Human Speech Perception in Noise: Can Large Language Models Paraphrase to Improve It?

Anupama Chingacham[1] **Miaoran Zhang**[1] **Vera Demberg**[1,2] **Dietrich Klakow**[1]
[1]Saarland University, Saarland Informatic Campus, Germany
[2]Max Planck Institute for Informatics, Germany
achingacham@lsv.uni-saarland.de

Abstract

Large Language Models (LLMs) can generate text by transferring style attributes like formality resulting in formal or informal text. However, instructing LLMs to generate text that when spoken, is more intelligible in an acoustically difficult environment, is an underexplored topic. We conduct the first study to *evaluate LLMs on a novel task of generating acoustically intelligible paraphrases* for better human speech perception in noise. Our experiments in English demonstrated that with standard prompting, LLMs struggle to control the non-textual attribute, *i.e.,* acoustic intelligibility, while efficiently capturing the desired textual attributes like semantic equivalence. To remedy this issue, we propose a simple prompting approach, *prompt-and-select*, which generates paraphrases by decoupling the desired textual and non-textual attributes in the text generation pipeline. Our approach resulted in a 40% relative improvement in human speech perception, by paraphrasing utterances that are highly distorted in a listening condition with babble noise at signal-to-noise ratio (SNR) -5 dB. This study reveals the limitation of LLMs in capturing non-textual attributes, and our proposed method showcases the potential of using LLMs for better human speech perception in noise.[1]

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; 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). Recent explorations on 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 (Chingacham et al., 2021, 2023). In other words, the enhanced speech perception with paraphrasing is driven by noise-robust acoustic cues.

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., 2022; Li et al., 2024). 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., 2023a). 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

[1]Our code and data are available at https://github.com/uds-lsv/llm_eval_PI-SPiN.

Proceedings of the 1st Human-Centered Large Language Modeling Workshop, pages 1–15

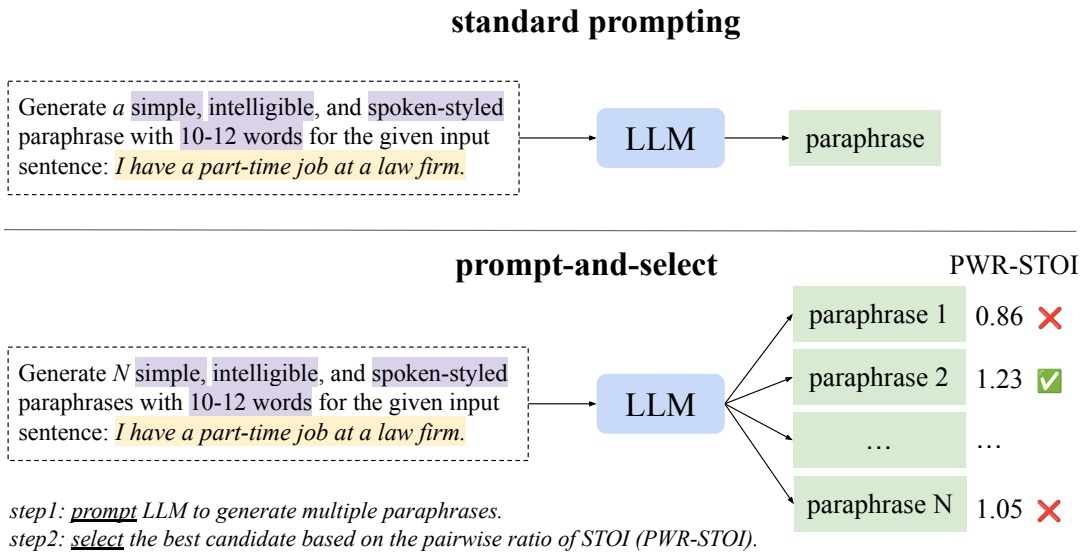


Figure 1: A schematic representation of the *prompt-and-select* and standard prompting approach to generate acoustically intelligible paraphrase in a noisy environment. A speech intelligibility metric, short-time objective intelligibility measure (STOI) is employed to select the paraphrase that is more likely to improve speech perception.

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' inherent capability to generate acoustically intelligible paraphrases*, without any model fine-tuning. Through standard prompting methods like zero-shot learning (ZSL), 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.

To effectively utilize LLMs for generating acoustically intelligible paraphrases, we propose a simple approach called *prompt-and-select*, which guides paraphrase generation by introducing the desired non-textual attribute in a post-processing step (see Figure 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. 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.[2]

[2]See definitions of babble noise and SNR in Appendix A.

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.

2 Related Work

Acoustic Intelligibility. Speech perception has been a long-standing research topic in speech science (Kalikow et al., 1977; Luce and Pisoni, 1998; McArdle and Wilson, 2008), which contributed towards a better understanding of human (*mis*)hearing. More specifically, several human perception experiments were conducted to investigate the intelligibility of speech tokens such as vowels (Pickett, 1957; Cutler et al., 2004), consonants (Weber and Smits, 2003; Jürgens and Brand, 2009), words in isolation (Luce and Pisoni, 1998; Clopper et al., 2010; Wilson and Cates, 2008), words in context (Kalikow et al., 1977; Uslar et al., 2011; Chingacham et al., 2021), especially in noisy environments. While several studies showcased the influence of linguistic characteristics such as predictability (Kalikow et al., 1977), syntactic complexity (Uslar et al., 2011; Carroll and Ruigendijk, 2013; van Knijff et al., 2018), and lexical features (Luce and

Pisoni, 1998; McArdle and Wilson, 2008), on the intelligibility of utterances in noise, there is limited explorations in utilizing the linguistic potential to improve acoustic intelligibility in noise.

We share the motivation to improve speech perception in noise using paraphrases with early studies (Cox and Vinagre, 2004; Nakatsu and White, 2006; Zhang et al., 2013; Chingacham et al., 2023). Nakatsu and White (2006) proposed to train a reranker to select the paraphrases that are more likely to sound natural, when synthesized. They generated multiple paraphrases for each sentence mainly by modifying the word order and replacing a few lexical units in the original sentence. On the other hand, Zhang et al. (2013) proposed an objective measure to distinguish the intelligibility difference among paraphrases that are of the same syntactic type, thereby restricting the type of sentential paraphrases. More recently, Chingacham et al. (2023) investigated the potential of improving intelligibility by considering a larger set of paraphrase types, which are generated using modern paraphrasing models. However, our work is distinct from theirs as we explore LLMs' inherent ability to generate acoustically intelligible paraphrases.

LLM Evaluation. Given the rapid growth of LLMs such as ChatGPT and GPT-4 (OpenAI, 2023), there has been a surge of research interest towards a holistic evaluation of their capabilities (Chang et al., 2024). Recent studies have attempted to examine their performance across diverse tasks, such as machine translation (Hendy et al., 2023; Zhu et al., 2023), text summarization (Yang et al., 2023; Pu and Demberg, 2023), etc; and also aspects of multilinguality (Lai et al., 2023b; Ahuja et al., 2023) and multimodality (Bang et al., 2023). Close to our work, there have been a few studies looking into the controllable generation ability of LLMs. Lai et al. (2023a) explore the potential of ChatGPT as a text-style transfer evaluator. Sun et al. (2023b) present a systematic study on ten controllable generation benchmarks. Notably, their control factors are derived from the language perspective (e.g., sentiment and number), whereas our work pioneers the investigation of the potential of LLMs as an acoustically intelligible paraphrase generator.

3 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 (PiN) dataset[3] (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. The better intelligibility of utterances can be attributed to both linguistic features like predictability (Kalikow et al., 1977), syntactic structure (Uslar et al., 2013), as well as acoustic features like the underlying sounds of the utterance (Luce and Pisoni, 1998). In the more recent investigations on the intelligibility difference among paraphrases (Chingacham et al., 2023), it was shown that the better intelligibility of s_2 in such high noise environments is mainly driven by the noise-robust acoustic cues that are defined by both the constituting sounds as well as the noise signal. 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

[3] See Appendix B for more samples.

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 generate the acoustic realization of a sentence, we used the Tacotron2 text-to-speech (TTS) system, which demonstrated performance on par with that of a professional voice talent (Shen et al., 2018). More specifically, we used the Tacotron2 model[4] pre-trained on the LJSpeech dataset by SpeechBrain (Ravanelli et al., 2021). Further, to create the noise-distorted signals, the clean audio signals underwent a noise-mixing procedure using an open-sourced tool, *audio-SNR*.[5] 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 text satisfies the desired outcome, we primarily relied on automatic metrics, which are discussed in detail in the following section.

4 Experimental Setup

Model. For all our experiments, we used *ChatGPT*[6] (Ouyang et al., 2022), which is one of the most popular LLMs. It has shown impressive performance on paraphrase generation with textual style attributes, while its ability on acoustically intelligible paraphrasing remains unclear. We adopt default parameters (temperature=1.0, top_p=1.0) for the API calls.

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).

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[7] to generate the phonemic transcript of a sentence. 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] https://huggingface.co/speechbrain/tts-tacotron2-ljspeech

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

[6] Version: gpt-3.5-turbo

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

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 1: Three prompts used in standard prompting, with an increasing level of detail in the task objective. Boldfaced words are task-specific keywords in the prompt statement.

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[8] (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.

5 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 two prompting methods that we employed and evaluated for the task.

5.1 Standard Prompting

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 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.

Results and Analysis Table 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

[8]Version: `distilgpt2`

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 2: An automatic evaluation of paraphrases generated by different prompts. 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.[9]

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.***

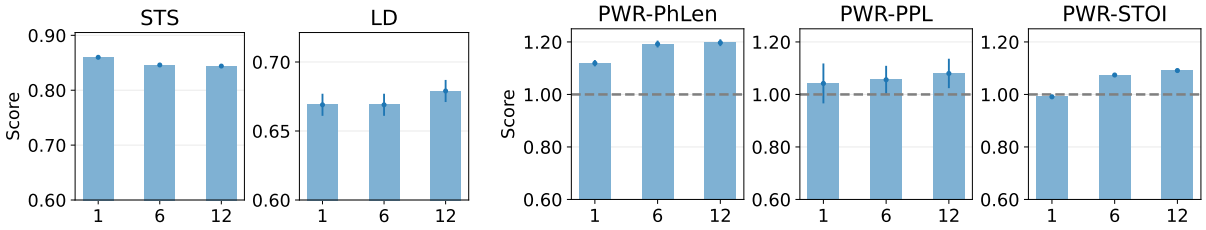
Table 3: A qualitative analysis of model-generated text, `{output}`, for a given `{input text}` under three standard prompts: $p_{zsl-low}, p_{zsl-med}, p_{zsl-high}$. The prompt $p_{zsl-high}$ generates several tokens that are irrelevant (boldfaced words) to the task objective.

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 as reflected by the *PWR*-STOI scores being not significantly different from 1.0. 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**.

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 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 standard prompting may not effectively elicit the model's ability to generate paraphrases with the intended non-textual attribute, which is beyond the model's comprehension.[10]

[9] See Appendix C for the absolute scores of different metrics.

[10] In Appendix D, we also conducted a preliminary study



(a) (b)

Figure 2: An automatic evaluation of the standard prompting ($n = 1$) and the proposed prompt-and-select ($n > 1$) approach. The *X*-axis is the number of candidates generated (n) and the *Y*-axis is the mean scores (with error bars at 95% confidence interval). The reference line in Fig. (b) 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), but it comes with a trade-off on semantic equivalence (STS).[11]

5.2 PAS: Prompt-and-Select

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 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.

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 2a, PAS showcased a high quality of paraphrase generation as indicated by high STS and high LD, similar to the standard prompting setup. Similarly, Figure 2b 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 2b 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 para-

on in-context learning, suggesting that demonstrations are not helpful in capturing the non-textual attribute.

[11] See Appendix C for the absolute scores of different metrics with varying numbers of candidates.

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 4: 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 ($p < 0.05$).

phrases. 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.

5.3 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 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., 2018) 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 4, 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 that are 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 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. For future work, we could consider two approaches to improve LLMs on this task: 1) fine-tuning LLMs with a large parallel dataset consisting of sentences and their corresponding intelligible paraphrases, and 2) incorporating the acoustic representation of the utterances to control the paraphrase generation.

Limitations

The proposed "prompt-and-select" approach relies on the efficacy of STOI scores to identify the best candidate which is more likely to be perceived correctly in noise. In other words, this approach requires a metric that accurately estimates the desired non-textual attribute. This could be a limitation for problems that require human annotations for candidate selection. Additionally, the current approach introduces an overhead in computation and inference time, due to multiple generations and STOI calculation that involves speech synthesis and noise-mixing procedure. Further investigations are required to study the trade-off between the benefits of paraphrasing and the cost of additional resources. Moreover, our study only evaluated ChatGPT, one of the representative LLMs, due to budget and resource constraints. We believe that a holistic evaluation covering more open-source models, such as Mistral (Jiang et al., 2023) and Llama 3 (Meta, 2024), will be beneficial to deepen our understanding of LLM capabilities.

Ethics statement

In this work, generative LLMs are evaluated for a new task without model fine-tuning. It is an impactful step to democratize LLMs for research facilities with limited data and computing resources. We conducted a human evaluation on Prolific, ensuring that all participants were paid (9 GBP) for their service, considering the recommended minimum wage per hour in the UK, in 2023. Also, we ensured to provide an inclusive environment for our participants in the perception experiment, providing non-binary options to mark their gender identity.

Acknowledgements

We would like to thank anonymous reviewers for their valuable feedback. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project-id 232722074 – SFB 1102.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Elron Bandel, Ranit Aharonov, Michal Shmueli-Scheuer, Ilya Shnayderman, Noam Slonim, and Liat Ein Dor. 2022. Quality controlled paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 596–609.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718,

Nusa Dua, Bali. Association for Computational Linguistics.

Daniel Bär, Chris Biemann, Iryna Gurevych, and Torsten 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)*, pages 435–440, Montréal, Canada. Association for Computational Linguistics.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Pratik Bhandari, Vera Demberg, and Jutta Kray. 2021. Semantic predictability facilitates comprehension of degraded speech in a graded manner. *Frontiers in Psychology*, 12.

Dan Bohus and Alexander I Rudnicky. 2008. Sorry, i didn't catch that! an investigation of non-understanding errors and recovery strategies. *Recent trends in discourse and dialogue*, pages 123–154.

Cedric Boidin, Verena Rieser, Lonneke van der Plas, Oliver Lemon, and Jonathan 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*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ivan Bulyko, Katrin Kirchhoff, Mari Ostendorf, and Julie Goldberg. 2005. Error-correction detection and response generation in a spoken dialogue system. *Speech Communication*, 45(3):271–288.

Rebecca Carroll and Esther Ruigendijk. 2013. The effects of syntactic complexity on processing sentences in noise. *Journal of psycholinguistic research*, 42(2):139–159.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Anupama Chingacham, Vera Demberg, and Dietrich Klakow. 2021. Exploring the Potential of Lexical Paraphrases for Mitigating Noise-Induced Comprehension Errors. In *Proc. Interspeech*, pages 1713–1717.

Anupama Chingacham, Vera Demberg, and Dietrich Klakow. 2023. A data-driven investigation of noise-adaptive utterance generation with linguistic modification. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 353–360. IEEE.

Cynthia G. Clopper, Janet B. Pierrehumbert, and Terrin N. Tamati. 2010. Lexical neighborhoods and phonological confusability in cross-dialect word recognition in noise. *Laboratory Phonology*, 1(1):65 – 92.

Martine Coene, Stefanie Krijger, Matthias Meeuws, Geert De Ceulaer, and Paul J Govaerts. 2016. Linguistic factors influencing speech audiometric assessment. *BioMed research international*, 2016.

Martin Cooke. 2006. A glimpsing model of speech perception in noise. *The JASA*, 119(3):1562–1573.

Stephen Cox and Lluis Vinagre. 2004. Modelling of confusions in aircraft call-signs. *Speech communication*, 42(3-4):289–312.

Anne Cutler, Andrea Weber, Roel Smits, and Nicole Cooper. 2004. Patterns of english phoneme confusions by native and non-native listeners. *The JASA*, 116(6):3668–3678.

Mangesh S Deshpande and Raghunath S Holambe. 2009. Speaker identification based on robust am-fm features. In *2009 Second International Conference on Emerging Trends in Engineering & Technology*, pages 880–884. IEEE.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE ICASSP - Volume 1*, ICASSP'92, page 517–520. IEEE Computer Society.

Lushan Han, Abhay L Kashyap, Tim Finin, James Mayfield, and Jonathan 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*, pages 44–52.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud,

Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Tim Jürgens and Thomas Brand. 2009. Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model. *The JASA*, 126(5):2635–2648.

Daniel N Kalikow, Kenneth N Stevens, and Lois L Elliott. 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The JASA*, 61(5):1337–1351.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023a. Multidimensional evaluation for text style transfer using chatgpt. *arXiv preprint arXiv:2304.13462*.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023b. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, 56(9):1–39.

Timothy Liu and De Wen Soh. 2022. Towards better characterization of paraphrases. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8592–8601, Dublin, Ireland. Association for Computational Linguistics.

Paul A Luce and David B Pisoni. 1998. Recognizing spoken words: The neighborhood activation model. *Ear and hearing*, 19(1):1.

Rachel McArdle and Richard H Wilson. 2008. Predicting word-recognition performance in noise by young listeners with normal hearing using acoustic, phonetic, and lexical variables. *Journal of the American Academy of Audiology*, 19(6):507–518.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

George A Miller. 1947. The masking of speech. *Psychological bulletin*, 44(2):105.

Crystal Nakatsu and Michael 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*, pages 1113–1120, Sydney, Australia. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

J. M. Pickett. 1957. Perception of vowels heard in noises of various spectra. *The JASA*, 29(5):613–620.

Dongqi Pu and Vera Demberg. 2023. ChatGPT vs human-authored text: Insights into controllable text summarization and sentence style transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1–18, Toronto, Canada. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Tim Schoof and Stuart Rosen. 2015. High sentence predictability increases the fluctuating masker benefit. *The JASA*, 138(3):EL181–EL186.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023a. Evaluating large language models on controlled generation tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore. Association for Computational Linguistics.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023b. Evaluating large language models on controlled generation tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore. Association for Computational Linguistics.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan 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*, pages 2195–2222, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE ICASSP*, pages 4214–4217.

Genichi Taguchi. 1986. *Introduction to quality engineering: designing quality into products and processes*.

Yan Tang and Martin Cooke. 2016. Glimpse-Based Metrics for Predicting Speech Intelligibility in Additive Noise Conditions. In *Proc. Interspeech 2016*, pages 2488–2492.

Verena Uslar, Esther Ruigendijk, Cornelia Hamann, Thomas Brand, and Birger Kollmeier. 2011. How does linguistic complexity influence intelligibility in a german audiometric sentence intelligibility test? *International Journal of Audiology*, 50(9):621–631.

Verena N Uslar, Rebecca Carroll, Mirko Hanke, Cornelia Hamann, Esther Ruigendijk, Thomas Brand, and Birger Kollmeier. 2013. Development and evaluation of a linguistically and audiologically controlled sentence intelligibility test. *The JASA*, 134(4):3039–3056.

Cassia Valentini-Botinhao and Mirjam Wester. 2014. Using linguistic predictability and the lombard effect to increase the intelligibility of synthetic speech in noise. In *Proc. Interspeech 2014*, pages 2063–2067.

Eline C van Knijff, Martine Coene, and Paul J Govaerts. 2018. Speech understanding in noise in elderly adults: the effect of inhibitory control and syntactic complexity. *International journal of language & communication disorders*, 53(3):628–642.

Andrew Varga and Herman JM 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*, 12(3):247–251.

Andrea Weber and Roel Smits. 2003. Consonant and vowel confusion patterns by american english listeners. In *15th International Congress of Phonetic Sciences [ICPhS 2003]*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Weston, Emily Dinan, and Alexander 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*, pages 87–92, Brussels, Belgium. Association for Computational Linguistics.

Richard H Wilson and Wendy 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*, 19(7):548–556.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37.

Mengqiu Zhang, Petko Nikolov Petkov, and W Bastiaan Kleijn. 2013. Rephrasing-based speech intelligibility enhancement. In *INTERSPEECH*, pages 3587–3591.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *Preprint*, arXiv:2304.04675.

A Definitions

Babble Noise. It is one of the most commonly occurring noise types in the real world (Miller, 1947). Typically, it is the noise that exists in a cafeteria or other crowded environments, wherein individuals engage in conversations in the backdrop of other conversations. The simultaneous speech produced by several individuals in the background masks the target speech and could hinder listening. The babble noise in the NOISEX-92 database that we use in this work is a recording of 100 people speaking in a canteen (Varga and Steeneken, 1993; Deshpande and Holambe, 2009).

Signal-to-Noise Ratio. To measure the noise level, a commonly used metric is the signal-to-noise ratio (SNR) (Taguchi, 1986). SNR represents the ratio of the power of a clean (undistorted) signal and a noise signal, which are combined to form the distorted signal. Simply put, it is a fraction of powers as defined in Equation (1). It is commonly measured on a logarithmic scale and referred to in units of decibels (dB), as defined in Equation (2). The power of a signal is the sum of the absolute squares of signal magnitudes averaged across the time domain.

$$SNR = \frac{P_{signal}}{P_{noise}} \tag{1}$$

$$\begin{aligned} SNR_{dB} &= 10\log_{10}(SNR) \\ &= 10\log_{10}(\frac{P_{signal}}{P_{noise}}) \end{aligned} \tag{2}$$

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).

B More Samples from the PiN Dataset

In Table 8, we provide more paraphrase triplets from the PiN dataset.

C Absolute Scores

We provide absolute scores for different evaluation metrics in Table 5 in addition to their pairwise ratios.

Prompt-ID	PhLen	PPL	STOI
$p_{zsl-low}$	50.67	159.95	0.570
$p_{zsl-med}$	42.08	165.56	0.569
$p_{zsl-high}$	46.68	193.85	0.577
$p_{pas_{(n=6)}}$	44.67	182.77	0.617
$p_{pas_{(n=12)}}$	44.88	184.52	0.627
p_{icl}	47.27	146.71	0.573
`{input text}`	38.02	236.65	0.577

Table 5: Absolute scores for utterance length (PhLen), linguistic predictability (PPL), and acoustic intelligibility (STOI) of `{input text}` and generated outputs by different prompts.

D 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 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 represents the exact prompt statement (p_{icl}) that we used for the in-context learning.

Results and Analysis As shown in Table 7, 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}) gener-

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: The prompt used for the in-context learning setup.

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

Table 7: An evaluation of the ICL setup. LLM fails to improve acoustic intelligibility (*PWR*-STOI < 1.0), though it learns to capture the demonstrated textual attributes like lexical deviation and predictability.

ated texts that are semantically more similar and lexically less divergent from the input sentences. More interestingly, the model also learned to optimize the desired 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.

Sentence_ID	Sentence
s_1	they give more information than opinions
s_2	they seem to give more of just the facts than opinions
s_3	they seem to give more facts than opinions
s_1	you don't hear much about it in the big ones
s_2	in the big ones you don't hear about it
s_3	you never hear about it really in the big ones
s_1	I think we talked for a good eight minutes about the subject
s_2	we talked for about eight minutes
s_3	I think we talked for about eight minutes
s_1	I like having people over for dinner
s_2	I enjoy having people over for dinner
s_3	if I have people over for dinner I like it to be
s_1	I studied every piece of material I could
s_2	I studied every part of the material
s_3	and studied every bit of material that I could study
s_1	I wanted to be a teacher at one time
s_2	at one point I wanted to be a teacher
s_3	I thought at one time I wanted to be a teacher
s_1	they never imagined it would be a hit
s_2	in fact, they never thought it would be a hit
s_3	they never expected it to be a hit
s_1	they want a lot more men to participate
s_2	they need more men to participate
s_3	they really looking for a lot more men to participate
s_1	we gave them about seven minutes
s_2	we gave them about seven minutes according to my watch
s_3	they were given seven minutes
s_1	you don't hear much about it in the big ones
s_2	in the big ones you don't hear about it
s_3	you never hear about it really in the big ones
s_1	at that stage of life you only have so much money left
s_2	you only have a limited amount of money left
s_3	you only have so much money left at that point in your life
s_1	I was angry that they were capable of doing that
s_2	I was mad that they could do that
s_3	I was just pissed as hell that they could do that

Table 8: A list of paraphrase triplets (s_1, s_2, s_3) from the PiN dataset. Sentences in each triplet are arranged in such a way that s_1 is acoustically less intelligible than s_2, and acoustically more intelligible than s_3, in a listening condition with babble noise at SNR -5 dB.