. Audio clipping is performed to avoid the glitches in noise-mixing.

2.3.2.2 *Phonemization*

A grapheme-to-phoneme (g2p) model is used for generating the phoneme sequence of a text, which is represented in words/graphemes. Based on the phonemic sounds

that exist in a language, IPA alphabets are used to represent the pronunciation of individual letters/alphabets. A grapheme-to-phoneme model is trained to predict the sequence of phonemes for a sequence of alphabets. This process of generating phonemic sequence is hereafter referred to as phonemization. Depending on the IPA representations that the model followed while training, the phonemic sequence of a word/sentence varies with different models. In addition to the phonemes, the model indicates the stress level of each phoneme with a digit next to the phoneme. The stress indicator associated with the phoneme is useful for detailed analysis of phones, however, we ignore this indicator as we concentrate only on the phoneme.

2.3.3 Data Modeling

In this section, we will discuss the data modeling techniques that we have used in chapters 4, 5, and 6.

2.3.3.1 *Linear regression models*

Linear regression models are used to explain/predict a *response* (*i.e.*, dependent) variable using *predictor* (*i.e.*, independent) variable(s). As the name suggests, *simple linear regression* explains the variance of a response variable with a single predictor variable. While a multi-variate regression model consists of more than one predictor variable. For example, consider a model to predict the complexity of a word j, using its lexical features such as word length, word familiarity, and, number of word senses.

Before delving into the details of the model, let us define the notations used to represent matrices and vectors here and elsewhere in the document. Following the conventional notations, the boldfaced uppercase letters (e.g., $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$) denote the multi-column matrices, and boldfaced lowercase letters (e.g., $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$) are used to represent vectors. Scalars are denoted by plain lowercase letters (e.g., x, y, z).

Let y^j be the scalar which represents the word complexity of j and $\mathbf{x}^j = [x_1^j, x_2^j \dots x_d^j]^T$ be the d-dimensional vector to represent its features. With an assumption that predictor variables x_i are linearly related to the response variable y, we can define a linear regression model as shown in Equation (2.9). This model has two types of parameters: intercept and slope. Together, they define a regression line (in case of simple regression) or a regression sub-space (in case of multiple regression) which maps the predictor variable(s) to the response variable. The *intercept* of

the model (β_0) indicates the predictor variable value when none of the features were given. The rest of the model parameters ($\beta_1, \beta_2 \dots \beta_d$) are referred to as the slopes of the regression sub-space with respect to the individual features. ϵ in Equation (2.9) refers to all errors that are not captured by the defined model. In other words, it captures all deviations in the expected model such as the measurement errors or those introduced by external features which are not considered in the model.

$$y^j = \beta_0 + \sum_{i=1}^{d} \beta_i . x_i^j + \epsilon \tag{2.9}$$

Since the actual relationship between the response and predictor variables is unknown, the model parameters in Equation (2.9) need to be estimated by fitting the model to an observed set of ($\mathbf{x}^j$, y^j) pairs. To this end, a prediction model is defined as shown in Equation (2.10). By comparing the observed and predicted response variables, parameter estimation attempts to reduce the error in prediction ($(y^j - \hat{y}^j)$, also called as *residuals*) by choosing appropriate values for $\hat{\beta}_i$. For instance, a dataset consisting of N pairs, i.e., $(\mathbf{X}, \mathbf{y}) = ((\mathbf{x}^1, y^1), (\mathbf{x}^2, y^2) \dots (\mathbf{x}^N, y^N))$, the overall error in prediction is usually measured as residual sum of squares (RSS) which is calculated as shown in Equation (2.11).

$$\hat{y}^j = \hat{\beta}_0 + \sum_{i=1}^{d} \hat{\beta}_i . x_i^j \tag{2.10}$$

Further, parameter estimation algorithms like Least Square Error (LSE) are used to reduce this error. LSE learns the model parameters by reducing the overall vertical distance between the actual and the predicted point to exhibit minimal error in prediction. More precisely, it will generate a set of equations by taking the partial derivative of Equation (2.11) with respect to individual parameters and equating them to zero.

$$\begin{aligned} RSS(\mathbf{y}, \hat{\mathbf{y}}) &= \sum_{j=1}^{N} (y^j - \hat{y}^j)^2 \\ &= \sum_{j=1}^{N} (y^j - (\hat{\beta}_0 + \sum_{i=1}^{d} \hat{\beta}_i . x_i^j))^2 \end{aligned} \tag{2.11}$$

In the case of a simple linear regression, this results in the following Equation 2.12.

$$\begin{aligned}\hat{\beta_0} &= \mu_y - \hat{\beta_1}\mu_x \\ \hat{\beta_1} &= \frac{\sum_{j=1}^{N}(x^j - \mu_x)(y^j - \mu_y)}{\sum_{i=1}^{N}(x^j - \mu_x)^2}\end{aligned} \tag{2.12}$$

For multiple linear regression, the estimated parameters can be represented in matrix form for better readability. That is when model predictions are viewed as a product of input features and the model parameters as shown in Equation (2.13), the calculation of RSS is simply a set of matrix operations. For better readability, a new multi-column matrix $\mathbf{X}_1 = [\mathbf{1};\mathbf{X}]$ is introduced here to consider the model intercept ($\hat{\beta_0}$) by concatenating a column vector of ones to $\mathbf{X}$. Thus the set of model parameters becomes a d+1-dimensional column vector $\hat{\beta} = [\hat{\beta_0}, \hat{\beta_1} \ldots \hat{\beta_d}]^T$.

This results in the parameter estimation as shown in Equation (2.14).

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}_1\hat{\beta} \\ RSS(\mathbf{y}, \hat{\mathbf{y}}) &= (\mathbf{y} - \mathbf{X}_1\hat{\beta})(\mathbf{y} - \mathbf{X}_1\hat{\beta})^T\end{aligned} \tag{2.13}$$

$$\hat{\beta} = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y} \tag{2.14}$$

The model parameters/coefficients associated with individual predictor variables indicate the change in a response variable with a unit change in the corresponding predictor. Because the model parameters are determined in the scale of individual predictor variables, it is important to normalize the scale of predictor variables. Model parameters are estimated based on the observed data pairs ($\mathbf{X}$,$\mathbf{y}$). It is important to perform statistical significance of model estimations as it is expected to exhibit a difference in estimation as and when the sample set changes slightly. Statistical software platforms like R (R Core Team, 2019) perform hypothesis testing with the null hypothesis that predictor variables have no effect on the response variable (i.e., $H_0 : \hat{\beta}_i = 0$). Thus the reported p-values indicate the significance of predictor variables in explaining the variance of the response variable.

In chapters 3 and 4, we utilize the model coefficients and their corresponding p-values to analyze the influence of different lexical and acoustic features on mishearing in noise. Further, the feature importance was studied by calculating the model fitness through a systematic addition (*forward selection*) or removal (*backward selection*) of predictor variables.

2.3.3.2 *SVMRank*

Support Vector Machines (SVM) is a well-established margin-based algorithm that was initially proposed for classification problems (Vapnik *et al.*, 1998). The core idea of SVM is to determine the hyperplanes that separate an n-dimensional data space into different parts. A *hyperplane* is a subspace that separates a space into two parts. In other words, it is the higher dimensional generalization of a line in 2D space or a plane in 3D space. The underlying assumption of SVM is that the data points are linearly separable. However, in most cases this doesn't hold and slack variables are introduced in the model to account for the violations of this assumption. But, before we delve into the details of non-linear SVM models, the notion of using such models for ranking needs to be discussed.

SVMs are also known as maximum margin classifiers, as they optimize the model parameters to maximize the margin. A *margin* is the shortest distance that's possible between a separating hyperplane and the data. Those data points that define the margin of the model are referred to *Support Vectors*.

To formalize the model, consider a binary classification problem with input variable $\mathbf{x} \in \mathbb{R}^n$ and an output variable $y \in \{1, -1\}$ which indicates the two classes. Thus the hyperplane of this model is defined as:

$$\mathbf{w}.\mathbf{x} + b \tag{2.15}$$

where $\mathbf{w}$ refers to the unit vector which is orthonormal to the hyperplane and b which indicates its distance from the origin. Based on the *sign* of the hyperplane equation, data points in different parts of a space are identified (i.e., points below the hyperplane are negative values and those above are positive values). However, in the case of n-class classification problem, the classification for each data point is performed by conducting $n - 1$ binary classifications. That is, SVM identifies a separating hyperplane for every pair of classes and the ensemble of binary classification outcomes is used to classify the data-point. The parameters $\mathbf{w}$ and b are optimized for a dataset $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots (\mathbf{x}_m, y_m)\}$ of m samples with the model objective of maximizing the margin as shown below:

$$\mathbf{w}, b = \underset{\mathbf{w},b}{\operatorname{argmax}} \min_{i=1,2,\ldots m} |\mathbf{w}.\mathbf{x}_i + b| \tag{2.16}$$

such that $y_i.(\mathbf{w}.\mathbf{x}_i + b) > 0$.

Similar to the multi-class classification problem, ranking is performed by con-

ducting multiple pairwise rankings and each pair-wise ranking is in turn a ternary classification. More precisely, the outcome of a pair-wise ranking is the ordering of items in each pair $p = ((\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j))$, indicating whether the rank of first element is better than the second, or the other way or, both are similar in their ranks ($\text{rank}(y_i) >_p \text{rank}(y_j) \in \{1, -1, 0\}$). y_i represents the relevance of an item in each pair which is unknown at the inference time. By considering a function f to predict the relevance of an item in each pair p and their corresponding feature vectors Φ_i and Φ_j, the core idea of ranking becomes the following two-way implication:

$$\text{rank}(y_i) >_p \text{rank}(y_j) \iff f(\Phi_i) >_p f(\Phi_j) \tag{2.17}$$

Assuming a linear function f (e.g., $\mathbf{w}.\Phi_i + b$) the implication becomes,

$$\text{rank}(y_i) > \text{rank}(y_j) \iff \mathbf{w}.(\Phi_i - \Phi_j) > 0 \tag{2.18}$$

2.3.4 Text Generation Models

In the past, sequence generation models like Hidden Markov Models (HMMs) and Recurrent Neural Networks (RNNs) were proposed for multiple text generation tasks like machine translation, text summarization, and paraphrase generation. In the last decade, with the introduction of new model architectures like encoder-decoder and transformers, text generation tasks have gained an incredible performance, surpassing their predecessors. More recently, Large Language Models (LLMs) pre-trained on humongous training data, showcases its incomparable capacity to perform several text generation tasks.

2.3.4.1 *Large Language Models*

The main objective of language models (LMs) is to learn the distribution of linguistic units such as words, phrases, sentences, and paragraphs of a language. That is, given a sequence of n linguistic tokens $x_{1:n} = (x_1, x_2, \ldots x_n)$, the goal of the language model is to estimate $p(x_{1:n})$.

By applying the chain rule of probability, a language model is decomposed into a model to predict the next word, given all previous words in the sequence (Manning and Schutze, 1999; Bengio *et al.*, 2000):

$$p(x_{1:n}) = p(x_1)\ p(x_2|x_1)\ \ldots p(x_n|x_1, x_2, ..., x_{n-1}) \tag{2.19}$$

$$= \prod_{i=1}^{n} p(x_i|x_{<i}) \tag{2.20}$$

$$= \sum_{i=1}^{n} log\ p(x_i|x_{<i}) \tag{2.21}$$

To train an LM, a dataset $D = \{x^1, x^2, x^3 \ldots x^{|D|}\}$ is required, as it represents a wide range of valid sequences of linguistic tokens in a language. Current state-of-the-art LMs (Radford *et al.*, 2019; Wei *et al.*, 2022a) have Transformer-based architectures (Vaswani *et al.*, 2017), which are represented by a huge set of parameters, Φ. Such models are more commonly referred to as **large language models** (LLMs) because of their large model capacity Φ and the enormous size of their training dataset D like GPT-2 trained on 40 GB of text to learn model parameters of size 1.5B. The model parameters are optimized to minimize the negative log-likelihood over D:

$$L(D) = -\frac{1}{|D|} \sum_{d=1}^{|D|} \sum_{i=1}^{d_n} log\ p_\Phi(x_i^d|x_{<i}^d) \tag{2.22}$$

where d_n denotes the number of tokens in the sequence x^d.

During inference, a new sequence of arbitrary length $y_{1:m}$ is generated by sampling each token from the learned distribution: $p_\Phi(y_0)$, $p_\Phi(y_1|y_{<1})$, $p_\Phi(y_2|y_{<2})$ $\ldots$ $p_\Phi(y_m|y_{<m})$. Because of the rich representational capabilities introduced by contextual embeddings (Vaswani *et al.*, 2017) and its exposure to a humongous amount of linguistic data, pre-trained LMs are widely used to build text-generation models.

2.3.4.2 *Fine-tuning LLMs*

The paraphrase generation task aims to generate a sequence of tokens $y_{1:m}$ for a given input sequence $x_{1:n}$, such that $y_{1:m}$ is semantically equivalent to $x_{1:n}$. To this end, a parallel dataset $D_p = \{(x^1, y^1), (x^2, y^2), (x^3, y^3) \ldots (x^{|D_p|}, y^{|D_p|})\}$ is used to fine-tune the model parameters and learn the distribution $p(y|x)$. Again, the chain rule of probability is applied to simplify the modeling objective:

$$p(y_{1:m}|x_{1:n}\ [\text{SEP}]) = \prod_{i=1}^{m} p(y_i|x_{1:n}\ [\text{SEP}]\ y_{<i}) \tag{2.23}$$

where [SEP] is a sequence of tokens that separates the input sequence ($x_{1:n}$) and output sequence ($y_{1:m}$).

[SEP] tokens are particularly useful at inference time, as it signal the model to perform task-specific text generation (like paraphrases). This process of re-training the model parameters with a task-specific dataset is commonly known as *model fine-tuning*. Similar to the pre-training phase, negative log-likelihood is the widely used objective function for fine-tuning.

$$L(D_p) = -\frac{1}{|D_p|} \sum_{d=1}^{|D_p|} \sum_{i=1}^{d_m} log\ p_\Phi(y_i^d|x_{1:n}^d\ [\text{SEP}]\ y_{<i}^d) \tag{2.24}$$

where d_m denotes the number of tokens in the sequence y^d.

As shown in the equation above, the fine-tuning loss is computed only for the output sequence tokens (ie., tokens after [SEP]). During inference, the learned distribution conditioned on the input sequence is used for sampling each constituting token in the output sequence: $p_\Phi(y_0|x_{1:n}\ [\text{SEP}])$, $p_\Phi(y_1|x_{1:n}\ [\text{SEP}]\ y_{<1})$, $p_\Phi(y_2|x_{1:n}\ [\text{SEP}]\ y_{<2})\ \ldots\ p_\Phi(y_m|x_{1:n}\ [\text{SEP}]\ y_{<m})$.

2.4 CONCLUSION

In this chapter, we discussed the existing literature on human misperceptions, especially in noisy environments. An ample amount of research in the past has systematically studied the linguistic and non-linguistic factors that can influence misperceptions in noisy environments. Additionally, researchers have also contributed toward a better understanding of human strategies that are employed to reduce the interlocutors' mishearing in adverse listening setups. We have also discussed some of the existing algorithmic solutions to mitigate mishearing in noise, as well as their limitations. With such rich research in the background, we begin our investigation of *using paraphrases to improve speech perception in noise*, which is elaborately discussed in the following chapters.

3 FORMALIZING SPEECH PERCEPTION ERRORS AND MITIGATION

Mishearing occurs when a listener incorrectly recognizes spoken words. Instances of mishearing occur quite commonly in everyday conversations as neither the listener nor the speaker confirms their interlocutor's perception after each dialogue exchange. Instead, they assume that their interlocutor correctly perceived words as spoken. Mishearing impedes listeners from comprehending the intended meaning of the spoken utterance and it may even lead to misunderstanding or communication breakdowns in extreme cases Grimshaw (1980).

This chapter focuses on providing a formal definition of the problem of mishearing. In Section 3.1, the distinction between mishearing and misunderstanding is discussed in detail. Further, in Section 3.2, we outlined the two different approaches to enhance the intelligibility of speech produced in noisy conditions. Overall, this chapter provides a concise description of the technical terms that will be discussed again in Chapters 4, 5, and 6.

3.1 SPOKEN LANGUAGE PROCESSING

The theory of **speech production** (Levelt *et al.*, 1999) explains that the process of generating spoken words to convey a message undergoes a series of steps starting with the preparation of a *concept/meaning* to be conveyed, followed by lexical selection to realize the meaning, encoding the morphological and phonological features of the selected lexical units and finally, the articulation step in which the sound waves are produced. The flow diagram in Figure 3.1 depicts a simplified view of the speech production theory discussed in Levelt *et al.* (1999). Thus the process of encoding a meaning into an acoustic signal undergoes two critical steps: (1) *text generation* and (2) *utterance generation*. Today, modern spoken dialogue systems (SDS) utilize a similar pipeline to produce dialogue responses by employing a dialogue manager (DM) and a Natural Language Generation (NLG) module to generate an intended response and then converting the generated text to an utterance using a Text-To-Speech (TTS)

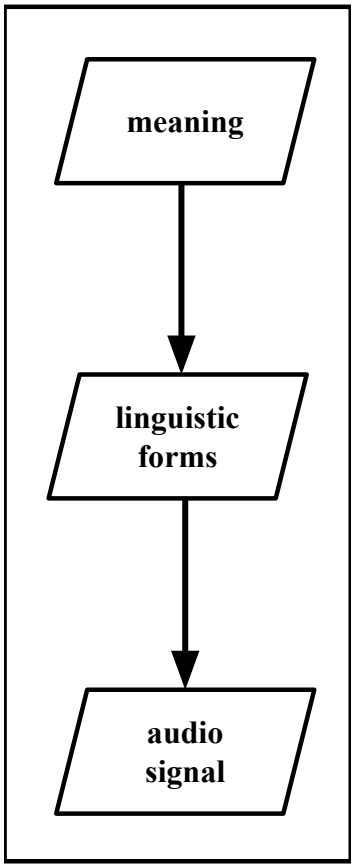


Figure 3.1: A simplified view of speech production theory proposed by Levelt *et al.* (1999)

system.

A speech production system can then be defined as a two-step mapping process: (1) encoding a *concept/meaning* to linguistic forms as defined by the text generation function **t** and (2) mapping linguistic forms to acoustic realizations by the utterance generation function **u**, as defined in Equations (3.1) and (3.2), respectively. Here, M is a set of K concepts, L is a set of J linguistic forms, and A is a set of I acoustic signals.

$$\mathbf{t}(M) = L, \text{ where } M = \{m_1, m_2, \ldots m_{\mathsf{K}}\} \text{ and } L = \{l_1, l_2, \ldots l_{\mathsf{J}}\} \tag{3.1}$$

$$\mathbf{u}(L) = A, \text{ where } L = \{l_1, l_2, \ldots l_{\mathsf{J}}\} \text{ and } A = \{a_1, a_2, \ldots a_{\mathsf{I}}\} \tag{3.2}$$

Neither **t** nor **u** is a one-on-one mapping function, which means that it is possible for **t** to map a meaning $m_k \in M$ to distinct logistic forms in L, as shown in Equation (3.3). In linguistics, this approach of representing a meaning/concept with V different linguistic forms is referred to as *paraphrasing*.

$$\mathbf{t}(m_k) = \{l_j, l_j^1, l_j^2 \dots l_j^V\} \subset L \tag{3.3}$$

In the above-given example, $l_j^1, l_j^2,$ and l_j^V are referred to as paraphrases as they differ in wordings but represent the same semantics. *Lexical paraphrases* (*i.e.*, synonyms) are a special category of paraphrases that exist among single words (rather than phrases or sentences). Similarly, **u** maps a linguistic form l_j to V different acoustic signals in A which utters the same linguistic form with varying acoustic characteristics like intensity, pitch, and duration.

$$\mathbf{u}(l_j) = \{a_i, a_i^1, a_i^2 \dots a_i^V\} \subset A \tag{3.4}$$

On the other hand, **Spoken Language Comprehension** (SLC) is the process of retrieving the meaning of an utterance from its audio signal. Understanding an utterance begins with the recognition of spoken tokens (Davis and Johnsrude, 2003). As observed in the extensive review by Weber and Scharenborg (2012), several speech perception models were proposed in the late 20th century to explain and replicate the human listeners' ability to perceive spoken words from an utterance. Models like Cohort (Marslen-Wilson and Welsh, 1978) and TRACE (McClelland and Elman, 1986) explained spoken word recognition with a series of processing steps like *activation* of multiple candidates in listener's mental lexicon and *selection* of candidates that best fit the input acoustic signal. This mapping process is referred to as either *speech perception* (or *spoken word recognition* when the given utterance is a single word). Merge is another word recognition model proposed in Norris *et al.* (2000), accounting for the close relatedness of speech perception and speech production. It is a bottom-up model that utilizes pre-lexical signals like acoustic/phonetic features to recognize spoken words. Figure 3.2 plots a bottom-up schema similar to the Merge model. Motivated by the earlier finding that speech comprehension results in separate activations of lexical recognition and meaning comprehension (Norris *et al.*, 2006), we distinctly identify speech perception as the preliminary step in the speech comprehension pipeline.

As depicted in Figure 3.2, decoding *meaning* from *speech signal* consists of two critical steps: (1) perception of an audio signal to recognize the underlying text (or

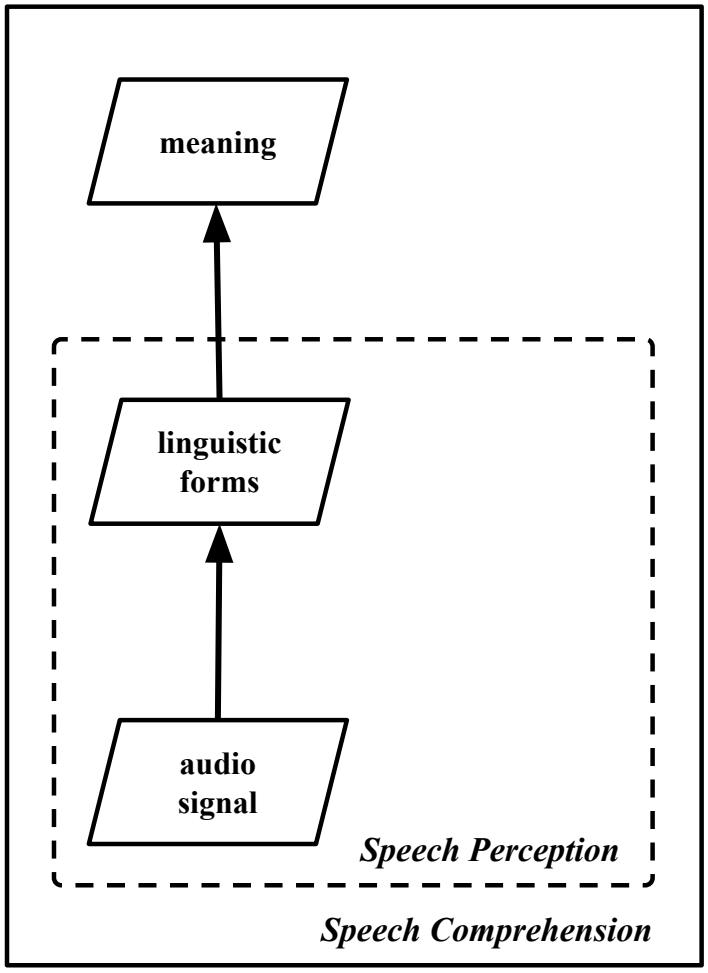


Figure 3.2: A speech comprehension pipeline involves speech perception, wherein the acoustic signals are first mapped to lexical items and then lead to the meaning of the utterance.

linguistic forms) and (2) retrieving a meaning from the recognized linguistic tokens.

More formally, let $\mathbf{p}$ be the perception function that maps an audio $a_i \in A$ to linguistic token(s) $l_j \in L$:

$$\mathbf{p}(A) = L \tag{3.5}$$

$\mathbf{p}$ is a many-to-one function as it is possible to perceive a linguistic form from two distinct audio signals. For example, consider a_i^1 and a_i^2, the two arbitrary acoustic realizations of the same word l_j, produced by two different voices of high pitch and low pitch, then speech perception defines

$$\mathbf{p}(a_i^1) = \mathbf{p}(a_i^2) = l_j \tag{3.6}$$

Speech comprehension comprises another decoding step in addition to **p** - the utterance meaning is extracted from the perceived linguistic tokens. Thus, a speech comprehension function **c**, maps an utterance audio signal a_i to a meaning $m_k \in M$ as defined below:

$$\mathbf{c}(A) = \mathbf{c}(\mathbf{p}(A)) = \mathbf{c}(L) = M \tag{3.7}$$

These definitions of speech perception and speech comprehension are further used in this thesis to define *errors in perception and comprehension*.

To begin with, let us consider a simple conversational setup with two speakers S_1 and S_2, in a noisy environment. As both of them have *no* speaking/hearing impairments, with native language proficiency, we assume that noise is the only factor that contributes to the mishearing instances. To define misperception, consider an instance of message exchange, where the speaker S_1 produced an utterance a_i with an intended meaning m_k (encoded in the linguistic form l_j).

However, because of the acoustic noise in the background, the interlocutor S_2 heard a distorted acoustic signal, $\tilde{a}_i$. S_2 performs speech comprehension as defined in Equation (3.7), by first decoding the $\tilde{a}_i$ to linguistic forms through perception $p(\tilde{a}_i)$. If the speech perception by S_2 results in the actual text l_j, then there is no mishearing/misperception. The listener is likely to receive the message as intended. However, any deviation in the actual spoken utterance and the perceived utterance is an indication of *mishearing*. In other words, mishearing occurs when

$$\Delta_{\mathbf{p}}(l_j,\ \mathbf{p}(\tilde{a}_i) \neq 0 \tag{3.8}$$

where $\Delta_{\mathbf{p}}$ is a function that measures the deviation in terms of linguistic forms. In this thesis, such instances are also referred to as *misperception/recognition errors*.

Similarly, comprehension errors are represented as

$$\Delta_{\mathbf{c}}(m_k,\ \mathbf{c}(\mathbf{p}(\tilde{a}_i)) \neq 0 \tag{3.9}$$

where $\Delta_{\mathbf{c}}$ captures the deviation in the semantics of the actual and perceived message, which are also called as *incomprehension/misunderstanding* in the literature. Equation 3.9 also highlights the need to reduce mishearing as the errors in speech perception are likely to be propagated to the comprehension steps as discussed in

prior research Grimshaw (1980); Wilson and Spaulding (2010); van Os *et al.* (2021).

Prior studies have employed different experimental designs to capture speech perception and speech comprehension. For instance, "listen and repeat" is a common speech perception task, where the participants are instructed to listen to audio signals and repeat what they have heard either by uttering or transcribing the utterance (Kalikow *et al.*, 1977; Uslar *et al.*, 2013; Wilson and Cates, 2008). On the other hand, speech comprehension is evaluated with secondary tasks like drag-and-drop experiments (Fontan *et al.*, 2015), or by using measures like working memory capacity (Wendt *et al.*, 2016).

The current work consists of only native listeners – individuals with high language proficiency. Thus, we assume that all instances of correct perception would lead to correct comprehension, as the experiment involves day-to-day conversations that are too easy to comprehend if perceived correctly. In other words, we simplify the speech comprehension defined in Equation (3.7) as:

$$\mathbf{c}(\mathbf{p}(A)) \approx \mathbf{p}(A) \approx L \tag{3.10}$$

Thus, the deviation in speech perception is equated to the deviations in speech comprehension $\Delta_p = \Delta_c = \Delta$; *i.e.*, misunderstanding occurs when linguistic tokens are not perceived as intended.

In the next section, we will define the two different approaches to reduce perception errors in noise.

3.2 ENHANCING SPEECH INTELLIGIBILITY IN NOISE

The main objective of enhancing the speech intelligibility is to reduce misperceptions ($min\ \Delta_{\mathbf{p}}$). One of the commonly explored approaches to enhance the intelligibility of synthesized speech is by leveraging the potential of acoustic modification to synthesize different variations ($\{a_i^1, a_i^2 \dots a_i^V\}$) of an intended audio signal a_i, which represents a meaning m_k, as illustrated in Figure 3.3. Thus, by modifying acoustic features like pitch, intensity, and speech rate, *different variations of an intended utterance* are generated for the same underlying text/linguistic form (l_j). Enhanced speech intelligibility is ensured by selecting the audio signal that is less likely to be misperceived in a listening condition.

In this work, we propose an alternative approach to generate a noise-robust acoustic signal that represents the intended meaning (m_k) – modify the linguistic

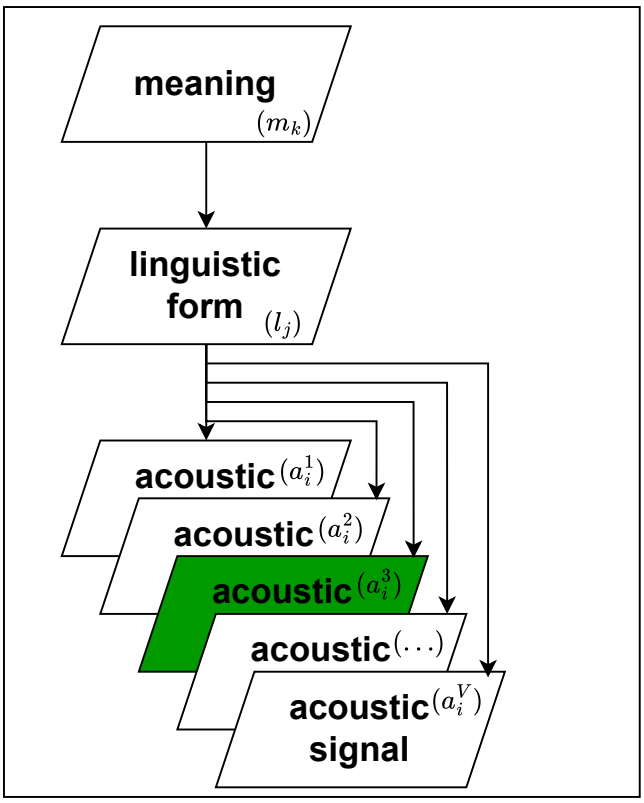


Figure 3.3: An existing approach to synthesize noise-robust speech, employing acoustic modification to generate different variations of an utterance. This approach focuses on selecting the utterance, which is likely to introduce minimal misperception (shaded in green), resulting in enhanced acoustic intelligibility.

form (l_j) to generate different utterances. As defined in Equation (3.3), modifying linguistic forms without altering the intended meaning, results in V paraphrases such as $l_j^1, l_j^2, l_j^3 \dots l_j^V$. Representing the intended message with different linguistic representations results in V different acoustic signals $a_i^1, a_i^2, a_i^3 \dots a_i^V$ corresponding to each paraphrase. As shown in Figure 3.4, paraphrasing results in different acoustic signals, even in the absence of acoustic modification. Further, selecting the linguistic representation that contributes to better perception in a listening environment, ensures noise-robust speech synthesis, without any signal distortions.

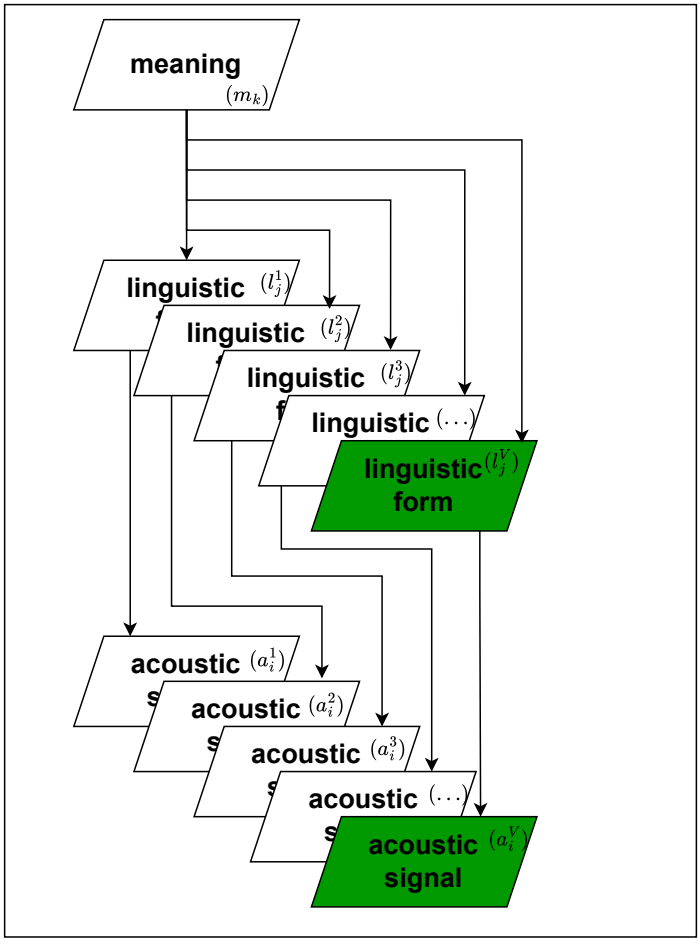


Figure 3.4: The proposed approach to synthesize noise-robust speech, employing **paraphrasing** to generate different representations of the underlying meaning of the intended message. This approach aims at selecting the linguistic form, which is likely to introduce minimal misperception (shaded in green), thereby enhancing the intelligibility of synthesized speech.

3.3 CONCLUSION

This chapter provided a formal definition of the fundamental concepts of the proposed approach – paraphrasing to improve speech perception in noise. However, it is not clear whether the different acoustic signals generated through simple paraphrasing methods like lexical replacement differ in their misperceptions under noisy listening conditions. Also, to use the proposed strategy in modern SDSs, it is im-

portant to understand why and to what extent paraphrasing is a useful strategy to enhance synthetic speech intelligibility in noise. More crucially, we need a systematic approach to measure misperceptions and determine whether a sentence is better than its paraphrase. The rest of the chapters in this thesis are focused on understanding the potentials as well as the limitations of paraphrasing to improve speech intelligibility in noise.

4 LEXICAL PARAPHRASES TO MITIGATE WORD MISPERCEPTION

Words exhibit varying recognition rates when subjected to noisy listening conditions. As we have seen in Chapters 2 and 3, in the past, several studies have demonstrated that word intelligibility in a noisy environment is significantly influenced by linguistic features like word familiarity and linguistic predictability, as well as by acoustic features such as the masking effect of the noise, the underlying sounds in the utterance, etc. However, this distinction of lexical units in terms of their noise-robustness is seldom leveraged in speech synthesis systems to combat noise-induced word recognition errors, produced by human listeners.

In this chapter, we set out to fill this gap by considering a specific linguistic operation - *lexical replacements with synonyms* - to improve word intelligibility in noise. Sections 4.1 and 4.2 outline the concept and the motivation to use lexical paraphrases for better speech perception in noise, by elaborately discussing an existing dataset. In Section 4.3, we explained the evaluation method used to study whether intelligibility is improved by replacing lexical units with their synonyms (*i.e.*, *lexical paraphrases*). Perception data of synonyms in noise environments is collected by conducting perception experiments with human subjects as elaborated in Section 4.4. Additionally, we investigated why certain words are better intelligible than their synonyms in noise, using modeling experiments, and the details are sketched out in Section 4.5. The main contributions of this work are as follows:

- We established preliminary evidence that lexical paraphrases are capable of reducing noise-induced comprehension errors.
- We created and published a new dataset[1] of synonyms in noise (with/without linguistic context) by conducting listening experiments with native listeners of English who have normal hearing (NH) thresholds.
- We analyzed the benefit of the mishearing mitigation strategy with lexical paraphrases under different noise levels with an explanation of the varying in-

[1]Dataset is publicly available here: https://github.com/SFB1102/A4-SynonymsInNoise.git

fluence of linguistic and acoustic factors that drive the intelligibility differences, for further research in this direction.

Overall, the experiments in this chapter become the foundational evidence of our claim that paraphrasing is an effective strategy for *improving word intelligibility in noisy listening conditions*.

4.1 INTRODUCTION

To begin with, we performed a sanity check to analyze the phonemic distance between the synonyms in English. This verification is important as synonyms that sound alike are *less likely* to introduce a significant difference in their intelligibility in noise. In other words, synonym pairs that are similar in pronunciation are equally likely to be misheard in noise, and choosing one over the other might not reflect a significant difference in comprehension errors. For this analysis, we considered all synonym pairs in a lexical corpus, WordNet (Fellbaum, 1998). The edit distance between the phonemic transcripts of two synonyms was calculated with a cost of one for insertion, deletion, and substitution of tokens. As portrayed in Figure 4.1, we found that about 65% of synonym pairs sound different by $4-8$ phonemes and it indicates that several words have a phonemically different synonym.

This observation is interesting as it suggests that some words can be more noise-robust than their synonyms, presumably because of their differences in the underlying sounds. However, whether this observed phonemic edit distance could contribute to differences in human perception in noisy environments is unclear without perception data. Thus to collect perception data of synonyms in different noise environments, we propose to conduct human listening experiments. Using the perception data, we systematically study the impact of lexical paraphrases on word misperception in noise. To this end, we identified the two critical research questions (RQs) of this work:

- RQ 1: Are lexical paraphrases different in their intelligibility under noise?
  - 1*a*: lexical paraphrases *without* any linguistic context
  - 1*b*: lexical paraphrases *with* linguistic context
- RQ 2: If there exists a difference in lexical intelligibility, why certain lexical units are better recognized than their synonyms?

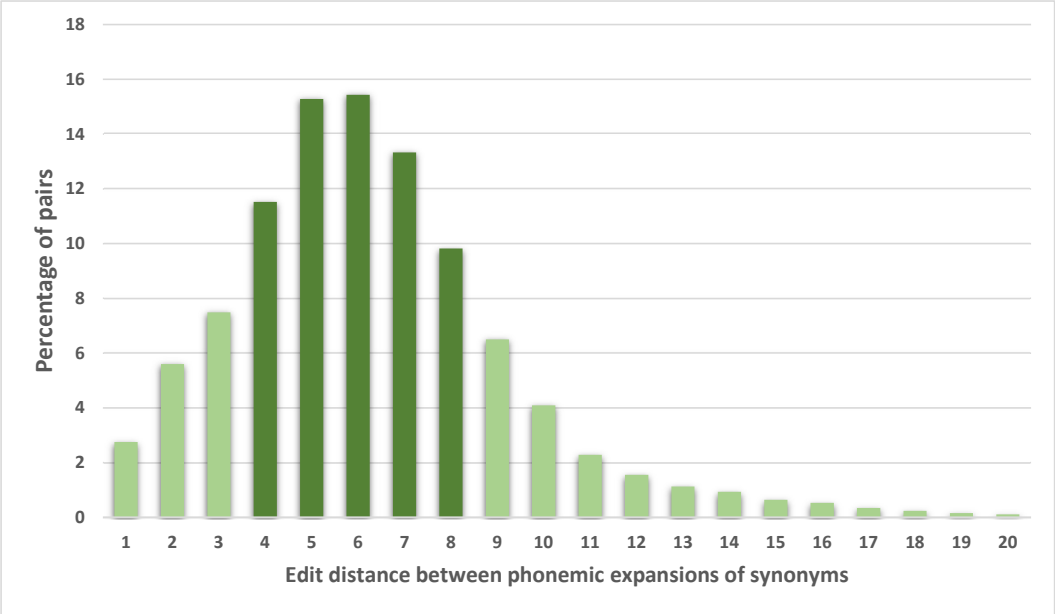


Figure 4.1: Levenshtein edit distance between phonemic transcripts of synonym pairs in the lexical corpus, WordNet. More than 80% of synonym pairs exhibited a difference of 4 phonemes, indicating their distinction in the underlying sounds.

4.1.1 Motivation

The proposed approach of reducing word misperception through paraphrasing is mainly motivated by the existing understanding that noise induces *slips-of-the-ear* (Bond, 1999). Slips-of-the-ear are attributed to mishearing wherein a word/phrase is incorrectly recognized as another spoken word(s). For example, perceiving the spoken word '*mouth*' as '*mouse*' is a slip-of-the-ear instance. This section discusses a preliminary analysis that we conducted with an existing dataset of human misperceptions of words in noise, before collecting a new dataset. This analysis is focused on verifying that word misperceptions vary under different noise environments and are influenced by lexical features.

Marxer *et al.* (2016b) published a large dataset of word misperception in English under three different types of maskers: stationary speech-shaped noise (SSN), four-talker babble noise (BAB4), and three-talker babble modulated noise (BMN3). Stimuli

utterances were generated with masker-specific SNR ranges like [−7 dB, −4 dB] for SSN, [−8 dB, −3 dB] for BMN3, and [−3 dB, +1 dB] for BAB4. Their choice of three masker types and corresponding noise levels were motivated by a similar study of word misperception in Spanish (Tóth *et al.*, 2015).

SSN: Stationary Speech-shaped Noise is generated by taking the long-term average of the speech spectrum, which was formed by combining multiple speech signals. This results in a noise signal, which has spectral properties almost identical to that of the input speech signals. It is a stationary noise. This masker type is commonly used to simulate an adverse listening environment.

BAB4: Babble noise is one of the most common real-world noise types, which occurs when two or more people converse in the background of a listening setup. It is more prevalent in public spaces like cafeterias, conference rooms, and parks. The more the number of speakers in the background, the more it is noisy and challenging for a listener to listen to the intended speech. Babble noise introduces a better possibility than SSN to capture glimpses of the clean speech, however with an increased risk that the listener may confuse the actual speech with the competing speech. In the case of the BAB4 noise condition, the noise signal was generated by mixing four continuous signals, which were created by concatenating randomly sampled words in the recorded speech material.

BMN3: BMN3 is a three-talker babble-modulated noise, which is generated by first estimating the envelopes of a three-talker babble noise signal and then, modulating an SSN carrier signal based on those estimated envelopes. BMN3 shares similar non-stationary properties with BAB4. Hence, their energetic masking effects are similar. However, BMN3 has no recognizable speech, just like the SSN signal. Thus BMN3 becomes slightly an easier noisy environment for perception, as it clubs the benefits of both BAB4 and SSN conditions: allows capturing glimpses like in the babble noise condition, but with a lesser impact of the competing speech as it is a nonverbal noise signal.

Slips-of-the-Ear varies with masker type. Stimuli words for each masker type varied with a little overlap. However, the average length (three phonemes) and the mean frequency (3.99) of targets are the same across all three masker types. After conducting several hours of perception experiments with a large pool of listeners (212), their dataset lists all those lexical units that exhibited a *consistent confusion* at a listening condition. Marxer *et al.* (2016b) defined consistent confusion as words that are always misheard as another word by a group of listeners in a listening condition.

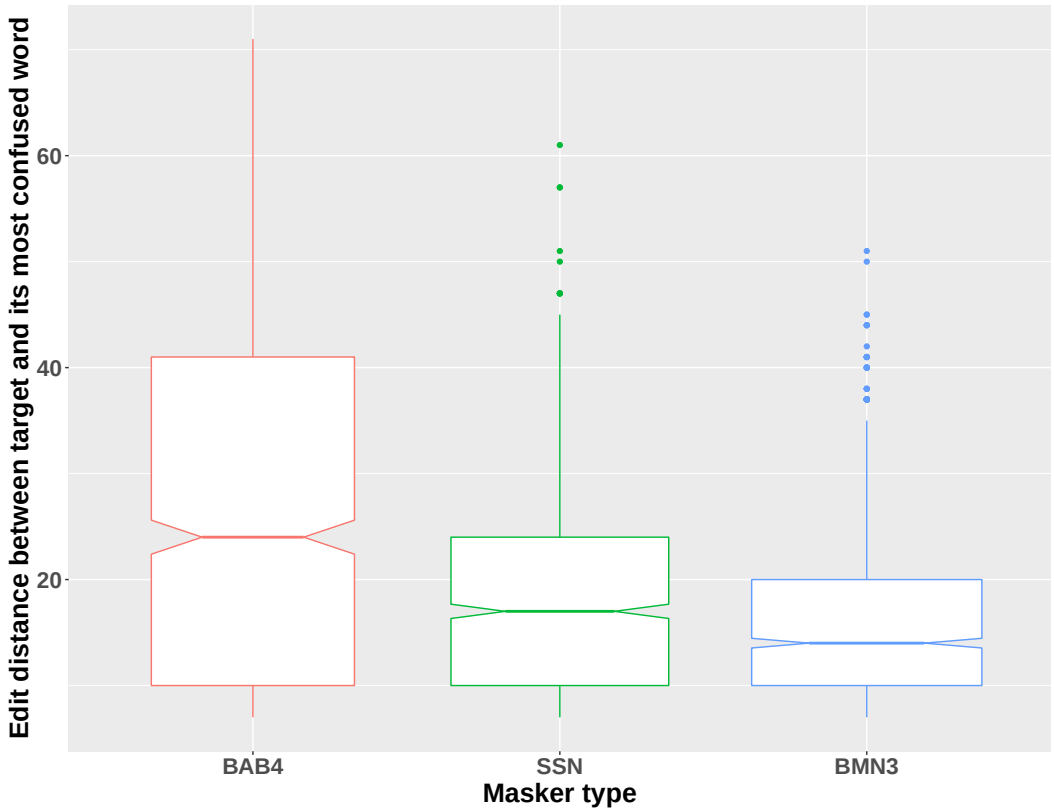


Figure 4.2: A demonstration of the varying word misperception distance under three maskers: BAB4 (four-talker babble noise), SSN (stationary speech-shaped noise), and BMN3 (three-talker babble modulated noise). The misperception distance is determined by calculating the edit distance between the phonemic transcripts of the target word and its most confused word published in the dataset (Marxer *et al.*, 2016b).

In the published dataset of word confusions in English, they reported only those word misperceptions that had a minimal consistency of 40% (i.e., at least 6 among 15 listeners misheard a word in the same way). The dataset also reported the edit distance between the phonemic transcripts of a target word and its most confused word to represent the distance/closeness in their sounds. Thus, slips-of-the-ear can otherwise be considered as the errors introduced by insertion, deletion, and/or substitution of speech tokens like phonemes in the target word.

Figure 4.2 reports the edit distance of a target word and its most confused word, calculated with a cost of seven for insertions and deletions, and ten for substitutions. A smaller value for the edit distance indicates that the confused word sounds a lot similar to the target word. A smaller edit distance also highlights that such misperceptions are possible even in the absence of noise, as they sound similar. On

the other hand, a large edit distance highlights that a target word is phonetically far away from its most confused word. One of the interesting observations from their work is the significant influence of the masker type on the edit distance between a target and its most confused word. For example, the consistent confusions in BAB4 are (phonetically) more distant compared to those in SSN and BMN3 noisy conditions. A higher edit distance indicates an increased challenge in predicting the misperceived word for a target word, as the search space is larger with more edit operations. This observation aligns with earlier findings that a masker-type influences the nature of misperception, as discussed in Section 2.1. Precisely it shows that compared to stationary noise environments, word recognition is better in the presence of non-stationary noise - glimpses of the intended speech are less available with a steady state background noise. It was also shown that the type of misperception varies with the masker type as showcased by different phoneme confusion matrices published for different noisy setups (Miller and Nicely, 1955; Pickett, 1957). We further extended their analysis on consistent confusion by calculating the mean edit distance at different SNRs of a noise type. For better interpretation, actual SNR values are rounded to their nearest integer. As portrayed in Figure 4.3, BAB4 is different from the other two masker types in two ways: (a) there is a larger spread of phonemic edit distance at BAB4, (b) noise level is a significant feature to determine the mean edit distance of consistent confusions *only* under BAB4 noisy condition. This indicates noise type as well as its SNR (in certain masker types) are critical for determining the the nature of word misperception. More precisely, under the BAB4 noise condition where the competing speech in the background is spoken words and recognizable speech, listeners are confused and wrongly perceive those words spoken in the background. Those misperceived words can be phonetically more distant from the actual spoken word. Additionally, we observed that when the SNR level is less for BAB4, misperceived words are phonetically closer to the target words, just like in SSN and BMN3 noise conditions.

Confusion Consistency is influenced by lexical features. In addition to the phonemic edit distance, the dataset also provides a measure of *confusion consistency* (ie., the number of participants who misheard a word as another one). We use this measure to evaluate whether the lexical features of a target word or its most confused word influence the consistency of misperception in noise. To this end, we conducted a simple linear regression analysis by fitting the data for each of the noise conditions separately. A complete model with five independent variables -

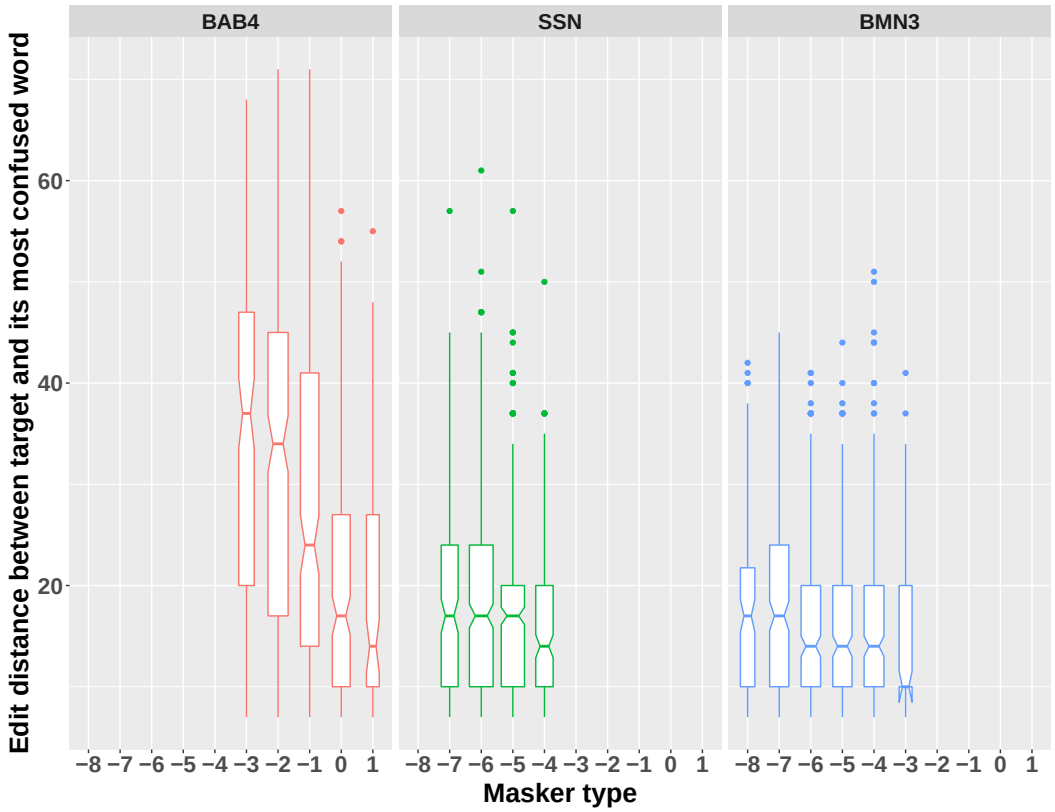


Figure 4.3: An illustration of how misperception varies in the presence of different types and levels of noise. An existing dataset of word confusions in English (Marxer *et al.*, 2016b) shows that, unlike the babble noise condition, in the presence of nonverbal signals (like SSN and BMN3), words are confused with another word that sounds similar.

length and frequency of a target word, length and frequency of the most confused word, and the phonemic edit distance between a target word and its most confused word - was considered to explain the dependent variable, confusion consistency. For selecting a best-fit model, we performed a step-wise variable selection algorithm implemented in the step function (Marhuenda *et al.*, 2014) in R(R Core Team, 2019).

Our models showed that the influence of lexical features on misperception consistency varied with the masker type. For instance, under SSN and BMN3, the consistency of misperception increased when a confused word sounded closer to its target word ($\beta = -0.004$; $p < 0.05$), indicating word misperception is driven by sound-alike words in the vocabulary. Additionally, we found that confusion consistency is better with longer target words (BMN3: $\beta = 0.012$, SSN: $\beta = 0.014$; $p < 0.05$). Intuitively, this observation aligns with the earlier finding that shorter words, which have high phonological neighborhood density (PND) introduce far

more options in the vocabulary to be misheard as another word and hence, lower confusion consistency. However, with longer words, the PND is less dense and there exist fewer options for misperception of a target word, leading to more consistent confusion among listeners. In contrast to the above observation, at the BAB4 noise condition, neither the length of a target word nor its edit distance was significant in determining the confusion consistency. Instead, the length of the most confused word determines the confusion consistency. This agrees with our earlier observation that word misperception under BAB4 is largely influenced by the words spoken in the background noise. More precisely, confusion consistency was observed to be better in BAB4, when the confused word is longer in length which in turn reflects the impact of longer words in the competing speech on the consistency of a word confusion.

In summary, this analysis demonstrates that word misperception in noisy environments is influenced by their linguistic characteristics as well as the masker type. The results suggest that some words are more likely to be misheard than others, in certain listening conditions. However, the existing dataset is not sufficient to study whether lexical replacement with synonyms is an impactful strategy to reduce mishearing in noise. This highlights the need for a dataset of spoken utterances of paraphrases annotated by a pool of listeners, reporting both the *correct and incorrect* instances of perception in noise.

4.2 LEXICAL PARAPHRASES

Natural languages provide the possibility of expressing an intended message in several different forms using different sets of lexical units and syntactic structures, which is often referred to as *paraphrases*. The potential of paraphrasing has been widely studied in the context of several language processing tasks such as style transfer, text summarization, text simplification, machine translation, and dialogue generation. However, the current study of utilizing paraphrases to improve word recognition is one of the initial attempts to improve comprehension under noise with linguistic variations. To illustrate the methodology used in the current study, consider the following incomplete utterance:

- *and he runs away scared and dives into the* _____ .

Given that, a speaker is given a choice to fill the _____ with a lexical unit that fits

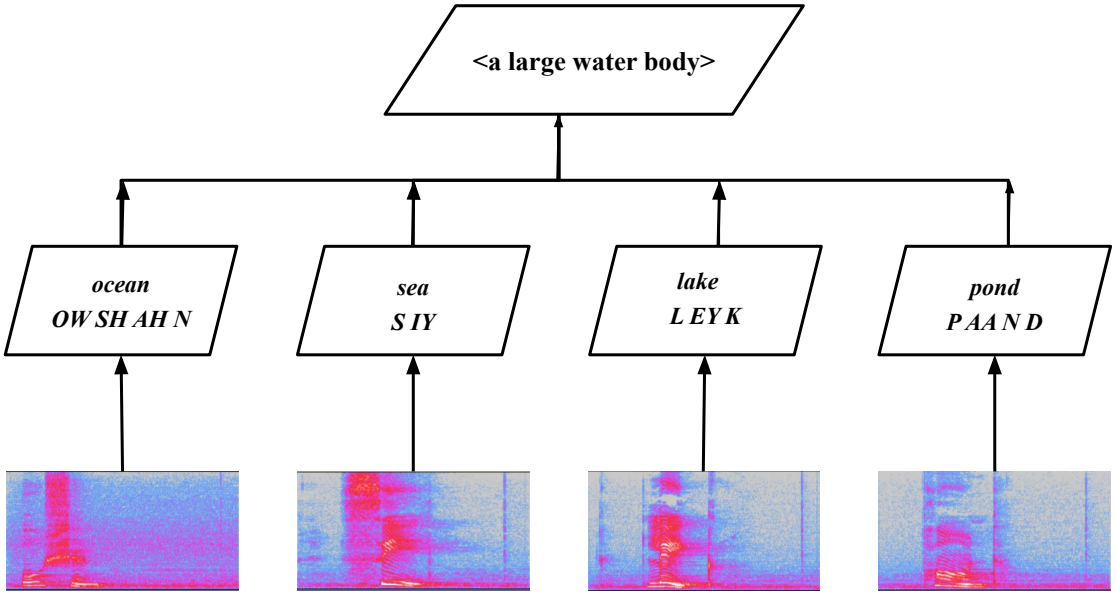


Figure 4.4: An illustration of the potential of synonyms to represent meaning with different acoustic realizations that vary in the underlying phonemes and acoustic characteristics.

in the linguistic context (*i.e.*, all words before the blank) and means ***<a large water body>***, lexical paraphrases in English provides multiple options like '**sea**', '**ocean**' and '**lake**. As shown in Figure 4.4, although all options have similar semantics, the underlying sounds (represented with their phonemic transcript and spectrogram) are different. Because of this difference, introducing a variation at the linguistic level indeed introduces a variation in the acoustic level. We hypothesize that for a listener in noisy environments, the speaker's choice of a lexical unit to represent ***<a large water body>*** is critical, as some lexical units are less likely to be perceived in noise due to the differing impact of noise on the underlying sounds. To validate our hypothesis, it is important to conduct a pair-wise comparison of noise-robustness among synonyms in noise. Additionally, it is also critical to study whether the perception difference among synonyms is introduced by the target words themselves or their interaction with the context.

In the following sections, we describe the experiments and data analysis that were conducted to validate the proposed strategy. To reduce the complexity of the experiment, the following assumptions were made:

- Although the noise in real-world conversational settings is highly unpredictable

and hard to model, we made an assumption that the noise (type and its level) in a listening environment is known in advance.

- Not all recognized words result in successful comprehension of the meaning. A listener's language proficiency is a critical factor for meaning comprehension. In the current study, we controlled for this effect by selecting native speakers as participants and familiar words for the stimuli. Thus, all instances of correct word recognition are considered instances of correct comprehension as well.

With those assumptions, let's look at the approaches we have taken to answer current research questions.

4.3 MEASURING HOW SYNONYMS INFLUENCE WORD RECOGNITION

In order to validate our proposed method of utilizing lexical paraphrases to mitigate mishearing, we need to study whether such variations introduced any recognition differences under a particular listening environment. To this end, it was essential to create a dataset comprising pairs of synonyms annotated with their (*mis*)perceived words by multiple listeners.

We conducted two sets of listening experiments with a large pool of listeners (125 in total) who have normal hearing thresholds: by presenting synonyms (a) without linguistic context and (b) with linguistic context, under different listening environments. Studying the recognition difference between synonyms in both these scenarios is critical as the first one without context demonstrates whether a background noise impacts the intelligibility of both synonyms in a pair equally or not. More precisely, if the recognition of synonyms in a pair under a listening environment is equally the same, then there is no real gain in recognition by performing the lexical replacement. Also, it is important to study whether such lexical replacements can introduce differences in synonym recognitions even in the presence of high-level signals such as linguistic contexts.

Utilizing the data collected from these experiment setups, the primary question that we explored was:

"Are synonyms significantly different in their recognition rates under noise? ".

Before we delve into the details of listening experiments and data modeling, let's begin this section by defining a measure to capture the recognition gain introduced by lexical replacements.

4.3.1 Word Intelligibility

The intelligibility of an utterance depends on how well its constituting spoken tokens are recognizable by its listeners. A highly intelligible utterance is one with fewer recognition errors. Hence, Speech Intelligibility (SI) is an inverse measure of recognition errors. Such measures have been traditionally used in clinical audiology to quantify human hearing capabilities. For instance, hearing tests such as Hearing in Noise Test (Soli and Wong, 2008) and Speech in Noise (Etymotic Research, 1993) utilize recognition rates to identify hearing loss, hearing aids fitting, and so on. Participants in those tests are usually instructed to 'listen and repeat' stimuli utterances. Speech Reception Threshold (SRT) is an alternative metric when such hearing tests are conducted with adaptive SNR levels (Taylor, 2003). SRT of an individual participant refers to the maximum noise level/SNR at which 50% of the speech material was recognizable. In literature, it is mainly used to compare listeners/pools of listeners and is seldom used to compare stimuli items. Hence, SRT is not identified as a suitable measure of intelligibility for the current study.

Human experiments of the type 'listen and repeat' were used in the past to study a broad range of SI-related questions such as the influence of predictability on word recognitions (Kalikow *et al.*, 1977), hearing ability differences between younger vs. older listeners (van Os *et al.*, 2021), L1 vs. L2 listeners in noise, fluctuation of Lombard effect with predictability, Lombard effect on L2 listeners (Cooke and Lecumberri, 2012) and others. In most of these studies, intelligibility differences between stimuli sets (eg: high vs. low predictable) or between listener sets (eg: L1 vs. L2) were analyzed by averaging sentence-level recognition rates over a set of listeners or stimuli items respectively. However, for the current study, the word intelligibility needs to be calculated at the stimuli level as the scores are further used for pair-wise comparison between synonyms. To this end, we defined a word intelligibility score as the mean recognition rate of a word among a pool of listeners. For consistency, we ensured that every stimulus was listened to by a fixed number of participants.

We refer to this metric as Human Recognition Score (HRS) and it is defined for each spoken word (s) in a particular listening environment as shown below:

$$\text{HRS}_w = \frac{\text{number of correct recognitions}_w}{\text{total number of listeners}_w} \tag{4.1}$$

To identify whether the perceived word is identical to the actual word, the edit distance between phonemic transcripts of the target and perceived words was

utilized. All listening instances with 0 phonemic edit distance are considered correct recognition. This ensured that homophones (eg: *sea* vs *see*) of the target words are identified as correct perception and minor spelling mistakes in transcriptions are rectified in the intelligibility score calculation. Since HRS captures the proportion of correctly recognized spoken words among all its listening instances, its value ranges between 0.0 to 1.0.

4.3.2 Gain in the Word Intelligibility

The HRS score is further used to measure whether a choice of lexical units among two synonyms (s_1, s_2) resulted in any intelligibility difference (Δ_p as described in Equation 3.8 and refer to Section 3.1, for more details.) by taking the absolute difference of their individual scores as defined in Equation (4.2). Hereafter, this measure is referred to as the **gain in word intelligibility**, as it represents the intelligibility improvement achieved by choosing the noise-robust lexical over its synonym.

$$\Delta_p = \text{diff.HRS}_{(s_1,s_2)} = abs(\text{HRS}_{s_1} - \text{HRS}_{s_2}) \tag{4.2}$$

The value of diff.HRS ranges from 0.0 to 1.0, indicating *no gain* to *maximum gain* in word intelligibility by choosing the noise-robust synonym. More precisely, this score reflects whether there exists any difference in synonyms intelligibility given that they were presented with the same linguistic context and the same acoustic background noise. The mean diff.HRS in each listening environment is calculated to analyze whether it is effective to utilize lexical paraphrases to reduce mishearing in noise.

4.4 LISTENING EXPERIMENTS

Human perception experiments were conducted to collect and create a dataset of pairs of synonyms along with their corresponding (*mis*)perceptions in noise. To the best of our knowledge, this dataset is the first of its kind[2]. The rest of this section is focused on solving the first research question discussed in Section 4.1. Initially, we studied the perception of synonyms without any linguistic context (RQ 1*a*). Then,

[2]The dataset is publicly available here: Synonyms-in-Noise (SiN)

synonyms with linguistic contexts were presented and their perception in noise was recorded for analyzing the influence of context on synonyms under noise (RQ 1*b*).

4.4.1 RQ 1*a*: Synonyms in noise, *without* context

For synonym replacement to be a promising approach, we need to test whether synonyms differ substantially concerning their intelligibility in noise. It only makes sense to attempt to replace one with the other if this is the case.

4.4.1.1 *Experimental setup*

In our experiment, synonyms were presented separately to different participants as spoken words in five different listening environments; clean (no noise), babble noise at SNR 0 & SNR -5, and white noise at SNR 0 & SNR -5.

Stimuli: Lexical items for this experiment were generated by selecting the most frequent words in a spoken corpus, Verbmobil (Wahlster, 1993). In order to make sure that these words can later be substituted reliably without changing the meaning of utterances, we further selected only those words that belong to a single synset in the lexical database WordNet (Fellbaum, 1998). A few examples of pairs are:

- *absolutely - perfectly*
- *film - movie*
- *eatery - restaurant*
- *usually - normally*

A set of 189 synonym pairs (265 unique words) were selected and split into multiple lists such that no two synonyms were presented to the same participants. Stimuli for this experiment consisted of spoken words which were synthesized using the Google Translate API (gTTS) (Durette, 2014: accessed July 30, 2020) and their noisy signals were generated by performing additive noise mixing with noise files retrieved from NOISEX-92 database (Varga and Steeneken, 1993). Additive noise mixing was performed using a python library named *audio-SNR*, in which distorted signals were generated by mixing noise signals at an arbitrary signal-to-noise ratio (SNR). SNR, which is defined in (2.3), is the ratio between clean and noise signal strengths. See Section 2.3.2.1 for more details on the noise-mixing procedure that we followed.

Design and procedure: The synthetic speech signals were categorized into multiple blocks and it was ensured each block was presented to five different listeners. Participants were instructed to write down what they heard after listening to each spoken word. In the study instructions, we asked participants to ensure a quiet environment and to use good-quality headphones in order to take part in the experiment. Also, the significance of these recommendations was highlighted to them by providing sample audio files and a warning that audio files will be played only once.

Participants: The single-word listening experiment was deployed on a crowd-sourcing platform, Prolific[3] using LingoTurk framework Pusse *et al.* (2016) with 75 native British English speakers (53 females and 22 males) with an average age of 31 (ranges from 18 to 49).

Analysis: Participants' responses were then utilized to calculate the HRS, as defined in (2.1) of each stimulus. Further diff.HRS was calculated for each synonym pair under different listening environments (as defined in (2.2)). We also calculated individual participants' performance for each listening environment to observe if any outliers exist. With the outlier analysis outcomes, we decided to include all participants for further analysis. From this experiment, we expect to find that synonyms' recognition would be significantly different under noise.

4.4.1.2 *Results and discussion*

Babble Noise. With the increase in noise (clean $\longrightarrow$ SNR 0 $\longrightarrow$ SNR -5), as expected, the average HRS reduced significantly from 0.93 to 0.81 and 0.57 ($p < 0.001$), as the increased masking effect of noise tampered word recognition. However, the *average recognition difference* between synonym pairs steadily increased from 0.09 to 0.28 ($p < 0.05$) and finally to 0.39 ($p < 0.05$). As presented in Figure 4.5, in both noisy environments, there is a significant increase in the number of synonym pairs that were distinct in their HRSs than those in the quiet listening environment.

The increased number of synonyms that differ substantially in intelligibility at SNR 0 and SNR -5 indicates that choosing a word over its synonym can introduce a significantly larger impact on noise-induced comprehension errors in noisy environments. At SNR 0, the average HRS of the more intelligible synonym in the pair was 0.97, while the average HRS of the harder-to-perceive synonym was 0.69; at SNR -5, the harder-to-perceive synonym had an average HRS of 0.37, while the

[3]https://www.prolific.co/

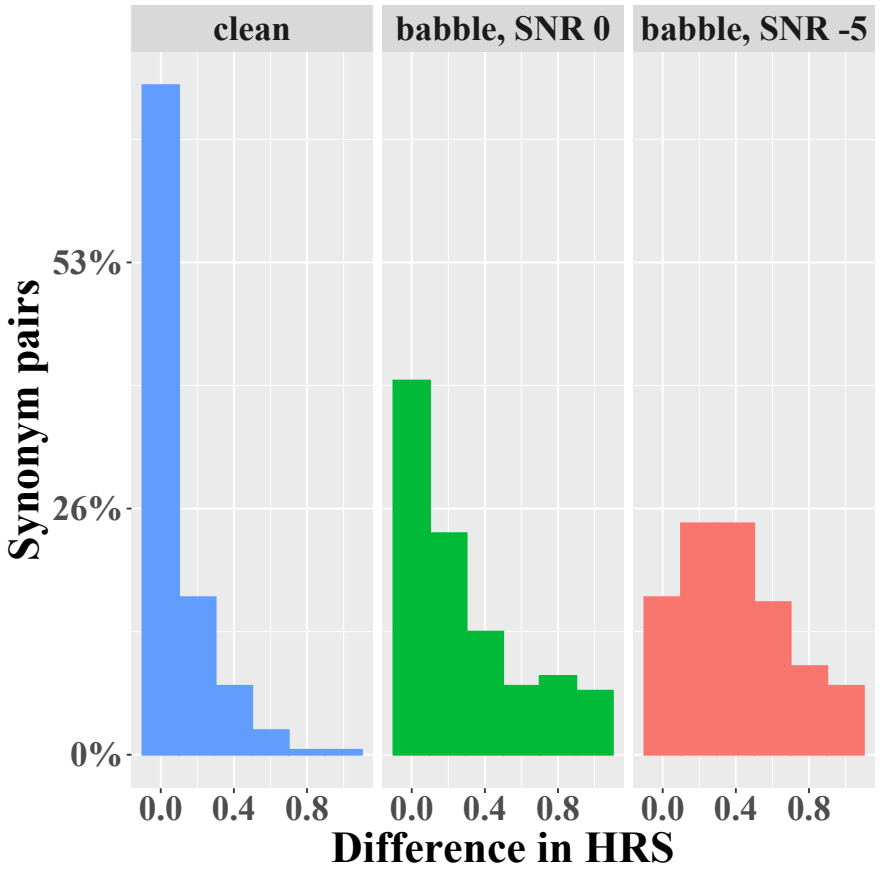


Figure 4.5: In the newly created dataset, Synonyms-in-Noise, the number of synonym pairs that were distinct in recognition significantly increased with an increase in *babble noise* level (clean $\longrightarrow$ SNR 0 $\longrightarrow$ SNR -5), when they were presented **without context** in different listening environments

more intelligible one in the pair had an average HRS of 0.77. These results reflect that the relative gain in intelligibility by choosing a lexical unit over its synonym which is at high risk of being misheard in babble noise is 40% at SNR 0 and 100% at SNR -5. Our experiment hence demonstrated that synonyms can differ substantially in intelligibility, especially at higher levels of babble noise. As an extension to this conclusion, we also studied the mishearing of synonyms in listening environments with a steady-state noise in the background.

White Noise. Similar to the babble noise condition, the presence of white noise in the background also introduced a significant reduction of overall word intelligibility of 0.93 at clean to 0.69 at SNR 0 and 0.52 at SNR -5 ($p < 0.001$), with an increasing noise level. At SNR 0, listeners experienced more difficulty in word recognition with

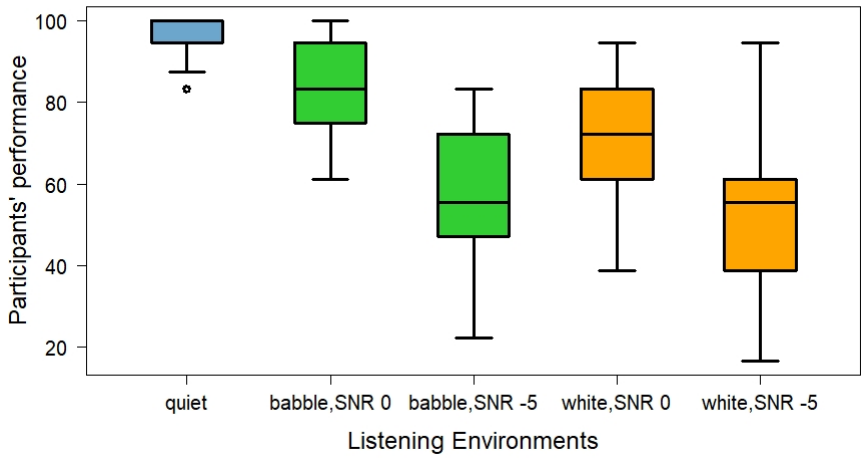


Figure 4.6: A comparison of word recognition performance under two different maskers. The recognition data of synonyms in isolation (of the Synonyms-in-Noise dataset) is used to plot the performance of participants under different listening environments. Human listeners were more challenged in white noise than in babble noise conditions.

white noise compared to babble noise, as observed in earlier work regarding the effect of different noise types on speech perception (Taitelbaum-Swead and Fostick, 2016). However, participants' performance was not significantly different with the masker type at SNR -5 condition, as depicted in Figure 4.6.

As depicted in Figure 4.7, the overall difference in HRS between synonym pairs increased steadily with reducing SNR of white noise, replicating the similar trend observed with babble noise condition. Similarly, with more noise in the background, the mean pairwise intelligibility difference increased from 0.09 (clean) to 0.30 (SNR 0) and 0.37 (SNR -5). Data collected from this experiment also assured us that appropriate choice of lexical paraphrases can improve intelligibility even in the presence of static noise.

Babble Noise vs. White Noise. We observed that lexical paraphrases are significantly different in their recognition in the presence of both stationary and non-stationary noise. However, it is not yet investigated whether a lexical unit that is more noise-robust (than its synonym) at babble noise condition remains as the more recognizable synonym for the other masker type - white noise. For this purpose, first, we ranked words in each lexical paraphrase pair based on their individual HRS under a noisy listening environment. Then, we calculated the percentage of

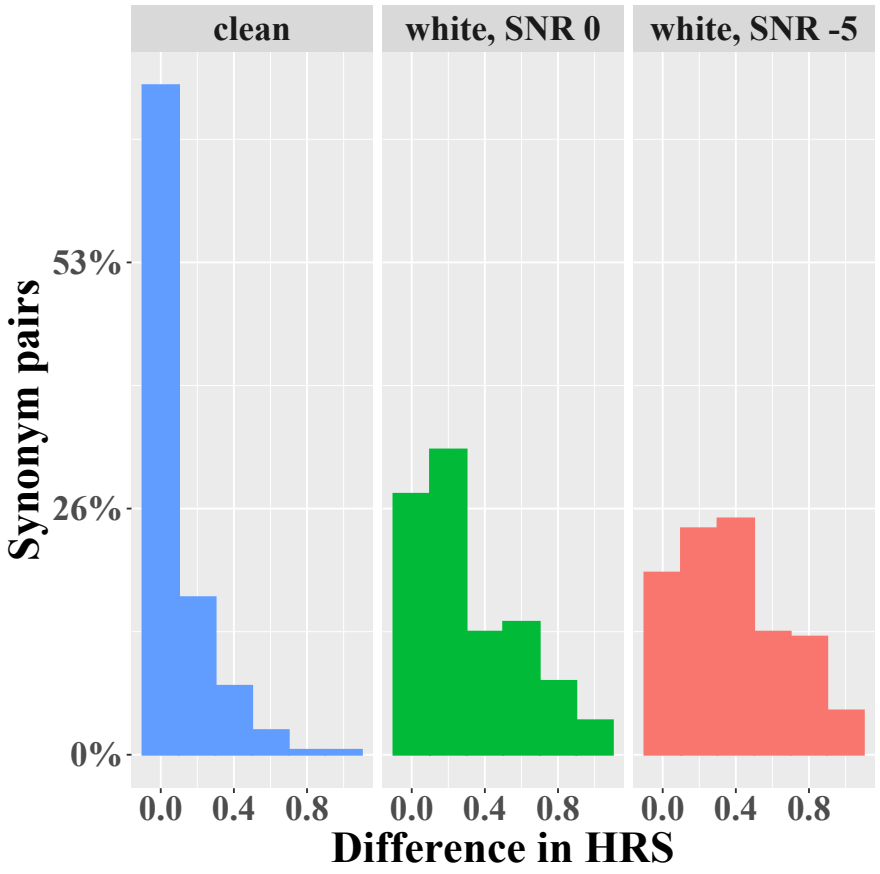


Figure 4.7: In the new dataset, Synonyms-in-Noise, the number of synonym pairs that were distinct in recognition significantly increased with an increase in *white noise* level (clean $\longrightarrow$ SNR 0 $\longrightarrow$ SNR -5), when they were presented **without context** in different listening environments

synonym pairs that agreed (*i.e.*, concordant pairs) on assigned ranks at a particular noise level of two masker types. As shown in Table 4.1, almost half of the synonym pairs disagreed (*i.e.*, discordant pairs) on their ranking at both SNR 0 and SNR -5. This indicates that the choice of a noise-robust lexical paraphrase is also influenced by the type of acoustic noise present in the background.

For example, consider the synonym pair *'ethnic'-'cultural'* at SNR 0 noise conditions. At white noise condition, we observed that the lexical unit *'ethnic'* / *EH TH N IH K* was better perceived than its synonym, *'cultural'* / *K AH L CH ER AH L*, and it was the reverse with babble noise. One possible explanation for such disagreements is presumably because of the interaction of noise with underlying sounds - the fricatives (like '*CH*') are hard to recognize in white noise conditions, while plosives (like '*TH*') are more misheard in babble noise - as demonstrated in earlier studies

-	% Agreement	Sample concordant pairs	Sample discordant pairs
SNR 0	51.85	*all-wholly* *doodle-scribble*	*finicky-picky* *excellent-splendid*
SNR −5	46.56	*absolutely-perfectly* *after-subsequently*	*section-department* *ethnic-cultural*

Table 4.1: A comparison of pairwise ranking of synonyms, under two masker types – babble noise and white noise. The agreement on ranking and a few samples of concordant and discordant pairs are reported.

on confusions of consonants in babble noise (Cutler *et al.*, 2004) and white noise (Phatak *et al.*, 2008). Similarly, past explorations on consonant misperceptions have also demonstrated that recognition of consonants at the initial position is more challenging than those at the final position (Weber and Smits, 2003; Cutler *et al.*, 2004). However, we also observed synonym pairs for which this explanation does not hold and we hypothesize that it might have been caused by presenting stimuli of spoken words rather than syllables or triphones, which were used in prior work.

To summarise, in this section, we observed that lexical replacement is a promising avenue for mitigating misperception in different noisy conditions. The next step is to test whether this effect also holds for words presented with linguistic context.

4.4.2 RQ 1*b*: Synonyms in noise, *with* context

In the previous section, we observed that synonyms in noise exhibit different noise-robustness. However, whether such differences in perception still exist in the presence of linguistic context is not yet investigated. This aspect is critical to be studied as prior work has demonstrated evidence that context influences word recognition. To this end, we designed a Short Utterance Listening (SUL) experiment to present synonyms with linguistic contexts.

4.4.2.1 *Experimental setup*

Participants of this listening experiment were asked to listen to noisy utterances at three listening setups: babble noise at SNR 5, 0, and -5. Unlike the noise levels in the single-word listening experiment, a low noise environment was considered instead of no noise as recognition in a quiet environment was close to a ceiling effect.

Stimuli: For generating stimuli for this experiment, initially a list of the top-most

500 frequent words from the Spoken BNC2014 corpus (Love *et al.*, 2017) and their synonyms from WordNet(Fellbaum, 1998) was created. Next, this list was filtered to identify synonym pairs such that both words from a pair semantically fit in a short utterance taken from the Spoken BNC2014 corpus. For example, for the synonym pair (*sea*, *ocean*), the following short utterances were used.

- *and he runs away scared and dives into the* ***sea***
- *and he runs away scared and dives into the* ***ocean***

This procedure resulted in 91 paired paraphrases, which were synthesized using gTTS. Subsequently, babble noise from NOISEX-92 (Varga and Steeneken, 1993) was added. For the noise mixing, the SNR was kept fixed across the target as well as its context.

Design and Procedure: We used the participants' transcription of what they heard to identify whether words were recognized correctly. Since the position of synonyms was not fixed across all utterances, participants were instructed to transcribe the whole utterance. To mark those words that they couldn't recognize in an utterance, they were informed to use '...' (3 dots) as a placeholder. Every stimulus was presented to six different participants in such a way that synonyms were not presented to the same participant.

Participants: Similar to the earlier experiment, participants were recruited from the crowd-sourcing platform, Prolific[4]. A total of 51 native British English speakers (36 females and 15 males) with an average age of 34 (ranging from 20 to 50) participated in this experiment.

Analysis: Participants' responses were processed to identify whether target words (i.e. synonyms that undergo the lexical replacement in pair paraphrases) were recognized or not. For each stimulus, we again calculated HRS (as defined in (4.1)). Further diff.HRS (as defined in (4.2)) was calculated for each paraphrase pair under different listening environments. From this experiment, we expect to find that synonyms' recognition would be significantly different, even when they were presented with a linguistic context in noise environments.

4.4.2.2 *Results and discussion*

Every lexical item in each synonym pair was then classified as either *less intelligible* or *more intelligible*, based on their HRS values. HRS_{min} and HRS_{max} are used to

[4]https://www.prolific.co/

refer to the HRS of the less intelligible and more intelligible synonyms in a pair. Then, we compared the recognition score difference at each paraphrase pair and analyzed the effect of a synonym replacement strategy under each noisy listening condition. Figure 4.8 summarizes the intelligibility differences of synonyms when they were presented with a linguistic context in noisy environments. The effect of replacing a target word with its synonym is evidently the largest for highly noisy environments. The mean difference in recognition score between a target and its synonym, at SNR -5 (0.37, $p < 0.001$) is significantly higher than in SNR 5 (0.15). However, the observed average difference at SNR 0 (0.21, $p = 0.10$) is not significantly different from that in SNR 5.

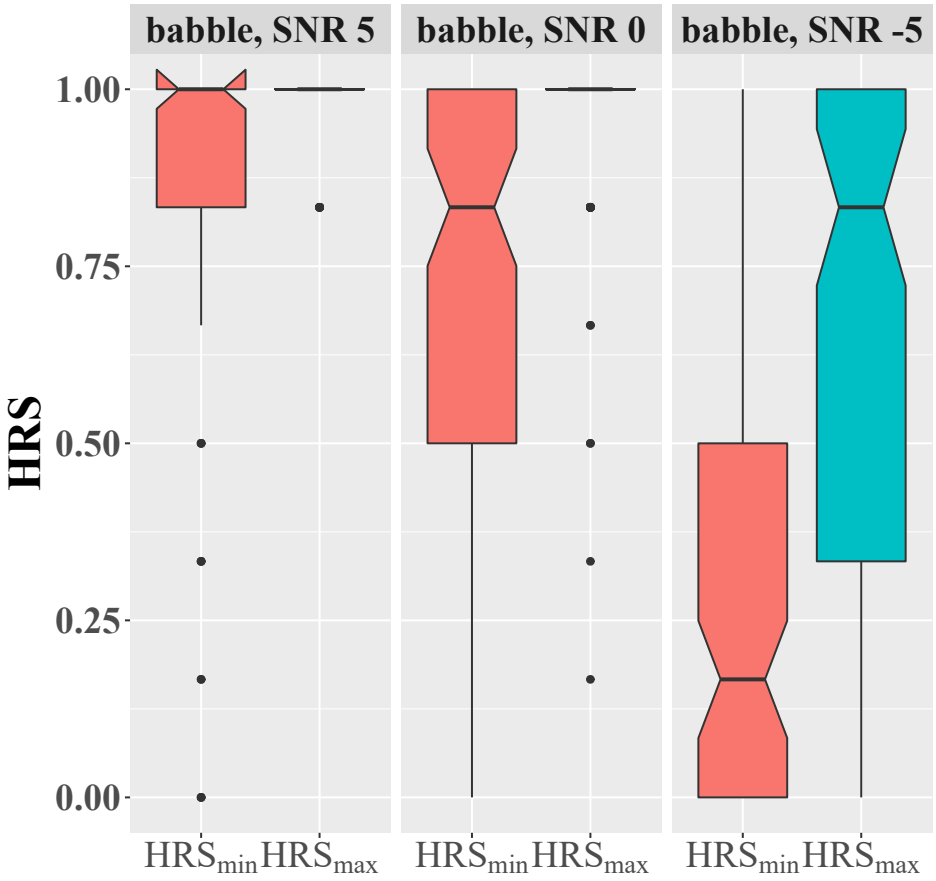


Figure 4.8: An overview of the distinction between two groups of synonyms – less intelligible (HRS_{min}) and more intelligible HRS_{max} – under different listening environments. Synonym pairs were presented to native English listeners **with linguistic context**.

This indicates that lexical replacement is most beneficial when there is a large

amount of noise. The average HRS_{min} (0.84) and average HRS_{max} (0.98) of synonyms at SNR 5 highlights that most of the target words were correctly recognized and this limits the scope of improvement that can be achieved by lexical replacement. At SNR 0, the average of HRS_{min} and HRS_{max} was 0.73 and 0.94 respectively. In contrast to these environments, the average of HRS_{min} and HRS_{max} at SNR -5 was 0.29 and 0.66 and this assured that at highly noisy environments, lexical replacement introduced a significant reduction in noise-induced comprehension errors. This observed distinction in synonyms' recognition, even when they were preceded by naturalistic context such as in everyday conversations (Love *et al.*, 2017) implicates the usefulness of this strategy for generating noise-robust utterances.

4.5 EXPLAINING THE GAIN IN WORD INTELLIGIBILITY

In this section, we address RQ 2 to study different factors that contribute to the better intelligibility of words over their synonyms in noise. The intelligibility difference among lexical paraphrases is modeled using their linguistic and acoustic cues. Findings from this modeling experiment explain the observed intelligibility gain, which is introduced by lexical paraphrases.

In Section 4.4, we analyzed the data of listening experiments. Both experiments have proven that lexical replacement can be a promising approach for improving spoken language comprehension in noisy environments. In order to automatically be able to choose the more intelligible synonym, it is necessary to classify word intelligibility automatically. In this section, we explore the extent to which computational measures can explain the variance of word recognition in noise. We used data collected from the Short Utterance Listening experiment for all models in this section.

Modeling word recognition in noise has been primarily studied from the perspective of understanding the nature of word misperception. Prior studies concentrated on identifying phoneme confusions under different noise and listening setups. Some of the microscopic confusion matrices were even used to model the probability of correct recognition of isolated words in noise. One such metric is FWNPR (Luce and Pisoni, 1998), where the frequency of the target word, its neighborhood density, and the frequency of its neighbors were considered for modeling word intelligibility in noise.

Previous work on the misperception of words with linguistic context provided

evidence that the predictability of the target item is a significant factor for its intelligibility in noise. We evaluated the following acoustic and linguistic features of a target word in context, by fitting a linear regression model (using its implementation in R (R Core Team, 2019)) to the data of three different noisy listening environments, separately.

(1) Linguistic predictability: A pre-trained LSTM-based language model(Merity *et al.*, 2018) was utilized to determine the predictability of a target word considering its left context in an utterance. The language model was trained on the transcription of a spoken corpus(Godfrey *et al.*, 1992) since stimuli utterances were taken from a speech corpus too. Log probability retrieved from this language model (hereafter referred to as *log.prob*), is used to represent how predictable a target word is given its linguistic context.

(2) Length of phonemic transcription: Number of phonemes is particularly an important lexical feature for word recognition, as several studies have shown that longer words (which have fewer neighbors that sound similar (Pisoni *et al.*, 1985)) are easier to recognize (Vitevitch, 2002; Vitevitch and Rodríguez, 2005). On the contrary, we can also find evidence in the literature that familiar words (which are usually shorter in length as per Zipf's law) are easier to recognize. However, whether the word familiarity favors its recognition in noise is not well documented in those studies. Thus to study such effects on word intelligibility in noise, the length of phonetic transcription (which was generated using a Grapheme-to-Phoneme(G2P) converter (Epp, 2018: accessed by July 30, 2020)) was used to represent this feature (hereafter referred to as *ph.len*).

(3) STOI: Short-Time Objective Intelligibility (STOI) (Taal *et al.*, 2010) measure is one of the classical SI metrics. By comparing temporal envelopes of clean and noisy speech it captures the mean correlation between the energy of clean and distorted time-frequency units over all frames and bands. STOI value ranges from -1.0 to 1.0, representing least intelligible to most intelligible audio signals. Three STOI values for each target word were then calculated by considering the clean and noisy signals from all 3 listening environments.

4.5.1 Results and Discussion

As a first analysis, the significance of the above-mentioned features for determining the HRS in noise was evaluated by fitting a linear regression model separately

for each of the listening environment data. At SNR 5, the model identified log probability ($\hat{\beta}$ = 0.031, SE = 0.01, t = 2.16, $p < 0.05$) as the only significant feature for explaining the variance in human recognition. At SNR 0, both log predictability ($\hat{\beta}$ = 0.05, SE = 0.02, t = 2.66, $p < 0.05$) and phoneme length ($\hat{\beta}$ = 0.04, SE = 0.02, t = 2.14, $p < 0.05$) were significant predictors of HRS.

However, for babble noise with SNR -5, we find that phoneme length ($\hat{\beta}$ = 0.06, SE = 0.03, t = 2.09, $p < 0.05$) and *STOI* ($\hat{\beta}$ = 0.06, SE = 0.03, t = 2.09, $p < 0.05$) are significant predictors of HRS, but not predictability. This difference between noise conditions may be due to difficulty with decoding the context: if the context cannot be fully understood, then it cannot be used effectively for predicting upcoming words.

HRS differences among synonyms. Next, we separate out overall effects of the predictability of a word from the difference in predictability between the two synonyms, in order to not only observe whether predictability as such is a significant predictor of HRS but also whether the difference in predictability between the synonyms makes a difference. Therefore, we encoded the response variable in terms of the difference between the HRS scores in a pair of synonyms by subtracting the HRS of the less intelligible synonym from the HRS of the more intelligible synonym. The resulting *diff.HRS* scores thus range between 0 and 1, with 0 indicating that there was no difference in intelligibility.

Furthermore, we used the variable *log.prob* to encode the predictability of the more intelligible word in the pair and separately encoded the difference between them in the variable *diff.log.prob*. Positive values of *diff.log.prob* thus mean that the synonym with higher HRS was also more predictable. Similarly, we separately encoded the word length of the better-recognized synonym in a pair as *ph.len*, and the difference in length to the other synonym as *diff.ph.len*. A positive coefficient value indicates that the lexical items with better perceptions are those words, which are longer in length than its synonym. In addition, for each listening environment, the intelligibility measure based on the acoustic features of the most recognized synonym in a pair was encoded as *STOI* and its difference with the other synonym as *diff.STOI*.

For the analysis, maximal models with all features were considered and the best fitting model (which has the lowest Akaike Information Criterion (AIC)) selection was performed using the step function in R (R Core Team, 2019). The maximal model at SNR 5 identified the difference in synonyms' predictability in context

	$\hat{\beta}$	**SE**	**t value**	**p-value**
Babble, SNR 5				
(Intercept)	-0.775	0.501	-1.546	0.126
log.prob	-0.034	0.014	-2.477	0.015 *
diff.log.prob	0.033	0.009	3.651	0.0 ***
ph.len	0.027	0.014	1.991	0.050 *
diff.ph.len	0.023	0.01	2.254	0.027 *
STOI	0.464	0.507	0.915	0.363
diff.STOI	-0.664	0.386	-1.722	0.089 .
Babble, SNR 0				
(Intercept)	-0.608	0.43	-1.415	0.161
log.prob	-0.045	0.015	-2.967	0.004 **
diff.log.prob	0.04	0.011	3.595	0.001 ***
ph.len	0.011	0.015	0.712	0.478
diff.ph.len	0.033	0.012	2.756	0.007 **
STOI	0.346	0.491	0.704	0.483
diff.STOI	-0.175	0.33	-0.531	0.597
Babble, SNR -5				
(Intercept)	1.134	0.449	2.523	0.014 *
log.prob	-0.018	0.019	-0.927	0.356
diff.log.prob	0.011	0.013	0.836	0.406
ph.len	-0.009	0.019	-0.473	0.637
diff.ph.len	0.025	0.015	1.749	0.084 .
STOI	-1.428	0.495	-2.887	0.005 **
diff.STOI	0.694	0.324	2.142	0.035 *

Table 4.2: An evaluation of the Human Recognition Score (HRS) difference between synonyms by fitting linear regression models to short utterance listening (SUL) experiment data of **babble noise at SNR** 5**, SNR** 0 **and SNR** -5.

as well as their difference in the number of phonemes as significant features in explaining the variance in diff.HRS; *diff.log.prob* ($\hat{\beta}$ = 0.03, SE = 0.01, $p < 0.001$) and *diff.ph.len* ($\hat{\beta}$ = 0.02, SE = 0.01, $p < 0.05$), see also Table 4.5.1 for a detailed report. As these predictors are in the direction of the response variable, it indicates that replacing a lexical unit with its synonym which has better predictability in a context leads to better recognition under a noisy environment in which the context is intelligible. Similarly, the model shows that there is a gain in recognition when a lexical unit is replaced by its synonym which has a greater number of phonemes. These observations are congruent with earlier studies on the effect of predictability

on word recognition (Kalikow *et al.*, 1977) and the reduction of confusion with longer words (Vitevitch, 2002; Vitevitch and Rodríguez, 2005). It is noteworthy that the difference in STOI was not significant in the maximal model and hence the best-fit model excluded acoustic-based features for explaining the variance in diff.HRS in a low-noise environment.

The model exhibited similar effects at a medium noisy environment (SNR 0) by identifying the significance of *diff.log.prob* ($\hat{\beta}$ = 0.04, SE = 0.01, $p <= 0.001$) and *diff.ph.len* ($\hat{\beta}$ = 0.03, SE = 0.01, $p < 0.01$) for explaining the variance in improved recognition. This reflects that under low/medium noisy environments, the gain in recognition introduced by lexical replacement is better explained by the improved predictability or the increased number of phonemes introduced by the replaced lexical item. However, the difference in the intelligibility of synonyms didn't have an effect on their recognition in such low/medium noisy environments.

In contrast, the model for SNR -5 showed that neither *diff.log.prob* nor *diff.ph.len* were significant predictors of the improvement in HRS through lexical replacement. Instead, it revealed that the replacement of a lexical unit with its more intelligible synonym can be predicted by the measure $STOI$ and *diff.STOI* ($\hat{\beta}$ = 0.69, SE = 0.32, $p < 0.05$). **This reflects that in highly noisy environments, choosing a lexical unit that has better noise-robust acoustic cues than its synonym can significantly improve its recognition.**

4.6 CONCLUSION

In this chapter, we investigated the potential of a new strategy of choosing noise-robust lexical paraphrases to mitigate comprehension errors that are caused by noise in listening environments. Listening experiments with human subjects were conducted to investigate whether the recognition of synonyms differs in an environment with babble/white noise in the background. We found that the potential impact of lexical replacement increased with an increase in the noise level (9% at clean, $\sim$ 28% at SNR 0, and $\sim$ 39% at SNR -5). Similar effects were also observed when synonyms were presented with the linguistic context in noisy listening setups.

Further investigation on the observed reduction in noise-induced comprehension errors by lexical replacement revealed that the intelligibility of a word in low and medium noise conditions is primarily driven by a word's predictability. On the other hand, in more noisy environments, the intelligibility of a word was mainly driven

by its acoustic features as captured by the STOI. This highlights that under high noise environments, the availability of top-down expectations like linguistic context, gets compromised and strong bottom-up signals like noise-robust acoustic cues are essential for better word perception. Thus, we conclude this chapter with a key observation that when an intended meaning needs to be realized as spoken words in very noisy environments, choosing noise-robust lexical paraphrases is a promising approach to improve comprehension.

For reducing word misperception in naturalistic dialogue environments, a speech synthesis strategy solely based on lexical paraphrases limits its scope of application. This leads to the need for a generic paraphrasing strategy and raises the question of whether modifying linguistic characteristics at a sentence level reflects any improvement in the overall utterance perception, under noisy listening setups. We will explore such questions in the next chapter, with a specific focus on sentential paraphrases and measuring its impact on intelligibility at a sentence-level.

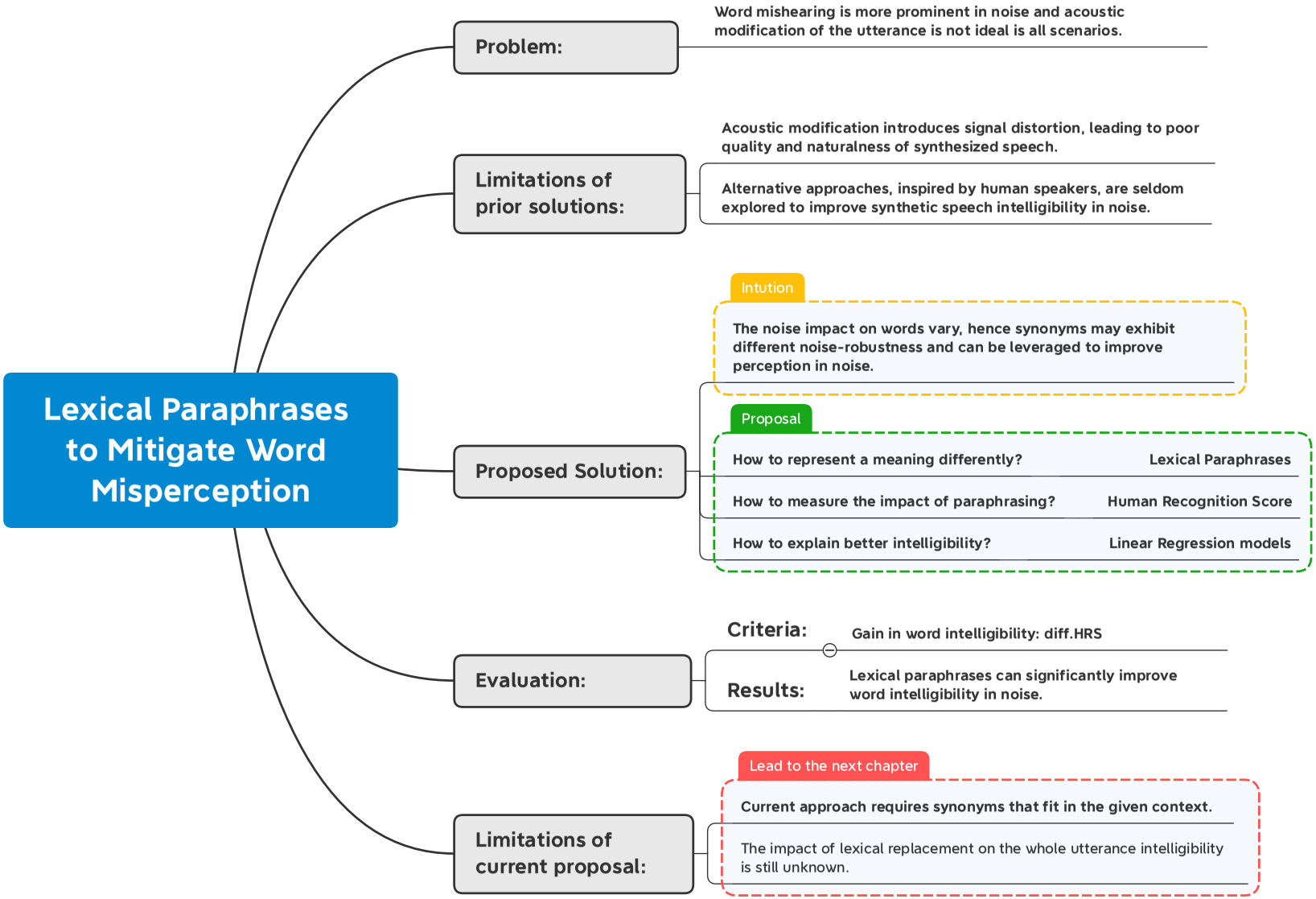


Figure 4.9: An overview of Chapter 4 that explores the potential of lexical paraphrases to mitigate word misperceptions in noise.

5 SENTENTIAL PARAPHRASES TO IMPROVE INTELLIGIBILITY

In the last chapter, we found that lexical paraphrases can significantly reduce word misperception in noise. However, it raises an important question of whether lexical replacements with synonyms can impact the overall utterance intelligibility - the difference in noise-robustness of synonyms is not comprehensive to determine the intelligibility difference between sentences, which consists of a word that undergoes the replacement as well as its linguistic context. Thus, measuring noise-robustness at the sentence level is critical and doubly so, when considering the application of mishearing mitigation strategies such as in Spoken Dialog Systems (SDS), which produce naturalistic utterances that are usually longer than single words. Additionally, an intelligibility improvement strategy solely based on lexical replacements is constrained by the availability of synonyms that can fit in a given linguistic context.

Hence, in this chapter, we shift our investigations toward sentence-level intelligibility (as described in Section 5.3), employing sentential paraphrases (see Section 5.2) to include different types of paraphrases in English (Lan *et al.*, 2017), like elaboration, word or phrase reordering, anaphora resolution, etc. The listening experiment that we conducted to create the perception data of sentential paraphrases in noise, is elaborately discussed in Section 5.4. Further, we proposed a working prototype for a noise-adaptive SDS by developing an intelligibility-aware paraphrase-pair ranking model (in Section 5.6), as it is critical to automatically identify the better intelligible linguistic form among the list of candidate paraphrases. The main contributions of the current work are the following:

- We demonstrated that the choice of linguistic forms to represent a message can indeed influence its utterance intelligibility in noise.
- We created a perception dataset of sentential paraphrases in noise [5]. The largest publicly available corpus of its kind.
- We outlined a schema for noise-robust speech synthesis, which introduces no signal distortion to synthesized utterances.

[5]Experiment data is publicly available here: Paraphrases-in-Noise (PiN)

Overall, the findings in this chapter demonstrate the potential of sentential paraphrases to *proactively* synthesize better intelligible speech in noise.

5.1 INTRODUCTION

Lexical replacements with noise-robust synonyms are capable of mitigating word misperception in noisy environments. However, the impact of such lexical replacements on the whole utterance intelligibility is still unknown: the difference in human recognition score (HRS) of synonyms *may not correlate* with the actual difference in the intelligibility of the whole sentence that underwent the lexical replacement. Measuring the impact of paraphrasing on the whole utterance intelligibility is also critical as prior studies have shown that speech perception is facilitated by top-down signals like linguistic cues and world knowledge, especially in adverse listening conditions where the bottom-up acoustic signals are compromised Kalikow *et al.* (1977); Schoof and Rosen (2015); Ward *et al.* (2017).

Consider the synonym pair (*sentiment*, *view*) that are equally likely to fit in a given linguistic context:

but probably not the actual ____ *behind it*

Filling the gap with one of the synonyms result in a pair of sentential paraphrases s_1 and s_2:

- s_1: *but probably not the actual* ***sentiment*** *behind it*
- s_2: *but probably not the actual* ***view*** *behind it*

Based on the listening experiment described in Section 4.5, we observed that under babble noise at SNR -5 dB, the lexical item *sentiment* exhibited higher perception (*i.e.*, high HRS) than the token *view*. However, it is still unknown whether the whole utterance intelligibility of s_1 is better than s_2. Though sentences s_1 and s_2 are only different by a single lexical item, the impact of lexical replacement needs to be measured at a sentence level – word(s) that underwent paraphrasing may or may not influence the perception of the whole sentence and the ultimate objective of paraphrasing is to improve the overall speech perception, rather than individual word intelligibility. This signifies the need to determine the *impact of paraphrasing on intelligibility by measuring noise-robustness at a sentence-level*, rather than the word-level noise-robustness, which was defined in (4.1).

Another limitation of the mishearing mitigation strategy proposed in Chapter 4 is the constrained lexical operation, which is associated with synonyms that fit in a given linguistic context. To this end, we propose to extend this simple mitigation strategy by considering sentential paraphrases, on top of lexical paraphrases, thereby including different paraphrase phenomena. As elaborately discussed in Bhagat and Hovy (2013), synonym replacement is just one of the many possibilities of creating sentential paraphrases that are distinct in surface form and equivalent in semantics (Lan *et al.*, 2017).

One of the main objectives of the current study is to analyze whether the human perception of sentential paraphrases is significantly different in noisy listening conditions – if the noise impact on sentential paraphrases is equally the same, then using one over the other is less likely to introduce any gain in intelligibility. Further, it is also important to analyze why certain sentences are more intelligible than their paraphrase if any difference exists. Prior experiments on sentence-level linguistic characteristics like syntactic structure (Carroll and Ruigendijk, 2013; Van Kuyk *et al.*, 2018) and word order (Uslar *et al.*, 2013) have demonstrated their influence on the average utterance intelligibility. However, pair-wise comparisons are rarely documented in the literature to analyze the benefits of paraphrasing on utterance intelligibility. Hence, we perform pairwise analysis to explain whether and to what extent sentential paraphrases can introduce an intelligibility-gain, in noise.

Finally, we propose an intelligibility-aware paraphrase ranking model, embedded in a spoken dialogue system (SDS), to generate noise-robust utterances using paraphrasing, as illustrated in Figure5.1.

The rest of this chapter is dedicated to discussing the following three research questions:

- RQ 1: Do sentential paraphrases differ from one another concerning how intelligible they are in noisy conditions?
- RQ 2: If so, what contributes to the observed intelligibility differences?
- RQ 3: Finally, how can we utilize paraphrasing to synthesize noise-robust utterances?

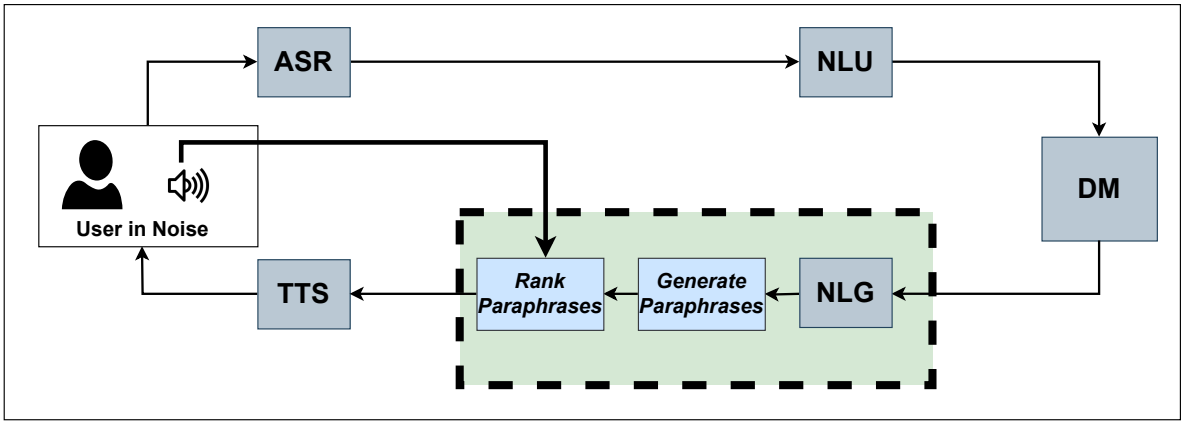


Figure 5.1: The proposed framework to generate *noise-adaptive system utterances*, using paraphrases. The framework propose to embed two new modules – paraphrase generation and intelligibility-aware ranking – to a traditional spoken dialogue system, which consists of automatic speech recognition (ASR), natural language understanding (NLU), dialogue manager (DM), natural language generation (NLG), and text-to-speech system (TTS).

5.2 SENTENTIAL PARAPHRASES

Paraphrases are phrases/sentences that represent a similar semantic meaning using different wordings. In Chapter 4, we have seen one of the common strategies of paraphrase generation: replacements of words in a sentence with their synonyms. However, paraphrasing techniques at a sentence level are not limited to synonyms. Instead, several linguistic modifications like the change of voice, person, specification, etc were also recognized as paraphrasing types in large paraphrase datasets in English (Bhagat and Hovy, 2013). A few examples of different types of sentential paraphrases are given below:

- Change of voice:
  - *Simone prepared the agenda.* ⇔ *The agenda was prepared by Simone.*
- Change of person:
  - *Vijay said, "I will stay home."* ⇔ *Vijay said that he would stay home.*
- General/Specific substitution:
  - *They are good with labradors.* ⇔ *They are good with dogs.*

Thus, to study the intelligibility differences between sentential paraphrases, it is important to consider different types of paraphrases in the experiment. Pairs of paraphrases can either be retrieved from a paraphrase corpus or it could be synthesized using a paraphrase generation model. Because of the unavailability of a paraphrase corpus for spoken data, stimuli for the listening experiment were created using a paraphrase generation model.

However, the notion comes with the difficulty that two different sentences rarely have the exact same meaning in all contexts, hence paraphrases, especially at the sentence level, typically only approximate the original meaning (Bhagat and Hovy, 2013). On the one hand, generating sentential paraphrases which are *exactly equivalent* in semantics leads to trivial patterns such as word order changes or minimal lexical substitutions among paraphrases (Madnani and Dorr, 2010). This however can mean that there is only a minimal difference in the effect of intelligibility in noise between such paraphrases. On the other hand, the generation of non-trivial paraphrases introduces better lexical/syntactic diversity, and may hence have larger effects on intelligibility, but this in turn, demands more scrutiny for semantic similarity (Dolan and Brockett, 2005).

In this chapter, we hence explore the effect of paraphrases that approximate semantic equivalence instead of strict semantic equivalence. To include a large variety of paraphrases, stimuli sentences were generated using a pre-trained text generation model (Rajauria, 2020: accessed by March 21, 2022; Zhang *et al.*, 2020) which was fine-tuned on several paraphrase datasets like Quora Question Pairs, PAWS (Zhang *et al.*, 2019) etc. For the input sentences to the paraphrasing model, we selected a list of short sentences (10-12 words) from the dialogue corpus Switchboard (Godfrey *et al.*, 1992). After paraphrase generation, we employed automatic filtering to select the top two paraphrases for each input sentence, based on a semantic similarity score (Zhang* *et al.*, 2020). This resulted in a list of paraphrase triplets (s_1, s_2, s_3), consisting of different paraphrase types formed by lexical replacements, changes in syntactic structure, reordering words, etc.

Since existing paraphrasing models lack the domain knowledge of spoken data, a manual selection was performed to ensure the quality of the generated paraphrases in terms of semantic equivalence. Every paraphrase triplet was converted to three pairs: (s_1, s_2), (s_2, s_3) and (s_1, s_3). Then, every paraphrase pair was verified for closeness in semantics. We identified about 300 triplets that exhibited approximate semantic similarity in all three pairs. Those triplets were randomly split into three groups of 100 (one for each listening environment). Hereafter, we refer to this dataset

Sentence-ID	Sentences
s_1	They seem to give more of just the facts than opinions.
s_2	They give more information than opinions.
s_3	They seem to give more facts than opinions.
s_1	You never hear about it really in the big ones.
s_2	You don't hear much about it in the big ones.
s_3	In the big ones you don't hear about it.
s_1	It was a very close game and hard fought game.
s_2	The game was close and hard fought.
s_3	It was a very close game.

Table 5.1: A few sample paraphrase triplets from the newly created Paraphrases-in-Noise (PiN) dataset.

Dataset	Total	SNR 5	SNR 0	SNR -5
PiN	900	300	300	300
PiN_{both}	332	104	123	105
PiN_{either}	596	195	205	196

Table 5.2: An overview of the number of paraphrase pairs per listening condition in PiN dataset. PiN_{both} and PiN_{either} are subsets of the PiN dataset, created based on human annotations.

as ***Paraphrases in Noise*** (PiN).

To ensure that the sentential paraphrases in the PiN dataset are indeed equivalent in semantics, two annotators were asked to label the pairs as 'paraphrase' only if they fit the definition of a 'quasi-paraphrase', i.e., *sentences or phrases that convey approximately the same meaning using different words* (Bhagat and Hovy, 2013). Around $\frac{2}{3}$ of PiN were identified as '*quasi paraphrase*' by at least one of the two annotators (hereafter referred to this subset as PiN_{either}) and $\frac{1}{3}$ by both annotators (hereafter referred to this subset as PiN_{both}). Table 5.2 shows details of these subsets per noise condition.

5.3 MEASURING HOW PARAPHRASES INFLUENCE INTELLIGIBILITY

In Chapter 4, we defined a measure called Human Recognition Score (HRS) in Equation (4.1), to calculate the noise-robustness of a lexical unit. Precisely, the HRS of a lexical unit represents the rate of correct recognition among all its listening instances

at a listening condition. Such lexical recognition scores were previously used in literature to define sentence-specific intelligibility scores to study the differences in sentence-specific characteristics like complexity (Uslar *et al.*, 2013). Similarly, earlier experiments have also used the percentage of correct keyword recognitions in an utterance to compare intelligibility differences among sentences, synthesized by different speech enhancement techniques (Cooke *et al.*, 2013b), or to compare the difference between speech perception and speech comprehension in noise (Fontan *et al.*, 2015).

We take a slightly different approach to measuring sentence-level intelligibility – instead of calculating the *exact recognition* of the lexical units in a target utterance, we *measure the deviation* of a perceived utterance from the target utterance. Specifically, we utilize an edit distance algorithm to determine whether the perceived utterance sounds similar to the actual utterance. A lower distance value indicates that the perception is better and the actual utterance is more noise-robust. Unlike HRS-based metrics, a sentence-level metric based on the edit distance captures the word order and repeated words in a sentence. Additionally, edit distance provides flexibility for calculating deviation at different granularities like words, characters, or phonemes (see subsection below for a detailed discussion).

We utilize a sentence-level intelligibility metric to perform pairwise comparisons among paraphrases, first, to determine whether their utterance intelligibility is different under noisy listening conditions. Further, we analyze the correlation between the difference in HRS of synonyms that underwent the lexical replacement and the difference in sentence-level intelligibility within their corresponding sentential paraphrases. The primary objective of that correlation analysis is to understand the significance of a sentence-level intelligibility score for comparing sentential paraphrases, which were created by a single lexical replacement in a sentence.

5.3.1 Sentence-level Intelligibility

We defined a measure called *sentence-level intelligibility* which captures the noise-robustness of an utterance (*ie.*, a sentence). This measure is motivated by earlier work on *slips of the ear*, where the proportion of listeners who misrecognized a word *w1* as another word *w2* is considered as the consistency of a confusion (Marxer *et al.*, 2016b). Similarly, to measure how well a target utterance (T) is perceived in a listening condition, the mean recognition rate of an utterance among a set of listeners

(with normal hearing thresholds) was calculated, as shown in (5.1). For this purpose, we ensured that every stimulus utterance was listened to by a set of n listeners.

$$\text{Sent-Int}(T) = \frac{1}{n}\sum_{i=1}^{n}\text{Recog-Rate}(T, P_i) \tag{5.1}$$

Further, the rate of recognition at each listening instance i is calculated by comparing the phonemic transcripts of the target (T) and perceived (P_i) utterances, as shown in (5.2). The phonemic transcripts T^{ph} and P_i^{ph} were generated using a Grapheme-to-Phoneme(G2P) converter (Kim, 2018: accessed by March 21, 2022). The stress markers in the phonemic transcript were ignored while comparing the transcripts. For string comparison, we used the Levenshtein distance (Lev), which calculates the minimum number of edits (i.e., deletions/substitutions/insertions of phonemes) required to change T^{ph} into P_i^{ph}.

$$\text{Recog-Rate}(T, P_i) = 1 - \frac{\text{Lev}(T^{ph}, P_i^{ph})}{\#phonemes_{T^{ph}}} \tag{5.2}$$

An equal cost (1) was assigned for all edit operations. The phoneme recognition rate was then calculated by first normalizing the edit distance by the number of phonemes in the target and then subtracting this value from 1. This ensures that the noise-robustness measure is not sensitive to the target utterance length. The intelligibility score for each utterance ranges between 0.0 (completely unintelligible) and 1.0 (completely intelligible), indicating that utterances with higher Sent-Int are better intelligible. As shown in the example reported in Table 5.3, the perceived utterance P_2 is closer to the target utterance (T) than P_1, which is also captured in their corresponding values for *Recog-Rate*.

We chose to calculate the Levenshtein distance between a target utterance and its perceived utterance in terms of phonemes, as defined earlier in (5.2). However, this measurement can be easily modified to calculate utterance intelligibility in terms of words or characters. Figure 5.2 depicts the correlation between intelligibility scores calculated using three different units: phonemes, characters, and words. With the perception data in the SUL dataset, we observed that all three intelligibility scores have a high Pearson correlation of above 0.9.

It should also be noteworthy that phoneme-based intelligibility is highly correlated to character-based intelligibility scores. This observation aligns with the earlier observation of high predictability in determining the spelling of English words from their pronunciation (Berndt *et al.*, 1987). On the other hand, we observed that the

Type	Transcripts
T	*you never hear about it really in the big ones*
T^{ph}	Y UW N EH V ER HH IY R AH B AW T IH T R IH L IY IH N DH AH B IH G W AH N Z
P_1	*it really is the big one*
P_1^{ph}	IH T R IH L IY IH Z DH AH B IH G W AH N
Recog-Rate	0.5
P_2	*he went about it really in the big one*
P_2^{ph}	HH IY W EH N T AH B AW T IH T R IH L IY IH N DH AH B IH G W AH N
Recog-Rate	0.7

Table 5.3: A sample target utterance (T) and its two perceived utterances (P_1 and P_2) are reported, along with their corresponding phonemic transcripts and recognition rates for each perceived utterance. As indicated by the recognition rates, perception is better at P_2 instance.

intelligibility scores based on words were relatively less, compared to those based on phonemes. This difference is primarily driven by minor morphological errors (eg: *calculation* vs. *calculate*) or spelling mistakes (eg: *practise* vs. *practice*), which results in a complete mismatch when measured in terms of words and a partial mismatch when the deviation is measured using phonemes. The current study is to determine the noise-robustness of an utterance. Therefore, we utilized the phoneme-based intelligibility measure as it represents the *similarity between the sounds* of the actual utterance and the perceived utterance.

5.3.2 Gain in Sentence-level Intelligibility

The Sent-Int score is further utilized to measure the difference in noise-robustness between two sentences: s_1 and s_2. When the paired sentences are sentential paraphrases, an absolute difference in the intelligibility of s_1 and s_2 indicates the impact of paraphrasing on the overall utterance intelligibility. Hereafter, this measure is referred to as *Sent-Int-Gain* and its value ranges from 0.0 (no gain) to 1.0 (maximum gain).

$$\text{Sent-Int-Gain}(S_1, S_2) = \mid \text{Sent-Int}(S_1) - \text{Sent-Int}(S_2) \mid \tag{5.3}$$

Sent-Int-Gain *vs.* diff.HRS: With noise-robustness measures defined at the word level (HRS) and sentence level (Sent-Int), at first, we analyzed whether better-

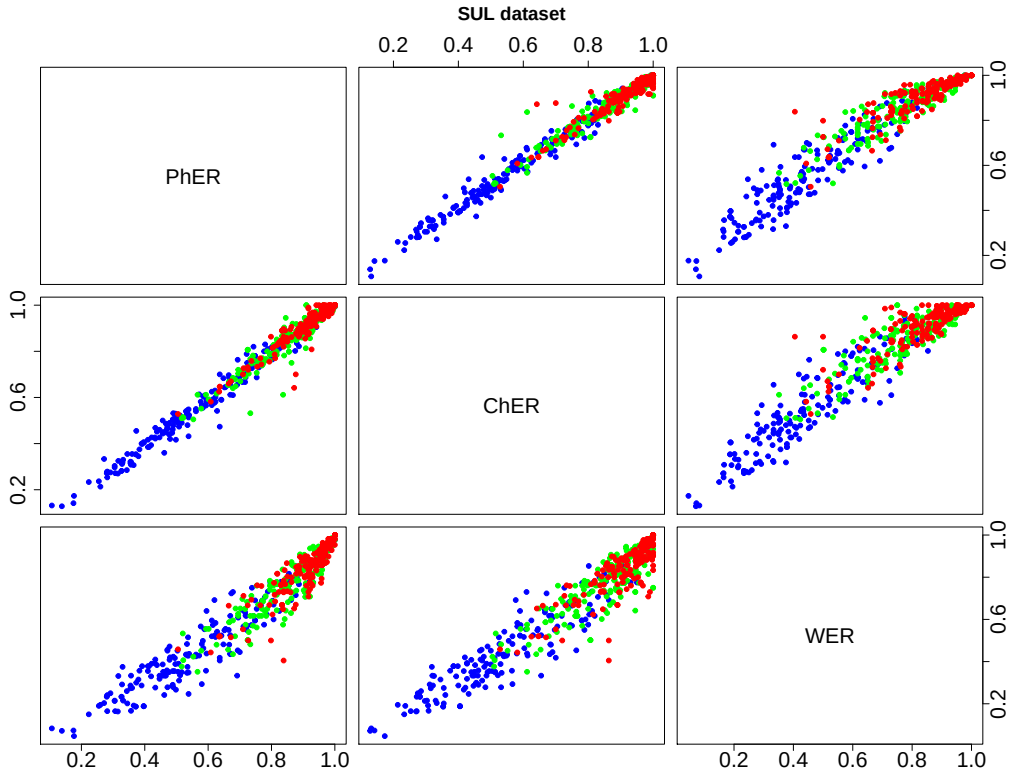


Figure 5.2: Correlation between three intelligibility scores, which are based on different tokens phonemes (PhER), characters (ChER), and words (WER). The plot entries are color-coded to distinguish the utterances of SUL dataset, perceived in the presence of babble noise at SNR 5 (red), SNR 0 (green), and SNR -5 (blue).

intelligible synonyms guarantee better utterance intelligibility, in noisy environments. To this end, we performed two different pair-wise rankings on the SUL dataset: sentences in each paraphrase pair were ranked, first based on their Sent-Int and then, based on the HRS of the synonyms that underwent the lexical replacement. This ranking is further utilized to study whether a lexical unit that is more noise-robust than its synonym leads to an utterance that is more intelligible than its sentence paraphrase (which is formed by lexical replacement with synonym).

Our results show that the correlation between the two rankings is relatively weak at all noise conditions: SNR 5 (r=0.33, $p < 0.05$), SNR 0 (r=0.32, $p < 0.05$), and SNR -5 (r=0.44, $p < 0.05$). It indicates that just comparing the intelligibility of lexical units that undergo the lexical replacement (*i.e.,* a word *vs.* its synonym) is not enough to determine the direction of sentence intelligibility, even when the rest of the words in the sentence remains the same - it could be possible that those lexical items that underwent the replacement are not critical for the overall utterance intelligibility.

Once again, this highlights the need to measure the impact of paraphrasing at the sentence-level, rather than just comparing the noise-robustness of lexical units that undergo lexical modification.

5.4 LISTENING EXPERIMENTS

Perception data of sentential paraphrases in noise is quintessential for studying their potential to improve utterance intelligibility. However, the availability of such perception data is very limited, owing to the limited prior explorations in this research direction. The SUL dataset described in Section 4.4.2 only consists of sentential paraphrase pairs which differ by a single lexical unit; excluding other paraphrasing styles to generate sentential paraphrases. Thus we conducted a perception experiment with paraphrase pairs in the PiN dataset, for a comprehensive study on perception differences among sentential paraphrases in noise.

5.4.1 Experimental Setup

This section describes the listening experiment setup that we followed to elicit (*mis*)perception of utterances in three different noisy listening conditions.

Stimuli: For audio stimuli creation, first, clean utterances for each sentence were synthesized using the speech synthesizer API of Google Translate (Durette, 2014: accessed July 30, 2020). Then, corresponding noisy utterances were generated by performing additive noise-mixing using the babble noise from the NOISEX-92 database (Varga and Steeneken, 1993) and the mixing tool audio-SNR (Sato, 2018: accessed July 6, 2022). To avoid priming effects, 15 stimuli lists with 60 sentences, were created while making sure none of the lists contained sentences that were paraphrases of each other. Every stimuli list consisted of utterances in all three noisy conditions.

Participants: The experiment was deployed on a crowd-sourcing platform, Prolific[6] using the LingoTurk framework (Pusse *et al.*, 2016). It was conducted with a group of 90 participants who are native speakers of British English, based in the UK. The group consisted of 60 women, 29 men, and one non-binary person. The average age of the cohort is 32.4 ranging from 19 to 50. Our experiment was not accessible to individuals who reported to have hearing difficulties.

[6]https://www.prolific.co/

Design and procedure: Every audio stimulus was presented to six different participants. Participants were asked to transcribe after listening to each utterance. Every participant had 60 listening instances (20 per noise level). Similar to the perception experiment described in our previous study (Section 4.4.2), participants were instructed to use a placeholder (...) to mark cases of completely unintelligible words. They were also encouraged to guess, when necessary.

Analysis: Experiment execution was followed by the intelligibility calculation (using Eq. 5.1) for all 900 utterances. The overall intelligibility in a listening environment was measured by averaging the sentence-level intelligibility across all utterances in a particular listening condition.

Using this score, we conducted several analyses. First, we compared the overall intelligibility at different noise levels for all three sets of paraphrase pairs: the full dataset, and the two subsets of stricter paraphrase pairs which were annotated as such by at least one or both the human judges. Further, we calculated the (absolute) difference in intelligibility for each paraphrase pair, in order to infer whether a variation in the linguistic form of a message introduces a gain in intelligibility, otherwise known as the Sent-Int-Gain as defined in (5.3). Sent-Int-Gain is a proposed metric to measure the impact of paraphrasing on speech perception, which was referred to as Δ_p in Section 3.1. The mean of this score represents the overall impact of paraphrasing on utterance intelligibility, ranging from 0.0 (no effect) to 1.0 (maximum effect), see Figure 5.3 below.

5.4.2 Results and Discussion

In this section, we analyzed the impact of paraphrasing on utterance intelligibility in noise, considering both quasi paraphrases (in PiN dataset) and strict paraphrases (in PiN_{both} and PiN_{either} datasets).

PiN Dataset. The overall intelligibility in each listening environment was calculated by averaging the sentence-level intelligibility scores across different sets of utterances. As expected, we observed a significant reduction ($p < 0.05$) in the overall intelligibility with an increase in the noise level, as shown in Table 5.4.

This indicates the impact of noise on utterance intelligibility, even with those individuals with NH thresholds. We noticed that listeners' ability to recognize utterances at SNR 0 is not as severely damaged as the SNR -5 condition. Reasons for this,

SNR 5	SNR 0	SNR −5
0.97	0.94	0.71

Table 5.4: An overview of overall utterance intelligibility in the PiN dataset. The overall intelligibility was reduced substantially, with an increase in the background noise level.

besides the lower effect of sound masking, could be that listeners also understand the context better and have more cognitive capacity available for generating predictions, which in turn help them to recognize the words.

RQ 1: Impact of paraphrasing: To study the impact of paraphrasing on utterance intelligibility, we looked at the intelligibility differences between paraphrase pairs under each listening environment. Figure 5.3 illustrates the intelligibility differences between paraphrase pairs for the PiN dataset. Further, we calculated the average intelligibility differences between paraphrases at each LE. This value is critical as it demonstrates whether the perception of paraphrases is different in a particular listening environment. We observed that at SNR 5 and SNR 0, most of the paraphrase pairs exhibit only a small difference in intelligibility. This is because of the ceiling effect of word recognition in listening environments with low noise levels. Although the mean intelligibility difference between paraphrases at SNR 0 (0.06, $p < 0.05$) is significantly above SNR 5 (0.04), their scores being close to 0.0 indicates the limited scope of paraphrasing to improve intelligibility under such less noisy environments. However, at SNR -5, we observed a mean intelligibility difference of 0.20 ($p < 0.05$) reflecting the increased number of pairs that are distinct in their utterance intelligibility, compared to both SNR 5 and SNR 0 noisy setups. This highlights that at a highly noisy condition, paraphrases are perceived with considerably different rates of recognition and thus, choosing one lexical realization of a message over its sentential paraphrase is crucial for the intelligibility of the message.

PiN$_{both}$ and PiN$_{either}$. Similar to earlier studies on sentential paraphrases, the outcome of annotating PiN paraphrase pairs highlighted that the notion of semantic equivalence at a sentence level is hard to define. As described in Section 5.2, the paraphrase pair annotation task that exhibited a moderate annotator agreement ($\kappa = 0.42$, $p < 0.05$), resulted in the formation of two subsets of stricter paraphrases: (a) PiN$_{both}$ and (b) PiN$_{either}$.

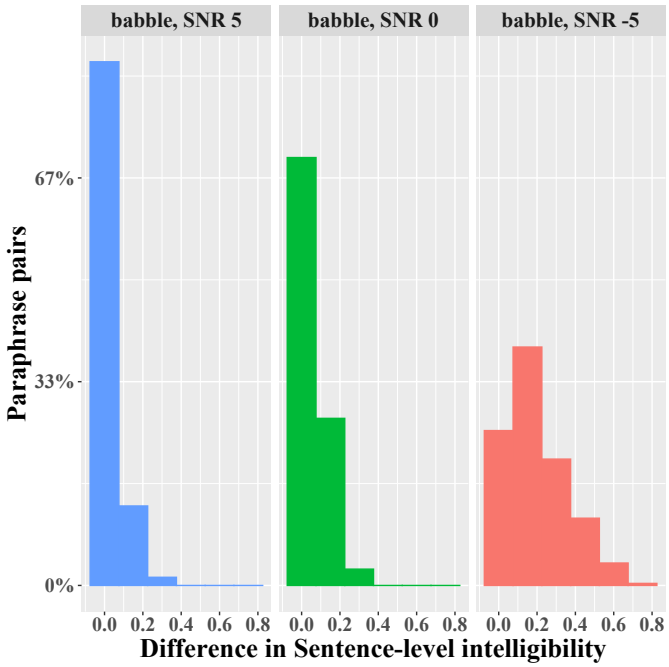


Figure 5.3: Difference in the sentence-level intelligibility of paraphrase pairs in **PiN** dataset. The number of paraphrase pairs that are distinct in their intelligibility significantly increased with an increase in the *babble noise* level (SNR 5 $\longrightarrow$ SNR 0 $\longrightarrow$ SNR -5).

For both subsets, we observed a steady increase in the mean intelligibility difference among paraphrase pairs, starting with 0.03 at SNR 5, to 0.06 ($p < 0.05$) at SNR 0 and 0.19 ($p < 0.05$) at SNR -5. As illustrated in Figures 5.4 and 5.5, the sets of stricter paraphrases also followed the trend similar to PiN dataset: the significant difference in intelligibility between paraphrases increased with an increase in the noise level. This observation also aligns with our earlier finding that the recognition difference between synonyms (*i.e.*, lexical paraphrases) increases with an increase in noise level (Chingacham *et al.*, 2021).

This repeated pattern highlights the significance of the choice of surface realizations (of a meaning) under noisy environments, as it influences the utterance intelligibility which in turn impacts the message comprehension. To assess the performance of an oracle surface-form selector, we labeled utterances in every paraphrase pair as either *more intelligible* or *less intelligible* by comparing their intelligibility scores (*ie.*, Sent-Int). Then, the relative gain in sentence-level intelligibility introduced by the oracle selector is calculated with the assumption that *more intelligible* paraphrase is always selected over its counterpart. This resulted in a relatively low gain in intelligibility at SNR 5 (2%) and SNR 0 (5%), compared to that of **an**

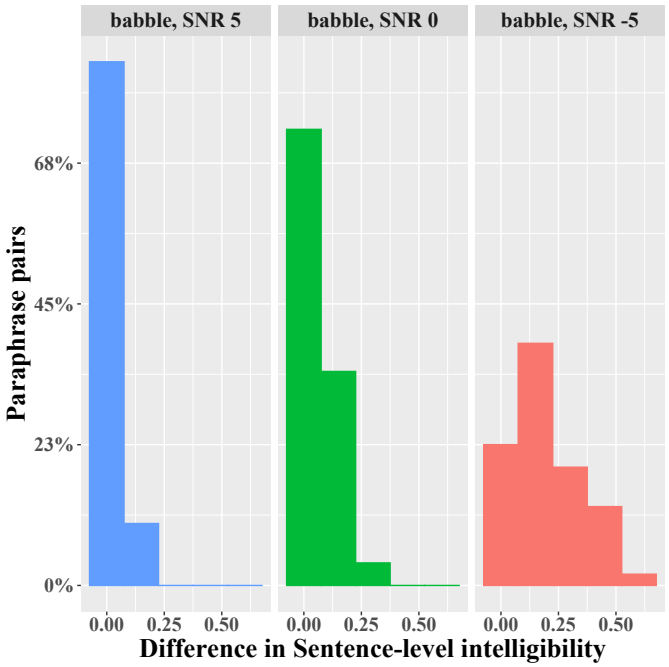


Figure 5.4: Difference in the sentence-level intelligibility of paraphrase pairs in **PiN**$_{both}$ dataset. The number of paraphrase pairs which are distinct in their intelligibility significantly increased with an increase in the *babble noise* level (SNR 5 $\longrightarrow$ SNR 0 $\longrightarrow$ SNR -5).

incredibly high intelligibility-gain of 33% at SNR -5. It leads us to the conclusion of this section that paraphrases can indeed introduce differences in sentence-level intelligibility, suggesting the possibility of improving utterance intelligibility by choosing a noise-robust sentential paraphrase.

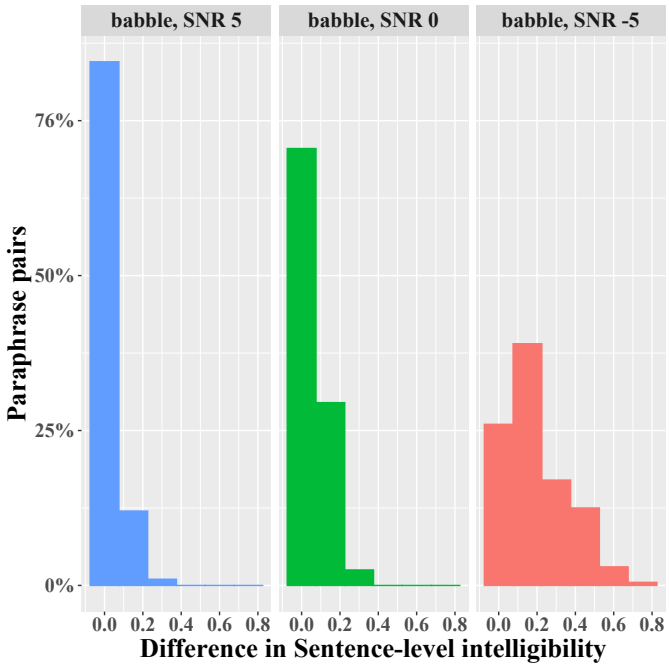


Figure 5.5: Difference in the sentence-level intelligibility of paraphrase pairs in **PiN**$_{either}$ dataset. The number of paraphrase pairs that are distinct in their intelligibility significantly increased with an increase in the *babble noise* level (SNR 5 $\longrightarrow$ SNR 0 $\longrightarrow$ SNR -5).

5.5 EXPLAINING THE INTELLIGIBILITY GAIN VIA PARAPHRASING

Prior work has shown that speech perception in noise is influenced by both acoustic and linguistic characteristics. However, there is little documentation in the literature to explain the features of sentential paraphrases that introduce a gain in intelligibility, in noise. To this end, we conducted modeling experiments to study the impact of paraphrasing on utterance intelligibility (RQ 2). For all experiments in this section, we utilized the implementation of linear regression models in the statistical software R (Version 3.6.1) (R Core Team, 2019). The following three features were used to define the models' dependent variables:

(1) Length of utterance: The length of utterance is an interesting correlate of intelligibility, as previous studies on speech modifications found that humans tend to shorten the utterance length to improve speech perception in noise (Cooke *et al.*, 2014). On the other hand, shorter words were found to be more confusing in noise (Luce and Pisoni, 1998). Similarly, in the last chapter, we observed that among synonym pairs, longer words were better perceived in noise, both with and without

linguistic context (for more details, see Section 4.5). The length of an utterance is represented by the number of phonemes (hereafter referred to as *phLen*).

(2) Linguistic predictability: Kalikow *et al.* (1977) showed that the predictability of a word influences its intelligibility in noise. The surprisal theory in language comprehension also demonstrates that the effort to process a word is inversely proportional to its predictability in context (Hale, 2001). Similarly, earlier studies observed that listeners' perceptual difficulties are influenced by high-level signals like linguistic predictability (Bhandari *et al.*, 2021; Coene *et al.*, 2016) and situational cues (Ward *et al.*, 2017). However, predictability may also lead to false hearing instances (Rogers *et al.*, 2012), where the listener is highly confident of the misheard utterance. It is also interesting to study this feature in environments where the actual context of a word in a sentence is (acoustically) *noisy* and different from the linguistic context. To represent how *hard to predict* an utterance *utt* $(u_0, u_1, u_2...u_t)$ is, we utilized the definition of perplexity, as stated in (5.4). For estimating the likelihood of a token from its preceding context, $p_\theta(u_i|u_{<i})$, a pre-trained dialog response generation model (Zhang *et al.*, 2020) was employed.

$$\text{PPL}(utt) = \exp\{-\frac{1}{t}\sum_{i=1}^{t}\log\, p_\theta(u_i|u_{<i})\} \tag{5.4}$$

Thus, utterances that are *less linguistically predictable are represented with high PPL scores*. Hereafter this feature is referred to as *ppl*.

(3) Speech Intelligibility: Speech Intelligibility (SI) metrics are widely used to perform speech enhancements and noise reductions. The Short-Time Objective Intelligibility (STOI) (Taal *et al.*, 2010) measure is one of the intrusive SI metrics, which requires the clean speech reference to estimate the intelligibility of a noisy speech. The STOI value ranges between -1.0 and 1.0, as it captures the mean correlation between the time-frequency units of the clean and the distorted signal. Higher STOI values indicate better audibility. *STOI* scores were generated using a Python module (Pariente, 2018: accessed July 30, 2020).

Analysis: Before modeling the gain in intelligibility induced by paraphrases, we studied the influence of the above-listed features on *sentence-level intelligibility* in noise. To this end, regression models were built separately for each noise level in the PiN dataset, by considering *Sent-Int* as the response variable. All models were fit to the data, after performing feature scaling with z-score normalization. Similarly, for modeling the *intelligibility-gain* in noise, we considered the *Sent-Int-Gain* as the response variable. We hypothesize that *the observed gain in sentence-level*

intelligibility can be explained by the relative difference in sentence-level features of paired paraphrases. For this purpose, first, we identified the *'more intelligible'* utterance in every paraphrase pair. Then, we calculated the feature difference between paraphrase pairs, with respect to the *'more intelligible'* utterance within each pair. The predictor variables of this model are referred to as *diff.phLen*, *diff.ppl* and *diff.STOI*. In addition to the PiN dataset, the two subsets of the dataset (as mentioned in Table 5.2) were also considered to model the intelligibility gain among stricter paraphrase pairs.

5.5.1 Results and Discussion

In this section, we discussed the outcomes of our modeling experiment, which were conducted to address RQ 2 – why certain sentences are better intelligible than their sentential paraphrases in noise.

Sentence-level intelligibility: Under all three noisy listening conditions, the selected three features exhibited only a weak correlation (< 0.15) with the dependent variable, Sent-Int. We also ensured the independence of features by observing a lower degree of multicollinearity (Variance Inflation Factor < 1.02) under all conditions. By considering the PiN pairs at SNR 5, we found that sentences which are shorter in length (*phLen*: $\hat{\beta} = -0.015$, $p < 0.05$) and linguistically more predictable (*ppl*: $\hat{\beta} = -0.007$, $p < 0.05$) are better perceived (refer Table 5.5.1 for details). However, *STOI* is not significant feature at SNR 5, which highlights the negligible effect of a masker on perception, at a low noise level.

Moreover, we observed a main effect of acoustic cues, at higher noise levels like SNR 0 (*STOI*: $\hat{\beta} = 0.015$, $p < 0.05$) and SNR -5 (*STOI*: $\hat{\beta} = 0.070$, $p < 0.05$), indicating the relevance of noise-robust acoustic cues on speech perception. Our models also showed a main effect of sentence predictability at higher noise levels: SNR 0 (*ppl*: $\hat{\beta} = -0.018$, $p < 0.05$) and SNR -5 (*ppl*: $\hat{\beta} = -0.049$, $p < 0.05$). This observation agrees with the earlier finding that linguistically predictable sentences are more intelligible in fluctuating noise conditions (Schoof and Rosen, 2015). At SNR -5, in addition to *STOI* and *ppl*, the model exhibited a main effect of *phLen* ($\hat{\beta} = -0.031$, $p < 0.05$), indicating better perception in noise with shorter utterances. Overall, models indicate that stronger cues in top-down (linguistics) as well as bottom-up (acoustics) signals, lead to better utterance intelligibility in noise. Next, we model the intelligibility-gain introduced by sentential paraphrases in noise.

Gain in sentence-level intelligibility: With stricter paraphrases in **PiN**$_{both}$ dataset, the model showed no impact of paraphrasing on intelligibility, at a less adverse

	Dependent variable: Sent-Int		
	Babble,SNR 5	Babble,SNR 0	Babble,SNR −5
scale(STOI)	−0.002	0.015***	0.070***
	(0.003)	(0.004)	(0.10)
scale(ppl)	−0.007***	−0.018***	−0.049***
	(0.003)	(0.004)	(0.010)
scale(phLen)	−0.015***	−0.007	−0.031***
	(0.003)	(0.004)	(0.010)
(intercept)	0.968***	0.938***	0.706***
	(0.003)	(0.004)	(0.010)
Observations	300	300	300
R^2	0.096	0.101	0.212
Adjusted R^2	0.087	0.092	0.204
Residual Std. Error (df = 296)	0.052	0.071	0.179
F Statistic (df = 3; 596)	10.49***	11.05***	26.5***
*p<0.1; **p<0.05; ***p<0.01			

Table 5.5: Modeling sentence-level intelligibility (Sent-Int) using linguistic and acoustic features of utterances in the **PiN dataset**. Model coefficients of features (with SEs in brackets) are reported, for all three noise levels.

listening setup (SNR 5). This outcome is expected, as we observed a ceiling effect in sentence-level intelligibility at a low noise level (see Section 5.4.2). However, the model exhibited a main effect of difference in acoustic-cues at both SNR 0 (*diff.STOI*: $\hat{\beta} = 0.014$; $p < 0.05$) and SNR −5 (*diff.STOI*: $\hat{\beta} = 0.066$; $p < 0.01$), indicating: (a) linguistic modifications can introduce utterances with better acoustic cues and (b) the observed intelligibility-gain at both SNR 0 and SNR −5, is mainly driven by noise-robust acoustic cues, which is captured in the *STOI* feature. See Table 5.5.1 for more details.

By modeling the intelligibility-gain among $\mathbf{PiN}_{either}$ pairs at SNR 5, we found that *diff.phLen* ($\hat{\beta} = -0.008$; $p < 0.05$) exhibits, a small but a significant effect on *Sent-Int-Gain* (refer Table 5.5.1). It indicates that paraphrases that are shorter in length than the given sentence, are better perceived in noise. This observation also states the impact of approximation of semantic equivalence on the intelligibility

	Dependent variable: Sent-Int-Gain		
	(SNR 5)	(SNR 0)	(SNR −5)
scale(diff.STOI)	−0.002	0.014**	0.066***
	(0.004)	(0.006)	(0.014)
scale(diff.ppl)	0.002	−0.005	−0.016
	(0.004)	(0.006)	(0.014)
scale(diff.phLen)	0.001	−0.008	−0.011
	(0.004)	(0.006)	(0.014)
(intercept)	0.030***	0.062***	0.196***
	(0.004)	(0.006)	(0.013)
Observations	104	123	105
R^2	0.005	0.058	0.207
Adjusted R^2	−0.025	0.034	0.184
Residual Std. Error	0.038 (df = 100)	0.064 (df = 119)	0.134 (df = 101)
F Statistic	0.157 (df = 3; 100)	2.449* (df = 3; 119)	8.800*** (df = 3; 101)
*p<0.1; **p<0.05; ***p<0.01			

Table 5.6: Modeling the *gain in sentence-level intelligibility* (Sent-Int-Gain) using linguistic and acoustic features of paraphrases in **PiN**$_{both}$ **dataset**. Model coefficients of features (with SEs in brackets) are reported, for all three noise levels.

gain. More precisely, compared to the PiN$_{either}$ subset, PiN$_{both}$ consists of more stricter paraphrases that are closer in semantics, but less different in their surface form. Hence, the approximation in semantic equivalence among the PiN$_{either}$ pairs introduced a main effect of utterance length on intelligibility gain. Similarly, at SNR −5, we observed the main effect of utterance length (*phLen:* $\hat{\beta} = -0.03$; $p < 0.05$), in addition to *STOI* ($\hat{\beta} = 0.04$; $p < 0.05$), representing that the gain in intelligibility being driven by both acoustic cues as well as the utterance length. However at SNR 0, we only found the main effect of STOI ($\hat{\beta} = 0.009$; $p < 0.05$) on intelligibility-gain.

Like the PiN$_{either}$ pairs, the paraphrases in the **PiN** dataset exhibited similar effects on the intelligibility-gain, at SNR 5 and SNR −5 (refer Table 5.5.1). However, at SNR 0, the paraphrase pairs in PiN dataset showed a main effect of *phLen* ($\hat{\beta} = -0.007$, $p < 0.05$) in addition to *STOI* ($\hat{\beta} = 0.007$, $p < 0.05$), once again indicating the significance of paraphrase type on the intelligibility-gain; the intelligibility difference

	Dependent variable: Sent-Int-Gain		
	(SNR 5)	(SNR 0)	(SNR −5)
scale(diff.STOI)	0.0002	0.009^{**}	0.044^{***}
	(0.003)	(0.004)	(0.011)
scale(diff.ppl)	0.002	−0.005	−0.014
	(0.003)	(0.004)	(0.011)
scale(diff.phLen)	-0.008^{**}	−0.006	-0.030^{***}
	(0.003)	(0.004)	(0.011)
(intercept)	0.034^{***}	0.058^{***}	0.197^{***}
	(0.003)	(0.004)	(0.010)
Observations	195	205	196
R^2	0.031	0.033	0.140
Adjusted R^2	0.016	0.018	0.126
Residual Std. Error	0.046 (df = 191)	0.060 (df = 201)	0.146 (df = 192)
F Statistic	2.049 (df = 3; 191)	2.257^{*} (df = 3; 201)	10.382^{***} (df = 3; 192)
*p<0.1; **p<0.05; ***p<0.01			

Table 5.7: Modeling the *gain in sentence-level intelligibility* (Sent-Int-Gain) using linguistic and acoustic features of paraphrases in **PiN**$_{either}$ **dataset**. Model coefficients of features (with SEs in brackets) are reported, for all three noise levels.

among semantically less strict paraphrase pairs are also driven by their difference in sentence length.

Additionally, we noticed that the predictability of paraphrases showed no significant effect on the intelligibility-gain, under all three noise levels. This observation is expected at SNR 5, as most of the utterances were perceived correctly. However, the absence of a predictability effect at higher noise levels indicates that the intelligibility gain is less influenced by the difference in linguistic cues, introduced by the paraphrases in this dataset. This could possibly be due to the limited variations in the linguistic structure of stimuli sentences, which are generated by a paraphrasing model.

Overall, we found that the intelligibility-gain in noise is mainly driven by paraphrases with noise-robust acoustic cues. Additionally, shorter paraphrases also improved intelligibility in noise, however this was mostly observed among para-

	Dependent variable: Sent-Int-Gain		
	(SNR 5)	(SNR 0)	(SNR −5)
scale(diff.STOI)	0.003	0.007	0.037***
	(0.003)	(0.004)	(0.009)
scale(diff.ppl)	0.001	−0.006	−0.013
	(0.003)	(0.004)	(0.008)
scale(diff.phLen)	−0.010***	−0.007	−0.032***
	(0.003)	(0.004)	(0.009)
(intercept)	0.035***	0.059***	0.199***
	(0.003)	(0.004)	(0.008)
Observations	300	300	300
R^2	0.040	0.028	0.123
Adjusted R^2	0.031	0.018	0.114
Residual Std. Error (df = 296)	0.048	0.062	0.146
F Statistic (df = 3; 296)	4.137***	2.874**	13.784***
*p<0.1; **p<0.05; ***p<0.01			

Table 5.8: Modeling the *gain in sentence-level intelligibility* (Sent-Int-Gain) using linguistic and acoustic features of paraphrases in **PiN dataset**. Model coefficients of features (with SEs in brackets) are reported, for all three noise levels.

phrases which are less strict in semantic equivalence. In other words, the additional effect of utterance length among less stricter paraphrases indicates, the trade-off between the intelligibility-gain introduced by paraphrases and the strictness in their semantic equivalence.

5.6 RANKING PARAPHRASE PAIR BASED ON INTELLIGIBILITY

In this final section, we investigate the potential of the aforementioned findings to identify the *more intelligible* utterance in a pair of paraphrases, under a particular listening condition. For this purpose, we considered the task of pairwise ranking to identify the paraphrase with better intelligibility. A pre-trained ranking model could be a potential solution to RQ 3, by utilizing it in spoken dialogue systems to select the linguistic representations that are more robust to the noise in a listening

environment. This enables an SDS to be more adaptive to the listening difficulties of their interlocutors in noisy listening environments.

In order to automatically choose the *more intelligible* utterance in a paraphrase pair, we trained an SVMRank model to perform pairwise ranking for each listening environment. For all our experiments, we utilized the model implementation by Joachims (2006). Every model was trained with 80% of the data and evaluated on the remaining 20%. Training models to rank sentences have already been explored in the past for different text processing tasks such as simplification (Vajjala and Meurers, 2014) and summarization (Madhuri and Kumar, 2019). However, our models differ from earlier ones, both with respect to the objective of ranking as well as the considerably small size of the dataset and feature set. To quantify the performance of the ranking model, we used the percentage of correctly ranked pairs among the test set, referred to as *pairwise ranking accuracy*. All ranking models were trained and evaluated repeatedly with ten different train/test splits of the PiN dataset; their mean values are reported in Table 5.9. We also performed an ablation study to quantify the influence of sentence-level feature(s) on ranking accuracy.

Baselines: For comparison, two baseline models, *uniform* and *majority* were considered. For each pair of paraphrases (s_1, s_2), the predicted pair ranking has three options– s_1 is *'more intelligible'* than s_2, s_2 is *'more intelligible'* than s_1, both s_1 and s_2 are equal in intelligibility. In the case of the *uniform* baseline, an equal probability is given to all possible pair rankings that exist in the training set. However for the *majority* baseline, the model always predicted the class which occurred the most in the training set. At SNR 5 and SNR 0, by sampling from a uniform distribution of all three ranking types, the *uniform* baseline achieved an accuracy of $\sim$ 33%. However at SNR -5, it achieved an accuracy of $\sim$ 50%, as all paraphrase pairs in the train set differed in intelligibility. Meanwhile the *majority* baseline, performed equal to or better than the *uniform* baseline.

Comparing the performance of ranking models with single features, we found that *phLen* is a better feature than *ppl* and *STOI* at all noise levels. For instance, at SNR 5, ranking paraphrase pairs in the PiN dataset solely based on their utterance length achieved a ranking performance of 53%, while ranking based on predictability and acoustic cues performed well only for 39% and 46%, respectively. Although *phLen* achieved better performance than both baselines at SNR 5 and SNR 0, it fails to perform better than the *uniform* baseline when the noise level is high. This indicates the necessity to consider other features for ranking. Including predictability in addition to the utterance length, ranking improved by 5% at both SNR 0 and

Feature(s)	SNR 5	SNR 0	SNR −5
STOI	46.0 +/- 2.5	49.0 +/- 3.7	53.0 +/- 3.1
ppl	39.0 +/- 2.5	52.0 +/- 3.1	55.0 +/- 3.1
phLen	**53.0 +/- 3.7***	59.0 +/- 3.7*	56.0 +/- 3.7
+ ppl	53.0 +/- 3.7*	**64.0 +/- 3.7***	61.0 +/- 3.1*
+ STOI	54.0 +/- 3.7*	60.0 +/- 4.3*	**67.0 +/- 3.7***
majority	43.0 +/- 3.7	48.0 +/- 5.0	46.0 +/- 3.7
uniform	33.0 +/- 3.7	32.0 +/- 3.1	51.0 +/- 2.5

Table 5.9: The intelligibility based pairwise ranking accuracy of models built with the **PiN** dataset. Each score is a mean over ten runs (+/- 95% CI). Scores with * are significantly better than both baselines. Bold-faced scores are the minimal models with considerably better accuracy at each noise level and the best score at SNR −5 is highlighted.

SNR −5. Interestingly, including STOI in addition to *ppl* and *phLen*, further improved the ranking performance by 6% at SNR −5. In other words, this model achieved a relative improvement of 31.37%, in comparison to the *uniform* baseline. It also highlights the earlier observation of a significant effect of noise-robust acoustic cues on the intelligibility-gain among paraphrases. Similarly, we observed this importance of STOI at SNR −5 being repeated for both subsets PiN_{both} and PiN_{either} by achieving a high accuracy of 70% and 66% respectively (as reported in Tables 5.10 and 5.11).

Feature(s)	SNR 5	SNR 0	SNR −5
STOI	33.0 +/- 6.8	53.0 +/- 5.0	52.0 +/- 5.6
ppl	32.0 +/- 6.8	46.0 +/- 5.6	46.0 +/- 7.4
phLen	46.0 +/- 3.7	56.0 +/- 2.5*	61.0 +/- 5.0*
+ ppl	47.0 +/- 3.7	61.0 +/- 3.7*	63.0 +/- 7.4*
+ STOI	44.0 +/- 3.1	**62.0 +/- 5.6***	**70.0 +/- 4.3***
majority	43.0 +/- 6.8	45.0 +/- 5.6	50.0 +/- 6.2
uniform	37.0 +/- 7.4	32.0 +/- 5.0	42.0 +/- 1.5

Table 5.10: The pairwise ranking accuracy of models which are trained to perform intelligibility-based ranking with **PiN**$_{both}$ dataset. Each score is a mean over 10 runs with a 95% confidence interval. Scores with * are statistically significantly better than both baseline models. Bold-faced scores are the minimal feature set with considerably better accuracy at each LE and the best score at SNR −5 is highlighted.

Feature(s)	SNR 5	SNR 0	SNR −5
STOI	45.0 +/- 3.1	51.0 +/- 5.6	51.0 +/- 4.3
ppl	38.0 +/- 5.6	49.0 +/- 3.1	55.0 +/- 6.2
phLen	52.0 +/- 3.7	57.0 +/- 3.7*	56.0 +/- 3.7
+ ppl	53.0 +/- 2.5	60.0 +/- 2.5*	59.0 +/- 4.3
+ STOI	52.0 +/- 3.7	**62.0 +/- 2.5***	**66.0 +/- 3.7***
majority	49.0 +/- 4.3	45.0 +/- 7.4	53.0 +/- 3.7
uniform	32.0 +/- 4.3	35.0 +/- 3.1	53.0 +/- 4.3

Table 5.11: The pairwise ranking accuracy of models which are trained to perform intelligibility-based ranking with **PiN**$_{either}$ dataset. Each score is a mean over 10 runs with a 95% confidence interval. Scores with * are statistically significantly better than both baseline models. Bold-faced scores are the minimal feature set with considerably better accuracy at each LE and the best score at SNR −5 is highlighted.

5.7 CONCLUSION

A majority of existing algorithmic solutions to synthesize noise-robust speech are driven by acoustic modifications. In this work, we explored the possibilities of utilizing linguistic modification – paraphrasing an utterance by modifying its constituting words and sentence structure – to improve better utterance intelligibility in noise. We conducted an extensive list of modeling experiments to first investigate whether the proposed strategy is useful at all, at different levels of babble noise. Our experiments

showed that by replacing a sentence with its noise-robust sentential paraphrase, a relative gain in the intelligibility of about 33% is achievable at highly noisy conditions like babble at SNR -5. Further experiments also demonstrated evidence that the observed intelligibility gain in noise is mainly driven by paraphrases with better acoustic cues. In addition to better acoustic cues, shorter paraphrases also improved intelligibility; however, this was sometimes linked to omissions of some aspects of the utterance meaning. Additionally, the current work introduced a new dataset of its kind – Paraphrases in Noise – which we created to capture the intelligibility differences between sentential paraphrases, in noisy listening conditions. We also developed an intelligibility-aware paraphrase ranking model, which could be further used in a traditional spoken dialogue system to generate noise-adaptive utterances. We believe that current findings provide better resources to further explore the possibility of controlling dialogue generation in SDS, with utterance intelligibility attributes. Figure 5.6 depicts a short summary of this chapter highlighting the proposed solution to improve utterance intelligibility in noise.

Two main limitations of the proposed intelligibility-aware ranking model are: (a) pair-wise ranking is performed, instead of list-wise ranking and (b) the ranking performance is below 70%. To improve the ranking model in terms of its scope and performance, the availability of labeled data is one crucial factor. However, a larger annotated dataset of sentential paraphrases demands further human listening experiment, which is both expensive and time-consuming. One other approach to handle the shortage of annotated data is to utilize automatic metrics, which facilitates an approximate identification of the better intelligible sentence among a list of paraphrases. In the next chapter, we adopt a similar approach to generating a large pseudo-parallel dataset, employing a proxy intelligibility metric.

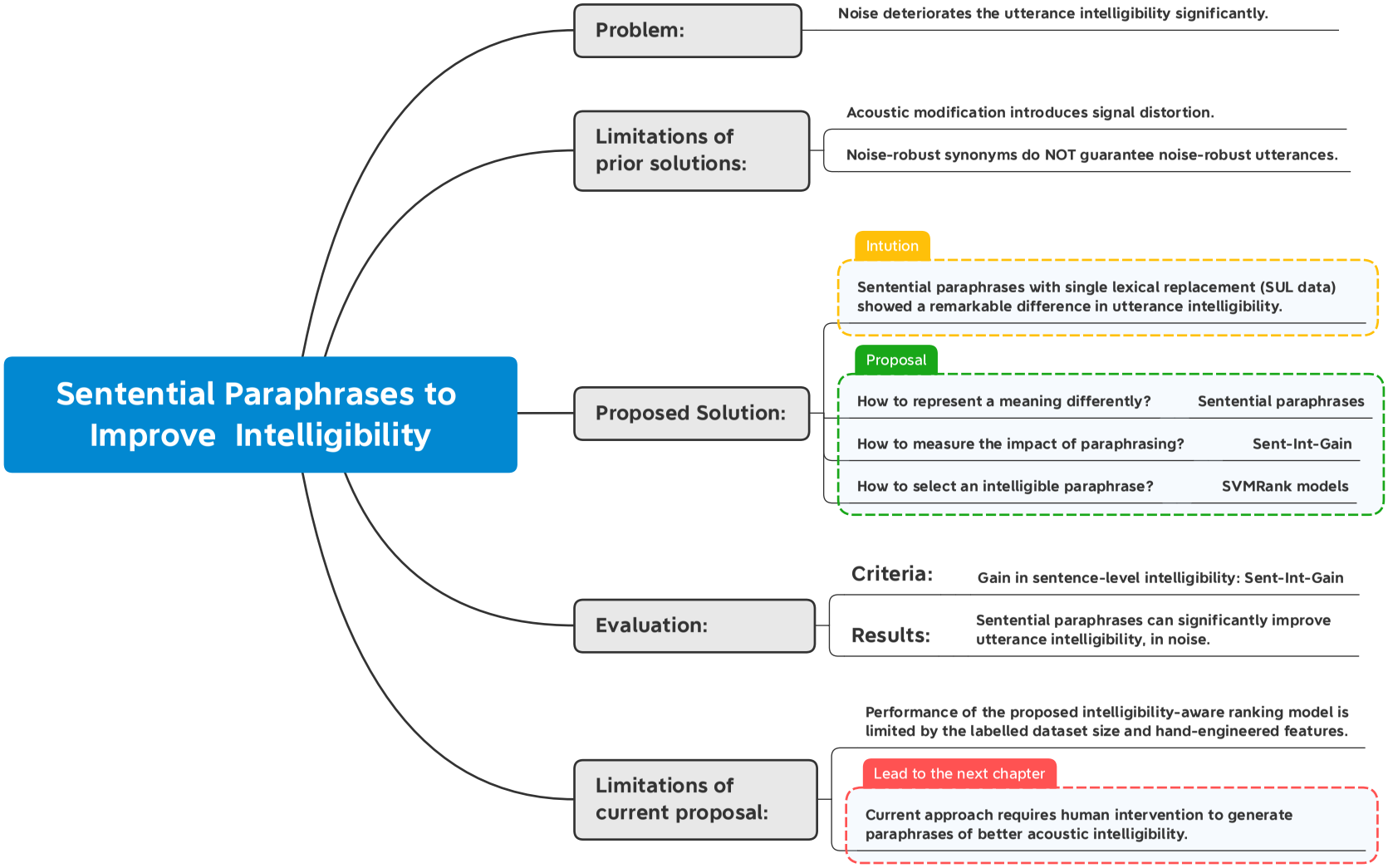


Figure 5.6: An overview of Chapter 5 that investigates the benefit of sentential paraphrases to improve utterance intelligibility in noise.

6 GENERATING ACOUSTICALLY INTELLIGIBLE PARAPHRASES

In the previous chapter, we found that sentential paraphrases can improve speech perception in noise. Replacing utterances with sentential paraphrases that have better acoustic cues, resulted in an overall intelligibility gain of 33%, in a highly noisy environment. However, human annotations were required to identify sentential paraphrases with better intelligibility. Thus, the remaining open problem is the automated generation of acoustically better intelligible paraphrases.

Hence, this chapter focuses on a novel text-generation task called PI-SPiN – paraphrasing to improve speech perception in noise – which is described in Section 6.2. Section 6.3 provides a detailed description of the models and metrics employed in this work. LLMs are the focal point of the current exploration, as they have exhibited incomparable performance on controlled text generation tasks. Section 6.4 is dedicated to showcasing the outcomes of using LLMs for generating acoustically intelligible paraphrases, with/without fine-tuning the model. Finally, in Section 6.5, a newly proposed approach for the PI-SPiN task is explained and evaluated. Overall, our main contributions are as follows:

- We conduct an elaborate study on the evaluation of LLMs on a novel task called PI-SPiN.
- Our results illustrate the weakness of standard textual prompting to control a non-textual attribute – acoustic intelligibility.
- Our proposed approach *prompt-and-select* is an effective solution to generate paraphrases that are more acoustically intelligible.

6.1 INTRODUCTION

Paraphrase generation is the task of rephrasing a sentence while retaining its meaning Bhagat and Hovy (2013). Humans perform paraphrasing in spoken conversations, to enable their listeners to perceive spoken messages as intended Bulyko *et al.* (2005);

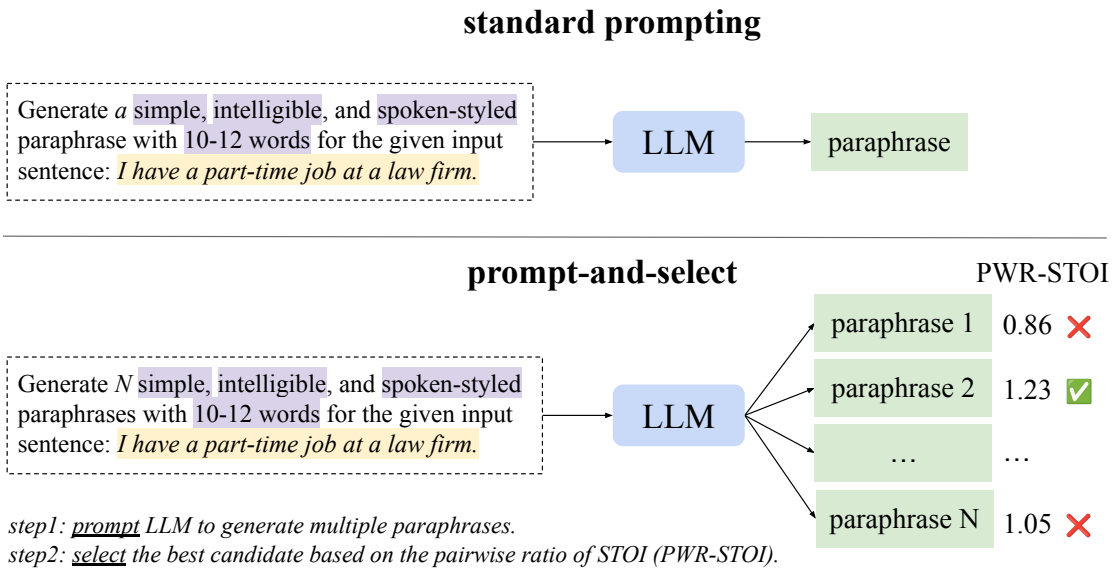


Figure 6.1: A schematic representation of the proposed approach, *prompt-and-select* and the standard prompting approach to generate acoustically intelligible paraphrase in a noisy environment. A speech intelligibility metric, short-time objective intelligibility measure (STOI) is employed to select the paraphrase that is more likely to improve speech perception.

Bohus and Rudnicky (2008). Motivated by human speech production strategies, paraphrasing has also been applied to speech synthesis systems, to enhance the quality, naturalness Nakatsu and White (2006); Boidin *et al.* (2009), and intelligibility of synthetic speech, especially in challenging acoustic conditions Zhang *et al.* (2013). Additionally, our investigations to explain why certain sentences are more intelligible than their paraphrases showed that the observed intelligibility gain in a noisy listening environment is attributed to the rephrasing, which introduces more acoustic cues that survived the masking effect of the noise (as discussed in Chapter 5).

The potential of paraphrasing is however, seldom used to build human-like spoken dialogue systems that are adaptive to human listeners' perception errors in noise, presumably due to the limited investigations to generate paraphrases that are acoustically more intelligible in a noise condition. Prior studies relied on human annotations to identify the ideal paraphrase among a set of candidates Nakatsu and White (2006); Zhang *et al.* (2013); Chingacham *et al.* (2023), with little discussion on generating intelligible paraphrases. This raises the question of *how to generate text that is semantically equivalent to and acoustically more intelligible than the given input sentence, for a noisy environment*. We refer to this task as **Paraphrase to Improve Speech Perception in Noise** (PI-SPiN).

This task is particularly interesting in the context of generative LLMs, which have shown incredible performance in natural language generation (NLG) tasks such as paraphrase generation and dialogue generation Radford *et al.* (2019); Wei *et al.* (2022b); Li *et al.* (2024a). Moreover, recent studies have demonstrated LLMs' capability to control text generation for a wide range of style attributes like sentiment, syntax, formality, and politeness Zhang *et al.* (2023); Sun *et al.* (2023). PI-SPiN differs from those controllable text generation problems, as it aims to generate text that satisfies the desired textual attributes (e.g., semantic equivalence), in addition to the non-textual attribute (*i.e.* acoustic intelligibility), which is hard to describe textually.

To explore the potential of LLMs in PI-SPiN, we proposed to *evaluate LLMs' capability to generate acoustically intelligible paraphrases*, without as well as with, model fine-tuning. Through standard prompting methods like zero-shot learning (ZSL) and in-context learning (ICL), we found that the model was able to capture textual attributes, while consistently struggling to improve acoustic intelligibility. We also observed that increasing the description of the desired non-textual attribute in the prompt only confuses the model, and it may even lead to a deterioration in textual attributes that were achievable otherwise. On the other hand, we found that when an LLM is fine-tuned using the downstream data, it adapts to generate paraphrases, while improving the acoustic intelligibility. This approach resulted in a small but noteworthy improvement of 2.4% in acoustic intelligibility.

Finally, to effectively utilize LLMs for generating acoustically intelligible paraphrases, we propose a simple approach called *prompt-and-select* (PAS), which guides paraphrase generation by introducing the desired non-textual attribute in a post-processing step (see Figure 6.1). It is a two-step process beginning with prompting the LLM to generate multiple candidates and then selecting the best candidate based on acoustic intelligibility, which is hard to capture in textual mode alone. We found that PAS leads to a relative improvement of 8.4% in acoustic intelligibility, compared to the standard prompting approach. By conducting a human evaluation with native English listeners, who have no hearing impairments, we verified that the LLM-generated paraphrases via *prompt-and-select* approach are indeed more intelligible than original sentences, in a listening environment with babble noise at SNR -5 dB.

6.2 PI-SPIN TASK DESCRIPTION

Typically, the paraphrase generation task focuses on generating text that is semantically equivalent to the given input text. However, the PI-SPiN task aims at generating text that is semantically equivalent to, as well as, acoustically more intelligible than the original input text, in an adverse listening condition.

For example, consider the following paraphrase triplet (s_1, s_2, s_3) from the Paraphrases-in-Noise dataset Chingacham *et al.* (2023):

s_1: *"i was raised in a generation we did need all those things."*

s_2: *"we did need all those things when i was a child."*

s_3: *"we did need all those things when i was young."*

s_1 is a sentence retrieved from a spoken corpus, while s_2 and s_3 are outcomes of a paraphrase generation pipeline. Though all sentences are semantically equivalent to each other, they exhibited a significant difference in acoustic intelligibility under noise. More precisely, when these sentences were uttered in a difficult listening condition with babble noise at an SNR of -5 dB, humans perceived s_2 with fewer errors in perception compared to s_1, while s_3 was perceived much worse than s_1. PI-SPiN aims to generate paraphrases (like s_2) that are likely to improve human speech perception in such noisy conditions.

Speech intelligibility in noise is better when sentences are simple Carroll and Ruigendijk (2013), shorter Coene *et al.* (2016), and linguistically more predictive Valentini-Botinhao and Wester (2014). However, the intelligibility of an utterance in noise is not only driven by its underlying text. The perception is also influenced by the acoustic cues that survived the masking effect of the background noise Cooke (2006). Hence, PI-SPiN is a text generation task, that involves both textual attributes like semantic equivalence and a non-textual attribute that captures the noise-robustness of an utterance.

To synthesize the acoustic realization of a sentence, we employed a text-to-speech (TTS) system Shen *et al.* (2018a). Further, to create the noise-distorted signals, the clean audio signals underwent a noise-mixing procedure using an open-sourced tool, *audio-SNR*.[7] The babble noise from the NOISEX-92 dataset Varga and Steeneken (1993) was mixed with clean audio at SNR-5 dB. To determine whether the generated

[7] `https://github.com/Sato-Kunihiko/audio-SNR`

text satisfies the desired outcome, we primarily relied on automatic metrics, which are discussed in detail in the following section.

6.3 EXPERIMENTAL SETUP

Models.

- **ChatGPT**[8], one of the most popular LLMs, is the model that we for all our experiments without fine-tuning. It is a generative model based on the transformer architecture, which is pre-trained on an extremely large and diverse dataset. In comparison to its prior model variations – GPT-2 and GPT-3 – ChatGPT(Ouyang *et al.*, 2022a) is trained on human feedback using reinforcement learning, in addition to the supervised fine-tuning on multiple NLP datasets. ChatGPT is a sibling model of InstructGPT (Ouyang *et al.*, 2022b) of $\approx 1.3B$ model size, which is significantly smaller than the originally proposed GPT model of $175B$ parameters. It is a proprietary product of OpenAI [9]; hence, the model is not publicly available and only its APIs are available (with a predetermined cost) to perform model inference and fine-tuning.

- **Llama 2** is an open-source language model, developed by Meta AI. The model is publicly available in different sizes (ranging from $7B$ to $70B$) and we chose to use the smallest model **llama2-7B-chat**, which is fine-tuned on conversation-styled data. All experiments with model fine-tuning were conducted with Llama 2.

LLMs have shown impressive performance on paraphrase generation with textual style attributes, while its ability on acoustically intelligible paraphrasing remains unclear.

Dataset. The evaluation dataset consists of 300 short sentences, which are spoken in a conversational scenario. The dataset is created by filtering out sentences with 10 to 12 words from the top 1000 lines of the speech corpus, Switchboard Godfrey *et al.* (1992).

[8]Version: `gpt-3.5-turbo`
[9]https://openai.com/blog/chatgpt

Metrics. Human evaluation is the gold standard for most text-generation tasks. However, human evaluation is expensive and time-consuming, which limits the scale of evaluation. Thus, we perform an automatic evaluation of the whole evaluation dataset and a human evaluation of a subset of the dataset. For automatic evaluation, we employed a range of metrics, which determine the semantic equivalence between the input and output texts, as well as, the linguistic and acoustic features that contribute to the acoustic intelligibility in noise.

1. Semantic equivalence.. Semantic Textual Similarity (STS) measures how similar two texts are in terms of their meaning. In the past, several STS scores were proposed Bär *et al.* (2012); Han *et al.* (2013). More recently, Zhang* *et al.* (2020) proposed BERTScore, which has shown encouraging results in correctly identifying the semantic equivalence/distance between two texts. For all our evaluations, the STS score is the BERTScore-f1 calculated using the distilled BERT model Sanh *et al.* (2019). The higher the STS value, the better the semantic equivalence between two texts.

2. Lexical deviation.. Lexical deviation (LD) shows to what extent two texts are similar or different in terms of their surface form. The difference in wording between the two texts is particularly interesting for paraphrase generation. Bandel *et al.* (2022) showed that the deviation in the linguistic forms of paraphrases is one of the critical factors that decides its quality – high-quality paraphrases exhibit high LD, as well as, high STS as they differ lexically, yet maintain the semantics. As defined in Liu and Soh (2022), we used the overlap in lexical tokens of the uncased lemmatized form of two texts to capture the lexical deviation between the input sentence and the model-generated paraphrase. The higher the LD score, the more difference in paraphrased wording.

3. Utterance length.. It is a textual attribute that influences acoustic intelligibility, as it was observed that shorter sentences introduce fewer misperceptions in noise Chingacham *et al.* (2023). Though paraphrases of shorter lengths are more likely to be perceived correctly, shorter paraphrases may risk missing some semantics of the original text. Hence, it is critical to evaluate utterance length along with semantic equivalence. To measure utterance length in terms of phonemes (*i.e.* PhLen), we used a grapheme-to-phoneme model[10] to generate the phonemic transcript of a sentence.

[10] `https://pypi.org/project/g2p-en/`

Further, to compare the length within each input-output pair, the *pairwise ratio of PhLen* is calculated by dividing the length of the model output by that of its input sentence (denoted as *PWR*-PhLen). Thus, when the model-generated text is similar to the input text, *PWR*-PhLen value is close to 1.0, while a value much less than 1.0 reflects that the model-generated text is considerably shorter than the original text.

4. Linguistic predictability.. Several studies in the past have shown that when lexical tokens are more predictable from the context, word misperceptions are less likely to occur Kalikow *et al.* (1977); Uslar *et al.* (2013); Valentini-Botinhao and Wester (2014); Schoof and Rosen (2015); Bhandari *et al.* (2021). Thus, we considered the perplexity (PPL) score determined by a pre-trained language model, GPT-2[11] Radford *et al.* (2019) to estimate the linguistic predictability of a sentence. To compare the linguistic predictability among input and output texts, the *pairwise ratio of the perplexity* is calculated by dividing the PPL of model-generated text by the input sentence PPL (denoted as *PWR*-PPL). Higher PPL scores indicate lesser linguistic predictability. Thus, a *PWR*-PPL value less than 1.0 indicates that the model-generated text is more predictable than the input text.

5. Acoustic Intelligibility.. The acoustic intelligibility of an utterance in a noisy environment is primarily driven by the acoustic cues that survived the energetic masking of the noise – utterances with better noise-robust acoustic cues are better perceived in noise Cooke (2006); Tang and Cooke (2016). We use the Speech Intelligibility (SI) metric, STOI Taal *et al.* (2010), to capture the acoustic intelligibility of an utterance. STOI is a non-textual attribute, as it measures the mean correlation of short-time envelopes between the clean and noisy audio signals of an utterance. The higher the STOI score, the higher the noise-robustness of an utterance. Similar to other pairwise ratios, the *pairwise ratio of STOI* (*PWR*-STOI) is calculated by dividing the STOI of model-generated text by the input text STOI. Thus, PI-SPiN aims at generating paraphrases with *PWR*-STOI values above 1.0 indicating that the model output is acoustically more intelligible than the input sentences.

All pairwise ratios range between 0.0 and $+\infty$, while STS and LD range between 0.0 and 1.0. For the evaluation, we report each of these metrics, averaging across the evaluation dataset.

[11] Version: `distilgpt2`

6.4 EVALUATING LLMS FOR PI-SPIN

In our experiments, an LLM is prompted to generate a paraphrase for each input sentence in the evaluation set with a prompt template: *{prompt prefix} + {input text}*. In the following section, we described three methods that we employed and evaluated for the task.

6.4.1 ZSL: Zero-shot Learning

In this setting, the model is prompted to generate an intelligible paraphrase given an input sentence in a zero-shot manner. As shown in Table 6.1, we investigate three types of prompts, which describe the desired attributes with different granularity: low ($p_{zsl-low}$), medium ($p_{zsl-med}$), and high ($p_{zsl-high}$). With the increasing number of task-specific tokens in the prompt, the task description is more detailed. Prompts are designed by including keywords like '*paraphrase*' and '*intelligible*' that represent the desired outcome. Additionally, a few more tokens like '*10-12 words*' and '*spoken-styled*' were used in the prompt to ensure that the generated paraphrase adheres to the length and style of input sentences. We hypothesize that with additional task-oriented tokens in the prompt, the model will steer the paraphrase generation by optimizing the intelligibility.

Prompt-ID	Prompt
$p_{zsl-low}$	Generate an **intelligible paraphrase** for the following input sentence: `{input text}`
$p_{zsl-med}$	Generate a **simple**, **intelligible**, and **spoken-styled paraphrase** with **10-12 words** for the following input sentence: `{input text}`
$p_{zsl-high}$	**For a noisy listening environment with babble noise at SNR -5**, generate a **simple**, **intelligible**, and **spoken-styled paraphrase** with **10-12 words**, for the following input sentence: `{input text}`

Table 6.1: Three prompts used in the zero-shot learning (ZSL) setup, with an increasing level of detail in the task objective. Bold-faced words are task-specific keywords in the prompt statement.

Results and Analysis. Table 6.2 summarizes the results of all three prompts that we used in standard prompting. We observed that ChatGPT can generate high-quality paraphrases as indicated by high scores for semantic equivalence and lexical deviation (*i.e.* STS and LD). More importantly, we found that the length of paraphrases generated by the prompt $p_{zsl-med}$ (PhLen = 42.08) is considerably shorter than those generated with the prompt $p_{zsl-low}$ (PhLen = 50.67), indicating the effectiveness of additional keywords in $p_{zsl-med}$ to control a textual attribute – length. However, the non-textual attribute, acoustic intelligibility (*i.e.* STOI) of model-generated paraphrases is not significantly different from their corresponding input sentences. Furthermore, paraphrases generated with a detailed task description in $p_{zsl-high}$, also resulted in a similar observation – **LLM struggles to improve the non-textual attribute while controlling textual attributes appropriately**.

Prompt-ID	STS↑	LD↑	*PWR*-PhLen↓	*PWR*-PPL↓	*PWR*-STOI ↑
$p_{zsl-low}$	0.852	0.699	1.343*	1.086	0.992
$p_{zsl-med}$	0.860	0.668	1.119*	1.042	0.991
$p_{zsl-high}$	0.837	0.729	1.250*	1.236*	1.005

Table 6.2: An automatic evaluation of paraphrases generated by different prompts in zero-shot learning. Pairwise ratios (*PWR*) significantly different from 1.0 ($p < 0.05$) are marked with an asterisk (*). They indicate the significant difference between the model-generated output and the input text.

Compared to $p_{zsl-low}$ and $p_{zsl-med}$, $p_{zsl-high}$ resulted in worse performance, indicated by considerably longer output texts despite prompting to control length (*PWR*-PhLen = 1.250) and output texts that are linguistically less predictive (*PWR*-PPL = 1.236). It is also reflected in a higher lexical deviation (LD = 0.723) at the expense of lower textual similarity between input and output (STS = 0.837). To have a deep understanding of its behavior, we conducted a qualitative analysis as shown in Table 6.3. We noticed that the **additional context of the non-textual attribute confused the model in understanding the task objective and resulted in model hallucination**. In sum, using zero-shot learning setup may not effectively elicit the model's ability to generate paraphrases with the intended non-textual attribute, which is beyond the model's comprehension.

Prompt-ID	Sentences	
—	`{input text}`	*I also have a part-time job at a law firm.*
$p_{zsl-low}$	`{output}`	*Additionally, I am employed part-time at a law firm.*
$p_{zsl-med}$	`{output}`	*I work part time at a law firm as well.*
$p_{zsl-high}$	`{output}`	***In a busy environment with loud chatter**, I work part-time at a law office.*
—	`{input text}`	*Is your father's cafe, one that you could cater out of?*
$p_{zsl-low}$	`{output}`	*Can you use your father's cafe as a location for catering purposes?*
$p_{zsl-med}$	`{output}`	*Could you use your father's cafe as a catering venue?*
$p_{zsl-high}$	`{output}`	*Can your dad's café be used for catering **in a noisy setting**?*
—	`{input text}`	*I believe that the cure for polio is already founded.*
$p_{zsl-low}$	`{output}`	*I hold the belief that the remedy for polio has already been discovered.*
$p_{zsl-med}$	`{output}`	*I think the polio cure has already been discovered.*
$p_{zsl-high}$	`{output}`	*I think polio's cure has already been discovered **in a noisy setting.***
—	`{input text}`	*I think you can do that for a little while.*
$p_{zsl-low}$	`{output}`	*I believe you have the capability to do that temporarily.*
$p_{zsl-med}$	`{output}`	*I believe you can manage that temporarily.*
$p_{zsl-high}$	`{output}`	*I believe you can manage it temporarily **amidst the loud chatter.***

Table 6.3: A qualitative analysis of model-generated text, `{output}`, for a given `{input text}` under three standard prompts: $p_{zsl-low}, p_{zsl-med}, p_{zsl-high}$. The prompt $p_{zsl-high}$ generates several tokens that are irrelevant (bold-faced words) to the task objective.

6.4.2 ICL: In-context Learning

Prior research has shown that LLMs can efficiently learn to control text generation with demonstrations and perform better than just providing a task description Brown *et al.* (2020). Thus for the in-context learning (ICL) setup, the input prompt is modified to include a set of exemplars that represent the desired model behavior. In

Prompt-ID	Prompt
p_{icl}	Look at the samples of a sentence and its intelligible paraphrase: 1. *I don't know if you are familiar with that. =>* *I have no idea if you're familiar with that.* 2. *what other long-range goals do you have besides college? =>* *Apart from college, what are your other long-term objectives?* 3. *I don't have access either. Although, I did at one time =>* *In the past, I had access, but currently, I don't.* 4. *Right now I've got it narrowed down to the top four teams. =>* *At this point, I've trimmed my options and picked 4 top teams.* 5. *prohibition didn't stop it and didn't do anything really. =>* *It continued despite the prohibition, which didn't accomplish anything.* Similarly, generate an **intelligible** paraphrase for the input sentence: `{input text}`

Table 6.4: The prompt used for generating intelligible paraphrase in *in-context learning* setup.

other words, to instruct the model to generate acoustically intelligible paraphrases in an ICL setting requires a set of sentences and their corresponding paraphrases that are acoustically more intelligible in a noise condition.

To provide the best in-context demonstrations, we created another set of 300 short sentences from the Switchboard corpus excluding those in the evaluation set. Then, their corresponding paraphrases were generated by prompting ChatGPT with $p_{zsl-med}$. Following speech synthesis and noise mixing with babble noise at SNR -5 dB, we identified the top 5 pairs that exhibited a larger pairwise difference in STOI scores. Further, the sentences within each pair were rearranged in such a way that the second sentence is always better intelligible than its paired paraphrase. Further, the sentences within each demonstration pair were concatenated with a token (eg: '=>') and embedded with $p_{zsl-low}$ for in-context learning. Table 6.4 represents the exact prompt statement (p_{icl}) that we used for the in-context learning.

Results and Analysis. As shown in Table 6.5, the model learned to generate paraphrases, similar to those given as examples. Compared to the zero-shot learning with minimal task description ($p_{zsl-low}$), the model in the ICL setup (p_{icl}) generated texts that are semantically more similar and lexically less divergent from the input sentences. More interestingly, the model also learned to optimize the desired

Prompt-ID	STS↑	LD↑	*PWR*-PhLen↓	*PWR*-PPL↓	*PWR*-STOI ↑
p_{icl}	0.872	0.627	1.250*	0.947	0.997

Table 6.5: An evaluation of the ICL setup. LLM fails to improve acoustic intelligibility (*PWR*-STOI < 1.0), though it learns to capture demonstrated textual attributes like lexical deviation and predictability.

textual attributes like length (*PWR*-PhLen) and linguistic predictability (*PWR*-PPL) of generated paraphrases, even in the absence of prompt tokens to explicitly control those features. Nevertheless, the **demonstrations are still not helpful in controlling the non-textual attribute**. We observed that the acoustic intelligibility scores of output sentences were *not significantly* different from their input sentences (*PWR*-STOI = 0.997). Once again, this shows the inability of the LLM to generate acoustically intelligible paraphrases, even though it captures textual attributes from the given exemplars.

6.4.3 SFT: Supervised Fine-tuning

Not all text-generation task objectives are achievable in zero-shot or in-context learning setups. A few tasks require model fine-tuning with task-specific data (Zhang *et al.*, 2022; Li *et al.*, 2024b). With supervised fine-tuning (SFT), the text generation objective of an LLM is optimized to satisfy the linguistic patterns that are captured in a fine-tuning dataset.

For the current task of generating acoustically intelligible paraphrases, SFT requires a parallel dataset, $D_p = \{(x^1, y^1), (x^2, y^2), (x^3, y^3) \dots (x^{|D_p|}, y^{|D_p|})\}$, which consists of sentences (x^i) and their corresponding paraphrases (y^i) that are acoustically more intelligible than x^i, for a (noisy) listening setup. LLM learns to optimize its parameters, when the model is prompted to generate y^i, for a given x^i. In other words, model parameters are optimized to reduce the overall cross-entropy loss of not generating the desired output text y^i, when the model is prompted to perform PI-SPiN. We used the prompt shown in Table 6.6 for the SFT.

One of the main challenges of fine-tuning large language models is the demand for high-compute resources and high-quality data. However, with recent advances in parameter-efficient fine-tuning (PEFT) like Low-Rank Adaptation (LoRA) (Hu *et al.*, 2022), fine-tuning LLMs is possible with limited compute resources. Chowdhury *et al.* (2022) is one of the recent studies that utilized LoRA to control a textual attribute

Prompt-ID	Prompt
p_{sft}	For the given input text, generate an acoustically better intelligible paraphrase with 10-12 words ###Input: x^i ###Response: y^i ######

Table 6.6: The prompt used for the supervised fine-tuning approach to generate acoustically intelligible paraphrase.

– novelty – while generating paraphrases. However, the approach still demands a large paraphrase corpus, annotated with acoustic intelligibility in different noise conditions. Considering the limitations of time and cost to conduct large-scale perception experiments, we proposed to use a pseudo-parallel dataset (PPD) for the model fine-tuning.

A Pseudo-Parallel Dataset. Unlike the human-annotated parallel dataset, PiN, which is discussed in Section 5.4.2, developing a PPD is significantly more scalable and flexible. We utilize the speech intelligibility metric, STOI, as a proxy intelligibility metric to determine the sentence that is more intelligible (y^i) within each pair of paraphrases. Technically, it is possible to build a pseudo-parallel dataset using any existing dataset of paraphrase pairs like QQP (Chen *et al.*, 2018), MRPC (Dolan and Brockett, 2005), or PPDB 2.0 (Pavlick *et al.*, 2015) Ḣowever, existing paraphrase datasets have limited spoken-styled paraphrase pairs as a majority of them are built based on text corpora.

For developing a PPD, first, we generated a list of short sentences ($10-12$ words) filtered from the spoken corpus SwitchBoard-Dialog Act (SWDA). The whole SWDA corpus consists of $\sim 12K$ short utterances. But, we limited to the first 3380 sentences for building the fine-tuning dataset. For each sentence in the filtered list, we used the ChatGPT[12] model to generate six paraphrases (using the prompt prefix $p_{pas(n=6)}$, which is discussed later in Section 6.5). This resulted in a total of $\sim 24K$ sentences (~ 7 times $3.4K$) and ~ 141960 permutation pairs (ie., $3.4K$ times $7P_2$) in the dataset.

Paraphrase generation was then followed by clean speech synthesis using a TTS model (Shen *et al.*, 2018b) and additive noise mixing with a specific noise condition of babble noise at SNR -5 dB. After noise-mixing, STOI scores were calculated for each

[12]Version: `gpt-3.5-turbo`

utterance in the dataset. Individual STOI scores were then used to calculate the *PWR*-STOI of each paraphrase pair. We ensured that the PPD consists of only pairs where the output paraphrase *is always more intelligible* than the original input sentence, by eliminating all pairs with a *PWR*-STOI below or equal to 1.0. Additionally, the *PWR*-STOI was used to generate a subset of the PPD, by filtering out pairs below a threshold value of 1.1. In other words, the PPD subset consists of paraphrase pairs, that exhibit a significant improvement in acoustic intelligibility.

Results and Analysis. To evaluate the efficiency of model fine-tuning on the task of generating intelligible paraphrases, firstly, we focus on the performance of the base model (m_{base}) – Llama2-7B-Chat. Similar to the outcomes of $p_{zsl-low}$, the (m_{base}) also resulted in generating high-quality paraphrases (*i.e.,* high LD and high STS) with *PWR*-STOI score not significantly different from 1.0. Additionally, we noticed that the generated paraphrases with m_{base} are linguistically less predictive than the input sentences (*PWR*-PPL = 1.699), indicating the linguistic style difference between the model input and output sentences. Table 6.7 provides a consolidated view of the base model as well as the three fine-tuned models (m_{PiN}, m_{PPD}, and $m_{\text{PPD}_{1.1}}$) developed using different subsets of datasets, PiN and PPD.

Model	#pairs	STS↑	LD↑	*PWR*-PhLen↓	*PWR*-PPL↓	*PWR*-STOI↑
m_{base}	-	0.845	0.690	1.121*	1.699*	1.002
m_{PiN}	300	0.915	0.357	0.950*	0.922	1.003
m_{PPD}	10K	0.888	0.516	1.050*	0.972	1.013*
$m_{\text{PPD}_{1.1}}$	4K	0.879	0.553	1.049*	1.025	**1.026***

Table 6.7: An automatic evaluation of LLM fine-tuning with human-annotated dataset (PiN) and a pseudo-parallel dataset (PPD). Pairwise ratios (*PWR*) significantly different from 1.0 ($p < 0.05$) are marked with an asterisk (*) and the best *PWR*-STOI is bold-faced.

m_{base} *vs.* m_{PiN}. m_{PiN} is created by fine-tuning m_{base} on the $\text{PiN}^{\text{SNR-5}}$ dataset, which consists of 300 paraphrase pairs annotated with sentence-level intelligibility (Sent-Int) scores. Thus, m_{PiN} is a model fine-tuned on a small set of human-annotated data. The fine-tuned model m_{PiN} is particularly interesting as it demonstrates the potential of model fine-tuning to capture the implicit linguistic characteristics of the fine-tuning dataset. For instance, $\text{PiN}^{\text{SNR-5}}$ dataset consists of paraphrase pairs that are less lexically divergent (LD = 0.43); hence, the model fine-tuned on this

dataset learned to generate paraphrases with less lexical deviation (LD = 0.357), compared to those generated by m_{base} (LD = 0.690). Similarly, fine-tuning has also guided the model m_{PiN} to generate text of shorter length (*PWR*-PhLen = 0.922), imitating the PiN dataset. However, fine-tuning with $\text{PiN}^{\text{SNR-5}}$ subset does not alter the non-textual attribute – STOI – even when the fine-tuning data demonstrates a significant difference in STOI scores between input and output sentences (*PWR*-STOI = 1.035). This highlights that fine-tuning with a small task-specific dataset did not help the LLM to control the non-textual attribute – acoustic intelligibility.

m_{PiN} and m_{PPD}. In comparison to m_{PiN}, the output text generated by m_{PPD} exhibits higher LD and slightly lower STS. However, m_{PPD}-generated paraphrases have a better semantic equivalence with input sentence, compared to the base model m_{base}. Interestingly, we found that both fine-tuned models, m_{PPD} and $m_{\text{PPD}_{1.1}}$, learned to generate text that is better intelligible than input sentences, as indicated by the *PWR*-STOI scores significantly above 1.0 ($p < 0.05$). Table 6.7 also demonstrates that models fine-tuned on the PPD generate more lexically diverse paraphrases than those by the m_{PiN} model. In comparison to the fine-tuned model m_{PPD}, paraphrases generated by the model $m_{\text{PPD}_{1.1}}$ exhibited a higher *PWR*-STOI on the evaluation set, indicating the benefit of LLM fine-tuning with samples that have a strong indication of the desired attribute – acoustic intelligibility. Overall, the findings from this experiment highlight the ability of LLM to learn to generate acoustically intelligible paraphrases, given fine-tuning data of sufficient size and quality. However, compared to the base models, the best fine-tuned model only resulted in a relative improvement of 2.4% in *PWR*-STOI.

6.5 PAS: PROMPT-AND-SELECT

In this section, we discuss our proposed post-processing approach to generate text that satisfies both textual and non-textual attributes.

Prior studies on dialogue generation Boidin *et al.* (2009); Nakatsu and White (2006); Weston *et al.* (2018) have demonstrated the utility of a simple yet effective pipeline of controlling text generation in two steps: first generating a candidate set of dialogues, and then selecting the best candidate based on the task requirement. Similarly, we proposed to decompose the current task into a two-step process: (1) **prompt** the LLM to generate multiple output texts that are semantically equivalent

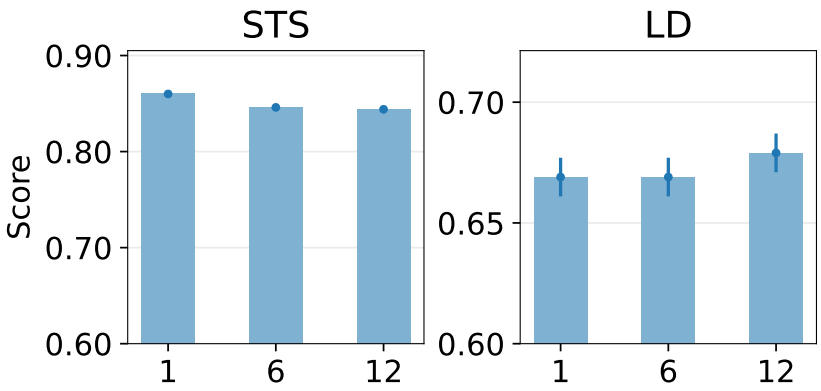


Figure 6.2: An automatic evaluation of the **quality of paraphrases** generated through standard prompting ($n = 1$) and the proposed prompt-and-select ($n > 1$) approach. n is the number of candidates generated in the first prompting step (marked on x-axis). The mean scores are reported with error bars (95% confidence interval). Increasing n, results in generating paraphrases with high lexical deviation and less semantic equivalence.

to the input text and (2) **select** the best candidate based on the acoustic intelligibility.

Our approach is similar to the *prompt-and-rerank* method proposed in Suzgun *et al.* (2022). However, our approach deviates from theirs mainly in two ways: (1) instead of using beam search at the decoding phase, we propose to utilize the potential of an LLM to generate multiple (n) candidates that exhibit the desired textual attributes and (2) the best candidate selection is based on a metric (*i.e. PWR*-STOI) that represents a non-textual attribute, which is not considered in prior studies.

For the first step of paraphrase generation, we perform zero-shot prompting with an appropriate task description, $p_{zsl-med}$. Thus, $p_{zsl-med}$ is the prompt that generates exactly one candidate and involves no selection; it is also referred to as $p_{pas(n=1)}$. However, to generate multiple paraphrases (eg: $n = 6$), the prompt statement can be simply modified to include the n value, as shown below

- *Generate **6** simple, intelligible, and spoken-styled paraphrases with 10-12 words for the given input sentence: {`input text`}*

Following the creation of the candidate set, STOI scores are calculated for all model-generated text as well as the input text, by first synthesizing the clean utterances and then mixing babble noise at SNR -5 dB. Finally, the candidate with the highest *PWR*-STOI is selected as the model output.

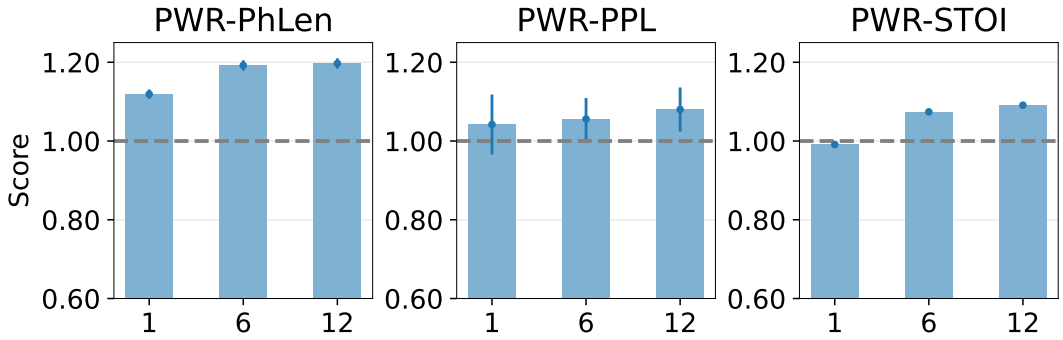


Figure 6.3: An automatic evaluation of the **pair-wise ratios of paraphrases** generated through standard prompting ($n = 1$) and the proposed prompt-and-select ($n > 1$) approach. n is the number of candidates generated in the first prompting step (marked on x-axis). The mean scores are reported with error bars (95% confidence interval). The reference line marks when the input text feature is the same as the output text feature. Increasing n improves the pairwise ratio of acoustic intelligibility (*PWR*-STOI).

Results and Analysis. We begin our analysis by comparing the results of standard prompting ($n = 1$) with the PAS approach, involving 6 candidates ($n = 6$). As shown in Figure 6.2, PAS showcased a high quality of paraphrase generation as indicated by high STS and high LD, similar to the standard prompting setup. Similarly, Figure 6.3 illustrates that other textual attributes like linguistic predictability (*PWR*-PPL = 1.056) and utterance length (*PWR*-PhLen = 1.192) of the PAS approach resulted in similar outcomes of the standard prompting method – output texts are a bit longer than input texts, while their linguistic predictability scores are similar. Importantly, compared to the standard prompting, the prompt-and-select approach yielded a noticeably high *PWR*-STOI ($\mu = 1.074$, $p < 0.05$), which is significantly above 1.0. This indicates that the model-generated text is considerably more intelligible than their corresponding input sentences in the given noise condition. We can see more clearly from Figure 6.3 that PAS ($n = 6$) leads to a relative improvement of 8.4% in *PWR*-STOI compared to the standard prompting ($n = 1$). Our findings suggest that **introducing the desired non-textual attribute in a post-processing step is a potential framework to generate desired text with multi-modal behavior**.

This raises a follow-up question of whether generating more candidates in the first step further improves the overall *PWR*-STOI of generated paraphrases. To this end, we modify the number of candidates (n) in the prompt statement to double the candidate pool size. We found that by increasing the candidate set, there is an improvement in acoustic intelligibility. However, when n is increased from 6 to 12,

there was only a limited improvement of 1.6% in *PWR*-STOI. On the other hand, we observed that textual attributes like linguistic predictability and lexical deviation are not significantly different under varying n values.

Interestingly, the pair-wise ratio of sentence length slightly increased, with more choices in the candidate selection; however, the overall *PWR*-PhLen in this approach is still below the standard prompting setup with no tokens to control length ($p_{zsl-low}$). Increasing n from 6 to 12 slightly reduced the overall semantic equivalence between the model input and output paraphrase. This indicates that the choice of n introduces a trade-off between the improvement in acoustic intelligibility (*PWR*-STOI) and the overall semantic equivalence (STS) and one has to choose n considering this trade-off between the gain in non-textual attribute and the need for semantic equivalence.

Subset	STS↑	LD↑	*PWR*-PhLen ↓	*PWR*-PPL ↓	*PWR*-STOI ↑	*PWR*-Sent-Int ↑
top_{30}	0.831	0.737	1.189*	1.428	1.22*	**1.70***
random_{30}	0.848	0.683	1.157*	1.314	1.07*	1.06

Table 6.8: The automatic and human evaluation of text generated with $p_{pas(n=6)}$. Evaluation on two subsets: top 30 pairs with highest *PWR*-STOI (top_{30}) and randomly selected 30 pairs (random_{30}). *PWR*-Sent-Int captures the pairwise ratio of human speech perception in noise. * marks values significantly above 1.0 ($p < 0.05$).

6.5.1 Human Evaluation

In addition to the evaluation with automatic metrics, we also conducted a human evaluation to verify whether the model output in the PAS setup (using $p_{pas(n=6)}$) is indeed more intelligible than their corresponding input sentences. For the human perception experiment, we created two subsets of the evaluation dataset of 300 pairs: random_{30} and top_{30}. random_{30} is a set of 30 pairs randomly selected from the evaluation dataset, while top_{30} is the top 30 input-output pairs that exhibited the larger improvements in STOI scores.

We followed the experiment design of our previous study Chingacham *et al.* (2023) to capture the human speech perception of an utterance in a (noisy) listening setup. After synthesizing the noisy utterances of each sentence using a TTS Shen *et al.* (2018a) and a noise-mixing tool (audio-SNR), participants were asked to listen and transcribe each sentence. Every utterance in the dataset was listened to by six different participants. For each listening instance, the edit distance between the

phonemic transcriptions of the actual and transcribed text is measured to determine the rate of correct recognition. Furthermore, the sentence-level intelligibility (Sent-Int) of each utterance is calculated by averaging the correct recognition rates exhibited by the six listeners.

The perception experiment was conducted with 24 native English listeners with no hearing impairments (14 females and 10 males; average age = 30.71). After data collection, we calculated the pairwise ratio of sentence-level intelligibility (*PWR*-Sent-Int) by dividing the Sent-Int scores of the output paraphrase by their corresponding input sentence. A mean score of *PWR*-Sent-Int significantly above 1.0 indicates that the model-generated paraphrase is significantly more intelligible than the input sentence, in a given listening condition.

Results and Analysis. As illustrated in Table 6.8, top_{30} items signify that the model-output paraphrases have considerably improved the human perception in a noisy listening condition. We observed that the overall human speech perception of model-output paraphrases (Sent-Int = 0.66) was considerably higher than the input sentences (Sent-Int = 0.47), introducing a 40% **relative gain in the overall intelligibility**. This is also reflected in the *PWR*-Sent-Int score that is significantly above 1.0.

We observed the *PWR*-Sent-Int of random_{30} is not significantly above 1.0, even though the *PWR*-STOI is significantly above 1.0. With further analysis of two subsets, we found that the mean STOI of input sentences in top_{30} ($\mu = 0.507$) is significantly less than random_{30} ($\mu = 0.561$). This means that random_{30} consists of sentences, which are already better intelligible in noise. Also, we observed a strong negative correlation ($r = -0.442, p < 0.001$) between the STOI of input sentences and the gain in intelligibility (*PWR*-Sent-Int), which highlighted the limited benefits of paraphrasing input sentences in random_{30}. However, top_{30} consists of all input sentences, which are more likely to benefit from paraphrasing in noisy listening conditions and they reflected the same in the human evaluation. We conclude with the observation PAS is a simple yet effective solution to alleviate the struggles of LLM to generate text with textual and non-textual attributes, without model fine-tuning.

6.6 CONCLUSION

In this work, we evaluate LLMs on acoustically intelligible paraphrase generation for better human speech perception in noise. Our results demonstrate the limitations of LLMs in controlling text generation with a non-textual attribute – acoustic intelligibility. To alleviate the struggles of LLMs in generating text that satisfies both textual and non-textual attributes, we proposed a simple yet effective approach called *prompt-and-select*. With human evaluation, we found that when the original utterances are highly prone to misperceptions in noise, prompt-and-select can introduce 40% of relative improvement in human perception. We hope the findings of this work inspire further explorations to control LLMs' text generation with different real-world context cues, thereby building more human-like agents. An overview of this chapter is represented in the following diagram. .

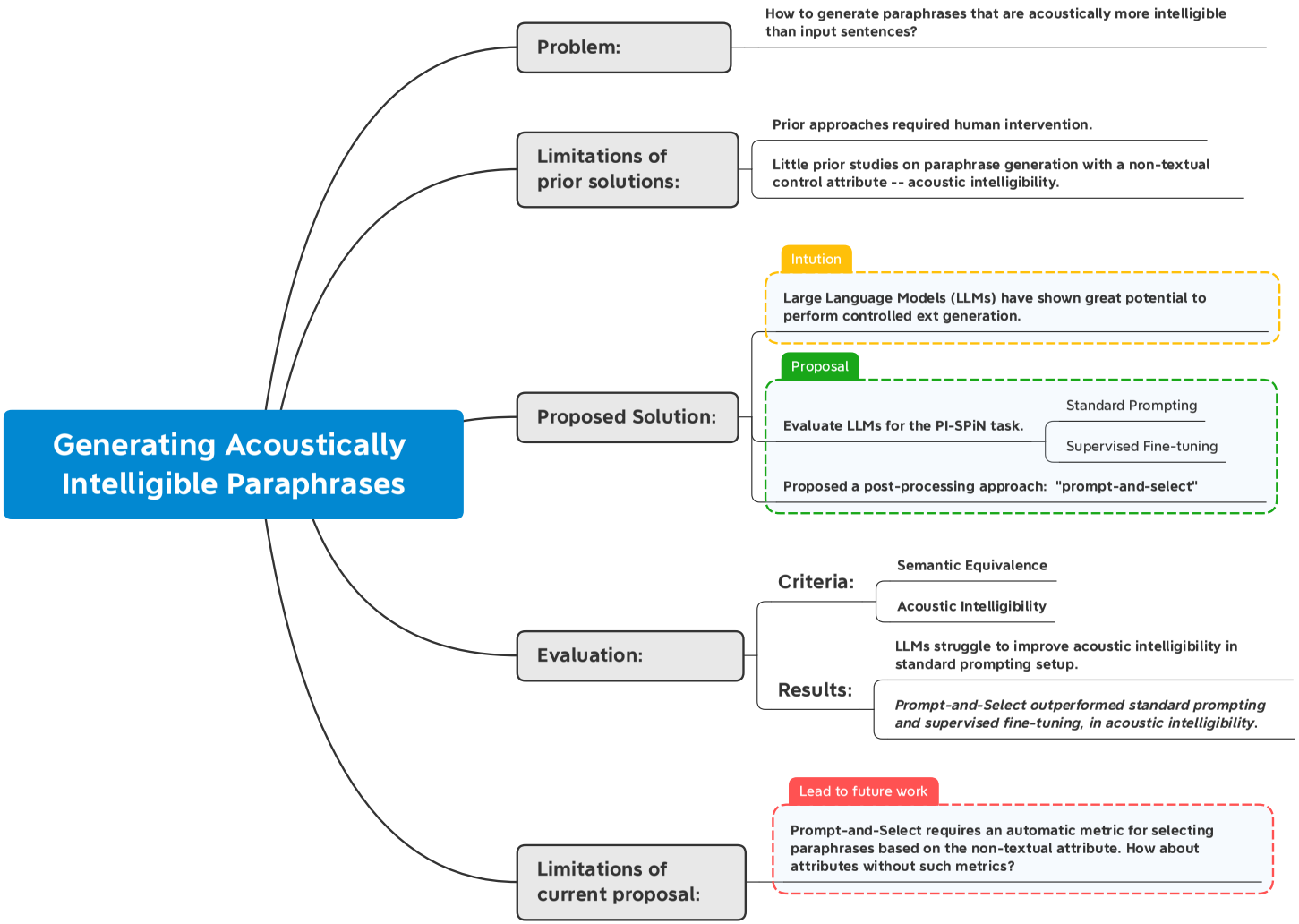


Figure 6.4: An overview of Chapter 6 that investigates the possibility of generating acoustically intelligible paraphrases using modern large language models.

7 CONCLUSION

The final chapter summarizes the main contributions of this thesis. In addition, it discusses the potential challenges and limitations of the current approach. The chapter concludes the dissertation by outlining some of the implications and future directions of this work.

7.1 SUMMARY OF CONTRIBUTIONS

This thesis contributes towards a novel approach of using paraphrases to improve the intelligibility of synthetic speech in noise, by conducting a series of empirical analyses, starting with investigations on whether and to what extent replacing words with lexical paraphrases improves word intelligibility in noise (Chapter 4), moving to more generic sentential paraphrases with rigorous analysis to understand why they can improve speech perception in noise (Chapter 5), and finally proposing new frameworks to generate acoustically intelligible paraphrases (Chapter 6) for building noise-adaptive spoken dialogue systems. Our contributions address the research objectives presented in Section 1.2.

Lexical paraphrases mitigate word misperceptions in noise. Existing approaches to enhance synthetic speech in noise are primarily focused on modifying acoustic characteristics. One of the main limitations of using acoustic modification is the signal distortions (as noted in several studies such as Langner and Black (2005); Anumanchipalli *et al.* (2010); Cooke and Lecumberri (2012); Valentini-Botinhao and Wester (2014)), which leads to compromised quality and naturalness of synthesized speech. In Chapter 4, we made a critical contribution to this problem, showcasing the potential of using an alternate approach – replacing words with lexical paraphrases – to improve word recognition in noise. We conducted listening experiments with human subjects (native listeners with no hearing impairments) and created new datasets of selected English synonyms and their corresponding perceptions in noisy environments. Based on human experiments, we observed that the selection of

a synonym (either in isolation or within a context) to represent a meaning, has a profound influence on mishearing, and choosing the more noise-robust synonym can introduce a relative gain in perception of about 37%. With further analysis of why certain words are more intelligible than their synonyms, we noted that the intelligibility gain is mainly driven by the synonyms with better noise-robust acoustic cues. The results of this experiment promised new avenues to improve synthetic speech intelligibility with linguistic modification, which introduces no signal distortions.

Paraphrases improve speech perception in noise. With such promising results, our logical next step was to extend the study to sentential paraphrases in noise. The outcomes of this study are imperative to understand the practical implications of using paraphrases in noise, such as noise-adaptive spoken dialogue systems. Linguistic modification purely based on synonyms and word intelligibility is not sufficient for using this approach in naturalistic conversational environments, in which the perception of the whole utterance is more important than individual word intelligibility. In addition, estimating sentence intelligibility is not trivial, as several high-level factors like the listener's world knowledge and language proficiency could impact speech perception, especially in noisy environments, where listeners utilize high-level linguistic cues to compensate for the degraded, noisy acoustic signals.

We conducted a large-scale perception experiment to create a new dataset of sentential paraphrases of selected utterances from a spoken corpus, annotated with corresponding human perception in noise. We ensured to include paraphrases of both syntactic and lexical variations, for better generalization of this approach. The experiment results showed that paraphrasing can significantly boost speech perception in highly noisy conditions. We built computational models to analyze the linguistic and acoustic characteristics of paraphrases that contribute to the observed intelligibility gain. Our models showed that the improved intelligibility is primarily attributed to sentential paraphrases that better survived the energetic masking. The results suggested the potential of using sentential paraphrases to build a noise-adaptive spoken dialogue system (SDS), by modifying the text-to-speech input sentence to its noise-robust paraphrase, for a given noise environment. We argue such adaptive systems can be built by performing either paraphrase selection or paraphrase generation, focusing on the non-textual attribute – acoustic intelligibility.

Generate acoustically intelligible paraphrases. With modern-day Large Language models (LLMs), rephrasing a sentence with fine-grained textual controls like politeness, sentiment, and language, results in high-quality output text, compared to other style transfer models. However, controlling the LLM text generation with a non-textual attribute, which is hard to describe in text is still an under-explored problem. The findings of this study discussed the limits as well as the potential of using LLMs for a text generation task that involves both textual control attributes (like semantic equivalence) and non-textual control attributes (like acoustic intelligibility). We analyzed the model's capability in learning the desired attributes without any fine-tuning as well as with fine-tuning. Our evaluation results showed that LLMs, in zero-shot and in-context learning setups, struggle to capture the desired non-textual attribute, even though it is efficient in controlling textual attributes.

We further showed that fine-tuning an LLM with a parallel dataset of paraphrases with different acoustic intelligibility has guided the model to generate paraphrases that are acoustically more intelligible. The resource-intensive model fine-tuning is not a desirable approach, considering the need to adapt the model for each new listening condition. We argued that acoustically intelligible paraphrases can be generated with LLMs, by separating the desired textual and non-textual attributes, in a zero-shot learning setup. Our proposed zero-shot learning framework – *prompt-and-select* – led to generating text that is semantically equivalent to and acoustically more intelligible than input sentences, outperforming fine-tuned models. The study highlights the existing gap in LLMs' capability to control text generation with non-textual attributes and suggests more investigations in the future to improve the multi-modality of LLMs.

7.2 LIMITATIONS AND CHALLENGES

In this work, we conducted rigorous investigations to showcase the under-explored potential of paraphrasing to improve synthetic speech intelligibility in noisy conditions. Nonetheless, the results must be interpreted with caution and the following limitations must be considered.

Human misperception data is collected through an online experiment. Conducting perception experiments on an online platform, rather than in a lab environment, has significantly helped in scaling up the experiment. However, an online experiment suffers from limited control over the environment of the experiment - the actual

environment of the listening experiment is not identical across all participants. Such differences may have impacted some listening instances more negatively, leading to misperceptions attributed to the listener's environment that is outside the listening experiment. In an attempt to reduce such variations, we instructed participants to choose a quiet environment for the listening environment and restricted their participation only through desktop systems.

The time and cost involved in collecting perception data is one of the main challenges in expanding the scope of this work. Although, the current listening setup – native English listeners in a listening environment with babble noise in the background – covers a large group of listeners, the findings of this work need to be evaluated in different listening conditions, for better generalization. Also, augmenting human perception data with simulated listeners like Automatic Speech Recognition (ASR) systems is limited at the moment, as much of the recent research on ASRs is focused on improving their noise-robustness, with little explorations of using ASRs to replicate human misperceptions in noisy environments.

Finally, the models that we employed to explain perception differences among paraphrases used few hand-engineered features, with no non-linearity. This modeling choice was taken to reduce the potential over-fitting issue that could arise in modeling with limited data. It is possible to improve the current model capacity with high-dimensional features, but it incurs a significant cost of creating a larger human (mis)perception data.

7.3 IMPLICATIONS AND FUTURE WORK

This work raises several important questions for future work. In this last section, we discussed some of this work's implications and future directions.

Phoneme-based paraphrase generation. In Chapter 4, we showed that recognition differences between synonyms are primarily attributed to their difference in noise-robust acoustic cues. The findings of this study validate our hypothesis that the difference in the constituting sounds (*ie.*, phonemes) of paraphrases can introduce an intelligibility difference among paraphrases in noisy conditions. However, it is not clear whether any specific set of phonemes or sequences of phonemes has guided such intelligibility differences. Although prior perception studies have demonstrated how phonemes influence misperceptions in noisy environments (compared to vowels,

consonants are more misrecognized in noise Cutler *et al.* (2004); fricatives are hard to recognize in white noise compared to fluctuating noise conditions (Phatak *et al.*, 2008)), their perception data were mainly based on mono-syllabic/nonsensical word recognition experiments. However, the noise influence on underlying sounds is difficult to analyze among conversational data, which consists of linguistic tokens and short sentences that are generally longer than a syllable or a triphone.

This raises a follow-up question of to what extent the perception of conversational speech in noise is influenced by *high-level cues* like linguistic predictability and situational contexts and *low-level signals* like noise-robust acoustic/phonetic cues. A detailed analysis of their interaction could shed more light on understanding human strategies to mitigate misperceptions in difficult listening environments. More investigation must be done to verify whether the strength of noise-robustness (*ie.*, STOI) is predictable from the phonemic sequence of a word or a sentence. We believe that building a model to predict the noise-robustness of a spoken utterance can be further extended to build paraphrasing models that can employ fine-grained controls on generating paraphrases with a desired set of phonemes. One can imagine that such controlled paraphrasing models would be useful for personalizing spoken dialogue systems to accommodate individual hearing difficulties.

Paraphrase selection based on meaning loss. In Chapters 4 and 5, we used an annotation task – *listen-and-transcribe* – to measure the perception loss of an actual (spoken) utterance and comparing with perceived utterances that are transcribed by listeners. Then, the perception loss is compared among a set of paraphrases to identify the *ideal paraphrase*, which has the minimum perception loss in a noisy environment. However, not all perception losses result in comprehension errors. As humans employ world knowledge and commonsense reasoning while comprehending speech, they auto-correct some of the misperceptions using linguistic and situational cues, limiting the occurrences of misunderstanding. Thus, one could replace perception loss with comprehension error, to select the ideal phrasing of a message that is less likely to introduce any wrong interpretations of a message in noise. Replacing perception error with comprehension error would also be useful in systematically comparing the benefits of linguistic modification with acoustic modification of utterances in noise.

Text generation controlled by non-textual attributes. In Chapter 6, we observed that pre-trained LLMs, without fine-tuning or a post-processing step, fail to para-

phrase with the desired non-textual attribute – acoustic intelligibility. However, such non-textual attributes are highly critical for conversations in real-world scenarios. For instance, attributes like cultural norms and situational demands influence how humans converse in daily life – using a term of endearment like "*Sweetheart*" or "*Sweetie*" to a stranger is accepted as polite manners in some parts of American culture, while it is considered inappropriate and derogatory in most of the East Asian cultures. In other words, human language processing is guided by the knowledge that an individual has learned and experienced through different life experiences; it is not restricted to just what one has read. Considering the prevalence of LLMs in modern natural language processing (NLP) pipelines, it raises an important question, to what extent the language generation be controlled by those non-textual attributes, which are hard to describe in a few sets of words or demonstrable examples? Enriching the conversational context with modalities outside textual data could be potential research in this direction.

Expanding the scope of the study. The current study has showcased that paraphrasing is an effective strategy to improve the intelligibility of synthetic speech, perceived by native English listeners (with no hearing impairments) in an adverse listening environment with babble noise. Although the findings in this study are promising evidence for developing a speech enhancement approach without any signal distortions, it would be interesting to expand the scope of this work to other languages, noise types, and listener groups. For instance, a speech enhancement approach without signal distortions maintains the quality/naturalness of synthetic speech and it would be particularly useful for noise-sensitive listener groups such as non-natives (Cutler *et al.*, 2004) and individuals with hearing aids, who suffer from a severe drop in perception, in the presence of background noise. Similarly, investigating the speech modification strategies among bilinguals (who could switch from one language to another) in noisy environments would be another interesting study to expand our knowledge of speech intelligibility enhancements with linguistic modification.

LIST OF FIGURES

LIST OF TABLES

BIBLIOGRAPHY

R. Albert Felty, A. Buchwald, T. M. Gruenenfelder, and D. B. Pisoni (2013). Misperceptions of spoken words: Data from a random sample of American English words, *The JASA*, vol. 134(1), pp. 572–585. Cited on pages 12 and 19.

G. K. Anumanchipalli, P. K. Muthukumar, U. Nallasamy, A. Parlikar, A. W. Black, and B. Langner (2010). Improving speech synthesis for noisy environments, in *Seventh ISCA Workshop on Speech Synthesis 2010*. Cited on pages 1 and 121.

E. Bandel, R. Aharonov, M. Shmueli-Scheuer, I. Shnayderman, N. Slonim, and L. E. Dor (2022). Quality Controlled Paraphrase Generation, in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2022*. Cited on page 104.

D. Bär, C. Biemann, I. Gurevych, and T. Zesch (2012). UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures, in **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012) 2012*. Cited on page 104.

Y. Bengio, R. Ducharme, and P. Vincent (2000). A neural probabilistic language model, *Advances in neural information processing systems*, vol. 13. Cited on page 29.

R. S. Berndt, J. A. Reggia, and C. C. Mitchum (1987). Empirically derived probabilities for grapheme-to-phoneme correspondences in English, *Behavior Research Methods, Instruments, & Computers*, vol. 19(1), pp. 1–9. Cited on page 78.

R. Bhagat and E. Hovy (2013). What is a paraphrase?, *Computational Linguistics*, vol. 39(3), pp. 463–472. Cited on pages 3, 73, 74, 75, 76, and 99.

P. Bhandari, V. Demberg, and J. Kray (2021). Semantic Predictability Facilitates Comprehension of Degraded Speech in a Graded Manner, *Frontiers in Psychology*, vol. 12. Cited on pages 87 and 105.

D. Bohus and A. I. Rudnicky (2008). Sorry, I didn't catch that! An investigation of non-understanding errors and recovery strategies, *Recent trends in discourse and dialogue*, pp. 123–154. Cited on page 100.

C. Boidin, V. Rieser, L. v. d. Plas, O. Lemon, and J. Chevelu (2009). Predicting how it sounds: Re-ranking dialogue prompts based on TTS quality for adaptive spoken dialogue systems, in *Tenth Annual Conference of the International Speech Communication Association 2009*. Cited on pages 100 and 113.

B. Bollepalli, L. Juvela, P. Alku, *et al.* (2019). Lombard speech synthesis using transfer learning in a Tacotron text-to-speech system, *Proc. Interspeech*. Cited on page 2.

D. Bonardo and E. Zovato (2007). Speech synthesis enhancement in noisy environments., in *Interspeech 2007*. Cited on page 1.

Z. S. Bond (1999). *Slips of the ear: Errors in the perception of casual conversation.*, Academic Press. Cited on pages 12 and 45.

I. Brons, R. Houben, and W. A. Dreschler (2014). Effects of noise reduction on speech intelligibility, perceived listening effort, and personal preference in hearing-impaired listeners, *Trends in hearing*, vol. 18. Cited on page 14.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei (2020). Language Models are Few-Shot Learners, in *Advances in Neural Information Processing Systems 2020*. Cited on page 108.

I. Bulyko, K. Kirchhoff, M. Ostendorf, and J. Goldberg (2005). Error-correction detection and response generation in a spoken dialogue system, *Speech Communication*, vol. 45(3), pp. 271–288. Cited on page 99.

R. Carroll and E. Ruigendijk (2013). The effects of syntactic complexity on processing sentences in noise, *Journal of psycholinguistic research*, vol. 42(2), pp. 139–159. Cited on pages 73 and 102.

Z. Chen, H. Zhang, X. Zhang, and L. Zhao (2018). *Quora question pairs*. Cited on page 111.

A. Chingacham, V. Demberg, and D. Klakow (2021). Exploring the Potential of Lexical Paraphrases for Mitigating Noise-Induced Comprehension Errors, in *Proc. Interspeech 2021*. Cited on page 84.

A. Chingacham, V. Demberg, and D. Klakow (2023). A Data-Driven Investigation of Noise-Adaptive Utterance Generation with Linguistic Modification, in *2022 IEEE Spoken Language Technology Workshop (SLT) 2023*. Cited on pages 100, 102, 104, and 116.

J. R. Chowdhury, Y. Zhuang, and S. Wang (2022). Novelty controlled paraphrase generation with retrieval augmented conditional prompt tuning, in *Proceedings of the AAAI Conference on Artificial Intelligence 2022*. Cited on page 110.

C. G. Clopper, J. B. Pierrehumbert, and T. N. Tamati (2010). Lexical neighborhoods and phonological confusability in cross-dialect word recognition in noise, *Laboratory Phonology*, vol. 1(1), pp. 65 – 92. Cited on page 19.

M. Coene, S. Krijger, M. Meeuws, G. De Ceulaer, and P. J. Govaerts (2016). Linguistic factors influencing speech audiometric assessment, *BioMed research international*, vol. 2016. Cited on pages 87 and 102.

M. Cooke (2006). A glimpsing model of speech perception in noise, *The JASA*, vol. 119(3), pp. 1562–1573. Cited on pages 16, 102, and 105.

M. Cooke (2009). Discovering consistent word confusions in noise, in *Tenth Annual Conference of the ISCA 2009*. Cited on page 19.

M. Cooke, M. L. Garcia Lecumberri, J. Barker, and R. Marxer (2019). Lexical frequency effects in English and Spanish word misperceptions, *The JASA*, vol. 145(2), pp. EL136–EL141. Cited on pages 12 and 19.

M. Cooke, S. King, M. Garnier, and V. Aubanel (2014). The listening talker: A review of human and algorithmic context-induced modifications of speech, *Computer Speech & Language*, vol. 28(2), pp. 543–571. Cited on pages 1, 2, 18, 19, and 86.

M. Cooke and M. L. G. Lecumberri (2012). The intelligibility of Lombard speech for non-native listeners, *Acoustical Society of America Journal*, vol. 132(2), p. 1120. Cited on pages 2, 53, and 121.

M. Cooke, C. Mayo, and C. Valentini-Botinhao (2013a). Intelligibility-enhancing speech modifications: the hurricane challenge., in *Interspeech 2013*. Cited on page 1.

M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang (2013b). Evaluating the intelligibility benefit of speech modifications in known noise conditions, *Speech Communication*, vol. 55(4), pp. 572–585. Cited on page 77.

S. Cox and L. Vinagre (2004). Modelling of confusions in aircraft call-signs, *Speech communication*, vol. 42(3-4), pp. 289–312. Cited on pages 13 and 20.

A. Cutler, A. Weber, R. Smits, and N. Cooper (2004). Patterns of English phoneme confusions by native and non-native listeners, *The JASA*, vol. 116(6), pp. 3668–3678. Cited on pages 13, 19, 60, 125, and 126.

M. H. Davis and I. S. Johnsrude (2003). Hierarchical processing in spoken language comprehension, *Journal of Neuroscience*, vol. 23(8), pp. 3423–3431. Cited on page 35.

W. B. Dolan and C. Brockett (2005). Automatically Constructing a Corpus of Sentential Paraphrases, in *Proceedings of the Third International Workshop on Paraphrasing (IWP2005) 2005*. Cited on pages 75 and 111.

K. Drager (2011). Speaker age and vowel perception, *Language and Speech*, vol. 54(1), pp. 99–121. Cited on page 15.

H. R. Dua (1990). The phenomenology of miscommunication, *Beyond Goffman*, pp. 113–139. Cited on page 1.

P. N. Durette (2014: accessed July 30, 2020). *Google Translate's text-to-speech API*. Cited on pages 55 and 81.

T. L. Eadie, H. Durr, C. Sauder, K. Nagle, M. Kapsner-Smith, and K. A. Spencer (2021). Effect of noise on speech intelligibility and perceived listening effort in head and neck cancer, *American Journal of Speech-Language Pathology*, vol. 30(3S), pp. 1329–1342. Cited on page 15.

R. Epp (2018: accessed by July 30, 2020). *BigPhoney, a python module*. Cited on page 64.

C. D. Etymotic Research (1993). *The SIN Test*, 61 Martin Lane, Elk Grove Village, IL 60007. Cited on page 53.

C. Fellbaum (1998). *WordNet: An Electronic Lexical Database*, Bradford Books. Cited on pages 44, 55, and 61.

L. Fontan, J. Tardieu, P. Gaillard, V. Woisard, and R. Ruiz (2015). Relationship between speech intelligibility and speech comprehension in babble noise, *Journal of Speech, Language, and Hearing Research*, vol. 58(3), pp. 977–986. Cited on pages 38 and 77.

M. G. Gaskell and W. D. Marslen-Wilson (1997). Integrating form and meaning: A distributed model of speech perception, *Language and cognitive Processes*, vol. 12(5-6), pp. 613–656. Cited on page 11.

J. J. Godfrey, E. C. Holliman, and J. McDaniel (1992). SWITCHBOARD: Telephone Speech Corpus for Research and Development, in *Proceedings of the 1992 IEEE ICASSP - Volume 1 1992*. Cited on pages 64, 75, and 103.

A. D. Grimshaw (1980). Mishearings, Misunderstandings, and other nonsuccesses in talk: A plea for redress of speaker-oriented bias, *Sociological inquiry*, vol. 50(3-4), pp. 31–74. Cited on pages 1, 12, 33, and 38.

J. Hale (2001). A Probabilistic Earley Parser as a Psycholinguistic Model, in *Second Meeting of the NAACL 2001*. Cited on page 87.

L. Han, A. L. Kashyap, T. Finin, J. Mayfield, and J. Weese (2013). UMBC_EBIQUITY-CORE: Semantic textual similarity systems, in *Second joint conference on lexical and computational semantics (* SEM), volume 1: Proceedings of the main conference and the shared task: Semantic textual similarity 2013*. Cited on page 104.

E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen (2022). LoRA: Low-Rank Adaptation of Large Language Models, in *International Conference on Learning Representations 2022*. Cited on page 110.

D.-Y. Huang, S. Rahardja, and E. P. Ong (2010). Lombard effect mimicking, in *Seventh ISCA Workshop on Speech Synthesis 2010*. Cited on page 2.

T. Joachims (2006). Training linear SVMs in linear time, in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining 2006*. Cited on page 93.

T. Jürgens, S. Fredelake, R. M. Meyer, B. Kollmeier, and T. Brand (2010). Challenging the speech intelligibility index: Macroscopic vs. microscopic prediction of sentence recognition in normal and hearing-impaired listeners, in *Eleventh Annual Conference of the ISCA 2010*. Cited on page 14.

T. Jürgens and T. Brand (2009). Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model, *The JASA*, vol. 126(5), pp. 2635–2648. Cited on page 19.

D. N. Kalikow, K. N. Stevens, and L. L. Elliott (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability, *The JASA*, vol. 61(5), pp. 1337–1351. Cited on pages 3, 17, 19, 20, 38, 53, 67, 72, 87, and 105.

K. P. . J. Kim (2018: accessed by March 21, 2022). *A Simple Python Module for English Grapheme To Phoneme Conversion*. Cited on page 78.

E. Kutlu (2023). Now you see me, now you mishear me: Raciolinguistic accounts of speech perception in different English varieties, *Journal of Multilingual and Multicultural Development*, vol. 44(6), pp. 511–525. Cited on page 15.

W. Lan, S. Qiu, H. He, and W. Xu (2017). A Continuously Growing Dataset of Sentential Paraphrases, in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing 2017*. Cited on pages 71 and 73.

B. Langner and A. W. Black (2005). Improving the understandability of speech synthesis by modeling speech in noise, in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. 2005*. Cited on page 121.

W. J. Levelt, A. Roelofs, and A. S. Meyer (1999). A theory of lexical access in speech production, *Behavioral and brain sciences*, vol. 22(1), pp. 1–38. Cited on pages 33, 34, and 127.

J. Li, T. Tang, W. X. Zhao, J.-Y. Nie, and J.-R. Wen (2024a). Pre-trained Language Models for Text Generation: A Survey, *ACM Computing Surveys*, vol. 56(9), pp. 1–39. Cited on page 101.

J. Li, T. Tang, W. X. Zhao, J.-Y. Nie, and J.-R. Wen (2024b). Pre-trained Language Models for Text Generation: A Survey, *ACM Computing Surveys*, vol. 56(9), pp. 1–39. Cited on page 110.

T. Liu and D. W. Soh (2022). Towards Better Characterization of Paraphrases, in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2022*. Cited on page 104.

S. P. López-Peláez and R. A. Clark (2014). Speech synthesis reactive to dynamic noise environmental conditions, in *Fifteenth Annual Conference of the ISCA 2014*. Cited on page 2.

R. Love, C. Dembry, A. Hardie, V. Brezina, and T. McEnery (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations, *International Journal of Corpus Linguistics*, vol. 22(3), pp. 319–344. Cited on pages 61 and 63.

P. A. Luce and D. B. Pisoni (1998). Recognizing spoken words: The neighborhood activation model, *Ear and hearing*, vol. 19(1), p. 1. Cited on pages 3, 7, 13, 17, 19, 63, and 86.

J. Madhuri and R. G. Kumar (2019). Extractive text summarization using sentence ranking, in *2019 International Conference on Data Science and Communication (IconDSC) 2019*. Cited on page 93.

N. Madnani and B. J. Dorr (2010). Generating phrasal and sentential paraphrases: A survey of data-driven methods, *Computational Linguistics*, vol. 36(3), pp. 341–387. Cited on page 75.

C. Manning and H. Schutze (1999). *Foundations of statistical natural language processing*, MIT press. Cited on page 29.

Y. Marhuenda, D. Morales, and M. del Carmen Pardo (2014). Information criteria for Fay–Herriot model selection, *Computational Statistics and Data Analysis*, vol. 70, pp. 268–280. Cited on page 49.

W. D. Marslen-Wilson and A. Welsh (1978). Processing interactions and lexical access during word recognition in continuous speech, *Cognitive psychology*, vol. 10(1), pp. 29–63. Cited on pages 11 and 35.

R. Marxer, J. Barker, M. Cooke, and M. L. Garcia Lecumberri (2016a). A corpus of noise-induced word misperceptions for English, *The JASA*, vol. 140(5), pp. EL458–EL463. Cited on pages 12 and 19.

R. Marxer, J. Barker, M. Cooke, and M. L. Garcia Lecumberri (2016b). A corpus of noise-induced word misperceptions for English, *The JASA*, vol. 140(5). Cited on pages 45, 46, 47, 49, 77, 127, and 128.

C. Mayo, V. Aubanel, and M. Cooke (2012). Effect of prosodic changes on speech intelligibility, in *Thirteenth Annual Conference of the International Speech Communication Association 2012*. Cited on page 2.

J. L. McClelland and J. L. Elman (1986). The TRACE model of speech perception, *Cognitive psychology*, vol. 18(1), pp. 1–86. Cited on pages 11 and 35.

S. Merity, N. S. Keskar, and R. Socher (2018). Regularizing and Optimizing LSTM Language Models, in *International Conference on Learning Representations 2018*. Cited on page 64.

G. A. Miller and P. E. Nicely (1955). An Analysis of Perceptual Confusions Among Some English Consonants, *The JASA*, vol. 27(2), pp. 338–352. Cited on pages 13, 17, and 48.

M. J. Munro (1998). The effects of noise on the intelligibility of foreign-accented speech, *Studies in Second Language Acquisition*, vol. 20(2), pp. 139–154. Cited on page 15.

C. Nakatsu and M. White (2006). Learning to Say It Well: Reranking Realizations by Predicted Synthesis Quality, in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics 2006*. Cited on pages 100 and 113.

D. Norris (1994). Shortlist: A connectionist model of continuous speech recognition, *Cognition*, vol. 52(3), pp. 189–234. Cited on page 11.

D. Norris, A. Cutler, J. M. McQueen, and S. Butterfield (2006). Phonological and conceptual activation in speech comprehension, *Cognitive Psychology*, vol. 53(2), pp. 146–193. Cited on page 35.

D. Norris, J. M. McQueen, and A. Cutler (2000). Merging information in speech recognition: Feedback is never necessary, *Behavioral and Brain Sciences*, vol. 23(3), pp. 299–325. Cited on page 35.

L. C. Nygaard and D. B. Pisoni (1998). Talker-specific learning in speech perception, *Perception & psychophysics*, vol. 60(3), pp. 355–376. Cited on page 15.

A. Ohashi and R. Higashinaka (2022). Adaptive natural language generation for task-oriented dialogue via reinforcement learning, *International Committee on Computational Linguistics*. Cited on pages 1 and 3.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe (2022a). Training language models to follow instructions with human feedback, in *Advances in Neural Information Processing Systems 2022*. Cited on page 103.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.* (2022b). Training language models to follow instructions with human feedback, *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730–27744. Cited on page 103.

M. Pariente (2018: accessed July 30, 2020). *pySTOI*. Cited on page 87.

R. Patel, M. Everett, and E. Sadikov (2006). Loudmouth: Modifying text-to-speech synthesis in noise, in *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility 2006*. Cited on page 1.

E. Pavlick, P. Rastogi, J. Ganitkevich, B. V. Durme, and C. Callison-Burch (2015). PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification, in *Association for Computational Linguistics 2015*. Cited on page 111.

S. A. Phatak, A. Lovitt, and J. B. Allen (2008). Consonant confusions in white noise, *The Journal of the Acoustical Society of America*, vol. 124(2), pp. 1220–1233. Cited on pages 13, 60, and 125.

J. M. Pickett (1957). Perception of Vowels Heard in Noises of Various Spectra, *The JASA*, vol. 29(5), pp. 613–620. Cited on pages 17, 19, and 48.

D. B. Pisoni, H. C. Nusbaum, P. A. Luce, and L. M. Slowiaczek (1985). Speech perception, word recognition and the structure of the lexicon, *Speech communication*, vol. 4(1-3), pp. 75–95. Cited on pages 17 and 64.

I. Pollack (1975). Auditory informational masking, *The Journal of the Acoustical Society of America*, vol. 57(S1), pp. S5–S5. Cited on page 16.

F. Pusse, A. Sayeed, and V. Demberg (2016). LingoTurk: managing crowdsourced tasks for psycholinguistics, in *Proceedings of the 2016 Conference of the NAACL: Demonstrations 2016*. Cited on pages 56 and 81.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing. Cited on pages 27, 49, 64, 65, and 86.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.* (2019). Language models are unsupervised multitask learners, *OpenAI blog*, vol. 1(8), p. 9. Cited on pages 30, 101, and 105.

A. Rajauria (2020: accessed by March 21, 2022). *PEGASUS fine-tuned for paraphrasing*. Cited on page 75.

J. Rennies, H. F. Schepker, C. Valentini-Botinhao, and M. Cooke (2020). Intelligibility-Enhancing Speech Modifications-The Hurricane Challenge 2.0., in *INTERSPEECH 2020*. Cited on page 1.

C. S. Rogers, L. L. Jacoby, and M. S. Sommers (2012). Frequent false hearing by older adults: the role of age differences in metacognition., *Psychology and aging*, vol. 27(1), p. 33. Cited on pages 14, 20, and 87.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, *ArXiv*, vol. abs/1910.01108. Cited on page 104.

K. Sato (2018: accessed July 6, 2022). *audio-SNR*. Cited on pages 24 and 81.

T. Schoof and S. Rosen (2015). High sentence predictability increases the fluctuating masker benefit, *The JASA*, vol. 138(3), pp. EL181–EL186. Cited on pages 72, 88, and 105.

J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, *et al.* (2018a). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) 2018*. Cited on pages 102 and 116.

J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, *et al.* (2018b). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) 2018*. Cited on page 111.

G. Skantze (2005). Exploring human error recovery strategies: Implications for spoken dialogue systems, *Speech Communication*, vol. 45(3), pp. 325–341. Cited on page 20.

S. D. Soli and L. L. Wong (2008). Assessment of speech intelligibility in noise with the Hearing in Noise Test, *International Journal of Audiology*, vol. 47(6), pp. 356–361. Cited on page 53.

J. Sun, Y. Tian, W. Zhou, N. Xu, Q. Hu, R. Gupta, J. Wieting, N. Peng, and X. Ma (2023). Evaluating Large Language Models on Controlled Generation Tasks, in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing 2023*. Cited on page 101.

M. Suzgun, L. Melas-Kyriazi, and D. Jurafsky (2022). Prompt-and-Rerank: A Method for Zero-Shot and Few-Shot Arbitrary Textual Style Transfer with Small Language Models, in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing 2022*. Cited on page 114.

C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech, in *2010 IEEE ICASSP 2010*. Cited on pages 22, 64, 87, and 105.

R. Taitelbaum-Swead and L. Fostick (2016). The effect of age and type of noise on speech perception under conditions of changing context and noise levels, *Folia Phoniatrica et Logopaedica*, vol. 68(1), pp. 16–21. Cited on pages 14 and 58.

Y. Tang and M. Cooke (2016). Glimpse-Based Metrics for Predicting Speech Intelligibility in Additive Noise Conditions, in *Proc. Interspeech 2016 2016*. Cited on page 105.

B. Taylor (2003). Speech-in-noise tests: How and why to include them in your basic test battery, *The Hearing Journal*, vol. 56(1), pp. 40–42. Cited on page 53.

M. A. Tóth, M. L. García Lecumberri, Y. Tang, and M. Cooke (2015). A corpus of noise-induced word misperceptions for Spanish, *The Journal of the Acoustical Society of America*, vol. 137(2), pp. EL184–EL189. Cited on page 46.

V. N. Uslar, R. Carroll, M. Hanke, C. Hamann, E. Ruigendijk, T. Brand, and B. Kollmeier (2013). Development and evaluation of a linguistically and audiologically controlled sentence intelligibility test, *The JASA*, vol. 134(4), pp. 3039–3056. Cited on pages 38, 73, 77, and 105.

S. Vajjala and D. Meurers (2014). Assessing the relative reading level of sentence pairs for text simplification, in *Proceedings of the 14th Conference of the EACL 2014*. Cited on page 93.

C. Valentini-Botinhao and M. Wester (2014). Using linguistic predictability and the lombard effect to increase the intelligibility of synthetic speech in noise, in *Proc. Interspeech 2014 2014*. Cited on pages 2, 102, 105, and 121.

K. J. Van Engen and J. E. Peelle (2014). *Listening effort and accented speech*. Cited on page 15.

E. C. van Knijff, M. Coene, and P. J. Govaerts (2018). Speech understanding in noise in elderly adults: the effect of inhibitory control and syntactic complexity, *International journal of language & communication disorders*, vol. 53(3), pp. 628–642. Cited on page 14.

S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks (2018). An evaluation of intrusive instrumental intelligibility metrics, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26(11), pp. 2153–2166. Cited on page 73.

M. van Os, J. Kray, and V. Demberg (2021). Mishearing as a Side Effect of Rational Language Comprehension in Noise, *Frontiers in Psychology*, vol. 12, p. 3488. Cited on pages 14, 38, and 53.

M. Van Os, J. Kray, and V. Demberg (2022). Rational speech comprehension: Interaction between predictability, acoustic signal, and noise, *Frontiers in Psychology*, vol. 13. Cited on page 20.

V. N. Vapnik, V. Vapnik, *et al.* (1998). Statistical learning theory. Cited on page 28.

A. Varga and H. J. Steeneken (1993). Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems, *Speech communication*, vol. 12(3), pp. 247–251. Cited on pages 16, 55, 61, 81, and 102.

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

M. S. Vitevitch (2002). Naturalistic and experimental analyses of word frequency and neighborhood density effects in slips of the ear, *Language and speech*, vol. 45(4), pp. 407–434. Cited on pages 7, 17, 64, and 67.

M. S. Vitevitch and E. Rodríguez (2005). Neighborhood density effects in spoken word recognition in Spanish, *Journal of Multilingual Communication Disorders*, vol. 3(1), pp. 64–73. Cited on pages 17, 64, and 67.

W. Wahlster (1993). Verbmobil, in *Grundlagen und anwendungen der künstlichen intelligenz 1993*, pp. 393–402, Springer. Cited on page 55.

L. Ward, B. G. Shirley, Y. Tang, W. J. Davies, *et al.* (2017). The effect of situation-specific non-speech acoustic cues on the intelligibility of speech in noise, in *INTERSPEECH 2017, 18th Annual Conference of the ISCA 2017*. Cited on pages 20, 72, and 87.

A. Warzybok, T. Brand, K. C. Wagener, and B. Kollmeier (2015). How much does language proficiency by non-native listeners influence speech audiometric tests in noise?, *International journal of audiology*, vol. 54(sup2), pp. 88–99. Cited on page 15.

A. Weber and O. Scharenborg (2012). Models of spoken-word recognition, *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 3(3), pp. 387–401. Cited on pages 11 and 35.

A. Weber and R. Smits (2003). Consonant and vowel confusion patterns by American English listeners, in *15th International Congress of Phonetic Sciences [ICPhS 2003] 2003*. Cited on pages 3, 13, 14, 17, 19, 60, and 127.

J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le (2022a). Finetuned Language Models are Zero-Shot Learners, in *International Conference on Learning Representations 2022*. Cited on page 30.

J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le (2022b). Finetuned Language Models are Zero-Shot Learners, in *International Conference on Learning Representations 2022*. Cited on page 101.

D. Wendt, T. Dau, and J. Hjortkjær (2016). Impact of background noise and sentence complexity on processing demands during sentence comprehension, *Frontiers in psychology*, vol. 7, p. 345. Cited on page 38.

J. Weston, E. Dinan, and A. Miller (2018). Retrieve and Refine: Improved Sequence Generation Models For Dialogue, in *Proceedings of the 2018 EMNLP Workshop SCAI:*

The 2nd International Workshop on Search-Oriented Conversational AI 2018. Cited on page 113.

E. Wilson and T. Spaulding (2010). Effects of noise and speech intelligibility on listener comprehension and processing time of Korean-accented English., *Journal of speech, language, and hearing research : JSLHR*, vol. 53 6, pp. 1543–54. Cited on page 38.

R. H. Wilson and W. B. Cates (2008). A comparison of two word-recognition tasks in multitalker babble: Speech Recognition in Noise Test (SPRINT) and Words-in-Noise Test (WIN), *Journal of the American Academy of Audiology*, vol. 19(7), pp. 548–556. Cited on pages 19 and 38.

H. Zhang, H. Song, S. Li, M. Zhou, and D. Song (2022). A survey of controllable text generation using transformer-based pre-trained language models, *ACM Computing Surveys*. Cited on page 110.

H. Zhang, H. Song, S. Li, M. Zhou, and D. Song (2023). A survey of controllable text generation using transformer-based pre-trained language models, *ACM Computing Surveys*, vol. 56(3), pp. 1–37. Cited on page 101.

J. Zhang, Y. Zhao, M. Saleh, and P. Liu (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, in *International Conference on Machine Learning 2020*. Cited on page 75.

M. Zhang, P. N. Petkov, and W. B. Kleijn (2013). Rephrasing-based speech intelligibility enhancement., in *INTERSPEECH 2013*. Cited on pages 3 and 100.

T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi (2020). BERTScore: Evaluating Text Generation with BERT, in *International Conference on Learning Representations 2020*. Cited on pages 75 and 104.

Y. Zhang, J. Baldridge, and L. He (2019). PAWS: Paraphrase Adversaries from Word Scrambling , in *Proc. of NAACL 2019*. Cited on page 75.

Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and W. B. Dolan (2020). DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations 2020*. Cited on page 87.