Assessment of Acoustic Markers of Conversational Difficulty

by

Menatalla Khaled Ellag

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2024

© Menatalla Khaled Ellag 2024

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Human conversations, one of the most complex behaviors, require the real-time coordination of speech production and comprehension, involving cognitive, social, and biological dimensions. There has been a rising need for laboratory and clinical assessments to evolve to capture the essence of everyday interactions. The cognitive demands of interactive conversation, which require listeners to process and store information while simultaneously planning their responses, often exceed those encountered in standard clinical tests. These assessments must encompass diverse contexts and participant groups, including varying hearing statuses, challenging listening environments such as background noise, the use of assistive devices that may alter the listening experience, and different conversation types such as relational versus transactional exchanges, dyadic versus group interactions, and face-to-face versus remote interactions.

This study consists of two investigations exploring how different conditions affect acoustic measures of speech production and conversational behavior. The first study was an extension of a study originally conducted for content analysis and participants' subjective rating questionnaires, focusing on hearing-impaired (HI) individuals. It examined the impact of face masks and remote microphones on communication dynamics. Four native English-speaking HI participants engaged in free-form conversations within small groups under a constant background noise of 55 dBA. Interestingly, the results showed that using remote microphones shortened floor-transfer offsets (FTOs) and extended conversation durations, suggesting improved communication. When participants did not wear a face mask, interpausal unit (IPU) durations were shorter with remote microphones than without, indicating easier communication. However, no significant difference was found between the two mask conditions, suggesting that face masks affect both speech perception and production by decreasing inhalation and exhalation volumes, thereby limiting the duration of utterances. Face masks are speculated to increase resistance to airflow, reducing subglottal pressure and consequently lowering fundamental frequency (F0). Despite no significant differences in articulation rate and floor transfer rate, the constant noise environment, presented at lower levels compared to previous studies, may have likely limited the potential for pronounced effects.

The second study involved normal-hearing (NH) individuals, investigating the effects of conversation type (free-form vs. task-based) and noise presence (70 dB SPL) on conversational dynamics. Dyadic interactions among NH participants were examined. Task-based conversations exhibited structured patterns with longer FTOs and higher floor transfer rates, while free-form conversations showed greater FTO variability, more frequent overlaps, longer IPUs, and increased pause durations and rates. Noise presence increased IPU durations and pause lengths but did not significantly alter floor-transfer rates or FTO variability. Both conversation types experienced increased articulation rates and speech levels in noise. Contrary to the expected change as part of the Lombard effect, the increase in articulation rates may be attributed to the noise acting as a stressor. Meanwhile, the increase in mean speech levels was less pronounced than expected, possibly due to the specific noise characteristics and the use of closed headphones.

These studies shine a light on the complexity of communicative interactions and the necessity of accounting for a wide spectrum of factors in experimental designs. The findings highlight the importance of considering both environmental conditions and conversation types when researching speech perception, production, and conversational dynamics. This research provides valuable insights for academic studies and the development of hearing-assistive technologies, emphasizing the need for assessments that reflect the varied nature of everyday communication.

Acknowledgments

Coming into this program, I was advised by many colleagues that the road ahead would be quite challenging due to the rigorous nature of academia. Yet, two years later, I stand here saying that it was a smooth sail.

I strongly believe this was due to having Dr. Ewen MacDonald as my supervisor. I am profoundly grateful to Dr. MacDonald, who has been not just an exceptionally knowledgeable supervisor but also an outstanding teacher and mentor. Your passion and dedication have made this journey enjoyable, making me feel like I was in an academic playground. Thank you for consistently making yourself available whenever I needed guidance and for being fully present, providing the time and attention essential for my growth as a student. Thank you for giving me the freedom to explore and learn from my mistakes, keeping my learning curve pointing upward all the time, and this thesis is the proof! I will always be grateful for this opportunity and your mentorship.

My deepest thanks go to my readers, Dr. Katherine White and Dr. Nima Maftoon, for their time, effort, and thoughtful feedback. Your insightful suggestions and comments have significantly contributed to the refinement of this thesis.

I would like to express my gratitude to Dr. Jinyu Qian, Dr. Ieda Ishida, Dr. Kathy Pichora-Fuller, and Dr. Lisa Dillon-Edgett for their valuable insights and advice on the Sonova experiment. I am thankful to Dr. Jinyu Qian for bringing the collaboration between Sonova and the Hearing Lab to life and to Dr. Ieda Ishida for her constant support and assistance over the past two years.

My gratitude extends to Kate Avison and Aric Quan for their assistance with the data collection, and to my lab mate Ben Masters for his help throughout this journey.

I am deeply appreciative of my family and friends for their unparalleled support and the continuous encouragement they provided, which fueled my perseverance.

Lastly, I am eternally grateful for the blessings and privileges that God has bestowed upon me, making this journey possible.

Dedication

This is dedicated to Khaled, Eman, and Ahmed—my father, mother, and brother—who mean the world to me.

Table of Contents

Author's Declaration **ii**

Abstract **iii**

Acknowledgments **v**

Dedication **vi**

List of Figures **xi**

List of Tables **xiii**

List of Abbreviations **xv**

1 Introduction **1**

1.1 Timing of Turn-Taking and Cues of Turn End 2

1.2 Factors Affecting Measures of Speech Production and Conversational Dynamics 4

1.2.1 Thesis Structure 7

2 The Impact of Remote Microphones and Face Masks on Group Conversations Between Hearing-Impaired Individuals 8

2.1 Introduction . . . 8

2.1.1 Effect of Face Masks on Communication . . . 10

2.1.2 Hearing Aids and Remote Microphone Technologies . . . 12

2.1.3 Changes in Behavior Associated with Conversational Difficulty . . . 12

2.2 Methods . . . 13

2.2.1 Participants . . . 13

2.2.2 Experimental Setup . . . 13

2.2.3 Recording Processing and Data Extraction . . . 14

2.2.4 Statistical Analysis and Visualization . . . 16

2.3 Results . . . 17

2.3.1 Rate of Floor Transfers . . . 17

2.3.2 Floor-Transfer Offset . . . 17

2.3.3 Interpausal Units . . . 20

2.3.4 Pause Duration . . . 23

2.3.5 Articulation Rate . . . 23

2.3.6 Fundamental Frequency . . . 24

2.3.7 Conversation Duration . . . 25

2.4 Discussion . . . 26

2.5 Conclusion . . . 30

3 The Effect of Task on Speech Production and Communication Behavior During Conversation in Healthy-Hearing Participants 32

3.1 Introduction . . . 32

3.2 Methods . . . 35
3.2.1 Participants . . . 35
3.2.2 Experimental Setup . . . 35
3.2.3 Procedure . . . 36
3.2.4 Recording Processing and Data Extraction . . . 36
3.2.5 Statistical Analysis . . . 38
3.3 Results . . . 39
3.3.1 Rate of Floor Transfers . . . 39
3.3.2 Floor-Transfer Offset . . . 40
3.3.3 Interpausal Units . . . 45
3.3.4 Pause Duration and Rate . . . 49
3.3.5 Articulation Rate . . . 54
3.3.6 Speech Levels . . . 55
3.3.7 Speaking Time . . . 57
3.4 Discussion . . . 58
3.4.1 Rate of Floor Transfers . . . 59
3.4.2 Floor-Transfer Offset . . . 60
3.4.3 Interpausal Units . . . 61
3.4.4 Pause Duration and Rate . . . 61
3.4.5 Speech Production . . . 63
3.4.6 Evaluation of Experimental Setup: Assessing the Adequacy of Communication Challenges . . . 64
3.5 Conclusion . . . 65

4 Thesis Summary and Future Work **66**

References **68**

APPENDICES **81**

A Appendix **82**

A.1 Supplementary Information 82

A.1.1 Rate of Floor Transfers 82

A.1.2 Floor-Transfer Offsets 83

A.1.3 Interpausal Units 84

A.1.4 Pause Duration and Rate 85

A.1.5 Articulation Rate 86

A.1.6 Speech Levels 87

A.1.7 Speaking Time 88

List of Figures

2.1 Illustration of experimental setup. 14
2.2 Schematic of communicative state, silence, and overlap classification. . . . 15
2.3 Distributions of FTO. 18
2.4 Estimated marginal means of FTO. 19
2.5 Distributions of the IPU durations. 21
2.6 Estimated marginal means of IPU duration. 22
2.7 Estimated marginal means of the fundamental frequency. 24
2.8 Estimated marginal means of group conversation duration. 25

3.1 Schematic of communicative state, silence, and overlap classification. . . . 37
3.2 Estimated marginal means for floor transfer rate. 40
3.3 Distributions of floor transfer offset. 41
3.4 Estimated marginal means of floor-transfer offset. 43
3.5 Estimated marginal means of floor-transfer offset interquartile range. . . . 44
3.6 Distributions of IPU durations. 46
3.7 Estimated marginal means of IPU duration. 47
3.8 Estimated marginal means of number of syllables per IPU. 48
3.9 Distributions of pause duration. 50

3.10 Estimated marginal means of pause duration. 51
3.11 Estimated marginal means of pause rate 53
3.12 Estimated marginal means of articulation rate. 55
3.13 Estimated marginal means of estimated speech levels. 56
3.14 Estimated marginal means of the proportion of time speaking 58

List of Tables

2.1 Fixed effects of GLMM of FTO. 20

2.2 Fixed effects of GLMM of IPU duration. 22

3.1 Fixed effects of GLMM of FTO. 43

3.2 Fixed effects of GLMM of IPU duration. 47

3.3 Fixed effects of GLMM of number of syllables per IPU. 49

3.4 Fixed effects of GLMM of pause duration. 51

A.1 Fixed effects of LMM of floor transfer rate. 82

A.2 Pairwise comparisons of floor transfer rates. 82

A.3 Pairwise comparisons of mean FTO. 83

A.4 Fixed effects of LMM of FTO IQR. 83

A.5 Pairwise comparisons of mean FTO IQR. 83

A.6 Pairwise comparisons of mean IPU duration. 84

A.7 Pairwise comparisons of mean number of syllables per IPU. 84

A.8 Fixed effects of GLMM of pause duration for Free-Form Conversation. . . . 85

A.9 Fixed effects of GLMM of pause duration for Task-Based Conversation. . . 85

A.10 Pairwise comparisons of mean pause durations. 85

A.11 Fixed effects of LMM of pause rates. 86
A.12 Pairwise comparisons of mean pause rates. 86
A.13 Fixed effects of LMM of articulation rates. 86
A.14 Pairwise comparisons of mean articulation rates. 87
A.15 Fixed effects of LMM of estimated speech levels. 87
A.16 Pairwise comparisons of mean speech levels. 87
A.17 Fixed effects of GLMM of the proportion of speaking time. 88
A.18 Pairwise comparisons of mean speaking time. 88

List of Abbreviations

F0 Fundamental Frequency

FTO Floor-Transfer Offset

HI Hearing-Impaired

IPU Interpausal Unit

NH Normal-Hearing

SNR Signal-to-Noise Ratio

Chapter 1

Introduction

Human conversations, considered one of the most complex and dynamic behaviors on Earth, require the ability to both comprehend and produce speech in real-time (Stangor and Walinga, 2014). This complex process involves cognitive, social, and biological dimensions, such as the coordination of vocal cords and the regulation of breath, throat, mouth, and tongue movements (Stangor and Walinga, 2014; Federmeier et al., 2020). Additionally, speech is intricately accompanied by vocal and gestural non-verbal signals that significantly influence meaning and emphasis, underscoring the multifaceted nature of effective communication (Argyle et al., 1970).

To accurately reflect real-world speech intelligibility, laboratory and clinical assessments must evolve. Developing tasks that more closely mimic everyday communication scenarios is crucial (Miles et al., 2022). The need for such representative tests arises from the tendency to overestimate real-world performance using traditional methods (Gifford et al., 2013). Moreover, the cognitive demands of interactive conversation, which require listeners to process and store information while simultaneously planning their responses, often exceed those encountered in standard clinical tests (Edwards, 2007; Sweetow and Henderson-Sabes, 2004). For hearing aid users to benefit fully, evaluations must occur not only in clinical settings but also reflect real-life conditions, ensuring that laboratory scenarios capture the essence of everyday interactions (Smeds et al., 2020).

Effective communication depends on the integration of verbal and non-verbal cues,

which are essential for managing turn-taking, maintaining connection, and conveying meaningful messages. Understanding these elements is particularly important in challenging communication settings, such as noisy environments or for individuals with hearing impairments. This thesis investigates these challenges by exploring how various conditions impact speech acoustic markers and conversational dynamics. The following sections elaborate on important metrics established in the literature to assess conversational difficulty.

1.1 Timing of Turn-Taking and Cues of Turn End

Turn-taking behaviors in conversations involve complex processes encompassing syntactic structure, intonation, semantic cues, and even physiological factors like breathing patterns. These elements interact dynamically to facilitate smooth and efficient communication among interlocutors. Sacks et al. (1974) emphasized the crucial nature of turn-taking in speech exchanges. They highlight that interlocutors tend to respond coherently and appropriately to each other, without leaving long gaps between contributions or engaging in simultaneous speech for extended periods. Moreover, they delineate various turn-constructional components, including sentential, clausal, phrasal, and lexical constructions, which speakers use to construct turns. These unit types allow for the projection of how long it will take to complete an instance of a particular unit type, contributing to the smooth flow of conversation.

Additionally, turn-allocation techniques can be categorized into those where the next turn is allocated by the current speaker selecting the next speaker and those where the next turn is self-selected. The turn-taking system in conversation inherently requires active listening to all utterances, irrespective of individual motivations such as interest or politeness. This listening is essential for potential speakers to analyze each utterance delivery, determining whether they are selected as the next speaker. Thus, participants must listen attentively to each utterance, potentially intending to speak next if selected. One plausible technique by which speakers comprehend the end of their interlocutor's turn was discussed by Magyari and de Ruiter (2012) in their button-press experiment. In this experiment, the precision of their anticipation of the length and conclusion of a speaker's turn was based on

predicting how it ends through anticipation of words and syntactic frames. De Ruiter et al. (2006) delved into the role of intonational contour and lexico-syntactic cues in anticipating the end of a turn. Their experiment reveals that participants are able to anticipate turn endings, relying on both intonation contour and lexico-syntactic information. However, the accuracy of anticipation decreases significantly when either cue is absent, highlighting the importance of syntactic structure in projecting turn endings.

Whether individuals are engaged in face-to-face conversations or communicating from different locations, such as during a telephone call, numerous studies have observed that there is a slight delay, known as the floor-transfer offset (FTO), between when one person finishes speaking and the next person begins. This interval typically falls within a range of about 200 milliseconds, based on various investigations (Aubanel et al., 2011; Ferreira, 1991; Foti and Roberts, 2016; Godfrey et al., 1992; Heldner and Edlund, 2010; Indefrey, 2011; Levinson and Torreira, 2015; Meyer, 2023; Roberts et al., 2015; Stivers et al., 2009). Interestingly, this timing pattern transcends linguistic and cultural boundaries suggesting that individuals prioritize maintaining socially appropriate timing for responses, even if it requires additional cognitive effort (Barthel and Sauppe, 2019; Stivers et al., 2009).

Other research efforts provide further insights into turn-taking dynamics. Riest et al. (2015) suggested that listeners utilize semantic information to predict the content of a speaker's turn, aiding in anticipation. Levinson and Torreira (2015) shed light on the intricacies of breathing patterns in speech production, indicating that participants often plan their responses during the preceding speaker's turn, with inhalation typically initiated shortly before the anticipated end of the current speaker's turn. During the preparatory phase for verbal expression, the coordination of various speech organs is essential. This includes the initiation of breathing necessary for speaking (Whalen and Kinsella-Shaw, 1997; Winkworth et al., 1995). Previous research has identified several time delays involved in this breathing process: first, a delay of approximately 140–320 milliseconds occurs from the moment the decision to inhale is made until the signal reaches the intercostal muscles (Draper et al., 1960); second, spontaneous dialogue typically involves inhalation times exceeding 500 milliseconds (McFarland, 2001). These findings collectively suggest a pre-speech latency of at least 500–800 milliseconds (Heldner and Edlund, 2010).

Research indicated that the process of preparing and delivering speech takes time, with

estimates ranging from at least 600 ms for a single word to over one second for longer utterances (Indefrey and Levelt, 2004; Magyari et al., 2014). This means that to respond verbally within 200 milliseconds after the other person stops speaking, one must anticipate their partner's speech content and begin formulating a response while the conversation is still ongoing. Failure to accurately interpret cues signaling the end of a turn could lead to premature or delayed responses, resulting in more variable turn-taking timings, FTO, without necessarily altering the average FTO duration (Barthel and Sauppe, 2019; Bögels et al., 2015; Bögels and Torreira, 2021; Corps et al., 2018; Gisladottir et al., 2015; Levinson and Torreira, 2015).

Investigations by Torreira et al. (2015) into conversational breathing patterns have highlighted distinctions between responses to short and long questions. Short responses often rely on residual lung air, while longer responses necessitate planned inhalation. Inhalation typically commences shortly (around 15 milliseconds) after the cessation of the interlocutor's question, potentially triggered just prior to the retrieval of the first word of the planned response. Thus, the data on breathing behavior suggests that the decision to inhale is a component of conceptual planning for the response. Moreover, the trigger for inhalation, usually occurring within the last few hundred milliseconds of the interlocutor's turn, is often based on an anticipation of the imminent completion of the speaker's turn. Hence, significant variability is observed in answer latencies, with shorter response times when preceded by an in-breath (Whalen and Kinsella-Shaw, 1997). This suggested that listeners may start speaking closer to the estimated end of the preceding turn, potentially influenced by predictions regarding the imminent conclusion of the current speaker's turn.

1.2 Factors Affecting Measures of Speech Production and Conversational Dynamics

The way by which speech is articulated has been confirmed by many researchers to differ by many factors including age, sex, hearing status, and listening environment (Bóna, 2014; Hartman and Danhauer, 1976; Jacewicz et al., 2010; Petersen et al., 2022; Winn and Teece, 2021; Yoon, 2004).

A study by Bóna (2014) conducted on young and older speakers found that speech rate (the number of syllables divided by total time, including pauses), articulation rate (the number of syllables divided by phonation time, excluding pauses), and pause frequency significantly differ across various types of conversations (i.e., narrative recall, three-person conversation, spontaneous narrative). The research showed that older participants exhibited temporal changes, including slower speech and articulation rates, as well as a higher frequency of pauses. This may be attributed to changes in cognitive function, aging of speech organs, and deterioration in hearing health. This was accompanied by other researchers' work that found that older adults tend to use more pauses, filler words (as an alternative to pauses), and disfluencies, which may all be related to challenges in lexical access (Bóna, 2014; Bortfeld et al., 2001; Schmitter-Edgecombe, 2000). It is well-established in the literature that the Lombard reflex, established by Etienne Lombard (1911), is observed in speakers altering their speech when noise is present in the background. To ensure effective communication despite background noise, speakers carefully gauge their speech by raising their intensity and fundamental frequency, as well as elongating segmental durations. This adaptation helps to enhance the signal-to-noise ratio (SNR), meeting both communicative needs and self-monitoring processes. The involuntary adjustment of speech level in response to noisy environments is referred to as the Lombard effect or Lombard reflex (Vainio et al., 2012). Early studies by Hanley and Steer (1949) or Summers et al. (1988) found that when talkers are subjected to masking noise in the background, their speech possesses altered segmental properties where they speak at a slower rate along with an increased utterance duration. Another study explored the impact of hearing impairments on speech rates and pauses and found that hearing-impaired (HI) speakers speak at a slower articulation rate, with higher pause duration and frequency (Yoon, 2004). This is accompanied by numerous research showing that hearing-impaired individuals prefer slower speech rates, which improves speech intelligibility by easing their listening effort (Gordon-Salant and Fitzgibbons, 1997; Lessa and Costa, 2013; Winn and Teece, 2021; Beechey et al., 2020).

Petersen et al. (2022) have explored a combination of the markers previously discussed by studying conversations between normal-hearing (NH) and older HI interlocutors in quiet and with 70 dBA babble background noise. This included observing changes in

inter-pausal units (IPU), which are chunks of connected speech that are longer than 180 ms, and surrounded by pauses (i.e., longer than 90 ms). They found that both NH and HI participants spoke more slowly in noise and that hearing aid amplification increased the articulation rate in HI individuals. When individuals speak slower (i.e., slower articulation rate), IPUs are expected to increase as the utterance duration also increases (Hanley and Steer, 1949; Petersen et al., 2022; Summers et al., 1988). Multiple research efforts have found that longer IPUs are associated with longer utterances which may be either filler words and disfluencies due to the time it takes to prepare an adequate response when subjected to speech in noise, or slower articulation rate (Beechey et al., 2018; Sørensen et al., 2021; Watson et al., 2019). This may be attributed to the cognitive resources that are recruited to spend extra time on understanding speech, and hence, delaying the planning of a response which usually happens before the other talker ends their turn. Hence, it becomes helpful for talkers to pick up the floor and use filler words until a complete response is prepared to be articulated (Sørensen et al., 2021). Moreover, Sørensen et al. (2020) studied the effect of noise on dyadic, task-based interactions in two different experiments and found that in a 70 dBA babble noise, interlocutors decrease their rate of floor transfers (FT).

Fundamental frequency, F0, is the frequency at which vocal folds vibrate to produce speech signals by modulating airflow through the glottis, propagating sound through the vocal tract (Hee Lee and Humes, 2012; Zhang, 2016b). Selective modification can be done to amplify or attenuate frequencies, which results in perceptible contrasts that are used to convey a wide spectrum of various linguistic sounds and meanings (Zhang, 2016b). Adult F0 ranges between 80 – 175 Hz for men and 160 – 270 Hz for women (Fouquet et al., 2016). Many previous research studies have shown talkers' mean F0 to be used as a measure of vocal effort (i.e., Lombard effect) which is also reflective of vocal fatigue. It was shown that F0 increased in challenging listening environments (e.g., high noise, far distance, interlocutor hearing status, etc.), which is one of many parameters used by speakers to adapt to challenging listening environments where they exert more effort to deliver their speech (Beechey et al., 2018; Hazan and Baker, 2011; Jessen et al., 2005; Liénard and Di Benedetto, 1999; Vlaj and Kacic, 2011). F0 may be affected by other factors that are not related to listening effort but speech production. One aspect is lung

volume and the depth of the inhalation and exhalation and its effect on subglottal pressure (i.e., provides the driving pressure for vibration of the vocal cord and glottal airflow) (Fuchs and Rochet-Capellan, 2021; Sundberg et al., 2013; Traser et al., 2020). Subglottal pressure is characterized by the relaxation pressure proportionally related to lung volume (Zhang, 2016a). Hence, when breathing volume decreases, it results in a decrease in the subglottal pressure and glottal leakage (Iwarsson et al., 1998; Zhang, 2016a). Lieberman (1966) found a direct consequence of this decrease in subglottal pressure, where it resulted in a decrease in F0, as well. Watson et al. (2003) also studied the impact of breathing on the speech's acoustic properties and found that lower lung volumes resulted in a steeper decrease in F0.

1.2.1 Thesis Structure

This thesis provides a comprehensive analysis of the impact of face masks and remote microphones on conversations involving hearing-impaired individuals, and the effect of tasks on speech production and communication behavior in healthy-hearing participants.

Chapter 2 explores the experiment conducted to study the impact of face masks and remote microphones on group conversations involving hearing-impaired individuals. This chapter details the methodology, data collection, and analysis of the speech and communicative behaviors observed under these variables.

Chapter 3 investigates the effect of tasks on speech production and communication behavior during conversations in healthy-hearing participants. This chapter details the methodology, data collection, and analysis of the speech and conversational dynamics for both conversation types in quiet and noise.

Chapter 4 summarizes the findings from the previous chapters and recommends ideas for future work.

Chapter 2

The Impact of Remote Microphones and Face Masks on Group Conversations Between Hearing-Impaired Individuals

2.1 Introduction

Conversations involve a highly dynamic interaction of speech comprehension and production (Petersen et al., 2022). To achieve the timing of turns typical in conversation, it is necessary for interlocutors to conduct processing related to speech perception, speech production, and predicting the timing of the end of a partner's turn (Levinson and Torreira, 2015).

Increased cognitive effort is required when speech is degraded by noise or hearing impairment. Preparing responses while simultaneously comprehending incoming speech is cognitively demanding because production and comprehension share the same neurological resources (Barthel et al., 2016; Menenti et al., 2011; Segaert et al., 2012). Real-time speech comprehension often relies on predicting the next word; when actual words do not match

expectations, additional cognitive resources are needed to support ongoing interpretation (Hagoort and Indefrey, 2014). According to the Ease of Language Understanding (ELU) model, when sensory input is degraded by hearing loss or noise interference, perceived phonological input is less likely to match stored representations in long-term memory, necessitating a shift from automatic to explicit processing to infer meaning (Rönnberg et al., 2008, 2013).

This increased effort can delay response preparation in conversations, as critical information may be perceived later, requiring reliance on top-down information to understand speech. As a result, speech planning may start later and be more variable in difficult listening conditions, leading to delayed and less precise turn-taking (Barthel et al., 2016). Noise can also capture the locus of attention, diverting it from the primary task (i.e., visual task material) to the background sound, thereby interrupting task performance (Sorqvist, 2015).

Recent studies have considered interactive conversation as a divided listening task and investigated changes in acoustic measures related to speech production and conversational behavior in response to more difficult communication conditions (Sørensen et al., 2020; Petersen et al., 2022). In these studies, both NH and HI interlocutors exhibited longer and more variable FTO durations, longer IPUs, and spoke louder. HI participants showed more variability, longer IPUs, and faster speech rates when using their hearing aids. In the presence of noise, NH individuals altered their conversational dynamics more when speaking to HI partners, suggesting that NH interlocutors adapt their speech to assist HI individuals by conversing at a positive SNR and slowing down their speech rate.

In a study by Ishida et al. (2022), groups of four aided hearing-impaired participants interacted in free conversation that flowed from an initial topic provided by a moderator. The conversations were collected in four conditions of varying difficulty (combinations of whether a remote microphone system was being used or not and whether the participants wore face masks or not) and transcriptions of the conversations were analyzed to examine the rate at which participants changed subjects/topics within the conversation. Thus, the experimental paradigm in that study had been developed with a focus on analyzing the content and eliciting conversation in a group setting where participants could freely adjust their hearing aids.

In the paradigms of Sørensen et al. (2020) and Petersen et al. (2022), participants conducted the Diapix task (a spot the difference task) in pairs. The dyadic task-oriented conversations elicited in these studies are likely to differ from those of free conversations. In the Diapix task, the content is likely to be more constrained and the style may be more transactional than relational. To reduce listening effort in more difficult communication situations, an interlocutor could engage in maladaptive strategies such as dominating a conversation (i.e., by holding the floor more than others) or withdrawing from speaking in the conversation (Demorest and Erdman, 1987). However, dyadic task-based conversations may provide little opportunity to engage in these strategies compared to free conversations in small groups. Thus the goal of the present study was to analyze the recordings of Ishida et al. (2022) to investigate whether the changes to speech production and conversational behavior associated with increased communication difficulty that has been observed for HI participants in task-based dyadic conversations are also observed in free conversation in small groups.

Based on previous research, conditions where participants wore face masks should have experienced more difficulty communicating due to the effect of the masks concealing visual cues and acoustically altering the speech signals. Conversely, remote microphones offer assistance that eases communication and listening efforts by improving the signal-to-noise ratio (SNR). Below we elaborate on how wearing a face mask and using the remote microphone can influence the difficulty experienced during the conversation.

2.1.1 Effect of Face Masks on Communication

In face-to-face interactions, visual signals, including hand gestures, are crucial for human communication (Kendon, 2004; McNeill, 1995). Speech perception is worse in auditory-only conditions compared to auditory-visual conditions (O'Neill, 1954). Visual cues become even more important in noisy environments, where speech perception suffers more under auditory-only conditions as noise increases (Sumby and Pollack, 1954). In such environments, talkers naturally enhance visual communication through changes in gestures and mouth movements to compensate for compromised auditory input (Trujillo et al., 2021).

Research shows that in noisy settings, speakers increase speech intensity, make larger

mouth movements, and use more pronounced gestures to help the listener follow the conversation despite auditory challenges (Kim et al., 2005; Davis et al., 2006; Junqua et al., 1999; Garnier et al., 2010). These visual cues are strategically designed to aid the listener by enhancing the visual aspects of speech, such as lip movements and hand gestures.

The integration of auditory and visual signals underscores the multi-modal nature of human communication. These adaptations indicate that the visual channel carries a significant communicative burden, especially in noisy environments, ensuring that the message is successfully conveyed (Perniss, 2018; Levinson and Holler, 2014; Trujillo et al., 2021). By relying on both speech and visual cues, speakers can adapt to noisy environments, highlighting the importance of considering both auditory and visual adaptations in studies of communication.

Given the challenges introduced by the COVID-19 pandemic, the impact of face masks on communication received increased interest. These and previous studies have observed several effects. Cloth face masks have been found to increase respiratory muscle effort, decrease ventilatory volume, and reduce respiratory rates at rest in healthy individuals (Engeroff et al., 2021; Marek et al., 2023; Ogawa et al., 2022). Numerous studies have observed that face masks acoustically filter speech, dampening higher frequencies (typically above 1 or 2 kHz). However, the specific frequency range and the degree of attenuation vary depending on the type of face mask being used (Badh and Knowles, 2023; Corey et al., 2020; Rahne et al., 2021). Reduced audibility of higher frequencies can make it more difficult to distinguish between some consonants, such as /t/,/d/,/k/,/z/, and/s/, hindering speech perception. Masks can occlude visual cues from lip and mouth movements, further hindering speech perception (Atcherson et al., 2017; Bernstein et al., 2022; Thibodeau et al., 2021).

Choi et al. (2022) observed that the extent of the decline in speech perception resulting from the presence of the mask (measured as the difference between speech perception scores in the no-mask–auditory-visual and mask–auditory-visual conditions) was notably higher for individuals with hearing impairment compared to those with normal hearing, but this difference was only significant at 0 dB SNR. McKenna et al. (2022) also found that face masks impose challenges in communication because of their impact on "spectral characteristics of speech, vowel space measures, and vocal effort". Furthermore, the benefit

provided by hearing aids, measured as the difference between speech perception scores without and with aids under different SNRs, was significantly higher at 0 dB and -4 dB SNR compared to -16 dB SNR in the mask–auditory-visual group. These results highlight the effectiveness of hearing aids in enhancing speech comprehension when participants wear masks in low-noise environments.

2.1.2 Hearing Aids and Remote Microphone Technologies

By amplifying sound, hearing aids can improve the audibility of speech. However, in noisy environments, hearing aids will also amplify background sounds. Remote microphones can be used to improve the signal-to-noise ratio (SNR). Sound levels decrease with distance from a sound source. Assuming the background noise is relatively diffuse (i.e., relatively constant across positions), placing a microphone closer to the source of interest will result in a sound mixture where the signal is higher than that from the microphones on a hearing aid, where the microphones are further away.

However, a remote microphone can only improve the received SNR of talkers near which it is placed. Thus, in complex environments with multiple talkers, the benefits of a remote microphone may be reduced (Wagener et al., 2018).

2.1.3 Changes in Behavior Associated with Conversational Difficulty

Previous research has shown that various conversational metrics change under conditions that increase communication difficulties, such as with background noise or hearing impairment (Sørensen et al., 2021, 2020; Petersen et al., 2022). Across studies, it has been observed that floor transfer offsets (FTOs), interpausal units (IPU), and pause durations tend to be longer when communication is more challenging. Further, the rate of floor transfers is typically lower in difficult conditions and a decrease in articulation rate has been observed in some cases. In noise, talkers increase vocal effort leading to increased average fundamental frequency (F0).

Building on these findings, we hypothesize that using a face mask will increase the difficulty of conversation, resulting in longer FTO, IPU, and pause durations, a lower rate of floor transfers, a slower articulation rate, and a higher F0. Conversely, using a remote microphone is expected to decrease conversational difficulty, leading to opposite effects. Finally, we hypothesize that easier conversations are likely to last longer than those where talkers are experiencing more difficulty interacting with each other.

2.2 Methods

2.2.1 Participants

The study by Ishida et al. (2022) involved 16 native English-speaking individuals (5 females, 11 males). All participants were older adults with hearing loss, grouped into groups of four (mean age = 71.5 years, sd = 6.3 years). The participants consisted of five females and 11 males. Participants were randomly assigned into groups of four, without consideration of sex, age, or degree of hearing loss.

2.2.2 Experimental Setup

Participants were seated at separate tables as illustrated in Figure 2.1, such that the minimum distance between participants was approximately 2 m. Two remote microphones were set up between pairs of participants such that each participant was approximately 1 m from the microphone nearest to them. A third remote microphone was placed on the table where the mediator sat. The three remote microphones were networked together in a manner that simulated the processing of wireless Roger^{TM} hearing aid microphones. Loudspeakers were used to play babble noise in the background at 55 dBA. The participants wore hearing aids and were free to adjust them at any time during the experiment to their preference. The order of the conversations for each condition was also randomized across the groups. The mediator provided a topic at the start of every conversation (i.e., "hobbies," "food," "seasons," and "pets") to initiate the conversation, and participants

were free to shift to other topics as they liked. Conversations ended when the mediator found no further input from all participants. Participants were instructed they could speak in any order.

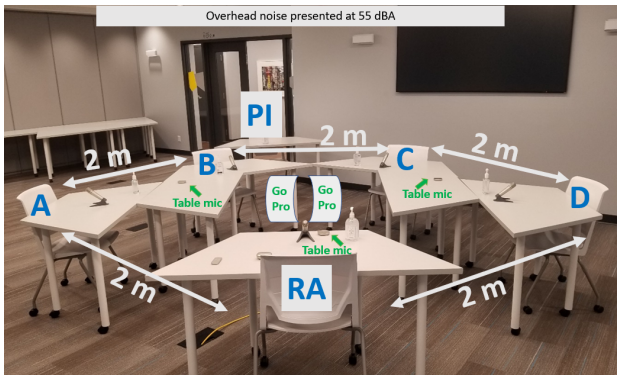


Figure 2.1: Experimental setup with two within-subject variables, Roger microphone and face mask (Ishida et al., 2022). The labeled components include participant labels (A-D), the principal investigator (PI), the research assistant (RA) who facilitated the conversation, and the camera used to record the experiment (Go Pro).

2.2.3 Recording Processing and Data Extraction

Using the transcriptions made for Ishida et al. (2022), the Montreal Forced Aligner (McAuliffe et al., 2017) was used to generate initial estimates of when each participant spoke in a conversation. The estimates were then reviewed and updated by hand and padded to include ∼15 ms before and after talkers' turns/floors. Notably, in cases of speaker overlaps or very short FTOs, padding was minimized or omitted to prevent potential interference or inaccuracies. Additionally, instances of standalone coughs, unidentifiable utterances, or brief laughter were excluded from labeling.

Using these adjusted timestamps, a four-channel recording, where each channel contained speech from only one talker, was created from the single-channel recording of the conversation. Subsequently, this four-channel recording was processed using a VAD script to determine when each participant spoke. The script calculated the power in 5 ms windows with a 4 ms step size. In each channel, a threshold of 30 dB below the maximum

observed power was used to generate a binary vector indicating the presence/absence of speech in each 5 ms window. This binary vector was further processed in a manner similar to Levinson and Torreira (2015). Intervals of silences shorter than 180 ms, which were assumed to correspond with stop consonants, were bridged (i.e., short intervals of 0 were set to 1). Intervals of speech activity that were shorter than 90 ms were set to 0. The processed binary vectors of speech activity from each participant were combined into a matrix. This binary matrix was then processed to identify and record details regarding IPUs, FTOs, pauses, and conversation duration according to the schematic in Figure 2.2.

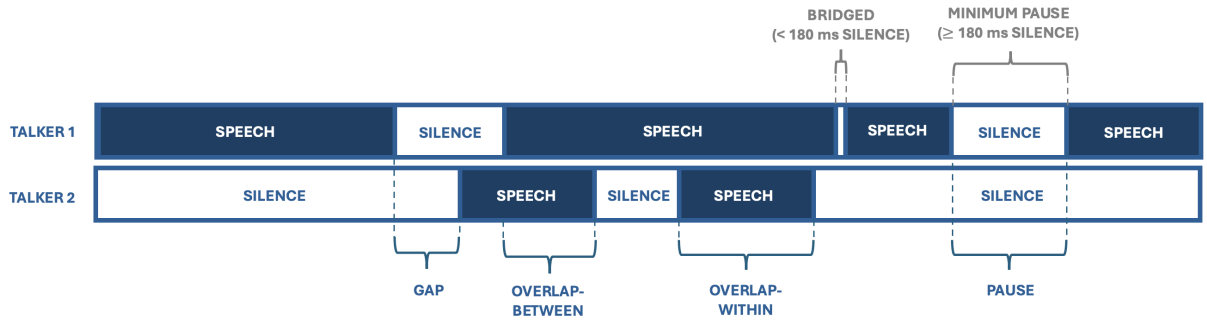


Figure 2.2: Schematic of communicative state, silence, and overlap classification, based on the procedure described by Heldner and Edlund (2010).

The detailed classification in Figure 2.2 is as follows:

1. **Gaps (FTOs):** The interval from when one talker that cedes the floor stops speaking to when the talker that takes over the floor starts talking. Positive FTOs manifest as a Gap (i.e., an interval of silence) while negative FTOs manifest as an Overlap-Between (i.e., there is an acoustic overlap of the two talkers during a floor transfer).

2. **Overlaps-Within:** Speech from one talker completely overlapped by the speech from the other without a floor transfer.

3. **Interpausal Units (IPUs):** Units of connected speech, including silences shorter than 180 ms.

4. **Pauses:** Silence from both talkers without a floor transfer.

5. **Turns:** Sequential IPUs produced by one talker, bounded by floor transfers.

2.2.4 Statistical Analysis and Visualization

Individual observations of each metric were organized and collected in CSV files. These data were analyzed using a systematic approach. For FTOs, IPU durations, and pause durations, the kernel density distributions were plotted using the geom_density function from the *ggplot2 R* package. This was then followed by performing Kolmogorov-Smirnov (KS) tests using the *ks.test* function to compare the distribution for the condition that was anticipated to be the most difficult (i.e., Mask: No Mic) with those from the other three conditions. This was then followed by fitting a generalized linear mixed model (GLMM) to the data. For other metrics, a linear mixed model (LMM) was fitted to the data. The *lme4* package in R was used to build the mixed-effects regression models for each variable of interest in the study using either the*glmer* or *lmer* functions. The fixed effects were mask (No Mask, Mask), microphone (No Mic, Mic), and their interaction, along with a random intercept of the group and/or talker. Hence, models were generally computed by the following equation: X $\sim$ Mask * Microphone + (1 | Group/Talker). For GLMMs, other predictors were included as fixed effects, and these will be described in the respective sections of the results. For LMMs, an ANOVA was conducted on the fitted model using the *anova* function in the *stats* package. For both GLMMs and LMMs, estimated marginal means of the fitted models were computed using the *emmeans* function in *emmeans* library in *R*.

2.3 Results

2.3.1 Rate of Floor Transfers

For each conversation, the rate of floor transfers (FT rate) was calculated by dividing the total number of floor transfers by the overall duration of the conversation in minutes. An LMM was fitted to this data using the following equation: FT Rate $\sim$ Mask * Microphone + (1 | Group). An ANOVA indicated that neither of the main effects (Mask: $F(1, 9) = 0.57$, $p = 0.47$; Microphone: ($F(1, 9) = 1.13$, $p = 0.32$), nor the interaction between Mask and Microphone ($F(1, 9) = 0.81$, $p = 0.40$) were statistically significant.

2.3.2 Floor-Transfer Offset

Kernel density distributions (computed using *geom_density* from the *ggplot2 R* package) of the FTOs pooled across all talkers in each of the four conditions are plotted in Figure 2.3. From the figure, compared to the other distributions, it appears that the distribution for the Mask:No Mic condition is the broadest (with the lowest peak) and is slightly shifted to the right.

KS tests were employed to compare the distributions of FTOs observed in the Mask:No Mic condition with the three other conditions. Both the No Mask:No Mic ($D = 0.1328$, $p < 0.01$) and No Mask:Mic ($D = 0.1227$, $p < 0.05$) conditions were found to be significantly different, but the Mask:Mic condition was not significantly different ($D = 0.0788$, $p = 0.21$).

As the observed distributions appear skewed, a gamma distribution was selected to be fitted to the data using a GLMM. However, gamma distributions can only be fitted to non-negative data. Thus, the data was pre-processed by adding a fixed value to all observations (i.e., shifting the distribution) and then removing any remaining negative observations. For a gamma distribution, the mean, variance, and skewness are related. Thus, the following steps were conducted to determine the shift to be applied to the data. First, the variance and skewness of the pooled distribution of all observed FTOs were calculated. Second, the mean of a gamma distribution with the same variance and skewness as that observed

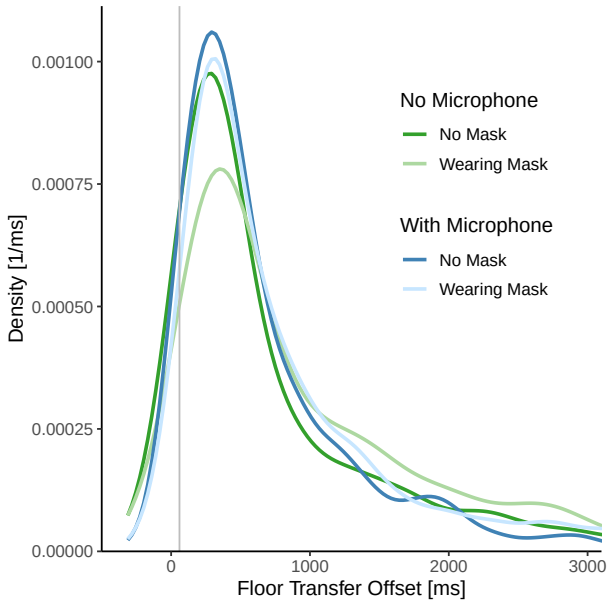


Figure 2.3: Distributions of the FTOs observed in each of the four conditions corresponding to the two experimental factors, Mask (lighter) vs No Mask (darker) and No Mic (green) vs. Mic (blue).

across all FTOs was computed and found to be 818.9 ms. Third, this was compared with the observed mean of all observed FTOs, which was found to be 879.9 ms. Finally, the difference between these means (-60ms) was used as the shift and added to all the FTO observations. Observations that remained negative after the shift (6.3% of the total number of observations) were removed and correspond with the region of the distribution to the left of the vertical line in Figure 2.3). When later calculating the estimated marginal means, a correction for this shift was applied.

A GLMM using a gamma family (with a log link) was fitted to the shifted data. Based on Roberts et al. (2015), the durations of the IPU immediately prior to the floor transfer (i.e., produced by the talker that ceded the floor) and the IPU immediately after the floor transfer (i.e., produced by the talker that took the floor) were included as predictors in the model. Thus, the GLMM was fitted using the following equation: FTO duration $\sim$ Mask * Microphone + Pre FT IPU Duration + Post FT IPU Duration + (1 |

Group/TalkerThatTookFloor).

The estimated marginal means of the FTO in each condition are plotted in Figure 2.4. The coefficients of the fixed effects are shown in Table 2.1. For the effect of No Mask, the coefficient was negative but not significantly different from zero. For the effect of Mic, the coefficient was negative and significantly different from zero. The coefficients for the interaction of No Mask and Mic, as well as the Pre-FT IPU duration were both positive but neither were significantly different from zero. However, the coefficient for the Post-FT IPU duration was positive and significantly different from zero.

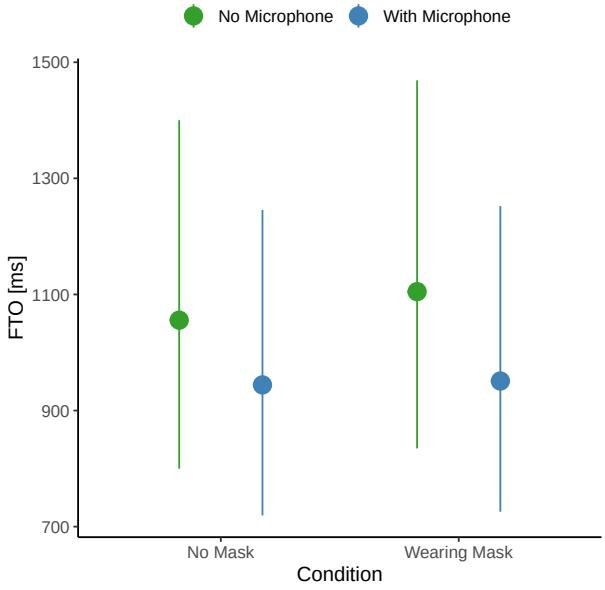


Figure 2.4: Mean FTO with no mask (left) and wearing mask (right). The estimated marginal means for the no microphone (green) and with microphone (blue) conversations are from the GLMM fitted to the data. The error bars indicate 95% confidence intervals.

Table 2.1: Fixed effects of the generalized linear mixed model of FTO. The values in the square brackets indicate 95% confidence intervals.

Intercept (Mask x No Mic)	$b_0 = -0.11$ [-0.43, 0.20]	$t(1416) = -0.71$, $p = 0.479$
No Mask	$b_1 = -0.05$ [-0.21, 0.11]	$t(1416) = -0.58$, $p = 0.561$
Mic	$b_2 = -0.16$ [-0.31, -7.49e-03]	$t(1416) = -2.06$, $p = 0.040$
No Mask x Mic	$b_3 = 0.04$ [-0.17, 0.25]	$t(1416) = 0.37$, $p = 0.711$
Pre FT IPU Duration	$b_4 = 0.01$ [-0.04, 0.06]	$t(1416) = 0.50$, $p = 0.619$
Post FT IPU Duration	$b_5 = 0.12$ [0.07, 0.16]	$t(1416) = 4.55$, $p < 0.001$

In summary, the use of a remote microphone resulted in a significant reduction in FTO duration, regardless of whether a face mask was worn. Additionally, longer post-FT IPU durations were associated with longer FTOs.

2.3.3 Interpausal Units

Kernel density distributions (computed using *geom_density* from the *ggplot2 R* package) of the IPU durations pooled across all talkers in each of the four conditions are plotted in Figure 2.5. From the figure, it appears that the IPU durations in both conditions with No Mic are broader and shifted to the right compared to the two conditions with Mic.

KS tests were employed to compare the distributions of IPUs observed in the Mask:No Mic condition with the three other conditions. The Mask:Mic condition was found to be significantly different ($D = 0.0474$, $p < 0.05$), while the No Mask:Mic condition was borderline ($D = 0.0476$, $p = 0.05$). The No Mask:No Mic condition was not significantly different ($D = 0.03575$, $p = 0.31$).

As the VAD processing applied to the recordings enforced a minimum duration of 90 ms, prior to fitting a GLMM, 90 ms was subtracted from all IPU observations. A GLMM using a gamma family (with a log link) was fitted to the shifted data using the following equation: IPU Duration $\sim$ Mask * Microphone + (1 | Talker). An initial fit was attempted using a random intercept of (1|Group/Talker) but this resulted in a singular fit.

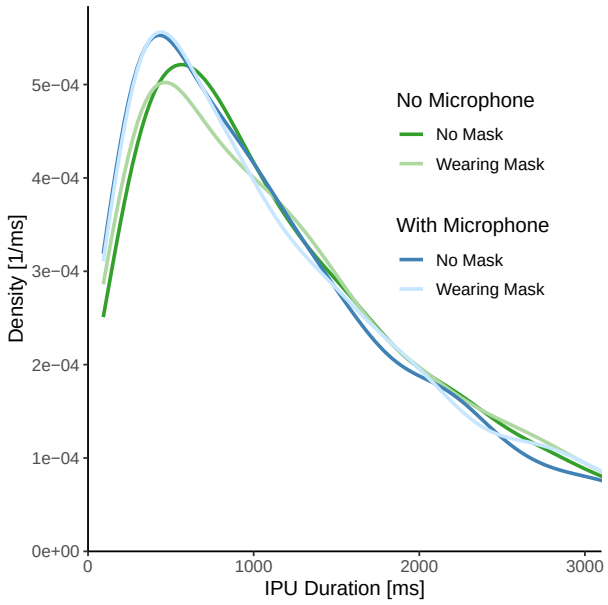


Figure 2.5: Distributions of the IPU durations observed in each of the four conditions corresponding to the two experimental factors, Mask (lighter) vs No Mask (darker) and No Mic (green) vs. Mic (blue).

The estimated marginal mean IPU durations were computed and a correction for the 90 ms shift was applied (see Figure 2.6). The coefficients of the fixed effects from the fitted model are given in Table 2.2. For the effect of No Mask, the coefficient was positive and significantly different from zero. For the effect of Mic, the coefficient was negative but not significantly different from zero. The coefficient of the interaction between No Mask and Mic was negative and significantly different from zero.

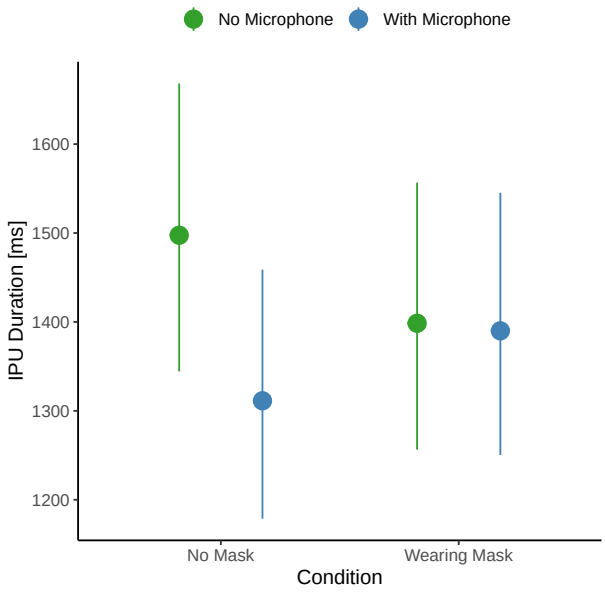


Figure 2.6: Mean IPU durations with no mask (left) and wearing mask (right). The estimated marginal means for the no microphone (green) and with microphone (blue) conversations are from the GLMM fitted to the data. The error bars indicate 95% confidence intervals.

Table 2.2: Fixed effects of the generalized linear mixed model of IPU duration. The values in the square brackets indicate 95% confidence intervals.

Intercept (Mask x No Mic)	$b_0 = 0.268$ [0.153, 0.382]	$t = 4.59$, $p < 0.0001$
No Mask	$b_1 = 0.072$ [0.005, 0.140]	$t = 2.12$, $p < 0.05$
Mic	$b_2 = -0.007$ [-0.070, 0.055]	$t = -2.34$, $p = 0.81$
No Mask x Mic	$b_3 = -0.135$ [-0.225, -0.044]	$t = -2.91$, $p < 0.01$

In summary, IPU durations were longest in the No Mask:No Mic condition and shortest in the No Mask:Mic condition. The IPU durations in both conditions where participants wore masks were similar to each other and different from both no-mask conditions.

2.3.4 Pause Duration

Kernel density distributions (computed using *geom_density* from the *ggplot2* *R* package) of the pause pooled across all talkers in each of the four conditions. Changes in pause durations in different conditions were not discernible.

KS tests were employed to compare the distributions of pauses observed in the Mask:No Mic condition with the three other conditions. The Mask:Mic condition was found to be statistically significant ($D = 0.0590$, $p = 0.022$), while the No Mask:Mic condition showed a trend ($D = 0.0515$, $p = 0.0745$). The difference with the No Mask:No Mic condition was not significantly different ($D = 0.0434$, $p = 0.25$).

As the VAD processing applied to the recordings enforced a minimum duration of 180 ms, prior to fitting a GLMM, 180 ms was subtracted from all pause observations. A GLMM using a gamma family (with a log link) was fitted to the shifted data using the following equation: Pause Duration $\sim$ Mask * Microphone + (1 | Talker). The coefficients of the fixed effects from the fitted model (with the intercept being the Mask:No Mic condition) showed that neither the effect of Mask nor the effect of Mic was significantly different from zero. Similarly, the coefficient of the interaction between No Mask and Mic was not significantly different from zero.

2.3.5 Articulation Rate

For each talker in each conversation, the articulation rate for each speaker was computed based on the estimated total number of syllables divided by the total phonation time. An LMM was employed to analyze this data using the formula: Articulation Rate $\sim$ Mask * Microphone + (1 | Group/Talker). An ANOVA conducted on the fitted model revealed that neither the main effects (Mask: $F(1, 54) = 0.0081$, $p = 0.93$; Microphone: $F(1, 45) = 0.0961$, $p = 0.76$) nor interaction ($F(1, 45) = 0.0599$, $p = 0.81$) were statistically significant.

2.3.6 Fundamental Frequency

For each conversation, the mean fundamental frequency (F0) for each talker was computed. An LMM was fitted to this data using the following equation: Mean F0 $\sim$ Mask * Microphone + Sex + (1 | Group/Talker). The estimated marginal means from the fitted model are plotted in Figure 2.7. An ANOVA conducted on the fitted model indicated that the main effects of Mask ($F(1, 45) = 5.9055$, $p < 0.05$) and Sex ($F(1, 12.223) = 35.5000$, $p < 0.001$) were significant. However, the main effects of Microphone ($F(1, 45) = 0.3190$, $p = 0.56$) and the interaction of Mask and Microphone ($F(1, 45) = 0.3192$, $p = 0.57$) were not statistically significant.

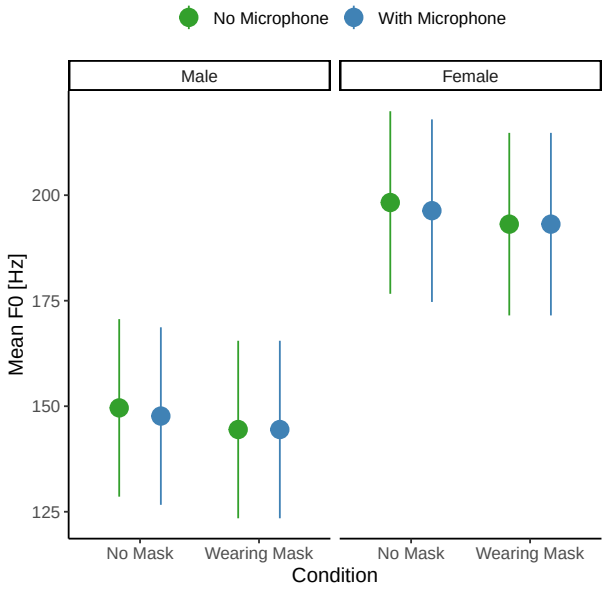


Figure 2.7: Mean F0 for male talkers (left) and female talkers (right) with no mask (left) and wearing mask (right). The estimated marginal means for the no microphone (green) and with microphone (blue) conversations are from the LMM fitted to the data. The error bars indicate 95% confidence intervals.

In summary, wearing a face mask led to a decrease in fundamental frequency for both male and female participants, whereas the use of a remote microphone had no effect.

2.3.7 Conversation Duration

An LMM was fitted to the duration of each conversation using the formula: Conversation Duration $\sim$ Mask * Microphone + (1 | Group). The estimated marginal means of the conversation duration in each condition is plotted in Figure 2.8. An ANOVA conducted on the fitted model indicated that the main effect of Microphone ($F(1, 9) = 5.2783$, $p = 0.05$) was statistically significant. The main effect of Mask ($F(1, 9) = 4.2709$, $p = 0.07$) and the interaction between Mask and Microphone ($F(1, 9) = 0.3042$, $p = 0.60$) were not statistically significant.

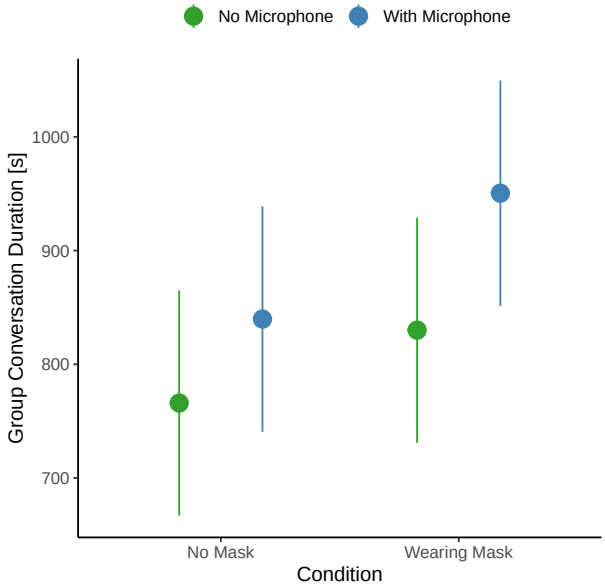


Figure 2.8: Mean group conversation duration with no mask (left) and wearing mask (right). The estimated marginal means for the no microphone (green) and with microphone (blue) conversations are from the LMM fitted to the data. The error bars indicate 95% confidence intervals.

In summary, the use of a remote microphone increased the duration of group conversations.

2.4 Discussion

The present study investigated acoustic measures of speech production and conversational behavior in recordings originally collected with the aim of analyzing conversation content. The recordings involved conversations between aided hearing-impaired talkers and were conducted in four conditions consisting of the combinations of wearing vs. not wearing face masks and using vs. not using a remote microphone. Previous studies have identified changes in speech production and conversation behavior in response to more challenging communication situations (Sørensen et al., 2021, 2020; Petersen et al., 2022): later and more variable timing of turn starts (i.e., a shift and increase in breadth of FTO distributions), longer IPU durations, a lower rate of floor transfers, slower articulation rates, and increased speech levels. However, these studies have typically investigated a large number of dyadic, task-based conversations. Thus, the goal of the present study was to investigate whether similar changes in behavior in response to increased difficulty are observed in a smaller number of free-form conversations amongst small groups that are more in line with real-life situations.

Compared to conversations where a remote microphone was used, the conversations without a remote microphone exhibited longer FTOs and lower overall conversation duration. These differences are consistent with changes that have been observed in response to more difficult communication situations. With regards to IPU duration, a more complex pattern involving an interaction between the factors of microphone and mask was observed. Further, compared to conversations where a face mask was worn, conversations without a mask exhibited a larger increase in mean F0, which is typically associated with higher vocal effort and would be expected in more difficult communication situations.

In the present study, the mean FTO was shorter in conversations where remote microphones were used but no effect of face mask was observed. In contrast to previous studies that have calculated median FTOs for each individual talker, the statistical analysis conducted here used a GLMM to model all of the observed FTOs, providing increased statistical power. Previous studies have compared the interquartile range of FTOs produced by individual talkers across conditions to examine changes in the breadth of the distribution. We have not presented a similar analysis here due to limited statistical power. Unlike

dyadic conversations, some talkers in a group setting may follow a conversation but contribute few turns. While it is tempting to attribute the lack of contribution to difficulty in following the conversation, some talkers may have taken only a few turns because they were not interested in the topic (e.g., they didn't have a pet) or did not feel they had anything further to add.

Across the conversations in this study, only 1% of the FTOs observed were negative (i.e., floor transfers where there is an acoustic overlap of speech between the two talkers), a much smaller proportion than that observed in other studies (e.g., 14% in Sørensen et al., 2021) This could be due to several factors: the nature of a free-form conversation compared to task-oriented conversation; the social acceptability of overlapping speech within this cohort of older hearing-impaired adults; and an effect of hearing impairment shifting the FTO distribution towards longer durations resulting in fewer observed negative FTOs. However, we think that hearing loss on its own is unlikely to explain this as 14% of the FTOs produced by the HI participants across all conditions in Sørensen et al. (2020) were negative.

Previous studies have consistently observed increases in IPU durations in more difficult communication conditions Sørensen et al. (2021, 2020); Petersen et al. (2022) investigated dyadic task-oriented conversations. These studies found that more challenging communication conditions led to increased durations of interpausal units (IPUs), which are utterances of connected speech. Longer IPUs are indicative of clearer and more intentional speech, allowing more time for interlocutors to plan and comprehend speech. Beechey et al. (2018) characterized this increased utterance duration as part of the "holding-the-floor" strategy, where speakers extend their turns to maintain clarity and ensure their message is understood in noisy environments.

In the present study, an interaction between the factors of mask and microphone was observed in the pattern of results for IPU duration. When participants did not wear a mask, IPU durations were shorter in the condition where remote microphones were used compared to the condition in which they were not. This pattern of results is consistent with the interpretation that the condition with the remote microphone was easier. However, the difference between the two conditions where participants wore a mask were not significantly different from each other. However, the mean IPU duration in these conditions was shorter

than the No Mask:No Mic condition and longer than the No Mask:Mic condition. A potential explanation for the observed interaction between the factors is that while face masks may affect speech perception, they also affect speech production. Face masks are known to decrease inhalation and exhalation volumes at rest (Ogawa et al., 2022; Engeroff et al., 2021; Marek et al., 2023). As speech sounds are typically generated by expelling air from the lungs, a reduction in the volume of air in the lungs would limit the duration of utterances of connected speech that could be produced. The similarity of IPU durations in the two mask conditions is consistent with this explanation.

Compared to the two conditions where participants wore masks, an increase in mean F0 was observed in the two conditions where talkers did not wear masks. Typically, an increase in mean F0 has been associated with increased vocal effort and speech levels (Beechey et al., 2018; Hazan and Baker, 2011; Jessen et al., 2005; Liénard and Di Benedetto, 1999; Vlaj and Kacic, 2011), and thus might be considered an acoustic marker of effort. However, wearing a mask will increase the resistance of airflow in and out of the respiratory tract (Marek et al., 2023; Engeroff et al., 2021). This would result in a reduction of subglottal pressure leading to a reduction in F0 (Lieberman, 1966). Thus, the pattern of observed differences in mean F0 observed across conditions is consistent with the influence of masks on speech production.

Although both articulation rate and the rate of floor transfers have been observed to decrease in more challenging conditions (Sørensen et al., 2020), we did not observe any significant differences in either measure across the conditions in the present study. While this could be attributed to a lack of statistical power, it could also be a result of the experimental design. According to Petersen et al. (2022), the FT rate and articulation rate are most affected by noise levels. In contrast with previous studies that varied noise levels across conditions (e.g., Sørensen et al., 2020; Petersen et al., 2022), in the present study, the background noise was held at a constant level that was much lower than that used in other studies. This speculation is supported by some participants noting in the questionnaire that while the background noise added a perceptible loudness that prompted them to invest more effort into concentration, it was not significantly stressful or uncomfortable (Ishida et al., 2022).

An underlying goal of many studies that have investigated changes in speech production

and conversational behavior across different conditions is to develop methods for evaluating hearing-assistive technologies using more realistic situations and tasks. The present study investigated free-form conversations between groups of four talkers and this conversational environment may be more realistic than previous studies that have used a dyadic task-based paradigm. However, the choice of which paradigm is best for evaluating hearing assistive technology is not clear. Task-oriented dialogue, such as that elicited using a Diapix task, exhibits a higher floor transfer rate, which may be advantageous when examining metrics involving FTOs, but results from paradigms closer to the typical conversation are likelier to generate results that can be more easily incorporated into hearing assistive technology. Another key choice in developing the paradigms is the number of participants that are involved in the conversation. In small groups like in the present study, it is easier for participants to withdraw and pretend to follow along. Many participants noted at various points in the questionnaire that they had difficulty hearing certain parts of the conversation (Ishida et al., 2022). However, they did not always request clarification or ask the speaker to repeat themselves. This is an important consideration when evaluating communication difficulties, as not every instance of difficulty may have been reported. The free-form nature of the conversation, combined with the fact that participants were strangers to one another, may have contributed to their reluctance to seek clarification.

Participants reported several important observations regarding the face mask in the questionnaire (Ishida et al., 2022). Some noted that they could hear certain individuals better than others, while others experienced the voice as mumbled or muffled. The absence of visible lip and facial movements also posed a challenge, with some participants finding that directionality was less clear, making it difficult to identify the speaker when the mouth was covered. These observations suggest that further investigations are needed to better assess communication difficulties when a factor of difficulty, such as a face mask, is introduced.

Nonetheless, most participants indicated that remote microphones improved the audibility of both their own voice and others' voices over background noise. Additionally, Ishida et al. (2022) findings showed that participants' subjective ratings suggested that remote microphone technology (RMT) significantly enhanced their cognitive and psychological experiences. This is further supported by our findings.

2.5 Conclusion

The present study aimed to explore how acoustic measures of speech production and conversational behavior are affected by different communication conditions in HI individuals, particularly when face masks and remote microphones are used. This investigation extends previous research on assessing communication difficulty in challenging listening environments by focusing on free-form conversations within small groups rather than task-based dyadic interactions. The results revealed that the use of remote microphones led to shorter FTOs and longer conversation durations, aligning with the notion that remote microphones facilitate easier communication. The interaction between microphone and mask conditions on IPU duration was complex, suggesting that while masks impact speech production, preventing us from seeing pronounced effects in IPUs, remote microphones might contribute to shorter IPUs that are indicative of easier communication.

An increase in mean F0 is typically associated with greater vocal effort. However, in this study, we observed a decrease in F0 when a face mask was worn. This can be explained by the fact that wearing a mask increases resistance to airflow, reducing subglottal pressure and consequently lowering F0. Thus, the pattern of differences in mean F0 across conditions is consistent with the influence of masks on speech production. While no significant differences were found in articulation rate and floor transfer rate across conditions, this might be attributed to the constant background noise level, which was lower than in other studies where noise was varied. These metrics are known to be influenced by noise levels, and the consistent noise environment in this study may have limited the potential for observing pronounced differences.

Future work is needed to better understand and characterize how factors such as the number of participants in a conversation and the nature of the conversation (e.g., free-form vs. task-oriented) influence acoustic measures of speech production and conversational behavior. These insights are crucial for developing effective hearing-assistive technologies and ensuring that evaluations reflect the complexity of everyday communication scenarios. Although testing in group conversations provides a more realistic context, dyadic interactions allow for a more balanced floor distribution where parties engage more equally. This structure enables a better assessment of listening effort and speech production, whereas, in

groups, some participants may dominate the conversation while others withdraw, resulting in unequal participation. This imbalance makes it more challenging to determine whether participants are experiencing difficulty or simply not engaging. Further research should explore these variables to enhance the real-world applicability of hearing-assistive device evaluations.

Chapter 3

The Effect of Task on Speech Production and Communication Behavior During Conversation in Healthy-Hearing Participants

3.1 Introduction

Speech and conversation are key elements of human communication, transferring information and conveying emotions between individuals. Depending on the aims of the participants, conversations can be broadly categorized into two types: transactional and relational.

Transactional communication is goal-oriented, focusing on achieving specific objectives, solving problems, and making decisions. When the purpose of communication is to achieve the speaker's goal by directing the listener's actions, it is oriented toward success, leading communicators to instrumentalize one another as a means to an end (Kang et al., 2017; Habermas, 1976). This type of communication is evident in customer service scenarios, where the primary goal is task completion and problem resolution (Froehle, 2006;

Diberardinis et al., 1979; Jarvenpaa and Leidner, 1998).

In contrast, relational communication emphasizes understanding and building relationships between the communicators. Here, the focus shifts from merely achieving goals to fostering mutual understanding and strengthening the connection (Habermas, 1976; Froehle, 2006; Coupland, 2000). This spontaneous, person-oriented communication is crucial for relationship-building, as it allows individuals to connect on a deeper emotional level (Coupland, 2000)

In many conversations, speakers must communicate in noisy environments such as in restaurants or at parties. In noise, talkers automatically raise their speech level, a response known as the Lombard effect (Lombard, 1911). Studies have observed that speakers typically increase their speech levels by about half the perceived increase in noise intensity Lane and Tranel (1971); Charlip and Burk (1969); Junqua et al. (1999); Davis et al. (2006). In addition to changes in level, changes in articulation rates have been observed in the presence of noise and interpreted as an adaptive strategy to maintain intelligibility (Charlip and Burk, 1969). Folk and Schiel (2011) describe how these adjustments align with Lindblom's H&H theory, which proposes a continuum between hyperarticulated (clear and slow) and hypoarticulated (fast and economical) speech. This theory suggests that the motor control of speech production balances maximal communication clarity with minimal effort, varying according to different contexts and situations (Lindblom, 1990). In noisy conditions, the Lombard effect prompts speakers to shift towards hyperarticulated speech to enhance communication clarity. Uchanski et al. (1996) supported this by showing that speech produced in noise tends to be clearer and more deliberate (i.e., clear speech), which makes it more intelligible when played back to listeners in a quiet setting compared to speech produced in quiet conditions (i.e., conversational speech). This enhanced clarity in noisy conditions likely arises because speakers instinctively adjust their speech to overcome the masking effects of noise, resulting in slower speech, exaggerated articulation, and reduced ambiguity in their utterances.

In challenging conditions, increased cognitive effort may be required to process the received noisy speech signals. According to the Ease of Language Understanding (ELU) model, a shift from automatic to explicit processing to infer meaning can occur when the sensory input is degraded and the perceived phonological input does not match stored rep-

resentations in long-term memory (Rönnberg et al., 2013, 2008). When speech perception requires more top-down processing, critical information may be perceived later than when automatic processing is sufficient. Thus, in more challenging conditions, speech planning may start later and be more variable, leading to delayed and less precise turn-taking in conversation (Barthel et al., 2016).

Several studies have investigated conversations in noise and have observed changes in both speech production and conversation behavior in more challenging conditions. Studies by Sørensen et al. (2021, 2020); Petersen et al. (2022) examined dyadic task-oriented conversations elicited using the Diapix task. In more challenging communication conditions, these studies have observed increases in the median and interquartile range of the distribution of floor transfer offsets (FTO; the interval from when the talker ceding the floor ends their turn to when the talker that takes the floor to start theirs) as well as an increase in the duration of interpausal units (IPU; utterances of connected speech). Thus, in noise, talkers started their turns later with more variable timing and held the floor for longer. Beechey et al. (2018) examined dyadic conversations elicited with a tangram task and observed changes in an expanded set of speech characteristics, including prosody, utterance length, and turn-taking behavior, which are not typically considered in studies of Lombard speech.

The results from the studies above suggest a set of acoustic markers that can be used to infer the difficulty experienced by talkers in a conversation. However, as these studies all elicited conversations using a task, the participants were likely engaged in transactional conversations. It is not clear that the observed changes in speech production and conversational behavior generalize to relational conversations. A preliminary study by Watson et al. (2019) elicited both free-form and task-oriented conversations in both quiet and noise from five pairs of native-Danish talkers. Similar changes in the FTO and IPU distributions were observed in noise for both types of conversations. However, the conversations were all conducted in English, so the participants were speaking in a second language, which has been observed to influence aspects of speech production and conversational behavior (Sørensen et al., 2021).

In the present study, we follow up on the work of Watson et al. (2019) to investigate the effects of noise and conversational type on various measures of speech production and

conversational behavior.

3.2 Methods

3.2.1 Participants

Ten mixed- and same-sex pairs participated in the study. Participants were recruited from the University of Waterloo community as friends and ranged in age from 18 to 22 years. All participants were screened to confirm normal hearing thresholds (pure-tone thresholds less than 25 dBHL at octave frequencies from 125 Hz to 4 kHz). All participants provided informed consent and the study was approved by the University of Waterloo Research Ethics Board (REB #45669). Participants were compensated for their time.

3.2.2 Experimental Setup

Participant pairs were seated in separate rooms and communicated using DPA 4088 directional headset microphones and Sennheiser HD 280 Pro around-the-ear, closed headphones. The setup was calibrated such that participants would hear each other as if they were seated 1.4 meters apart. This was achieved by adjusting the gain of the headset microphones to match that of the iSEMcon EMX-7150-CF/MP reference microphone placed approximately 1.4 meters away from each participant.

In the conditions where the participants conversed in the presence of background noise, a 10-minute noise file was presented through their headphones at 70 dB SPL.

The noise file was created using sound texture synthesis from a 1-minute clip of the multi-talker babble used in the SPIN test (Bilger et al., 1984). The code used to conduct the sound texture synthesis was based on the study conducted by McWalter and McDermott (2018). Briefly, the sound texture synthesis algorithm involves two steps. In the first step, the algorithm takes an input signal and calculates a wide range of statistical estimates based on a model of peripheral auditory processing. In the second step, these statistics are

imposed on a Gaussian white noise using an iterative process. The final result is an output signal of arbitrary length that has the same statistical properties as the input signal.

3.2.3 Procedure

Participants engaged in two types of conversational tasks: task-based and free-form. For the task-based interaction, participants completed the DiapixUK task (Baker and Hazan, 2011). This task involved pairs of almost identical cartoon illustrations, with each participant viewing only one version. The objective was to collaboratively identify 10 differences between the two illustrations. Recordings were concluded after participants identified 10 differences or six minutes, whichever occurred first. For the free-form conversations, participants were instructed to choose a general topic of mutual interest (e.g., "sports", "career", "travel", etc.) and engage in unrestricted conversation for six minutes. They were instructed that they could freely change topics within a conversation as they liked.

Both task-based and free-form conversations were conducted in both a quiet condition and a 70 dB SPL noise condition. Three conversations were recorded for each combination of conversation type and background condition. For the task-based conversation, A different set of pictures was used for each recording.

Before testing began, participants received training with a DiapixUK illustration to familiarize themselves with the task. This training session was supervised by the experimenter to ensure participants understood the task requirements.

For each participant pair, the order of noise conditions and DiapixUK pictures used was randomized. Participants took breaks after completing a pair of conversations (i.e., one free-form and one task-based conversation). Across the experiment, a total of 120 conversations were recorded (10 pairs x 4 conditions x 3 replicates).

3.2.4 Recording Processing and Data Extraction

The recordings were processed in MATLAB to automatically segment speech from each talker based on classification schemes by Heldner and Edlund (2010) and Levinson and

Torreira (2015). First, an energy-based Voice Activity Detector (VAD) was applied to each talker's channel. This detector used 5 ms window segments with a 1-ms overlap, computing the root-mean-square (RMS) of each segment. Given that the dynamic range of speech is about 30 dB, the VAD threshold was set to 30 dB below the peak power detected throughout the conversation (Amlani et al., 2002). Using this threshold, windows were categorized as either speech (labeled as 1) or silence (labeled as 0).

Following Heldner and Edlund (2010), silent segments shorter than 180 ms were bridged (i.e., were relabeled as having speech present) to avoid the silences in stop consonants being categorized as pauses. Segments of speech shorter than 70 ms were assumed to be non-speech sounds (e.g., coughs, accidental bumping of the headset microphone, etc.) and relabeled as silence (Sørensen et al., 2021). After generating binary speech/silence vectors for each talker, a conversational state algorithm classified segments into distinct categories as detailed in Figure 3.1.

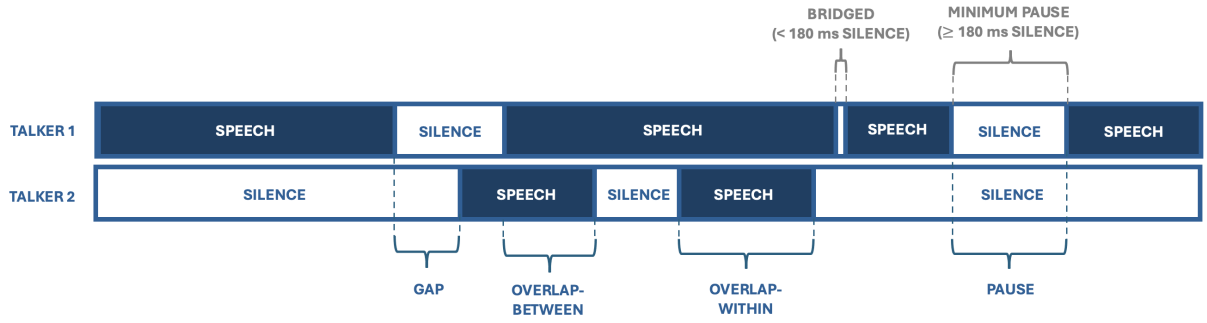


Figure 3.1: Schematic of communicative state, silence, and overlap classification, based on the procedure described by Heldner and Edlund (2010).

The detailed classification in Figure 3.1 is as follows:

1. **FTOs:** The interval from when one talker that cedes the floor stops speaking to when the talker that takes over the floor starts talking. Positive FTOs manifest as a

Gap (i.e., an interval of silence) while negative FTOs manifest as an Overlap-Between (i.e., there is an acoustic overlap of the two talkers during a floor transfer).

2. **Overlaps-Within:** Speech from one talker completely overlapped by a speech from the other without a floor transfer.

3. **Interpausal Units (IPUs):** Units of connected speech, including silences shorter than 180 ms.

4. **Pauses:** Silence from both talkers without a floor transfer.

5. **Turns:** One or more IPUs and pauses produced by a talker that is bounded by floor transfers.

To detect syllable nuclei, a Praat script de Jong et al. (2021) was employed using the default parameters. This script identifies syllable boundaries based on intensity fluctuations, where boundaries are marked 6 dB below the syllable peak. In cases where the intensity dip between syllables falls between 2 dB and 6 dB, boundaries are adjusted to maintain syllable coherence. Voicedness criteria are applied to exclude unvoiced peaks, such as surrounding voiceless consonants with high intensity. Detected syllables are segmented into intervals in the TextGrid and stored in an array extracted through a PRAAT-MATLAB interface. The articulation rate for each talker was then computed by dividing the total number of estimated syllables by the talker's total phonation time, as determined by the VAD.

For each talker in each conversation, speech levels were estimated based on the average RMS of the recording when silences were removed. This was then converted to dB SPL based on the digital gain and calibration of the system.

3.2.5 Statistical Analysis

Either a linear mixed model (LMM) or a generalized linear mixed model (GLMM) was fitted to each metric in *R*, using the *lme4* and *lmerTest* packages. The models were fitted using Restricted Maximum Likelihood (REML) for LMMs and Maximum Likelihood (ML)

for GLMMs. The general formula used, which varied depending on the variable, was: X $\sim$ Condition * Conversation Type + (1 | Talker) + (1 | Replicate).

For LMMs, an ANOVA was conducted on the fitted model using the *anova* function. Marginal means of the fixed effects were then back-transformed from the model coefficients using the *emmeans* package. Confidence intervals for the model coefficients were estimated using the *lmerTest* package, followed by one-sided t-tests using Satterthwaite's method.

Post-hoc comparisons of estimated marginal means were conducted using the *contrast* function. This allowed t-tests to use the Kenward-Roger approximation for degrees of freedom and the multivariate t distribution (mvt) for controlling multiple comparisons.

3.3 Results

3.3.1 Rate of Floor Transfers

For each conversation, the rate of floor transfers (FT) was calculated by dividing the total number of floor transfers by the overall duration of the conversation. A linear mixed model was fitted to the data using the following equation: FT Rate $\sim$ Condition * Conversation Type + (1 | Pair) + (1 | Replicate). The estimated marginal means of the FT rate for the fixed effects of Condition and Conversation Type are plotted in Figure 3.2. The parameters of the fixed effects are shown in Table A.1. The main effect of Conversation Type ($F(1, 105) = 334.86$, $p < 0.0001$) was found to be statistically significant. However, the main effect of Condition ($F(1, 105) = 0.63$, $p = 0.43$) and the interaction of Condition $\times$ Conversation Type ($F(1, 105) = 0.47, p = 0.50$) were not statistically significant. The results of the pairwise comparison of the means can be found in Table A.2. Only the pairwise comparisons for Task Quiet vs. Free Quiet and Task Noise vs. Free Noise were found to be significant ($p < 0.0001$).

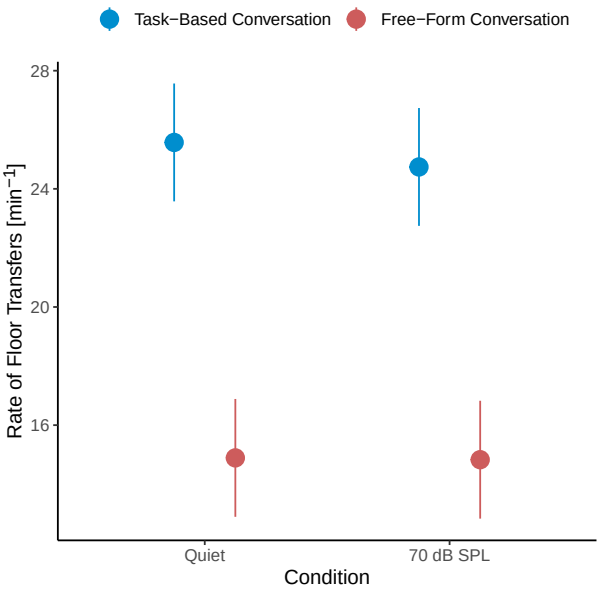


Figure 3.2: Mean rate of floor transfers in quiet (left) and 70 dB SPL noise (right). The estimated marginal means for the task-based (blue) and free-form (red) conversations are from the linear mixed model fitted to the data. The error bars indicate 95% confidence intervals.

In summary, the rate of floor transfers was lower in free-form conversations compared to task-based conversations. However, the rate of floor transfers did not significantly vary between quiet and noisy conditions for both conversation types.

3.3.2 Floor-Transfer Offset

The kernel density distributions of FTOs (computed using *geom_density* from the *ggplot2* R package) are plotted in Figure 3.3. In the figure, the FTO distributions for the free-form conversations appear broader (with a lower peak) and shifted leftwards compared to task-based conversations. Hence, free-form conversations have a higher number of overlaps (negative FTOs). However, within each conversation type, there appears to be little difference between the quiet and noise conditions.

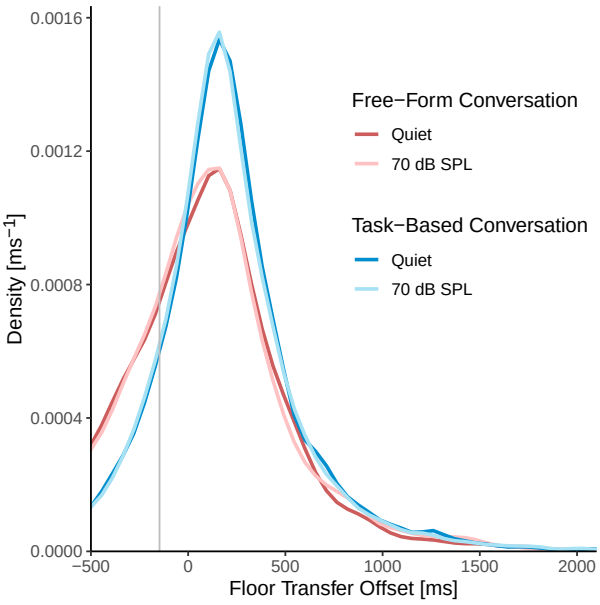


Figure 3.3: Distributions of floor transfer offsets (FTO) produced by talkers during free-form (red) and task-based (blue) conversations in quiet (dark) and 70 dB SPL noise (light). The vertical gray line represents the distribution truncation (where observations to the left of the line were removed) when fitting a generalized linear mixed model with a gamma distribution.

As the FTO distributions appear skewed, a generalized linear mixed model using a gamma distribution was fitted to the data. Since the gamma distributions only include positive-valued observations, the FTO distributions were shifted to the right by adding a fixed value to every observation. All observations that remained negative after the shift were truncated. For a gamma distribution, the mean can be estimated from the variance and skewness. The observed variance and skewness of the pooled distribution of all observed FTOs were calculated. A gamma distribution with the same variance and skewness as the pooled distribution of all observed FTOs had a mean of 293.6 ms. The mean of all the observed FTOs was 146.4 ms. The difference between these means is 147.2 ms. This value was added to each FTO observation and any values that remained negative after this adjustment (20.4% of all FTOs) were excluded from the analysis. This

percentage represents the portion of the distribution left of the vertical line in Figure 3.3. When calculating the estimated marginal means, a correction for this shift was applied.

A generalized linear mixed model using a gamma family (with a log link) was fitted to the data using the following equation: FTO $\sim$ Condition * Conversation Type + Pre FT Number of Syllables + Post FT Number of Syllables + Pre FT Articulation Rate + Post FT Articulation Rate + (1 | TalkerThatTookFloor) + (1 | Replicate). The Pre- and Post-FT Number of Syllables are the number of syllables estimated in an IPU before and after the floor transfer, respectively. The Pre- and Post-FT Articulation Rates are the number of syllables estimated in the preceding and following IPU divided by that IPU's duration, respectively. The estimated marginal means of the FTOs for the fixed effects of Condition and Conversation Type are plotted in Figure 3.4. The parameters of the fixed effects are shown in Table 3.1. The results of the pairwise comparison of the means can be found in Table A.3. The Task Quiet vs. Free Quiet comparison ($p < 0.001$) was significant, while the Free Quiet vs. Free Noise ($p = 0.06$) comparison showed a trend for significance. All other pairwise comparisons were not significant.

For the effect of conversation type, the coefficient is negative and significantly different from zero. The coefficient for the effect of Condition is negative but not significantly different from zero. However, the coefficient of the interaction of condition and conversation type is positive and significantly different from zero. The coefficient of the number of syllables preceding and following the floor transfer are negative and positive, respectively, and both are significantly different from zero. The coefficients for the articulation rate preceding and following the floor transfer were both negative, however, only the articulation rate following the floor transfer is significantly different from zero.

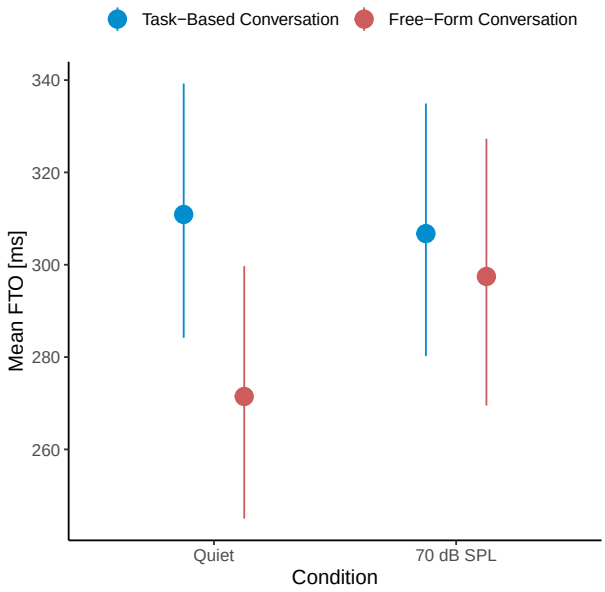


Figure 3.4: Mean FTO in quiet (left) and 70 dB SPL noise (right). The estimated marginal means for the task-based (blue) and free-form (red) conversations are from the generalized linear mixed model fitted to the data. The error bars indicate 95% confidence intervals.

Table 3.1: Fixed effects of the generalized linear mixed model of FTO. The values in the square brackets indicate 95% confidence intervals.

Intercept (Task-Based in Quiet)	$b_0 = 6.13$ [6.07, 6.19]	$t = 199.87$, $p < 0.0001$
70 dB SPL	$b_1 = -0.01$ [-0.04, 0.03]	$t = -0.50$, $p = 0.62$
Free-Form	$b_2 = -0.09$ [-0.13, -0.05]	$t = -4.03$, $p < 0.0001$
Pre FT Number Syllables	$b_3 = -0.07$ [-0.08, -0.05]	$t = -8.74$, $p < 0.0001$
Post FT Number Syllables	$b_4 = 0.01$ [-0.00, 0.03]	$t = 1.94$, $p = 0.05$
Pre FT Articulation Rate	$b_5 = -0.01$ [-0.02, 0.01]	$t = -1.23$, $p = 0.22$
Post FT Articulation Rate	$b_6 = -0.07$ [-0.08, -0.05]	$t = -8.98$, $p < 0.0001$
70 dB SPL x Free-Form	$b_7 = 0.07$ [0.01, 0.13]	$t = 2.24$, $p = 0.03$

The interquartile range (IQR) of the FTOs produced by each talker in each conversation

was computed as a measure of the breadth of the FTO distributions. A linear mixed model was fitted to this data using the following equation: FTO IQR $\sim$ Condition * Conversation Type + (1 | TalkerThatTookFloor). The estimated marginal means of the FTO IQR for the fixed effects of Condition and Conversation Type are plotted in Figure 3.5. The parameters of the fixed effects are shown in Table A.4. The main effect of Conversation Type ($F(1, 217) = 71.86$, $p < 0.0001$) was found to be statistically significant. Neither the main effect of Condition ($F(1, 217) = 0.39$, $p = 0.53$) nor the interaction of Condition $\times$ Conversation Type ($F(1, 217) = 0.33$, $p = 0.57$) was found to be statistically significant. The results of the pairwise comparison of the means can be found in Table A.5. Only the pairwise comparisons for Task Quiet vs. Free Quiet and Task Noise vs. Free Noise were significant (both with $p < 0.0001$).

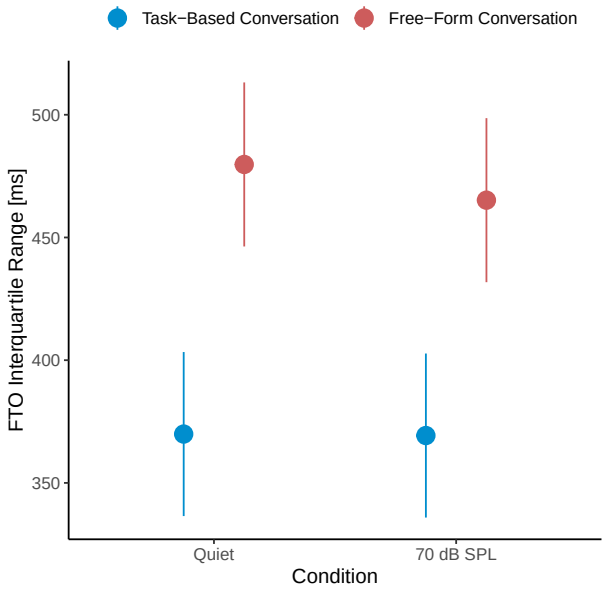


Figure 3.5: FTO interquartile range in quiet (left) and 70 dB SPL noise (right). The estimated marginal means for the task-based (blue) and free-form (red) conversations are from the linear mixed model fitted to the data. The error bars indicate 95% confidence intervals.

In summary, FTO distributions were more broad and skewed towards shorter durations

during free-form conversations compared to task-based ones. Hence, FTOs were significantly shorter with more overlaps observed during free-form conversations compared to task-based interactions. However, the presence of noise did not show any observed effect on FTO durations in either conversation type. Moreover, the FTO IQR was significantly larger for free-form conversations compared to task-based ones, yet noise did not induce significant changes in either case.

3.3.3 Interpausal Units

Interpausal units (IPUs) are defined by segments of continuous speech that are separated by 180 ms silence intervals. The IPUs do not include overlaps-within, which are speech segments that take place entirely within another speaker's turn(i.e., with no floor transfer; see Figure 2.2). The kernel density distributions of IPU duration (computed using *geom_density* from the *ggplot2 R* package) are plotted in Figure 3.6. In the figure, the IPU duration distributions for the free-form conversations appear to be broader (with a lower peak) and shifted to the right compared to task-based conversations. Additionally, there is a slight decrease in the peak between quiet and noise conditions, where noise appears to have longer IPU durations for both conversation types.

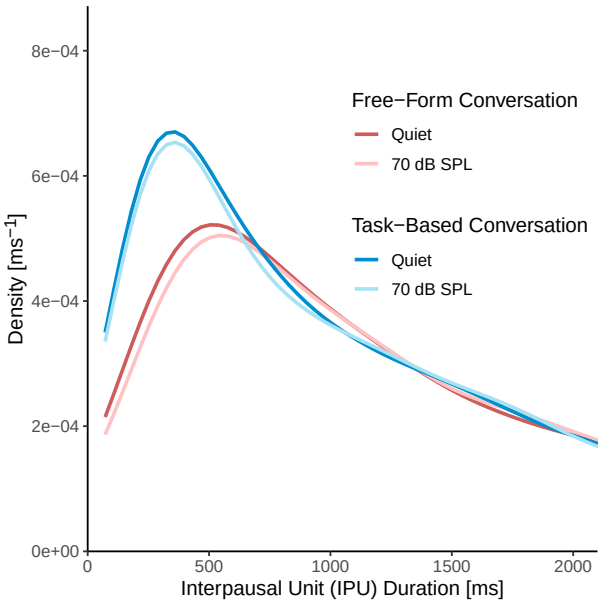


Figure 3.6: Distributions of IPU durations produced by talkers during free-form (red) and task-based (blue) conversations in quiet (dark) and 70 dB SPL noise (light).

During the VAD processing, the minimum threshold for IPU durations was set to 70 ms, where IPUs shorter than 70 ms were considered non-speech segments. Therefore, this 70 ms offset was subtracted from all IPU durations before fitting a generalized linear mixed model. When calculating the estimated marginal means, 70 ms was added to account for this offset.

A generalized linear mixed model using a gamma family (with a log link) was fitted to the data using the following equation: IPU Duration $\sim$ Condition * Conversation Type + (1 | Talker) + (1 | Replicate). The estimated marginal means of the IPU durations for the fixed effects of Condition and Conversation Type are plotted in Figure 3.7. The parameters of the fixed effects are shown in Table 3.2. The results of the pairwise comparison of the means can be found in Table A.6. All pairwise comparisons were significant ($p < 0.05$).

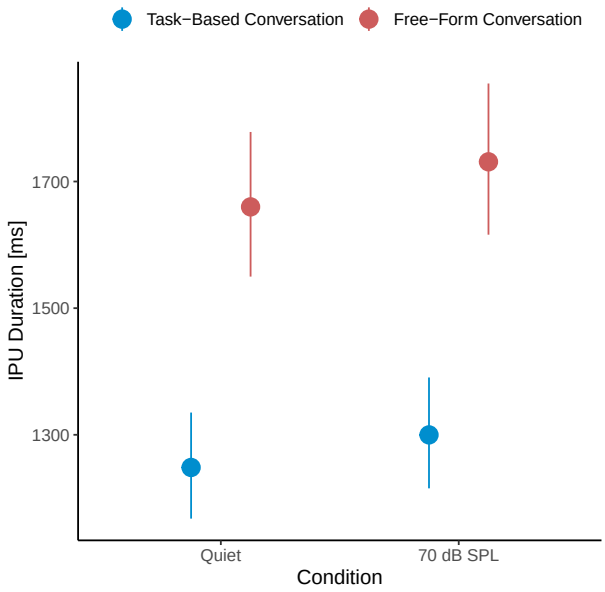


Figure 3.7: Mean IPU duration in quiet (left) and 70 dB SPL noise (right). The estimated marginal means for the task-based (blue) and free-form (red) conversations are from the generalized linear mixed model fitted to the data. The error bars indicate 95% confidence intervals.

Table 3.2: Fixed effects of the generalized linear mixed model of IPU duration. The values in the square brackets indicate 95% confidence intervals.

Intercept (Task-Based in Quiet)	$b_0 = 7.07$ [7.00, 7.14]	$t = 195.21$, $p < 0.0001$
70 dB SPL	$b_1 = 0.04$ [0.01, 0.07]	$t = 2.67$, $p = 0.01$
Free-Form	$b_2 = 0.30$ [0.27, 0.33]	$t = 17.88$, $p < 0.0001$
70 dB SPL x Free-Form	$b_3 = 0.00$ [-0.05, 0.05]	$t = 0.04$, $p = 0.96$

For the effect of noise, the coefficient is positive and significantly different from zero. The coefficient for the effect of conversation type is positive and significantly different from zero. However, the coefficient of the interaction of condition and conversation type is not significantly different from zero.

To better understand the changes in articulation rate with conversation type and noise conditions, further processing was run to compute the number of syllables per IPU. A generalized linear mixed model using a Poisson distribution (with a log link) was fitted to the data using the following equation: Num Syllables $\sim$ Condition * Conversation Type + (1 | Talker) + (1 | Replicate). The estimated marginal means of the number of syllables per IPU for the fixed effects of Condition and Conversation Type are plotted in Figure 3.8. The parameters of the fixed effects are shown in Table 3.3. The results of the pairwise comparison of the means can be found in Table **??**. All pairwise comparisons were significant ($p < 0.0001$).

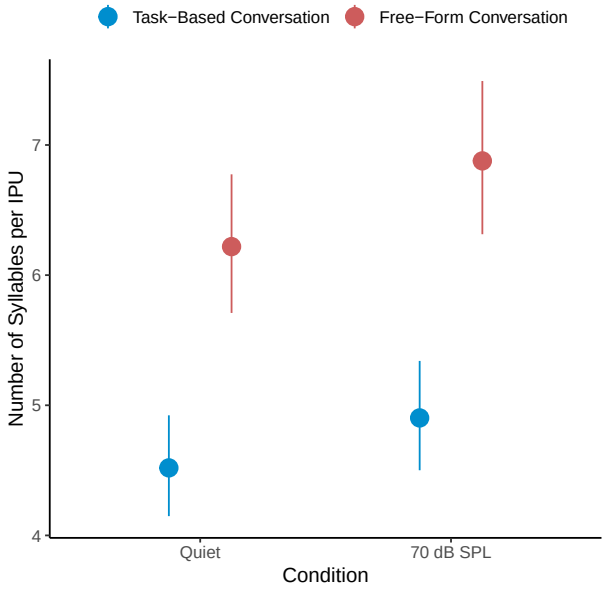


Figure 3.8: Mean number of syllables per IPU in quiet (left) and 70 dB SPL noise (right). The estimated marginal means for the task-based (blue) and free-form (red) conversations are from the generalized linear mixed model fitted to the data. The error bars indicate 95% confidence intervals.

Table 3.3: Fixed effects of the generalized linear mixed model of number of syllables per IPU. The values in the square brackets indicate 95% confidence intervals.

Intercept (Task-Based in Quiet)	$b_0 = 1.51$ [1.42, 1.59]	$z = 34.53$, $p < 0.0001$
70 dB SPL	$b_1 = 0.08$ [0.07, 0.10]	$z = 10.24$, $p < 0.0001$
Free-Form	$b_2 = 0.32$ [0.30, 0.33]	$z = 40.57$, $p < 0.0001$
70 dB SPL x Free-Form	$b_3 = 0.02$ [-0.00, 0.04]	$z = 1.75$, $p = 0.08$

The coefficients for both, the noise condition and conversation type, are positive and significantly different from zero. The coefficient for the interaction of 70 dB SPL x Free-Form, however, was not significantly different from zero.

In summary, IPU durations were longer for free-form conversations compared to task-based conversations, and there were also more syllables per IPU in free-form conversations. Additionally, the presence of noise resulted in an increase in both IPU duration and the number of syllables per IPU for both types of conversations.

3.3.4 Pause Duration and Rate

The kernel density distributions of pause duration (computed using *geom_density* from the *ggplot2 R* package) are plotted in Figure 3.9. In the figure, the pause duration distributions for the free-form conversations appear to have a higher peak compared to task-based conversations. Additionally, there is a decrease in the peaks between quiet and noise conditions, where noise distributions appear to have broader distributions for both conversation types.

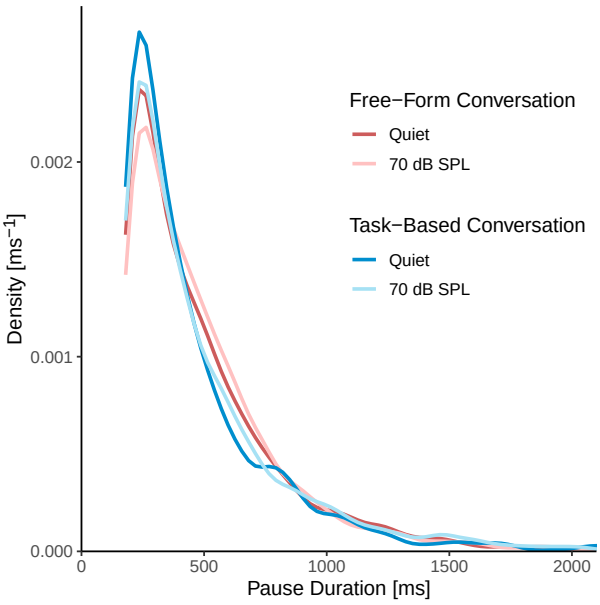


Figure 3.9: Distributions of pause duration produced by talkers during free-form (red) and task-based (blue) conversations in quiet (dark) and 70 dB SPL noise (light).

During the VAD processing, the minimum threshold for a pause duration was set to 180 ms, where pauses shorter than 180 ms were considered part of stop consonant productions. Therefore, this 180 ms offset was subtracted from all observed pause durations before fitting a generalized linear mixed model. The offset was added during the estimated marginal means calculation to compensate for the applied shift.

A generalized linear mixed model using a gamma family (with a log link) was fitted to the data using the following equation: Pause Duration $\sim$ Condition * Conversation Type + Pre-Pause IPU Duration + Post-Pause IPU Duration + (1 | Talker). The estimated marginal means of the pause duration for the fixed effects of Condition and Conversation Type are plotted in Figure 3.10. The parameters of the fixed effects are shown in Table 3.4. The results of the pairwise comparison of the means can be found in Table A.10. The Free Quiet vs. Free Noise comparison ($p < 0.001$) was significant, while all other pairwise comparisons were not significant.

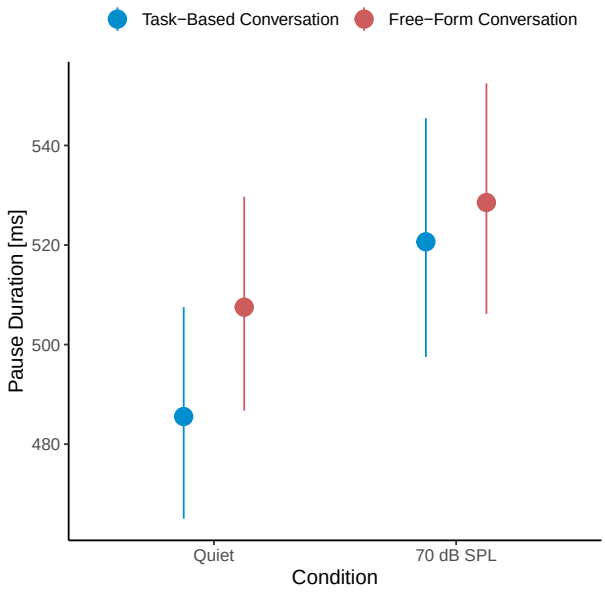


Figure 3.10: Mean pause duration in quiet (left) and 70 dB SPL noise (right). The estimated marginal means for the task-based (blue) and free-form (red) conversations are from the generalized linear mixed model fitted to the data. The error bars indicate 95% confidence intervals.

Table 3.4: Fixed effects of the generalized linear mixed model of pause duration. The values in the square brackets indicate 95% confidence intervals.

Intercept (Task-Based in Quiet)	$b_0 = 5.79$ [5.73, 5.86]	$t = 173.75$, $p < 0.0001$
70 dB SPL	$b_1 = 0.06$ [0.00, 0.12]	$t = 2.09$, $p = 0.04$
Free-Form	$b_2 = -0.07$ [-0.13, -0.01]	$t = -2.16$, $p = 0.03$
Pre-Pause IPU Duration	$b_3 = -0.10$ [-0.12, -0.08]	$t = -9.18$, $p < 0.0001$
Post-Pause IPU Duration	$b_4 = -0.03$ [-0.05, -0.01]	$t = -2.43$, $p = 0.02$
70 dB SPL x Free-Form	$b_5 = 0.05$ [-0.04, 0.13]	$t = 1.03$, $p = 0.30$

For the effect of noise condition, the coefficient is positive and significantly different from zero. Furthermore, the coefficient for the effect of free-form conversation is negative

and significantly different from zero. The coefficient for the interaction of 70 dB SPL x Free-Form, however, is not significantly different from zero. The coefficients for the pauses preceding and following IPU were both negative and significantly different from zero.

To further understand the Pre-Pause IPU Duration and Post-Pause IPU Duration for each of the conversation types, further GLMM analysis was conducted (refer to A.8 and A.9). For Free-Form Conversations, the coefficients for both Pre-Pause IPU Duration and Post-Pause IPU Duration are negative and significantly different from zero. In contrast, for Task-Based Conversations, while the coefficient for Pre-Pause IPU Duration is also negative and significantly different from zero, the coefficient for Post-Pause IPU Duration is not significantly different from zero.

For each talker in each conversation, the pause rate was calculated by dividing the count of all pauses by the sum of the durations of all IPUs and pauses produced by that talker in that conversation. A linear mixed model was fitted to the data using the following equation: Pause Rate $\sim$ Condition * Conversation Type + (1 | Talker). The estimated marginal means of the pause rate for the fixed effects of Condition and Conversation Type are plotted in Figure 3.11. The parameters of the fixed effects are shown in Table A.11. The main effect of Condition ($F(1, 217) = 6.25$, $p = 0.01$) and the main effect of Conversation Type ($F(1, 217) = 29.13$, $p < 0.0001$) were found to be statistically significant. However, the interaction of Condition $\times$ Conversation Type was not statistically significant. The results of the pairwise comparison of the means can be found in Table A.12. Pairwise comparisons for Task Quiet vs. Free Quiet and Task Noise vs. Free Noise were both significant ($p < 0.0001$), while other comparisons were not significant.

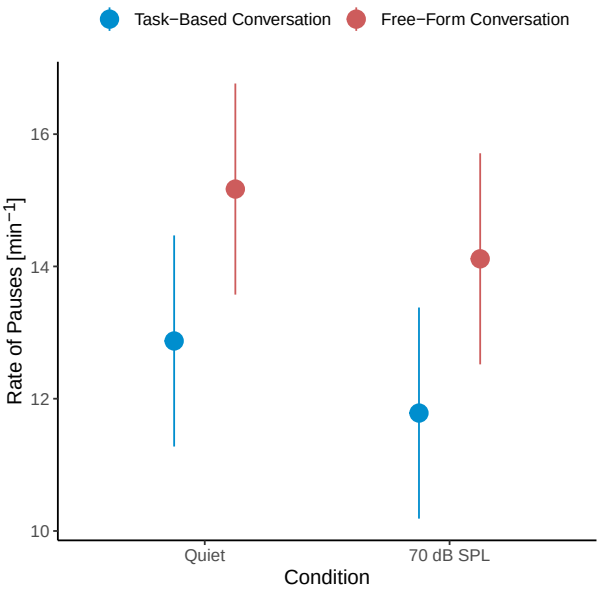


Figure 3.11: Mean pause rate (defined by the total number of pauses divided by the sum of the durations of all IPU and pauses produced by a talker during a conversation) in quiet (left) and 70 dB SPL noise (right). The estimated marginal means for the task-based (blue) and free-form (red) conversations are from the generalized linear mixed model fitted to the data. The error bars indicate 95% confidence intervals.

For the effect of noise condition, the coefficient is negative and shows a trend for significance. Furthermore, the coefficient for the effect of free-form conversation is positive and significantly different from zero. The interaction of 70 dB SPL x Free-Form, however, is not significantly different from zero.

In summary, pause durations were longer for free-form conversations compared to task-based conversations. Similarly, pause rates were higher for free-form conversations. For free-form conversations, longer pauses are associated with shorter IPU durations both before and after the pause. For task-based conversations, longer pauses were linked to shorter IPU durations before the pause, but not with the IPU duration after the pause. The presence of noise resulted in an increase in pause duration in both conversation types but did not result in a change in pause rate for either.

3.3.5 Articulation Rate

For each conversation, the articulation rates for each speaker were computed by calculating the ratio of the total number of syllables uttered to the overall speaking time for that talker. The overall speaking time was derived by summing the durations of all IPUs produced by the talker during the conversation.

A linear mixed model was fitted to the data using the following equation: Articulation Rate $\sim$ Condition * Conversation Type + (1 | Talker). The estimated marginal means of the articulation rates for the fixed effects of Condition and Conversation Type are plotted in Figure 3.12. The parameters of the fixed effects are shown in Table A.13. The main effect of Condition ($F(1, 217) = 38.27$, $p < 0.0001$) was found to be statistically significant. The main effect of Conversation Type ($F(1, 217) = 8.93$, $p < 0.01$) was also statistically significant. However, the interaction of Condition $\times$ Conversation Type was not statistically significant. The results of the pairwise comparison of the means can be found in Table A.14. All pairwise comparisons were significant ($p < 0.05$), except for Task Quiet vs. Free Quiet.

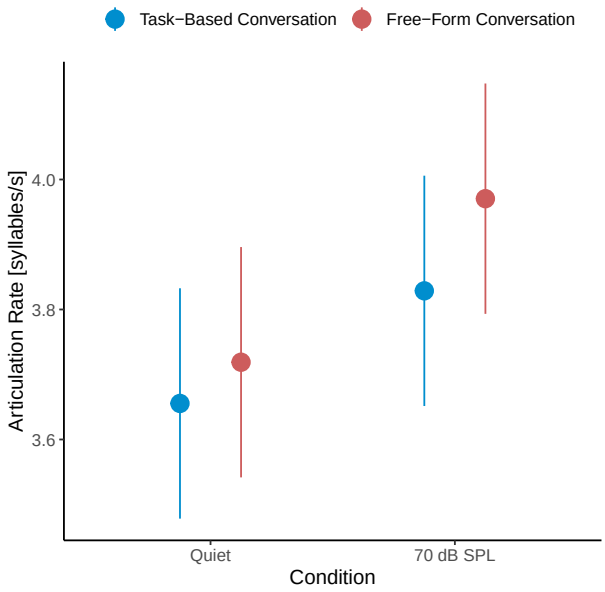


Figure 3.12: Mean articulation rates (defined by the number of syllables produced per unit of time the talker spoke) in quiet (left) and 70 dB SPL noise (right). The estimated marginal means for the task-based (blue) and free-form (red) conversations are from the linear mixed model fitted to the data. The error bars indicate 95% confidence intervals.

For the effect of noise condition, the coefficient is positive and is significantly different from zero. However, the coefficients for the effect of free-form conversation and the interaction of 70 dB SPL x Free-Form are not significantly different from zero.

In summary, articulation rates varied between conversation types, with faster rates observed in free-form conversations. Additionally, the presence of noise led to an increase in articulation rates for both types of conversations.

3.3.6 Speech Levels

During each conversation, the estimated average speech levels were computed at a distance of 1.4 meters from each speaker. A linear mixed model was fitted to the data using the following equation: Speech Level $\sim$ Condition * Conversation Type + (1 | Talker) + (1 |

Replicate). The estimated marginal means of the mean speech levels for the fixed effects of Condition and Conversation Type are plotted in Figure 3.13. The parameters of the fixed effects are shown in Table A.15. The main effect of Condition ($F(1, 215) = 289.40$, $p < 0.0001$) and main effect of Conversation Type ($F(1, 215) = 7.70$, $p < 0.01$) were found to be statistically significant. The interaction of Condition x Conversation Type ($F(1, 215) = 8.27$, $p < 0.01$) was also statistically significant. The results of the pairwise comparison of the means can be found in Table A.16. All pairwise comparisons were significant ($p < 0.001$), except for Task Noise vs. Free Noise.

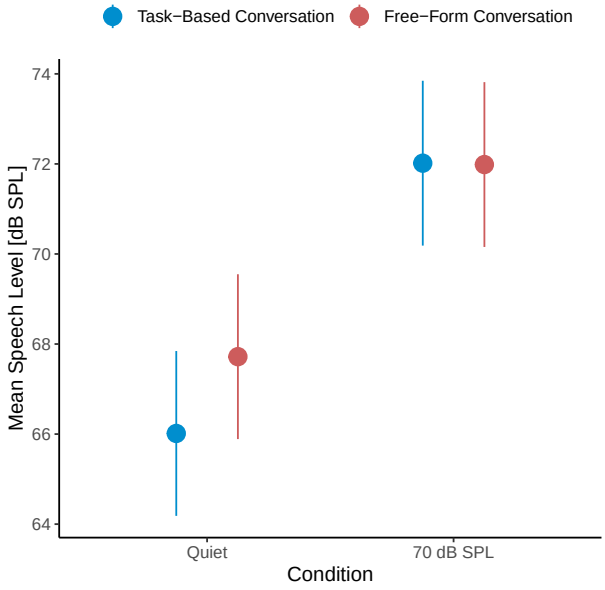


Figure 3.13: Mean speech levels of talkers in quiet (left) and 70 dB SPL noise (right). The estimated marginal means for the task-based (blue) and free-form (red) conversations are from the linear mixed model fitted to the data. The error bars indicate 95% confidence intervals.

For the effect of noise condition, the coefficient is positive and is significantly different from zero. The coefficient for the effect of free-form conversation is positive and significantly different from zero. The interaction of 70 dB SPL x Free-Form shows a negative coefficient and is significantly different from zero.

In summary, the presence of noise resulted in an increase in mean speech levels for both conversation types. Mean speech levels did vary significantly between both types of conversations in quiet where the mean speech levels were higher for the free-form conversations, but did not vary when noise was present.

3.3.7 Speaking Time

The proportion of speaking time was determined by adding up the duration of all IPUs a talker produced during a conversation and dividing this sum by the total length of that conversation. A generalized linear mixed model using a beta distribution (glmmTMB with beta_family) was fitted to the data using the following equation: Proportion of Time Speaking $\sim$ Condition * Conversation Type + (1 | Talker) + (1 | Replicate). The estimated marginal means of the proportion of speaking time for the fixed effects of Condition and Conversation Type are plotted in Figure **??**. The parameters of the fixed effects are shown in Table **??**. For the effects of noise condition, conversation type, and the interaction of 70 dB SPL x Free-Form, all coefficients were not significantly different from zero. The results of the pairwise comparison of the means can be found in Table A.18. All pairwise comparisons were not significant.

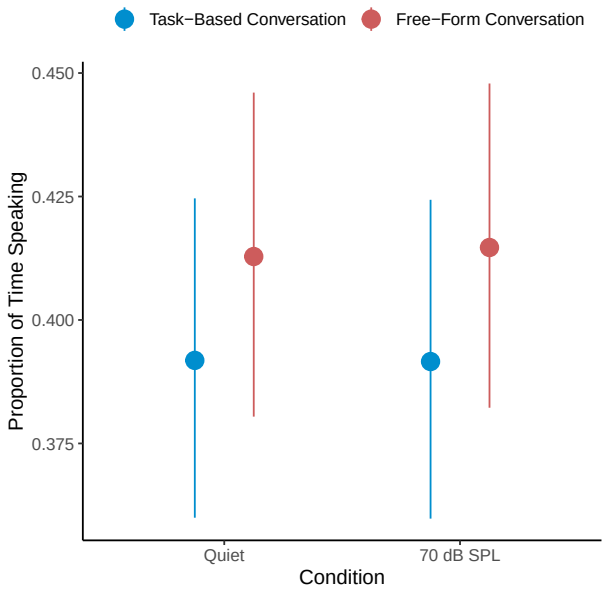


Figure 3.14: Mean proportion of time talkers produced speech during a conversation in quiet (left) and 70 dB SPL noise (right). The estimated marginal means for the task-based (blue) and free-form (red) conversations are from the linear mixed model fitted to the data. The error bars indicate 95% confidence intervals.

In summary, the proportion of time talkers spent talking was consistent across different conversation types and unaffected by the presence of noise.

3.4 Discussion

The purpose of the experiment was to study the impact of the conversation type and the presence of noise on various acoustic and conversational dynamics markers among native English-speaking, young, normal-hearing individuals. The familiarity between talkers is known to influence speech perception, with listeners demonstrating better sentence recognition when speech is produced by a familiar talker compared to a stranger (Buntrock et al., 2021). This familiarity aspect is particularly relevant in our study, as participants were

friends. This design choice eliminated potential social factors related to not knowing the person, such as the need to adapt to unfamiliar speaking styles and voices, thus allowing for a more controlled examination of the impact of conversation type and noise.

In general, our findings suggest that while conversation type significantly influenced the explored parameters, the presence of noise did not uniformly alter these variables as expected. Task-based conversations featured longer average FTOs and higher rates of floor transfers. Free-form conversations exhibited greater variability in FTOs (with a larger proportion of floor transfer exhibiting an acoustic overlap of the two talkers), longer IPUs with a higher number of syllables per IPU, longer pauses, a higher pause rate, increased articulation rate, and elevated speech levels in quiet. The presence of noise resulted in longer IPU durations and pauses, as well as an increase in the number of syllables per IPU, articulation rate, and speech levels, but did not have an effect on the FTO distributions.

3.4.1 Rate of Floor Transfers

A lower rate of floor transfers was observed in free-form compared to task-based conversations. This is likely due to the inherent nature of task-based interactions, which may require participants to engage in more structured turn-taking patterns. Task-based exchanges often involve setting cooperative goals, necessitating frequent floor transfers to exchange information efficiently and achieve task objectives. In contrast, free-form exchanges may be less goal-oriented, allowing for longer turns with fewer floor transfers.

The rate of floor transfers did not significantly change between quiet and noise conditions for either conversation type. This suggests that the 70 dB SPL noise presented during the experiment may not have been sufficient to disrupt natural conversation flow significantly. Therefore, we speculate that the noise level used might not have been high enough to challenge communication dynamics, allowing participants to maintain their usual conversational patterns without needing to adapt to noise-induced challenges.

3.4.2 Floor-Transfer Offset

The presence of broader and more skewed floor-transfer offset (FTO) distributions towards shorter durations in free-form conversations may reflect the less structured and more fluid nature of these interactions. Free-form exchanges are inherently more spontaneous, leading to more variable FTOs. In such conversations, interlocutors often diverge to various topics with different sentiments, which can influence the timing and formation of reactions. This variability may also stem from the nature of the topics being discussed. In free-form conversations among friends, overlaps are more frequent and serve as a cooperative device signifying enthusiasm rather than an interruption (Tannen, 2005). Hönemann (2023) describes these overlaps as instances where participants provide supplementary information or confirmatory feedback without claiming the floor. This collaborative approach fosters a dynamic exchange of ideas and signals interest, thereby encouraging continuous dialogue compared to task-based conversations, which may be more structured.

Watson et al. (2019) observed longer IPUs and a decrease in median FTO during free-form conversations, leading to a higher rate of overlaps. This trend may reflect a prioritization of social interaction quality over the speed of information transmission in relational information. In contrast, task-based interactions such as the Diapix task may emphasize minimizing overlaps to ensure precise information exchange and task accuracy. Participants may intentionally extend their FTOs in these structured, transactional conversations as a way to aim for clearer communication. Alternatively, the longer average FTO may be related to more question-answer interactions that require some visual search (e.g., "Do you see the bird in the upper-right corner?").

Neither the FTO durations nor the FTO IQR were significantly altered by the presence of noise in either type of conversation. This suggests that participants were able to maintain their natural turn-taking patterns effectively despite the noise. It was anticipated that the noise used in this study would have similar effects to those of the ICRA noise used in (Sørensen et al., 2021).

3.4.3 Interpausal Units

The difference in IPU durations between conversation types is likely due to the nature of the Diapix task used to elicit the task-based conversation. Compared to the free-form conversations, a larger proportion of the IPUs in the task-oriented conversations were shorter than 500 ms, indicating that a large number of very short utterances of only one or two syllables were produced. Heldner et al. (2011) suggests brief responses (such as "yes," "uh...," "yup," "huh...") promote efficient communication, contributing to quicker turn-taking in task-based conversations. This is further supported by the significantly lower number of syllables per IPU in task-based interactions. This aligns with the work of Watson et al. (2019), who observed longer utterance durations in free conversations, highlighting the impact of conversational context on speech dynamics.

The increase in IPU durations in both types of conversation in the presence of noise suggests that noise impacts the structuring of speech across different interactional contexts. Longer IPUs are consistent with clearer and more intentional speech, and provide more time for interlocutors to plan and understand speech. Beechey et al. (2018) described the increased utterance duration as part of the "holding-the-floor" strategy, where speakers extend their turns to maintain clarity and ensure their message is understood in noisy conditions. Our findings align with the results of Watson et al. (2019) and Beechey et al. (2018), who observed increased utterance duration in noisy conditions for free-form and task-based conversations, respectively. Hence, it further confirms that challenging listening environments can affect the temporal structuring of speech in various conversational contexts.

3.4.4 Pause Duration and Rate

Our results show an increase in both pause duration and pause rate in free-form conversations compared to task-based ones. This finding is expected due to the inherent nature of free-form conversations, which are more fluid and allow speakers to take longer turns and use more words. In free-form conversations, if talkers take longer turns, they may need more time to consider their thoughts, opinions, or responses and plan their speech appro-

priately. This ongoing cognitive effort likely results in more frequent pauses and longer pause durations as a strategy to plan their next utterance.

In contrast, the task-based conversations elicited in this study involved more concise descriptions and responses. While some visual search must be conducted, the content of the conversations is limited based on the images. Further, an efficient strategy for conducting the task is to start in one location and move systematically around the picture comparing objects. If one talker pauses in order to have more time to plan the description of an object, the other talker may jump in to describe what they see. Consequently, the pauses in task-based conversations are shorter and less frequent.

The presence of noise increased the pause duration while decreasing the pause rate. A pause, defined as a silent interval lasting at least 180 ms, may occur either because a talker needs additional time to plan their utterance or to provide an opportunity for a floor transfer to their partner (which may not be taken if the talker continues speaking). In the present study, both scenarios are plausible explanations for the increased pause duration. As noise levels rise, both interlocutors face greater challenges that may lead to longer pauses due to the increased time required for speech planning. Alternatively, a talker might stop talking to provide an opportunity for their partner to take the floor. In this case, a talker would then need to wait a certain amount of time, likely related to the mean FTO of their partner before resuming their speech. In more difficult conditions, talkers may require more time to consider and plan their speech. Thus, one would expect an increase in pause rate in the noise conditions, but the opposite was observed. We speculate that here, the noise acted as a stressor but did not pose a substantial challenge to the communication. Thus, in response to a stressor, talkers aim to convey their message as quickly as possible leading to fewer pauses and faster articulation rates.

In both task-based and free-form conversations, shorter pre-pause IPU durations were associated with longer pause durations. For free-form conversation, longer post-pause IPU durations were associated with shorter pause durations, but no relationship was observed for the task-based conversations. We speculate that pauses that follow short IPUs reflect the need for a talker to consider and plan a response. In both the task-based and free-form conversations, talkers may have employed short filler words to indicate that they were considering a response. However, as the content of the conversations was not analyzed,

this remains a conjecture. Compared to short utterances, producing longer utterances of connected speech requires larger lung volumes. Thus, the association of short pauses preceding longer IPUs may be related to talkers taking short breaths before continuing with an extended utterance. As the IPUs were shorter in the task-based conversations, talkers may not have needed to pause to breathe. Further, the more rapid rate of floor transfers (i.e., shorter turns) would also provide more opportunities to breathe while a partner was taking their turn.

3.4.5 Speech Production

Our study aimed to build upon past research to explore the influence of noise and conversation type on speech production, particularly articulation rate and speech levels. The observed effects align, for the most part, with the literature on the impact of noise, where interlocutors increase their vocal effort in challenging environments to enhance speech intelligibility, as part of the Lombard effect.

Our findings indicate that articulation rates varied between conversation types, with free-form interactions involving faster rates. Additionally, the presence of noise increased articulation rates in both types of conversations. This partially aligns with Watson et al. (2019), who found that the articulation rate was faster in free-form conversations. However, the presence of noise did not influence the articulation rate in their study. Conversely, Sørensen et al. (2021) found that NH talkers increased their articulation rate when communicating with another NH talker in the presence of 70 dBA babble noise.

The increased articulation rate in free-form conversations may be attributed to the spontaneous and informal nature of such exchanges, where speakers tend to use more familiar words and enjoy longer turns to share information expansively. In contrast, task-based conversations are more structured and goal-oriented, leading interlocutors to focus on accuracy and clarity to complete the task efficiently. This focus may result in slower and more deliberate speech, with less talking in general to ensure precise information transfer.

The increase in articulation rate in the presence of noise may be due to various reasons. Sørensen et al. (2020) suggested that the noise acts as a stressor, causing speakers to talk

more rapidly with fewer pauses. This approach does not necessarily improve intelligibility but is an effort to communicate more quickly. This is further supported by other research findings, which suggest that the increased articulation rate in noise may be an attempt to counteract the masking effects of the noise.

In the present study, we observed the Lombard effect, where speakers increase their speech levels in the presence of noise. While this response can be attributed to the need for an adaptive strategy to improve the SNR and counteract the masking effect of noise, it is not clear the extent to which this is driven by the perception of one's own voice vs. an adaptation to the needs of a conversational partner.

Interestingly, the mean speech levels in quiet were higher in free-form conversations compared to task-based ones, a difference not observed by Watson et al. (2019). This difference may be due to the more dynamic and enthusiastic nature of free-form conversations. However, in the presence of noise, conversation type did not affect speech levels.

3.4.6 Evaluation of Experimental Setup: Assessing the Adequacy of Communication Challenges

Based on previous studies, we anticipated that noisy conditions would lead to an increase in FTOs due to increased communication difficulty. As the effect of noise on FTO has been replicated in several studies, we speculate that the lack of an effect observed in the present study might be due to either the characteristics of the noise used or the experimental setup. The sound texture noise utilized in our experiment, presented at 70 dB SPL, might not have been as effective a masker as the ICRA noise presented at 70 dBA in Sørensen et al. (2021). In addition, participants in the present study wore closed headphones, which likely reduced the level of air-conducted auditory feedback of their own voice. As a consequence, talkers likely produced higher speech levels than they would have if they were wearing open headphones or not wearing headphones at all. Thus, their partners received higher speech levels, resulting in a higher SNR, reducing the difficulty in understanding the speech in noise.

To investigate this, at the time of writing, a follow-up study is being conducted using

the same ICRA noise as Sørensen et al. (2021), presented at 70 dBA and with participants wearing open headphones.

3.5 Conclusion

This study investigated how conversation type and noise presence impact various acoustic and conversational dynamics among native English-speaking, young, normal-hearing individuals. Our findings reveal that conversation type significantly influenced the many metrics that have been suggested as markers of conversational difficulty.

Compared to free-form conversations, task-based conversations produced narrower FTO distributions with a larger mean, more floor transfers, fewer overlaps, shorter IPUs, shorter pause durations, and fewer pauses. These findings support the idea that free-form interactions are inherently more spontaneous and less structured, allowing for greater conversational flexibility.

The presence of noise did not affect conversational dynamics as initially hypothesized. While noise led to increased IPU durations and pause lengths, it did not significantly alter floor-transfer rates or the variability of FTOs. The observed increase in articulation rates in the noisy conditions was likely because the noise acted as a stressor, prompting faster speech. Both conversation types saw increases in mean speech levels in noisy conditions, consistent with the Lombard effect where speakers elevate their vocal effort to enhance intelligibility. However, the magnitude of this increase was not as pronounced as in previous studies, potentially due to the specific characteristics of the noise used and the experimental setup. This suggests that the 70 dB SPL noise used might not have been sufficiently disruptive to affect turn-taking patterns as anticipated. Additionally, the use of closed headphones may have influenced speech levels, reducing the perceived impact of noise on communication.

Future research can explore these dynamics in interactions between NH and HI participants, as well as among HI participants themselves, in triadic or group conversations, and face-to-face settings, which will provide a wider perspective on communication behaviors across different contexts and participant groups.

Chapter 4

Thesis Summary and Future Work

This thesis investigates the effects of communication conditions on speech production and conversational behavior across two studies involving HI and NH individuals.

Chapter 2 examined the impact of face masks and remote microphones on group conversations among HI individuals. The motivation for this study was to assess the possibility of extracting useful information from an existing study that was not originally designed for our purposes, which aimed to evaluate the effect of face masks and remote microphones in near real-life scenarios. In a study by Ishida et al. (2022), groups of four aided HI participants interacted in free-form conversations initiated by a moderator-provided topic. Conversations were collected under four conditions (combinations of using face masks and remote microphones) and were used for content analysis and participants' subjective ratings. The experiment was designed to focus on content analysis and conversation elicitation in a group setting where participants could freely adjust their hearing aids. Interestingly, our findings revealed that remote microphones lead to shorter FTOs and longer conversation durations, indicating enhanced communication. The interaction between microphone and mask conditions on IPU duration is complex, suggesting that while masks impact speech production, remote microphones contribute to shorter IPUs, indicative of easier communication. The study also finds that wearing a face mask decreases the F0 which we speculate is due to an increased airflow resistance, which in return reduces the subglottal pressure. No significant differences in articulation rate and floor transfer rate across

conditions are observed, likely due to the constant background noise level.

The primary motivation for the study in **Chapter 3** emerged from comparing the findings of Chapter 2 to those of previous studies. The paradigm used in the study in Chapter 2 differs along several factors compared to previous studies that have examined FTOs in noise. In the study presented in Chapter 3, we investigated how conversation type (task-based vs. free-form) and noise presence (quiet vs. 70 dB SPL) affect speech production and conversational dynamics among normal-hearing individuals. The results indicate that conversation type significantly influences several metrics that have been proposed for inferring effort. Task-based conversations exhibit longer FTOs and higher rates of floor transfers, while free-form conversations show greater FTO variability, more frequent overlaps, longer IPUs, along with increased pause durations and rates. Contrary to our initial hypotheses, the presence of noise did not significantly alter floor-transfer rates or FTO variability. Both conversation types see increases in speech levels in noisy conditions, consistent with the Lombard effect, although the magnitude of this increase was not as pronounced as expected. Additionally, the articulation rate increased in noisy conditions, which might be attributed to the noise acting as a stressor, prompting faster speech. This suggests that the specific characteristics of the noise used and the experimental setup may have influenced the outcomes.

Future research should extend and build upon these findings by exploring interactions between participants with different hearing statuses (e.g., effects of hearing impairment) and interactions with more than two participants, such as triadic or group interactions, with different conversational contexts. These interactions could be studied both face-to-face and remotely to provide a more comprehensive understanding of the factors affecting conversational dynamics. This exploration is crucial for advancing the development of effective hearing-assistive technologies and ensuring evaluations reflect the complexity of everyday communication scenarios.

At the time of the thesis writing, a follow-up study to the study in Chapter 3 is being conducted using a 70 dBA ICRA noise and open headphones to further refine the findings related to noise impact on communication behaviors.

References

Amlani, A. M., Punch, J. L., and Ching, T. Y. C. (2002). Methods and Applications of the Audibility Index in Hearing Aid Selection and Fitting. *Trends in Amplification*, 6(3):81–129.

Argyle, M., Salter, V., Nicholson, H., Williams, M., and Burgess, P. (1970). The Communication of Inferior and Superior Attitudes by Verbal and Non-verbal Signals*. *British Journal of Social and Clinical Psychology*, 9(3):222–231.

Atcherson, S. R., Mendel, L. L., Baltimore, W. J., Patro, C., Lee, S., Pousson, M., and Spann, M. J. (2017). The effect of conventional and transparent surgical masks on speech understanding in individuals with and without hearing loss. *Journal of the American Academy of Audiology*, 28(1):58–67.

Aubanel, V., Cooke, M., Villegas, J., and Lecumberri, M. L. G. (2011). Conversing in the presence of a competing conversation: effects on speech production. In *Interspeech 2011*, pages 2833–2836, ISCA. ISCA.

Badh, G. and Knowles, T. (2023). Acoustic and perceptual impact of face masks on speech: A scoping review. *PLOS ONE*, 18(8):e0285009.

Baker, R. and Hazan, V. (2011). DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods*, 43(3):761–770.

Barthel, M. and Sauppe, S. (2019). Speech Planning at Turn Transitions in Dialog Is Associated With Increased Processing Load. *Cognitive Science*, 43(7).

Barthel, M., Sauppe, S., Levinson, S. C., and Meyer, A. S. (2016). The Timing of Utterance Planning in Task-Oriented Dialogue: Evidence from a Novel List-Completion Paradigm. *Frontiers in Psychology*, 7.

Beechey, T., Buchholz, J. M., and Keidser, G. (2018). Measuring communication difficulty through effortful speech production during conversation. *Speech Communication*, 100:18–29.

Beechey, T., Buchholz, J. M., and Keidser, G. (2020). Hearing Impairment Increases Communication Effort During Conversations in Noise. *Journal of Speech, Language, and Hearing Research*, 63(1):305–320.

Bernstein, L. E., Auer, E. T., and Eberhardt, S. P. (2022). During Lipreading Training With Sentence Stimuli, Feedback Controls Learning and Generalization to Audiovisual Speech in Noise. *American Journal of Audiology*, 31(1):57–77.

Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., and Rzeczkowski, C. (1984). Standardization of a Test of Speech Perception in Noise. *Journal of Speech, Language, and Hearing Research*, 27(1):32–48.

Bögels, S., Magyari, L., and Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific Reports*, 5(1):12881.

Bögels, S. and Torreira, F. (2021). Turn-end Estimation in Conversational Turn-taking: The Roles of Context and Prosody. *Discourse Processes*, 58(10):903–924.

Bóna, J. (2014). Temporal characteristics of speech: The effect of age and speech style. *The Journal of the Acoustical Society of America*, 136(2):EL116–EL121.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., and Brennan, S. E. (2001). Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech*, 44(2):123–147.

Buntrock, M. S., Barker, B. A., Gurries, M. M., and Barrett, T. S. (2021). The Effect of Talker Familiarity on Sentence Recognition Accuracy in Complex Noise. *Experimental Psychology*, 68(1):49–55.

Charlip, W. S. and Burk, K. W. (1969). Effects of noise on selected speech parameters. *Journal of Communication Disorders*, 2(3):212–219.

Choi, J. H., Choi, H. J., Kim, D. H., Park, J. H., An, Y.-H., and Shim, H. J. (2022). Effect of face masks on speech perception in noise of individuals with hearing aids. *Frontiers in Neuroscience*, 16.

Corey, R. M., Jones, U., and Singer, A. C. (2020). Acoustic effects of medical, cloth, and transparent face masks on speech signals. *The Journal of the Acoustical Society of America*, 148(4):2371–2375.

Corps, R. E., Crossley, A., Gambi, C., and Pickering, M. J. (2018). Early preparation during turn-taking: Listeners use content predictions to determine what to say but not when to say it. *Cognition*, 175:77–95.

Coupland, J. (2000). *Small Talk*. Pearson Education, Harlow, UK.

Davis, C., Kim, J., Grauwinkel, K., and Mixdorff, H. (2006). Lombard speech: auditory (a), visual (v) and AV effects. In *Speech Prosody 2006*, pages 252–0, ISCA. ISCA.

de Jong, N. H., Pacilly, J., and Heeren, W. (2021). PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically. *Assessment in Education: Principles, Policy & Practice*, 28(4):456–476.

De Ruiter, J. P., Mitterer, H., and Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3):515–535.

Demorest, M. E. and Erdman, S. A. (1987). Development of the Communication Profile for the Hearing Impaired. *Journal of Speech and Hearing Disorders*, 52(2):129–143.

Diberardinis, J., Tocci, J. A., and O'Brien, K. (1979). The Impact of Task-Oriented Communication and Visual Impairment on Competitive Groups. *Group & Organization Studies*, 4(2):229–238.

Draper, M. H., Ladefoged, P., and Whitteridge, D. (1960). Expiratory Pressures and Air Flow During Speech. *BMJ*, 1(5189):1837–1843.

Edwards, B. (2007). The Future of Hearing Aid Technology. *Trends in Amplification*, 11(1):31–45.

Engeroff, T., Groneberg, D. A., and Niederer, D. (2021). The Impact of Ubiquitous Face Masks and Filtering Face Piece Application During Rest, Work and Exercise on Gas Exchange, Pulmonary Function and Physical Performance: A Systematic Review with Meta-analysis. *Sports Medicine - Open*, 7(1):92.

Federmeier, K. D., Jongman, S. R., and Szewczyk, J. M. (2020). Examining the Role of General Cognitive Skills in Language Processing: A Window Into Complex Cognition. *Current Directions in Psychological Science*, 29(6):575–582.

Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language*, 30(2):210–233.

Folk, L. and Schiel, F. (2011). The lombard effect in spontaneous dialog speech. In *Interspeech 2011*, pages 2701–2704, ISCA. ISCA.

Foti, D. and Roberts, F. (2016). The neural dynamics of speech perception: Dissociable networks for processing linguistic content and monitoring speaker turn-taking. *Brain and Language*, 157-158:63–71.

Fouquet, M., Pisanski, K., Mathevon, N., and Reby, D. (2016). Seven and up: individual differences in male voice fundamental frequency emerge before puberty and remain stable throughout adulthood. *Royal Society Open Science*, 3(10):160395.

Froehle, C. M. (2006). Service Personnel, Technology, and Their Interaction in Influencing Customer Satisfaction*. *Decision Sciences*, 37(1):5–38.

Fuchs, S. and Rochet-Capellan, A. (2021). The Respiratory Foundations of Spoken Language. *Annual Review of Linguistics*, 7(1):13–30.

Garnier, M., Henrich, N., and Dubois, D. (2010). Influence of Sound Immersion and Communicative Interaction on the Lombard Effect. *Journal of Speech, Language, and Hearing Research*, 53(3):588–608.

Gifford, R. H., Dorman, M. F., Skarzynski, H., Lorens, A., Polak, M., Driscoll, C. L. W., Roland, P., and Buchman, C. A. (2013). Cochlear Implantation With Hearing Preservation Yields Significant Benefit for Speech Recognition in Complex Listening Environments. *Ear & Hearing*, 34(4):413–425.

Gisladottir, R. S., Chwilla, D. J., and Levinson, S. C. (2015). Conversation Electrified: ERP Correlates of Speech Act Recognition in Underspecified Utterances. *PLOS ONE*, 10(3):e0120068.

Godfrey, J., Holliman, E., and McDaniel, J. (1992). SWITCHBOARD: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 517–520. IEEE.

Gordon-Salant, S. and Fitzgibbons, P. J. (1997). Selected Cognitive Factors and Speech Recognition Performance Among Young and Elderly Listeners. *Journal of Speech, Language, and Hearing Research*, 40(2):423–431.

Habermas, J. (1976). What is universal pragmatics? In Cooke, M., editor, *On the Pragmatics of Communication*, pages 21–104. MIT Press, Cambridge, MA, USA.

Hagoort, P. and Indefrey, P. (2014). The Neurobiology of Language Beyond Single Words. *Annual Review of Neuroscience*, 37(1):347–362.

Hanley, T. D. and Steer, M. D. (1949). Effect of Level of Distracting Noise upon Speaking Rate, Duration and Intensity. *Journal of Speech and Hearing Disorders*, 14(4):363–368.

Hartman, D. E. and Danhauer, J. L. (1976). Perceptual features of speech for males in four perceived age decades. *The Journal of the Acoustical Society of America*, 59(3):713–715.

Hazan, V. and Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *The Journal of the Acoustical Society of America*, 130(4):2139–2152.

Hee Lee, J. and Humes, L. E. (2012). Effect of fundamental-frequency and sentence-onset differences on speech-identification performance of young and older adults in a competing-talker background. *The Journal of the Acoustical Society of America*, 132(3):1700–1717.

Heldner, M. and Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Heldner, M., Edlund, J., Hjalmarsson, A., and Laskowski, K. (2011). Very short utterances and timing in turn-taking. In *Interspeech 2011*, pages 2837–2840, ISCA. ISCA.

Holler, J., Kendrick, K. H., Casillas, M., and Levinson, S. C., editors (2016). *Turn-Taking in Human Communicative Interaction*. Frontiers Research Topics. Frontiers Media SA.

Hönemann, A. (2023). Cooperative aspects in a multi-talker interview scenario. In *Proceedings of the 13th International Conference of Nordic Prosody*, pages 140–146. Sciendo.

Indefrey, P. (2011). The spatial and temporal signatures of word production components: A critical update.

Indefrey, P. and Levelt, W. (2004). The spatial and temporal signatures of word production components. *Cognition*, 92(1-2):101–144.

Ishida, I., Pichora-Fuller, K., Dillon Edgett, L., Voss, S. C., Paccioretti, D., Trusler, A., Liu, Y., Cui, E., and Qian, J. (2022). The effect of a Remote Microphone Technology on the group listening experiences of older adults. In *International Hearing Aid Research Conference*.

Iwarsson, J., Thomasson, M., and Sundberg, J. (1998). Effects of lung volume on the glottal voice source. *Journal of Voice*, 12(4):424–433.

Jacewicz, E., Fox, R. A., and Wei, L. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *The Journal of the Acoustical Society of America*, 128(2):839–850.

Jarvenpaa, S. L. and Leidner, D. E. (1998). Communication and Trust in Global Virtual Teams. *Journal of Computer-Mediated Communication*, 3(4):0–0.

Jessen, M., Koster, O., and Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech, Language and the Law*, 12(2):174–213.

Junqua, J.-C., Fincke, S., and Field, K. (1999). The Lombard effect: a reflex to better communicate with others in noise. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, pages 2083–2086. IEEE.

Kang, I., Han, S., and Lee, J. (2017). Task-Oriented and Relationship-Building Communications between Air Traffic Controllers and Pilots. *Sustainability*, 9(10):1770.

Kendon, A. (2004). *Gesture Visible Action as Utterance*. Cambridge University Press.

Kim, J., Davis, C., Vignali, G., and Hill, H. (2005). A visual concomitant of the Lombard reflex . In *Auditory-Visual Speech Processing*, pages 17–22, British Columbia, Canada.

Lane, H. and Tranel, B. (1971). The Lombard Sign and the Role of Hearing in Speech. *Journal of Speech and Hearing Research*, 14(4):677–709.

Lessa, A. H. and Costa, M. J. (2013). The impact of speech rate on sentence recognition by elderly individuals. *Brazilian Journal of Otorhinolaryngology*, 79(6):745–752.

Levinson, S. C. and Holler, J. (2014). The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651):20130302.

Levinson, S. C. and Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6.

Lieberman, P. (1966). *Intonation, Perception, and Language*. PhD thesis, MIT, Cambridge.

Liénard, J.-S. and Di Benedetto, M.-G. (1999). Effect of vocal effort on spectral properties of vowels. *The Journal of the Acoustical Society of America*, 106(1):411–422.

Lindblom, B. (1990). Explaining Phonetic Variation: A Sketch of the H&H Theory. In *Speech Production and Speech Modelling*, pages 403–439. Springer Netherlands, Dordrecht.

Lombard, E. (1911). Le signe de l'elevation de la voix. *Annals maladiers oreille, Larynx, Nez, Pharynx*, 37:101–119.

Magyari, L., Bastiaansen, M. C. M., de Ruiter, J. P., and Levinson, S. C. (2014). Early Anticipation Lies behind the Speed of Response in Conversation. *Journal of Cognitive Neuroscience*, 26(11):2530–2539.

Magyari, L. and de Ruiter, J. P. (2012). Prediction of Turn-Ends Based on Anticipation of Upcoming Words. *Frontiers in Psychology*, 3.

Marek, E.-M., van Kampen, V., Jettkant, B., Kendzia, B., Strauß, B., Sucker, K., Ulbrich, M., Deckert, A., Berresheim, H., Eisenhawer, C., Hoffmeyer, F., Weidhaas, S., Behrens, T., Brüning, T., and Bünger, J. (2023). Effects of wearing different face masks on cardiopulmonary performance at rest and exercise in a partially double-blinded randomized cross-over study. *Scientific Reports*, 13(1):6950.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech 2017*, pages 498–502, ISCA. ISCA.

McFarland, D. H. (2001). Respiratory Markers of Conversational Interaction. *Journal of Speech, Language, and Hearing Research*, 44(1):128–143.

McKenna, V. S., Kendall, C. L., Patel, T. H., Howell, R. J., and Gustin, R. L. (2022). Impact of Face Masks on Speech Acoustics and Vocal Effort in Healthcare Professionals. *The Laryngoscope*, 132(2):391–397.

McNeill, D. (1995). Hand and Mind. In *Advances in Visual Semiotics*. DE GRUYTER, Berlin, Boston.

McWalter, R. and McDermott, J. H. (2018). Adaptive and Selective Time Averaging of Auditory Scenes. *Current Biology*, 28(9):1405–1418.

Menenti, L., Gierhan, S. M. E., Segaert, K., and Hagoort, P. (2011). Shared Language. *Psychological Science*, 22(9):1173–1182.

Meyer, A. S. (2023). Timing in Conversation. *Journal of Cognition*, 6(1).

Miles, K., Beechey, T., Best, V., and Buchholz, J. (2022). Measuring Speech Intelligibility and Hearing-Aid Benefit Using Everyday Conversational Sentences in Real-World Environments. *Frontiers in Neuroscience*, 16.

Ogawa, T., Koike, J., and Hirano, Y. (2022). The effects of cloth face masks on cardiorespiratory responses and VO2 during maximal incremental running protocol among apparently healthy men. *Scientific Reports*, 12(1):22292.

O'Neill, J. J. (1954). Contributions Of The Visual Components Of Oral Symbols To Speech Comprehension. *Journal of Speech and Hearing Disorders*, 19(4):429–439.

Perniss, P. (2018). Why We Should Study Multimodal Language. *Frontiers in Psychology*, 9.

Petersen, E. B., MacDonald, E. N., and Josefine Munch Sørensen, A. (2022). The Effects of Hearing-Aid Amplification and Noise on Conversational Dynamics Between Normal-Hearing and Hearing-Impaired Talkers. *Trends in Hearing*, 26.

Rahne, T., Fröhlich, L., Plontke, S., and Wagner, L. (2021). Influence of surgical and N95 face masks on speech perception and listening effort in noise. *PLOS ONE*, 16(7):e0253874.

Riest, C., Jorschick, A. B., and de Ruiter, J. P. (2015). Anticipation in turn-taking: Mechanisms and information sources. *Frontiers in Psychology*, 6(FEB).

Roberts, S. G., Torreira, F., and Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: a corpus study. *Frontiers in Psychology*, 6.

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., Dahlström, Ö., Signoret, C., Stenfelt, S., Pichora-Fuller, M. K., and Rudner, M. (2013). The Ease

of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience*, 7.

Rönnberg, J., Rudner, M., Foo, C., and Lunner, T. (2008). Cognition counts: A working memory system for ease of language understanding (ELU). *International Journal of Audiology*, 47(sup2):S99–S105.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. Technical Report 4.

Schmitter-Edgecombe, M. (2000). Aging and Word-Finding A Comparison of Spontaneous and Constrained Naming Tests. *Archives of Clinical Neuropsychology*, 15(6):479–493.

Segaert, K., Menenti, L., Weber, K., Petersson, K. M., and Hagoort, P. (2012). Shared Syntax in Language Production and Language Comprehension–An fMRI Study. *Cerebral Cortex*, 22(7):1662–1670.

Smeds, K., Gotowiec, S., Wolters, F., Herrlin, P., Larsson, J., and Dahlquist, M. (2020). Selecting Scenarios for Hearing-Related Laboratory Testing. *Ear & Hearing*, 41(Supplement 1):20S–30S.

Sørensen, A. J. M., Fereczkowski, M., and MacDonald, E. N. (2021). Effects of Noise and Second Language on Conversational Dynamics in Task Dialogue. *Trends in Hearing*, 25.

Sørensen, A. J. M., Macdonald, E. N., and Lunner, T. (2020). Timing of turn taking between normal-hearing and hearing-impaired interlocutors. In *International Symposium on Auditory and Audiological Research*, pages 37–44, Nyborg, Denmark. ISAAR.

Sorqvist, P. (2015). On interpretation and task selection: the sub-component hypothesis of cognitive noise effects. *Frontiers in Psychology*, 5.

Stangor, C. and Walinga, J. (2014). Communicating with Others: The Development and Use of Language. In *Introduction to Psychology*, chapter 10. BCCampus, 1st canadian edition edition.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., and Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

Sumby, W. H. and Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, 26(2):212–215.

Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Stokes, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, 84(3):917–928.

Sundberg, J., Scherer, R., Hess, M., Müller, F., and Granqvist, S. (2013). Subglottal Pressure Oscillations Accompanying Phonation. *Journal of Voice*, 27(4):411–421.

Sweetow, R. W. and Henderson-Sabes, J. (2004). The case for LACE. *The Hearing Journal*, 57(3):32–35.

Tannen, D. (2005). *Conversational style: analyzing talk among friends*. Oxford University Press, New York, new edition edition.

Thibodeau, L. M., Thibodeau-Nielsen, R. B., Tran, C. M. Q., and Jacob, R. T. d. S. (2021). Communicating During COVID-19: The Effect of Transparent Masks for Speech Recognition in Noise. *Ear & Hearing*, 42(4):772–781.

Torreira, F., Bögels, S., and Levinson, S. C. (2015). Breathing for answering: the time course of response planning in conversation. *Frontiers in Psychology*, 6.

Traser, L., Burk, F., Özen, A. C., Burdumy, M., Bock, M., Blaser, D., Richter, B., and Echternach, M. (2020). Respiratory kinematics and the regulation of subglottic pressure for phonation of pitch jumps – a dynamic MRI study. *PLOS ONE*, 15(12):e0244539.

Trujillo, J., Özyürek, A., Holler, J., and Drijvers, L. (2021). Speakers exhibit a multimodal Lombard effect in noise. *Scientific Reports*, 11(1):16721.

Uchanski, R. M., Choi, S. S., Braida, L. D., Reed, C. M., and Durlach, N. I. (1996). Speaking Clearly for the Hard of Hearing IV: Further Studies of the Role of Speaking Rate. *Journal of Speech, Language, and Hearing Research*, 39(3):494–509.

Vainio, M., Aalto, D., Suni, A., Arnhold, A., Raitio, T., Seijo, H., Järvikivi, J., and Alku, P. (2012). Effect of noise type and level on focus related fundamental frequency changes. In *Interspeech 2012*, pages 671–674, ISCA. ISCA.

Vlaj, D. and Kacic, Z. (2011). The Influence of Lombard Effect on Speech Recognition. In *Speech Technologies*. InTech.

Wagener, K. C., Vormann, M., Latzel, M., and Mülder, H. E. (2018). Effect of Hearing Aid Directionality and Remote Microphone on Speech Intelligibility in Complex Listening Situations. *Trends in Hearing*, 22.

Watson, P. J., Ciccia, A. H., and Weismer, G. (2003). The relation of lung volume initiation to selected acoustic properties of speech. *The Journal of the Acoustical Society of America*, 113(5):2812–2819.

Watson, S., Sørensen, A. J. M., and MacDonald, E. N. (2019). The effect of conversational task on turn taking in dialogue. In *International Symposium on Auditory and Audiological Research (ISAAR)*, pages 61–68, Nyborg. The Danavox Jubilee Foundation.

Whalen, D. and Kinsella-Shaw, J. (1997). Exploring the Relationship of Inspiration Duration to Utterance Duration. *Phonetica*, 54(3-4):138–152.

Winkworth, A. L., Davis, P. J., Adams, R. D., and Ellis, E. (1995). Breathing Patterns During Spontaneous Speech. *Journal of Speech, Language, and Hearing Research*, 38(1):124–144.

Winn, M. B. and Teece, K. H. (2021). Slower Speaking Rate Reduces Listening Effort Among Listeners With Cochlear Implants. *Ear & Hearing*, 42(3):584–595.

Yoon, M. S. (2004). Speech Rate and Pause Characteristics of Adults with Hearing Impairment. *Commun Sci Disord*, 9(1):15–29.

Zhang, Z. (2016a). Mechanics of human voice production and control. *The Journal of the Acoustical Society of America*, 140(4):2614–2635.

Zhang, Z. (2016b). Respiratory Laryngeal Coordination in Airflow Conservation and Reduction of Respiratory Effort of Phonation. *Journal of Voice*, 30(6):7–760.

APPENDICES

Appendix A

A.1 Supplementary Information

A.1.1 Rate of Floor Transfers

Table A.1: Fixed effects of the linear mixed model of floor transfer rate. The values in the square brackets indicate 95% confidence intervals.

Intercept (Task-Based in Quiet)	$b_0 = 25.57$ [23.70, 27.44]	$t(15.90) = 27.19$, $p < 0.0001$
70 dB	$b_1 = -0.83$ [-2.38, 0.72]	$t(105.0) = -1.04$, $p = 0.30$
Free-Form	$b_2 = -10.68$ [-12.23, -9.13]	$t(105.0) = -13.42$, $p < 0.001$
70 dB x Free-Form	$b_3 = 0.77$ [-1.43, 2.96]	$t(105.0) = 0.68$, $p = 0.50$

Table A.2: Pairwise comparisons of floor transfer rates. The values in square brackets indicate standard error. The p values have been adjusted using mvt to compensate for multiple comparisons.

Task Quiet vs. Task Noise	$\Delta M = 0.83$ [0.80]	$t(105.0) = 1.04$, $p = 0.67$
Free Quiet vs. Free Noise	$\Delta M = 0.06$ [0.80]	$t(105.0) = 0.08$, $p = 1.00$
Task Quiet vs. Free Quiet	$\Delta M = 10.68$ [0.80]	$t(105.0) = 13.42$, $p < 0.0001$
Task Noise vs. Free Noise	$\Delta M = 9.91$ [0.80]	$t(105.0) = 12.46$, $p < 0.0001$

A.1.2 Floor-Transfer Offsets

Table A.3: Pairwise comparisons of mean FTO. The tests were performed on the log scale. The values in the second column are the ratio of means and square brackets indicate standard error. The z score and corresponding p value are given in the third column. The p values have been adjusted using mvt to compensate for multiple comparisons.

Task Quiet vs. Task Noise	1.009 [0.018]	0.503, $p = 0.96$
Free Quiet vs. Free Noise	0.942 [0.024]	-2.387, $p = 0.06$
Task Quiet vs. Free Quiet	1.094 [0.024]	4.030, $p < 0.001$
Task Noise vs. Free Noise	1.021 [0.023]	0.930, $p = 0.74$

Table A.4: Fixed effects of the linear mixed model of FTO IQR. The values in the square brackets indicate 95% confidence intervals.

Intercept (Task-Based in Quiet)	$b_0 = 369.90$ [337.26, 402.54]	$t(50.6) = 22.23$, $p < 0.0001$
70 dB	$b_1 = -0.60$ [-34.15, 32.95]	$t(217.0) = -0.04$, $p = 0.97$
Free-Form	$b_2 = 109.85$ [76.30, 143.40]	$t(217.0) = 6.40$, $p < 0.0001$
70 dB x Free-Form	$b_3 = -13.95$ [-61.40, 33.50]	$t(217.0) = -0.58$, $p = 0.57$

Table A.5: Pairwise comparisons of mean FTO IQR. The values in square brackets indicate standard error. The p values have been adjusted using mvt to compensate for multiple comparisons.

Task Quiet vs. Task Noise	$\Delta M = 0.6$ [17.2]	$t(217) = 0.04$, $p = 1.0000$
Free Quiet vs. Free Noise	$\Delta M = 14.6$ [17.2]	$t(217) = 0.85$, $p = 0.80$
Task Quiet vs. Free Quiet	$\Delta M = -109.8$ [17.2]	$t(217) = -6.40$, $p < 0.0001$
Task Noise vs. Free Noise	$\Delta M = -95.9$ [17.2]	$t(217) = -5.59$, $p < 0.0001$

A.1.3 Interpausal Units

Table A.6: Pairwise comparisons of mean IPU duration. The tests were performed on the log scale. The values in the second column are the ratio of means and square brackets indicate standard error. The z score and corresponding p value are given in the third column. The p values have been adjusted using mvt to compensate for multiple comparisons.

Task Quiet vs. Task Noise	0.96 [0.015]	-2.67, $p = 0.03$
Free Quiet vs. Free Noise	0.96 [0.017]	-2.49, $p = 0.04$
Task Quiet vs. Free Quiet	0.74 [0.012]	-17.88, $p < 0.0001$
Task Noise vs. Free Noise	0.74 [0.013]	-17.77, $p < 0.0001$

Table A.7: Pairwise comparisons of mean number of syllables per IPU. The tests were performed on log scale. The values in the second column are the ratio of means and square brackets indicate standard error. The z score and corresponding p value are given in the third column. The p values have been adjusted using mvt to compensate for multiple comparisons.

Task Quiet vs. Task Noise	0.922 [0.00734]	-10.235, $p < 0.0001$
Free Quiet vs. Free Noise	0.904 [0.00676]	-13.457, $p < 0.0001$
Task Quiet vs. Free Quiet	0.727 [0.00572]	-40.568, $p < 0.0001$
Task Noise vs. Free Noise	0.713 [0.00540]	-44.655, $p < 0.0001$

A.1.4 Pause Duration and Rate

Table A.8: Fixed effects of the generalized linear mixed model of pause duration for Free-Form Conversation. The values in the square brackets indicate 95% confidence intervals.

Intercept (Free-Form in Quiet)	$b_0 = 5.79$ [5.72, 5.85]	$t = 179.21$, $p < 0.0001$
70 dB	$b_1 = 0.06$ [0.01, 0.11]	$t = 2.13$, $p = 0.03$
Pre-Pause IPU Duration	$b_2 = -0.07$ [-0.10, -0.05]	$t = -5.50$, $p < 0.0001$
Post-Pause IPU Duration	$b_3 = -0.05$ [-0.07, -0.02]	$t = -3.66$, $p < 0.001$

Table A.9: Fixed effects of the generalized linear mixed model of pause duration for Task-Based Conversation. The values in the square brackets indicate 95% confidence intervals.

Intercept (Task-Based in Quiet)	$b_0 = 5.71$ [5.62, 5.80]	$t = 126.00$, $p < 0.0001$
70 dB	$b_1 = 0.10$ [0.03, 0.17]	$t = 2.88$, $p = 0.004$
Pre-Pause IPU Duration	$b_2 = -0.17$ [-0.21, -0.13]	$t = -8.16$, $p < 0.0001$
Post-Pause IPU Duration	$b_3 = 0.00$ [-0.04, 0.04]	$t = 0.04$, $p = 0.97$

Table A.10: Pairwise comparisons of mean pause durations. The tests were performed on the log scale. The values in the second column are the ratio of means and square brackets indicate standard error. The z score and corresponding p value are given in the third column. The p values have been adjusted using mvt to compensate for multiple comparisons.

Task Quiet vs. Task Noise	0.940 [0.028]	-2.09, $p = 0.12$
Free Quiet vs. Free Noise	0.897 [0.031]	-3.19, $p < 0.01$
Task Quiet vs. Free Quiet	1.072 [0.034]	2.16, $p = 0.10$
Task Noise vs. Free Noise	1.023 [0.034]	0.69, $p = 0.87$

Table A.11: Fixed effects of the linear mixed model of pause rates. The values in the square brackets indicate 95% confidence intervals.

Intercept (Task-Based in Quiet)	$b_0 = 12.87$ [11.32, 14.42]	$t(31.4) = 16.44$, $p < 0.0001$
70 dB	$b_1 = -1.09$ [-2.28, 0.10]	$t(217.0) = -1.80$, $p = 0.07$
Free-Form	$b_2 = 2.30$ [1.11, 3.48]	$t(217.0) = 3.79$, $p < 0.001$
70 dB x Free-Form	$b_3 = 0.04$ [-1.64, 1.71]	$t(217.0) = 0.04$, $p = 0.97$

Table A.12: Pairwise comparisons of mean pause rates. The values in square brackets indicate standard error. The p values have been adjusted using mvt to compensate for multiple comparisons.

Task Quiet vs. Task Noise	$\Delta M = 1.09$ [0.607]	$t(217) = 1.797$, $p = 0.23$
Free Quiet vs. Free Noise	$\Delta M = 1.05$ [0.607]	$t(217) = 1.739$, $p = 0.25$
Task Quiet vs. Free Quiet	$\Delta M = -2.30$ [0.607]	$t(217) = -3.787$, $p < 0.001$
Task Noise vs. Free Noise	$\Delta M = -2.33$ [0.607]	$t(217) = -3.845$, $p < 0.001$

A.1.5 Articulation Rate

Table A.13: Fixed effects of the linear mixed model of articulation rates. The values in the square brackets indicate 95% confidence intervals.

Intercept (Task-Based in Quiet)	$b_0 = 3.66$ [3.48, 3.83]	$t(24.5) = 24.48$, $p < 0.0001$
70 dB	$b_1 = 0.17$ [0.08, 0.27]	$t(217.0) = 3.57$, $p < 0.001$
Free-Form	$b_2 = 0.06$ [-0.03, 0.16]	$t(217.0) = 1.31$, $p = 0.19$
70 dB x Free-Form	$b_3 = 0.08$ [-0.06, 0.21]	$t(217.0) = 1.14$, $p = 0.26$

Table A.14: Pairwise comparisons of mean articulation rates. The values in square brackets indicate standard error. The p values have been adjusted using mvt to compensate for multiple comparisons.

Task Quiet vs. Task Noise	$\Delta M = -0.17$ [0.0486]	$t(217) = -3.57$, $p < 0.01$
Free Quiet vs. Free Noise	$\Delta M = -0.25$ [0.0486]	$t(217) = -5.18$, $p < 0.0001$
Task Quiet vs. Free Quiet	$\Delta M = -0.06$ [0.0486]	$t(217) = -1.31$, $p = 0.50$
Task Noise vs. Free Noise	$\Delta M = -0.14$ [0.0486]	$t(217) = -2.92$, $p = 0.01$

A.1.6 Speech Levels

Table A.15: Fixed effects of the linear mixed model of estimated speech levels. The values in the square brackets indicate 95% confidence intervals.

Intercept (Task-Based in Quiet)	$b_0 = 66.01$ [64.25, 67.78]	$t(22.8) = 74.65$, $p < 0.0001$
70 dB	$b_1 = 6.00$ [5.17, 6.84]	$t(215.0) = 14.06$, $p < 0.0001$
Free-Form	$b_2 = 1.71$ [0.87, 2.54]	$t(215.0) = 4.00$, $p < 0.0001$
70 dB x Free-Form	$b_3 = -1.74$ [-2.92, -0.56]	$t(215.0) = -2.88$, $p < 0.01$

Table A.16: Pairwise comparisons of mean speech levels. The values in square brackets indicate standard error. The p values have been adjusted using mvt to compensate for multiple comparisons.

Task Quiet vs. Task Noise	$\Delta M = -6.00$ [0.427]	$t(215) = -14.06$, $p < 0.0001$
Free Quiet vs. Free Noise	$\Delta M = -4.27$ [0.427]	$t(215) = -10.00$, $p < 0.0001$
Task Quiet vs. Free Quiet	$\Delta M = -1.71$ [0.427]	$t(215) = -4.00$, $p < 0.001$
Task Noise vs. Free Noise	$\Delta M = 0.03$ [0.427]	$t(215) = 0.07$, $p = 1.00$

A.1.7 Speaking Time

Table A.17: Fixed effects of the generalized linear mixed model of the proportion of speaking time. The values in the square brackets indicate 95% confidence intervals.

Intercept (Task-Based in Quiet)	$b_0 = -0.44$ [-0.58, -0.30]	$z = -6.34$, $p < 0.0001$
70 dB	$b_1 = -0.00$ [-0.13, 0.13]	$z = -0.02$, $p = 0.99$
Free-Form	$b_2 = 0.09$ [-0.04, 0.22]	$z = 1.34$, $p = 0.18$
70 dB x Free-Form	$b_3 = 0.01$ [-0.17, 0.19]	$z = 0.09$, $p = 0.93$

Table A.18: Pairwise comparisons of mean speaking time. The tests were performed on log odds ratio scale. The values in the second column are the ratio of means and square brackets indicate standard error. The z score and corresponding p value are given in the third column. The p values have been adjusted using mvt to compensate for multiple comparisons.

Task Quiet vs. Task Noise	1.00 [0.065]	0.02, $p = 1.00$
Free Quiet vs. Free Noise	0.99 [0.064]	-0.12, $p = 1.00$
Task Quiet vs. Free Quiet	0.92 [0.060]	-1.34, $p = 0.47$
Task Noise vs. Free Noise	0.91 [0.059]	-1.48, $p = 0.39$