Enhancing Naturalness in LLM-Generated Utterances through Disfluency Insertion

Anonymous ACL submission

Abstract

Disfluencies are a natural feature of spontaneous human speech but are typically absent from the outputs of Large Language Models (LLMs). This absence can diminish the perceived naturalness of synthesized speech, which is an important criteria when building conversational agents that aim to mimick human behaviours. We show how the insertion of disfluencies can alleviate this shortcoming. The proposed approach involves (1) fine-tuning an LLM with Low-Rank Adaptation (LoRA) to incorporate various types of disfluencies into LLM-generated utterances and (2) synthesizing those utterances using a text-to-speech model that supports the generation of speech phenomena such as disfluencies. We evaluated the quality of the generated speech across two metrics: *intelligibility* and *perceived spontaneity*. We demonstrate through a user study that the insertion of disfluencies significantly increase the perceived spontaneity of the generated speech. This increase came, however, along with a slight reduction in intelligibility.

1 Introduction

Disfluencies are an intrinsic component of spontaneous speech, often manifesting as hesitations, fillers ('uh,' 'like'), or repeated words and phrases. They play an important role in human speech communication and are a common occurrence across all types of non-scripted dialogues (Lickley and Bard, 1998; Yaruss and Quesal, 2002). Previous studies have highlighted that disfluent speech is often seen as more spontaneous and natural (Kampf, 2022). Pauses and hesitations have also been shown to serve as markers of spontaneous thought, enhancing the perceived naturalness of conversation (Saryazdi et al., 2021). Incorporating disfluencies into the outputs of conversational agents may therefore improve their perceived "naturalness".

This perception of naturalness is an important consideration when building conversational systems designed to emulate human behaviours in various settings. Such systems are often described as virtual *avatars* and are employed in applications such as educational tools, gaming, and healthcare support (Chheang et al., 2024; Qin and Hui, 2023; Yan and Alterovitz, 2024). Previous studies have shown that factors like appearance, speech, lip-sync, and a strong sense of presence are important for user engagement and effectiveness with these avatars (Hassan et al., 2022; Salehi et al., 2022), along with the quality of speech synthesis (Mattheyses and Verhelst, 2015). Avatars have also been employed for the purposes of training human professionals to handle sensitive dialogues, such as delivering bad news in healthcare settings (Andrade et al., 2010) or conducting investigative interviews with children who have experienced abuse (Pompedda et al., 2015; Dalli, 2021; Baugerud et al., 2021). As emotional stress tends to decrease speech fluency (Buchanan et al., 2014), disfluencies are highly frequent in those stressful and challenging dialogues, and an avatar developed for such training purposes should seek to reproduce such conversational behaviours.

This paper presents an approach that artificially inserts disfluencies in non-disfluent utterances, and passes on the results to a speech synthesis model that is well-suited for the generation of spontaneous speech. Our contributions are as follows:

- Fine-tuning an LLM with LoRA to introduce disfluencies in dialogue utterances.
- A user study to evaluate the effectiveness of the generated disfluencies by converting the disfluent text into audio clips for assessment.

2 Background

Shriberg (1994) showed that disfluencies in speech are not random but follow specific patterns. Disfluencies can be categorized into typical and atypical forms. Typical disfluencies include hesitations, fillers like "um" and "uh" which often signal pauses for thought or plan future speech. Atypical disfluencies include conversational phenomena such as repetition, substitution, insertion, and speech errors (Yaruss and Quesal, 2002).

Only a few previous works have explored the generation of disfluencies in spoken utterances. To our knowledge, Marge et al. (2010) presented the first system attempting to artificially inserting disfluencies in the outputs of a dialogue system. Although their motivation is similar to ours, their insertion mechanism relied on a set of heuristic rules instead of a fine-tuned LLM. They also did not explicitly assess the intelligibility of the synthesized utterances as part of their study.

Yang et al. (2020) proposed a Planner-Generator model for generating natural and diverse disfluent texts, with the goal of creating synthetic data for training disfluency detection models. Their approach consists of two components: a Planner, which identifies optimal locations for inserting disfluent segments, and a Generator, which creates the corresponding disfluent phrases based on the Planner's predictions. Their experiments demonstrated state-of-the-art performance for disfluency detection on the Switchboard corpus.

Marie (2023) trained a disfluency generation model using the base version of the T5 model (Raffel et al., 2020). Their base model was fine-tuned for few-shot learning on the Fisher corpus, using fluent-disfluent parallel data subsets of varying sizes and the full set of available utterances. The work presented in this paper extends this approach along several dimensions. In addition to using a larger, decoder-only model for the disfluency generation, we also provide a detailed human evaluation of how the insertion of disfluencies affects their perceived spontaneity and intelligibility.

3 Methods

We outline below the dataset and fine-tuning method employed to insert disfluencies in non-disfluent utterances, as well as the text synthesis model converting those into speech.

3.1 Dataset

We used the Switchboard dataset including disfluency annotations (Zayats et al., 2019). The Switchboard corpus was developed by transcribing telephonic conversations (Godfrey et al., 1992), and disfluencies were initially hand-annotated in the Penn Treebank release (Marcus et al., 1999). This corpus was subsequently refined by researchers to correct transcription errors (Mantha, 2000). Zayats et al. (2019) employed the cleaned-up, disfluency-tagged dataset, and re-annotated with BIO tagging to account for reparandum and correction spans. We also used the NXT Switchboard Corpus (Calhoun et al., 2010), focusing on annotations about pitch contours, pauses, and other acoustic features. Our primary focus was on their annotations of silent pauses. These pauses often signal hesitation, planning, or shifts in dialogue.

Dataset	**Train**	**Test**
No. Sentences	32490	3610
Avg No. of tokens in fluent utterance	24.28	24.15
Avg No. of tokens in disfluent utterance	33.08	32.84
Total No. fluent tokens	789K	87K
Total No. disfluent tokens	1075K	119K
Rate of disfluency (%)	24.5%	23.9%

Table 1: Statistics of the training set used for fine-tuning the LLM and the test set used for evaluation.

The data from Zayats et al. (2019) was used to extract instances of atypical disfluencies, while Calhoun et al. (2010) provided instances of typical disfluencies. Table 1 shows an overview of both the training and test set. The rate of disfluency is defined as the ratio of disfluent tokens to the total number of tokens in an utterance.

3.2 Model Training

We fine-tuned the **Llama-2-7b-chat-hf** with LoRA (Hu et al., 2021) for an efficient and targeted adaptation process. We chose Llama-2-7b to strike a balance between accuracy and computational efficiency. The fine-tuning setup included a maximum sequence length of 200 tokens, which represents a moderately long input that optimized the memory demands. For the LoRA-specific parameters, we set the LoRA rank (r) to 32 and the scaling factor (alpha) to 64, with a dropout rate of 0.1. The training was conducted with a batch size of 2 and gradient accumulation steps of 4, effectively simulating a larger batch size to stabilize training on limited computational resources. We adopt a learning rate of 2×10^{-4}. The model was fine-tuned to generate disfluencies like repetition, substitution, insertions, and speech pauses, as well as typical disfluencies such as filled and silent pauses.

3.3 Text-to-speech

Disfluencies are a speech phenomena. We use existing text-to-speech models to convert the disfluent utterances generated by the fine-tuned LLM into audio form. To select a TTS model for generating realistic disfluent speech, we conducted a pilot study in which we evaluated audio samples generated by three TTS models: Bark TTS model developed by Suno-AI (Suno-AI, 2023), Tortoise TTS (Betker, 2023) and OpenAI TTS (OpenAI, 2023). The evaluation focused on the prosody, intonation, and in particular the pronunciation of disfluencies. An important consideration was how the text-to-speech rendered false starts such as "Th- they" or "B- birthday", which are particularly frequent in Switchboard – and hence in the disfluent utterances produced by the fine-tuned LLM.

Among the three models, Bark TTS was deemed most effective at synthesizing these disfluencies into speech. While all three TTS models did at times produce errors, such as buzzing sounds and prolong disruptive pauses in longer utterances, Bark TTS was found to perform better overall. To ensure the audio clips generated with Bark for the user study were of high quality, we regenerated samples when noisy artifacts such as buzzing sounds were detected in the audio outputs.

4 Automatic Evaluation

To assess the performance of our fine-tuned large language model (LLM) at inserting disfluencies, we first compare the disfluent utterances generated by the model with the actual utterances in the Switchboard test set, in which disfluencies and pauses are explicitly transcribed (Meteer et al., 1995). This comparison is conducted using both BLEU and BERTScore (Zhang et al., 2019) metrics to evaluate lexical and semantic alignment. Table 2 summarizes the evaluation results for the fine-tuned LLM. The BLEU score suggests substantial lexical overlap between the generated outputs and reference transcriptions. The high BERTScores (precision, recall and F_1) also demonstrates the model ability to maintain a large degree of semantic consistency between the generated and reference texts. The average disfluency rate for generated disfluent text is 29.1%, which is slightly higher than found in training data, shown in Table 1.

Metric	Score
BLEU Score	0.5524
BERTScore Precision	0.9287
BERTScore Recall	0.9370
BERTScore F1	0.9327

Table 2: Evaluation of the fine-tuned LLM on the task of replicating disfluencies in the Switchboard test set.

5 User Study

To assess the effectiveness of our approach, we conducted a user evaluation study aimed at systematically understanding listeners' perceptions of fluent versus disfluent audio conditions. These conditions were designed to simulate real-life conversational scenarios, allowing us to observe how variations in speech fluency impact listener experience.

5.1 Study Design

In our user study, participants were asked to listen to 10 audio clips, five clips each for disfluent and fluent audios, of simulated conversations. Separate sets of conversation content were used to distinguish between the disfluent and fluent audio conditions. We used ChatGPT to generate ten fictive conversations based on real-life scenarios for human evaluation services, each with 5 turns of dialogues. We inserted disfluencies into five of these conversations using our fine-tuned language model, leaving the remaining five fluent. Both the fluent and disfluent conversations were converted into audio clips via the BARK TTS model. Conversations and detailed instructions can be found in the appendix. Participants were provided with the task to assess each audio recording based on two dimensions:

1. **Intelligibility**: How clear and comprehensible was the spoken conversation?
2. **Spontaneous versus scripted**: Did the conversation sound natural and unrehearsed, as if it were happening in real-time without prior planning? Or did the dialogue sound scripted in advance, similar to dialogue in a movie or play where lines are memorized and performed?

The study participants were asked to rate each clip on intelligibility and spontaneity on a 5-point Likert scale (from worst to best). The study involved 41 participants. They accessed the user study through an online questionnaire filled from their mobile phones (17), laptops (15), tablets (6), or desktop computers (3). 32 participants used headphones. Participation was voluntary, and each participant received a gift card as compensation.

5.2 Results

The average intelligibility ratings from the participants was 4.42 ± 0.79 for fluent utterances and 4.23 ± 1.04 for disfluent utterances. Spontaneity

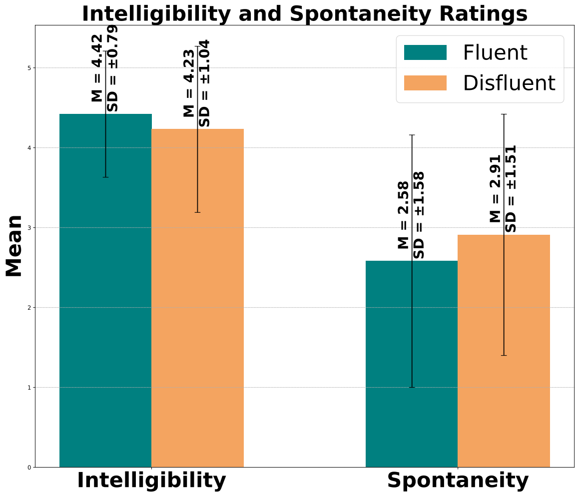

Figure 1: Bar-plot (mean values and 95% conf. intervals) of user ratings on intelligibility and spontaneity.

Metric	Test	Statistic	p-value	GV
Intelligibility	T-Test	$t = 2.03$	0.043	-
	Mixed E.	$z = 2.809$	0.005	0.420
Spontaneity	T-Test	$t = -2.14$	0.033	-
	Mixed E.	$z = -2.615$	0.009	0.805

Table 3: Statistical analysis of perceived intelligibility and spontaneity ratings. GV=Group Variance.

scored an average of 2.58 ± 1.58 for fluent utterances and 2.91 ± 1.51 for disfluent utterances. Figure 1 presents the mean and standard deviation scores for each metric under each condition. Fluent utterances demonstrated higher intelligibility and lower spontaneity compared to disfluent utterances.

The statistical significance of those results was assessed using both independent two-sample T-tests as well as mixed-effects models, with results shown in Table 3. For intelligibility, the t-statistic was 2.03 (p = 0.043), and for spontaneity, the t-statistic was -2.14 (p = 0.033). Both p-values, below the α level of 0.05, indicate statistically significant differences. As participants rated multiple clips in both conditions, we also derived a mixed effect analysis to account for within-subject variability and provide more accurate estimates of condition effects. For intelligibility, the mixed-effects model yielded a z-value of 2.809 (p = 0.005), confirming a statistically significant effect of the condition, consistent with t-test results. The group variance for intelligibility was 0.420, indicating moderate variability across groups. For spontaneity, the model showed a z-value of -2.615 (p = 0.009), again demonstrating a significant effect, with a group variance of 0.805, reflecting greater between-group variability.

6 Discussion

The two questions were sought to answer are whether *intelligibility decrease when we add the disfluencies to the fluent spoken utterances* and whether *perceived spontaneity increase when we add the disfluencies to the fluent spoken utterances*.

The main findings from this user study are that disfluencies significantly enhance the perceived spontaneity of speech, while leading to a slight reduction in intelligibility. The first effect is, however, stronger than the second, as evidenced by the results in Table 3. The statistical results obtained with the mixed-effect model also highlighted the variability within participants' responses. Greater variability in spontaneity ratings suggests that perceptions of speech realism are more subjective and shaped by individual expectations of what constitutes a spontaneous interaction.

As shown in Table 2, the fine-tuned LLM effectively inserts disfluencies while preserving the content of the original utterance. High BERTScore and BLEU scores confirm its outputs align closely with reference transcriptions in the test set. Despite the sporadic nature of disfluencies in spontaneous speech, the model reliably mimics those in the Switchboard dataset. The disfluency rate (29.1%) slightly exceeds the training data (24.5%), indicating a tendency to overproduce disfluencies, which could be adjusted in future work.

7 Conclusion

This paper showed that (1) disfluencies can be automatically inserted into non-disfluent utterances using a fine-tuned LLM, and (2) this insertion enhances the perceived spontaneity of synthetic speech. More precisely, the ratings obtained from a user study revealed a trade-off: utterances in which disfluencies were inserted were deemed as more spontaneous by the participants, but also scored slightly lower on the intelligibility scale.

Those findings are particularly beneficial for the design of LLM-powered conversational avatars, such as those employed to train human professionals to handle "difficult" and stressful dialogues such as investigative interviews, in which disfluencies are particularly prevalent. Future work will focus on fine-tuning disfluency rates for specific scenarios, such as simulating stress or emotional speech in sensitive conversations.

8 Limitations

The outputs of the fine-tuned LLM for inserting disfluencies has a slightly higher disfluency rate compared to the Switchboard training data, indicating a tendency to generate disfluencies slightly more often than in natural human speech. This discrepancy could affect the realism of synthesized conversations in certain contexts. While lexical and semantic alignment metrics (BLEU and BERTScore) demonstrated the fine-tuned model's effectiveness, these measures did not assess the quality of inserted disfluencies, such as their contextual appropriateness and placement.

Although Bark TTS was the most effective TTS model for our study, it occasionally introduced artifacts like buzzing sounds or prolonged pauses, requiring manual intervention to regenerate samples in extreme cases. Additionally, the user study was limited to a small set of fictitious conversations generated by ChatGPT to simulate real-life scenarios. While representative, these scenarios may not fully capture the diversity of real-world conversational contexts where disfluencies occur. Both of these limitations could have influenced participants' perceptions during the study, potentially affecting the results.

9 Ethical Considerations

The use of language models to introduce disfluencies into fluent utterances presents some ethical challenges. One significant concern is the potential for deception: by mimicking human behavior, these models create speech that "sounds more human," potentially tricking the listener into believing they are interacting with a human speaker when they are not. Although our overarching objective is to develop avatars for professional training purposes in which participants are well aware that they are interacting with a virtual agent, deploying these models in contexts where users are not informed of the artificial nature of their interlocutor could lead to ethical concerns about trust, consent, and manipulation.

Additionally, fine-tuning such models risks amplifying biases inherent in the training data, potentially associating disfluency patterns with specific linguistic or socio-demographic groups. There is also the risk that artificial disfluencies might be exploited to manipulate perceptions of a speaker's credibility or emotional state, posing risks in sensitive domains such as legal or professional contexts.

To mitigate these risks, we advocate for clear disclosure policies in any deployment of such models that simulate human speech. Users should be informed whenever they are interacting with an artificial agent, ensuring transparency and preserving trust. Additionally, we are committed to the responsible release of our fine-tuned model and aggregated data, which will only occur after peer review and with accompanying guidelines to minimize potential misuse.

References

Allen D Andrade, Anita Bagri, Khin Zaw, Bernard A Roos, and Jorge G Ruiz. 2010. Avatar-mediated training in the delivery of bad news in a virtual world. *Journal of palliative medicine*, 13(12):1415–1419.

Gunn Astrid Baugerud, Miriam S Johnson, Ragnhild Klingenberg Røed, Michael E Lamb, Martine Powell, Vajira Thambawita, Steven A Hicks, Pegah Salehi, Syed Zohaib Hassan, Pål Halvorsen, et al. 2021. Multimodal virtual avatars for investigative interviews with children. In *Proceedings of the 2021 ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*, pages 2–8.

James Betker. 2023. Better speech synthesis through scaling. *arXiv preprint arXiv:2305.07243*.

Tony W Buchanan, Jacqueline S Laures-Gore, and Melissa C Duff. 2014. Acute stress reduces speech fluency. *Biological psychology*, 97:60–66.

Sasha Calhoun, Jean Carletta, Jason M Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language resources and evaluation*, 44:387–419.

Vuthea Chheang, Shayla Sharmin, Rommy Márquez-Hernández, Megha Patel, Danush Rajasekaran, Gavin Caulfield, Behdokht Kiafar, Jicheng Li, Pinar Kullu, and Roghayeh Leila Barmaki. 2024. Towards anatomy education with generative ai-based virtual assistants in immersive virtual reality environments. In *2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*, pages 21–30. IEEE.

Kevin Charles Dalli. 2021. Technological acceptance of an avatar based interview training application: The development and technological acceptance study of the avbit application. *Digitala Vetenskapliga Arkivet*.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Syed Zohaib Hassan, Pegah Salehi, Ragnhild Klingenberg Røed, Pål Halvorsen, Gunn Astrid Baugerud, Miriam Sinkerud Johnson, Pierre Lison, Michael Riegler, Michael E. Lamb, Carsten Griwodz, and Saeed Shafiee Sabet. 2022. Towards an ai-driven talking avatar in virtual reality for investigative interviews of children. In *Proceedings of the ACM Workshop on Games Systems (GameSys)*, page 9–15.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ksenia Kampf. 2022. Perceived naturalness of typical disfluent speech and speech of adults who stutter. *The Journal of the Acoustical Society of America*, 151(4):A45–A46.

Robin J Lickley and Ellen G Bard. 1998. When can listeners detect disfluency in spontaneous speech? *Language and speech*, 41(2):203–226.

Vishwanath Mantha. 2000. *Mississippi State, Mississippi August 2000*. Ph.D. thesis, Mississippi State University.

MP Marcus, B Santorini, MA Marcinkiewicz, and A Taylor. 1999. Treebank–3 (tech. rep.). *Philadelphia: Linguistic Data Consortium*.

Matthew Marge, Joao Miranda, Alan W Black, and Alexander Rudnicky. 2010. Towards improving the naturalness of social conversations with dialogue systems. In *Proceedings of the SIGDIAL 2010 Conference*, pages 91–94.

Benjamin Marie. 2023. Disfluency generation for more robust dialogue systems. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11479–11488, Toronto, Canada. Association for Computational Linguistics.

Wesley Mattheyses and Werner Verhelst. 2015. Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, 66:182–217.

Marie Meteer et al. 1995. Dysfluency annotation stylebook for the savitchboard corpus. Technical report.

OpenAI. 2023. Open text-to-speech.

Francesco Pompedda, Angelo Zappalà, and Pekka Santtila. 2015. Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality. *Psychology, Crime & Law*, 21(1):28–52.

Hua Xuan Qin and Pan Hui. 2023. Empowering the metaverse with generative ai: Survey and future directions. In *2023 IEEE 43rd international conference on distributed computing systems workshops (ICDCSW)*, pages 85–90. IEEE.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Pegah Salehi, Syed Zohaib Hassan, Saeed Shafiee Sabet, Gunn Astrid Baugerud, Miriam Sinkerud Johnson, Pål Halvorsen, and Michael A. Riegler. 2022. Is more realistic better? a comparison of game engine and gan-based avatars for investigative interviews of children. In *Proceedings of the ACM Workshop on Intelligent Cross-Data Analysis and Retrieval (ICDAR)*, page 41–49.

Raheleh Saryazdi, Daniel DeSantis, Elizabeth K Johnson, and Craig G Chambers. 2021. The use of disfluency cues in spoken language processing: Insights from aging. *Psychology and Aging*, 36(8):928.

Elizabeth Ellen Shriberg. 1994. *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis, Citeseer.

Suno-AI. 2023. Bark text-to-speech.

Nicholas Yan and Gil Alterovitz. 2024. A general-purpose ai avatar in healthcare. *arXiv preprint arXiv:2401.12981*.

Jingfeng Yang, Diyi Yang, and Zhaoran Ma. 2020. Planning and generating natural and diverse disfluent texts as augmentation for disfluency detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1450–1460, Online. Association for Computational Linguistics.

J Scott Yaruss and Robert W Quesal. 2002. Academic and clinical education in fluency disorders: An update. *Journal of fluency disorders*, 27(1):43–63.

Vicky Zayats, Trang Tran, Richard Wright, Courtney Mansfield, and Mari Ostendorf. 2019. Disfluencies and human speech transcription errors. *Interspeech*.

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

A Appendices

Analysing of Simulated Conversations

Thank you for participating in our study! Your feedback is valuable in understanding how effectively and realistically human speech is perceived in various contexts. In this survey, you will listen to several audio recordings of simulated conversations and evaluate them based on specific criteria.

Objectives

Your task is to assess each audio recording based on the following aspects:

1. **Intelligibility**: How clear and comprehensible is the conversation?
2. **Spontaneous versus scripted**:
   1. **Spontaneousness**: This refers to how natural and unrehearsed the conversation sounds, as if it were happening in real time without prior planning.
   2. **Scripted**: This refers to how planned and rehearsed the conversation sounds, similar to dialogue in a movie or play where lines are memorized and performed.

Instructions

1. **Listen Carefully**: For each audio clip provided, please listen to the ***entire*** recording at least once.
2. **Evaluation**: AAfter listening to each audio clip, you will be asked to rate various statements about the clip on a scale of 1 to 5. For intelligibility, a rating of 1 indicates very poor intelligibility, while a rating of 5 indicates excellent intelligibility. For spontaneity versus scripted, a rating of 1 means very spontaneous, whereas a rating of 5 means very scripted.
3. **Technical Requirements**: Ensure you are in a quiet environment and use good quality headphones or speakers for the best listening experience.
4. **Confidentiality**: All responses will be kept confidential and used only for the purposes of this study.

Duration

The entire task is expected to take approximately 15 minutes to complete.

Consent and Participation

Your participation in this study is entirely voluntary, and you may withdraw at any time without penalty. By concluding the evaluation survey, you consent to the use of your responses for research purposes and get a 100 NOK gift card.

Figure 2: Instructions to the user study's participants

Conversation No	Conversation Data
Conversation 1	Speaker 1: Our solution will significantly improve your efficiency. Speaker 2: I'm not sure if this fits within our budget. Speaker 1: We offer a proven track record of success. Speaker 2: Can you provide some case studies? Speaker 1: This investment will pay off in the long run. Speaker 2: What is the expected ROI? Speaker 1: Our team will provide full support during implementation. Speaker 2: What kind of support do you offer? Speaker 1: Let's move forward with this proposal. Speaker 2: I'll need to discuss this with my team.
Conversation 2	Speaker 1: It's been ages since we last met! Speaker 2: I know, time flies! Speaker 1: How have you been? Speaker 2: I've been good, just busy with work. Speaker 1: We should plan a trip together. Speaker 2: That sounds like a great idea, where should we go? Speaker 1: Do you remember our school days? Speaker 2: Of course, those were some of the best days. Speaker 1: Let's not wait too long to catch up again. Speaker 2: Absolutely, let's make it a regular thing.
Conversation 3	Speaker 1: We need to finalize this deal soon. Speaker 2: We're also keen to close this deal promptly. Speaker 1: Our terms are quite clear and non-negotiable. Speaker 2: We can discuss some flexibility in certain areas. Speaker 1: We expect a long-term partnership. Speaker 2: We're committed to a sustainable partnership. Speaker 1: Can we agree on a mutual benefit clause? Speaker 2: Yes, that seems reasonable and beneficial. Speaker 1: What is your final offer? Speaker 2: Here's our final offer, let's seal the deal.
Conversation 4	Speaker 1: Our goal is to inspire change and innovation. Speaker 2: We need to provide concrete data and examples. Speaker 1: We must focus on the big picture. Speaker 2: Details matter when explaining our strategy. Speaker 1: Every challenge is an opportunity. Speaker 2: What are the specific challenges and solutions? Speaker 1: Think about the impact we can make. Speaker 2: We should quantify the potential impact. Speaker 1: Let's engage the audience with our vision. Speaker 2: Our presentation should balance vision with facts.
Conversation 5	Speaker 1: We need to complete this project by Friday. Speaker 2: Understood, I'll prioritize my tasks accordingly. Speaker 1: Everyone must focus on their assigned tasks. Speaker 2: I'll make sure to stay on track. Speaker 1: Report your progress at the end of each day. Speaker 2: I'll update you with my progress daily. Speaker 1: Ensure all deliverables meet our quality standards. Speaker 2: I'll double-check my work for quality. Speaker 1: Let's work together to meet our deadline. Speaker 2: I'll collaborate with the team to ensure success.

Table 4: Fluent conversational data used in the user study

Conversation No	Conversation Data
Conversation 1	Speaker 1: I'm i'm um worried about um m um my child's perf um performance in math. Speaker 2: I understand your concern and let's discuss a a plan. Speaker 1: What what can we do um to improve the erm their grades? Speaker 2: We can provide addi additional practice material and t tutoring sessions. Speaker 1: Are there any additional um re um resources we can use at um at home? Speaker 2: Yes, I I'll send you those some online resources that can can be very helpful. Speaker 1: Is my child erm is participating in class? Speaker 2: Your child is quite en engaged, but we can en encourage more participation. Speaker 1: How can we erm how can we work together erm to um to support my ch um child's learning? Speaker 2: Communication is key let's set up reg regular updates and support.
Conversation 2	Speaker 1: I've been um ha um having these headaches for for weeks. Speaker 2: Let's let's go through your your symptoms in in detail first. Speaker 1: Could it be um could it be something serious um? Speaker 2: It's it's too early to to say, but but we'll we'll investigate thoroughly Speaker 1: Should I get some erm some tests done um. Speaker 2: Yes, we we'll start with with a a few basic tests and and go from there. Speaker 1: What what can I do to to alleviate the the pain? Speaker 2: I'll give you some medication and I'll suggest some lifestyle changes. Speaker 1: Is there a specific um rea reason for these erm these symptoms? Speaker 2: It could be due stress or other factors but we'll find find out
Conversation 3	Speaker 1: Can you um can you erm tell me about your um your previous work experience? Speaker 2: I've I've worked in in similar roles for f five years with with great results. Speaker 1: What makes you a um a good fit for for this role. Speaker 2: I I have the skills and and passion that align with with this company's val values. Speaker 1: How do you um how do you handle tight erm tight deadlines? Speaker 2: I prioritize prioritize tasks and and stay focused under under pressure. Speaker 1: Why do you want erm why do you want to to work with us? Speaker 2: I admire this this company's vision a and want to to contribute to it. Speaker 1: Where erm where do you see yourself in um in five years. Speaker 2: I see myself growing with this company and growing within this company and and taking on more respons responsibilities.
Conversation 4	Speaker 1: How do I im um improve my leadership skills? Speaker 2: Take on m more responsibilities and and seek feedback regularly. Speaker 1: What what should I focus on in in my career development. Speaker 2: Focus on learning continuous learning and networking. Speaker 1: Can you erm can you give me feedback on erm on my recent project. Speaker 2: You did well but there's you're there's room for improv improvement in communication . Speaker 1: How do you handle erm con erm workplace conflicts? Speaker 2: Approach conflicts with with a a calm and solution oriented mindset. Speaker 1: What's the what's the erm best advice y erm you've received in in your career? Speaker 2: Always be open to to learning and and never be afraid to to ask questions.
Conversation 5	Speaker 1: Have you um have you reviewed the the latest project report? Speaker 2: Yes, I- I think we need a a more innova- innovative approach. Speaker 1: The data suggests that a more traditional method is um is effective um. Speaker 2: But creativity could g- give us an edge. Speaker 1: Are there any erm any risks associated with erm you- um your approach ? Speaker 2: Some some, but nothing we can n't handle. Speaker 1: We need to erm con- we need to consider the erm the budget constraints as well erm Speaker 2: I- I believe we can we can be cost- effective and creative. Speaker 1: I'll need to see a to see a detailed plan before. Speaker 2: I'll i'll draft something and and send it over.

Table 5: Disfluent conversational data used in the user study