Measuring Interactivity in Audio Teleconferencing using Conversation Analysis

Karis L. Bailey

A thesis submitted for the degree of

Doctor of Philosophy

School of Computer Science and Electronic Engineering

University of Essex

July 2020

Abstract

The aim of this research is targeted directly at determining the quality of multi-party teleconferencing with or without the presence of spatialised audio and induced audio transmission delay. The objectives of the work are to explore, in an interdisciplinary manner, the combination of engineering metrics and work from the field of Conversation Analysis (CA) to fulfil this aim. Specifically, techniques are developed to identify conversation attributes using objective measures of speech performance to help determine the quality of experience and ease of conversational flow within a multi-party teleconference scenario. This thesis will also consider CA methods in order to explore further the interactional behaviour displayed during multi-party teleconferencing in relation to recognised group interactions. The assessment of the impact of transmission delay within a multi-party teleconferencing context is an important element of the work, along with developing task selection and design for conversation-based testing. This work has been achieved through the implementation of a series of conversation-based subjective tests which required over 100,000 data points to be manually reviewed.

This work will also consider the design of a more effective subjective measurement/reference system than possible with current techniques, such as the mean opinion score (MOS) based subjective testing standard. The new testing methodologies are the first to show objective differences in multi-party speech interaction in the presence of differing delays and monophonic/spatialised presentations. In particular the work shows that higher transmission latencies have significantly different conversational parameters when compared to normal face-to-face meetings. However, when spatialisation is added, the effects of this latency are reduced. In conclusion this thesis demonstrates that CA can be used to objectively measure differences in multi-party teleconferencing with different degradations and presentation methods in ways that other, current techniques, cannot.

Acknowledgements

First and foremost, I would like to start by thanking Dr Martin Reed for his continued encouragement, support and guidance throughout the duration of this work and my time at the University of Essex. Without his exceptional guidance, patience and caring nature, this work would simply have not been possible. His help, suggestions and confidence in my ability throughout the testing times of this study were so greatly appreciated. If but a small amount of his vast knowledge and expertise has passed onto me, for that I am truly grateful.

I would also like to thank Peter Hughes, Ian Kegel and everyone at BT who made this project possible and for that I would like to express my deepest gratitude. In addition, I would like to thank Dr Mike Hollier and Dr Ludovic Malfait at Dolby Laboratories for their continued interest in this work; their advice and suggestions throughout this study proved invaluable. Also, I would like to acknowledge all the help and support from Dr Rebecca Clift from the department of Language and Linguistics at the University of Essex. Her guidance and allowing me to attend her CA lectures proved instrumental in this study. Also, without the sponsorship a BT/EPSRC CASE award this thesis would not have been possible, for that I am truly thankful.

Also, a sincere thank you to all the departmental staff and students from the Computer Science and Electronic Engineering department at the University of Essex, along with some of my family and volunteers who took part in the subjective testing phases of this work; their efforts have not gone unappreciated. They have allowed me to take part in some truly unique research and none of this would have been possible without them.

Finally, I would like to thank my family for their encouragement throughout my time at University and for always providing a safe retreat in which for me to frequently come home to. I would also like to thank my Grandmother Carmel and Sharon Donn for always being interested in my progress and being proud of my achievements. Also, a special thank you must go to my boyfriend John for getting me through the last six months of this project. Thank you for all your love and support, I could not have done it without you. A very special thank you must also go to my mother Lindsay and father Lloyd. Without their tireless efforts to raise myself and my twin brother Mitchell, whilst providing a stable and loving home, I fear I may never have achieved my goals in life. They have selflessly given to me more than I could ever have asked for.

This work is dedicated in loving memory of my irreplaceable nanny Rose and aunts Junie and Sharon, to you all I dedicate this thesis. When your world is turned upside down and life has forever changed, all I can do is strive to make you all proud.

Table of Contents

List of Figures

List of Tables

Abbreviations

CA – Conversation Analysis

HRTF – Head-Related Transfer Function

ISDN – Integrated Services Digital Network

ITU – International Telecommunications Union

MOS – Mean Opinion Score

MUSHRA – MUltiple Stimuli with Hidden Reference and Anchor

PAMS – Perceptual Analysis Measurement System

PCM – Pulse Code Modulation

PESQ – Perceptual Evaluation of Speech Quality

QoE – Quality of Experience

SEM – Standard Error of Mean

TCU – Turn Constructional Unit

TRP – Transition Relevance Place

VAD – Voice Activity Detection

VoIP – Voice over Internet Protocol

Chapter 1

1. Introduction

This thesis researches the fundamental components of speech and conversation in order to use this to discern the quality of multi-party teleconference conversations using parameters that represent features in the conversation. The study of these parameters will assist in forming a greater understanding of how effective conversation is conducted when in a multi-party teleconference type environment and how behaviours and interactions are affected by external factors. The main hypothesis of this thesis is that: conversational features can be used to discriminate between good and poor quality multi-party teleconference conversations. This hypothesis is tested through extensive subjective tests and through analysis of appropriate conversation features in an interdisciplinary manner that combines engineering metrics and work from the field of Conversation Analysis (CA) [1]. With this in mind, previously explored methods used for subjective testing of high-quality systems employed over a voice over IP (VoIP) network and methods for obtaining objective results from a CA approach in previous works [1, 2, 3, 4] will be expanded upon. This will assist in the development of an effective method for assessing quality of experience (QoE) for multi-party audio teleconferencing. This work will consider elements such as the impact of delay and audio quality in a multi-party context and aims to develop an effective reference system to enable the evaluation of user performance/measure QoE given variation in delay within a system. A key focus of this work will also be upon the design of the conversational tasks that have been used within the subjective testing phases. These tasks have been tailored to fit various meeting scenarios, whilst at the same time allowing their sensitivity to delay to be varied. This will allow for measurements on the impact of delay across a number of conversation and task styles to be monitored.

Additionally, the work plans to take an interdisciplinary approach combining CA (traditionally a “soft” science) with objective engineering measurements to better understand the social and conversational interactions which take place and thus enabling better classification of the obtained audio data. The integration of the two fields is a new area, with little previous work detailing interactional behaviours within multi-party teleconferencing using CA. In addition, much work on teleconferencing with the presence of transmission delay does not consider the group dynamic of multi-party situations and how this effects the conversation as end-to-end transmission delay increases. The work merges CA and readily available digital signal processing techniques to semi-automatically evaluate high-level speech parameters (specifically detecting speech). However, it is acknowledged that much CA work is highly manual at present, as automatic detection of small interactional and conversational behaviours are beyond the capabilities of most signal processing which is available today. Therefore, it is foreseeable that

most of the CA work used will only be semi-automated. The alternative route which this research could have taken would be to explore in depth a highly automated, heavily digital signal processing approach which would have focused on improving and creating precision signal processing techniques. In contrast, an exclusively CA route could have been followed, which would have represented a language and linguistic approach to this work. However, a fusion of both fields seemed the appropriate route given background knowledge and previous research. This provides the scope for this work to refine previous methods of CA [1, 2, 3, 4] and to explore more sophisticated techniques to measure conversation performance.

1.1 Outline of Area of Study and Problems to be Investigated

In order to fully understand the complex interactions associated with multi-party teleconferencing we must first identify and understand the key features and components of conversation and communication behaviours over a mediated channel. One method used to measure this is via CA tools, which currently are limited in number. Additionally, the few that are available are restrictive as they have been found to be time consuming and ineffective when large quantity of data are required to be processed. Without any existing tools to carry out the work of this thesis, the existing tools were extended to ease the processing. However, in the absence of existing work there was still much manual processing required for this thesis work. The result of the painstaking analysis of the (approx.) 103,000 conversation data points generated in this thesis work could be used in future work to aid a more automated approach using machine learning.

Initial experiments with two-party audio teleconferencing [4] have shown that CA may provide some insight into Quality of Experience (QoE), but further work would need to be undertaken to highlight any correlation between common speech parameters found in both mediated and non-mediated forms of communication. The opportunity to take a soft focus language and linguistics approach to this work is utilised to identify certain speech attributes or patterns which may help in categorising what constitutes more natural, effective communication. Three basic needs of communication and conversation of any type, be it within a teleconference or everyday situation, could be considered as:

- **Audibility** – the need to hear what is being said.
- **Comprehensibility** – the need to be able to understand what is said and to be confident that you in return are understood.
- **Interaction** – the need to interact with other participants.

All three of these basic needs should be able to be fulfilled to have good communication. A selection of tasks designed around these three basic needs have been designed and implemented in an attempt to record or track all of these features. Task selection within the subjective testing phases was critical as there were many different types of task to consider. There is the possibility that one task may not suit or

satisfy all the requirements therefore a combination of three different style tasks have been trialled in this thesis in order to determine if user performance is greatly determined by the type of task being used. Three categories of task were recognised for possible use during the testing phase:

- **Strategic Tasks** – where transmission delay is unlikely to affect overall performance, consisting of creative and descriptive tasks.
- **Interactive Group Tasks** – delay is much more apparent within this type of task, it could be time constrained and have possible rewards or incentives for fastest to complete the task. This type of task is categorised as being highly sensitive to transmission delay.
- **False Situation Tasks** – implemented to specifically pick out certain speech parameters from conversations, e.g. multiple conversations happening within a conference type call therefore promoting frequent talker overlap etc.

The approach taken within this work will consider the above mentioned subjective testing methods and implementations in order to evolve a specific methodology suitable for this body of work. The structure of the testing phases will consist of refining the tasks and their design through vigorous preliminary tests to accommodate the above-mentioned categories and variations in meeting styles. In Chapter 4, it is shown how data obtained from the subjective tests will undergo voice activity detection (VAD) to provide basic on/off speech detection; from this stage on custom developed code was used to aid the manual labelling of speech parameters within the recordings. This will provide the means for further in-depth analysis of the audio to extract CA parameters from the corpus of audio.

1.2 Thesis Aims

The principal aims of this thesis are to consider the measurement of quality in multi-party audio based teleconferencing and how combining CA knowledge, along with the fundamentals of speech and conversation, can improve upon current measurement methods. This work will consider known impacting external factors associated with teleconferences, such as transmission delay and audio quality, to assess and gain a greater understanding into how they affect group conversations and interactions over a mediated platform. The subjective testing phases of this body of work set out to compare varying scenarios with differing levels of transmission delay and how the presence, or absence, of spatial audio can have an impact on intelligibility and overall conversation quality. The comprehensive testing in Chapter 5 will also consider face-to-face speech as a reference of performance. During the testing phases not only will subjective responses be recorded from participants, but also important objective CA measures from the recorded group conversations will be gathered via speech analysis tools to determine how delay and/or audio quality can help or hinder multi-party teleconferences. The results gathered from

the testing phases will highlight any differences between the tested conditions and should provide clear evidence as to which performs most closely to face-to-face “normal” speech, in both subjective and objective respects. This thesis will also question whether using standard subjective testing methods such as the MOS 5-point scale is suitable for rating high-quality teleconferencing audio systems with impairments. In addition, it will consider which, if any, CA parameters extracted from the data set would be able to be automatically recognised or detected in future work, to assist with automatically identifying whether conversation quality is good or poor.

It should be noted that this study has been constrained by the project specification set out by British Telecom (BT), who sponsored this work through an EPSRC CASE award. Certain criteria were adhered to; specific areas of interest in relation to audio quality with spatialised audio being a key interest for its ability to allow for audible separation of participants in a teleconferencing environment. Also, the transmission delays levels were discussed closely with BT to assist in selecting appropriate levels of delay to explore based on industry and widely known “acceptable” levels of delay. This was in conjunction with experimental ideas and tests of how far a system could be pushed to see the effects it would have on multi-party conversations/interactions and how participants behaviour would be affected and potentially adapted. The project has been entitled “Measuring Interactivity in Audio Teleconferencing using Conversation Analysis.” However, in contrast to the restrictions placed on the project by BT sponsorship, it also gave unique access to the professional grade measurement facility at BT Adastral Park as well as a plethora of background expertise on subjective testing.

1.3 Dissemination and Publication

Due to an ongoing patent application, this work is not yet available for publication at this current time. However, a draft paper to IEEE Access is ready for submission when the patent process is finalised:

Karis L. Bailey, Martin J. Reed, Rebecca Clift, Ian Kegel, Peter Hughes, “Conversation Analysis for Measuring Interactive Audio Quality of Experience,” IEEE Access in draft for submission.

Ian Kegel, Karis Bailey, Martin Reed, Peter Hughes, Worldwide Patent WO2019122343, “Managing Streamed Audio Communication Sessions,” filed by BT, published June 2019 [5]. This patent derived directly from this thesis work (see diagram Figure 2 in the patent). The patent describes a method of stopping False Starts in teleconference meetings by utilising information from CA. However, the evaluation of the technique described in the patent is for future work as described in Chapter 6.

ITU-T P.1305 (07/2016) SERIES P: TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS (Telemeeting assessment), “Effect of delays on telemeeting quality.”[6] The proposal for using CA for this

type of testing came directly from this thesis work and was presented to the ITU via the author's industrial supervisor Peter Hughes. The testing methodology proposed in Annex A of this standard came directly from the proposals in this thesis work (see Chapter 3) and the problem statement regarding delays and "False Starts" as shown in Chapter 4 came directly from this thesis work.

1.4 Thesis Structure

Chapter 2 provides background information and reviews material on the subject area; it starts by primarily concentrating on existing research undertaken on both *mediated* and *unmediated* (face-to-face) communication. It focuses on behaviours and conversation parameters examined throughout the body of research incorporating the field of CA and knowledge gained from this area. This chapter also discusses literature associated with the basics of telephony, codecs, conversational attributes and audio signal processing. The chapter also visits ITU standards used for subjective and objective testing methods/models, where various methods are discussed in detail. It concludes with a discussion on speech analysis tools and methods used to obtain specified speech parameters from recorded conversational audio.

Chapter 3 provides information on the experimental methodology and background for the first testing phase of this study. The chapter is broken down into sections that discuss each feature and component used towards the testing stage. It highlights subjective testing methods, experiment design methods, task design and transmission delay levels that were deployed.

Chapter 4 introduces the results acquired from the first phase of experimental tests. It begins by discussing VAD processing of the audio and software used to assist during this stage. In addition, this chapter also details CA parameters and techniques used to analyse the audio data in greater depth. This is followed by a detailed look into the subjective results from participants and also the objective results gathered from speech analysis of the recordings made during the tests.

Chapter 5 presents the results acquired from the second phase of experimental tests. It begins by discussing the refinements to the test and task design and outlines specific details relating to transmission delay levels under test, along with the presence of spatial and face-to-face as additional features of the test. It provides further information on the modified experimental methodology and gives background to the second testing stage of this study. The chapter concentrates on the key changes to the testing phase and how elements from the first phase of testing have been improved with greater emphasis on test and task design to establish a solid and reliable data set. It concludes by giving an in-depth look at the key objective CA results captured during the testing phase with discussion of each speech parameter presented.

Chapter 6 focuses on the expected impact of the research and proposes some recommendations to be taken into account for any further testing strategies within this area. The chapter also discusses research outcomes and patents associated with this body of work, in addition to possible further work in line with this area of study. This chapter also explores some hypothesis related to transmission delay induced multi-party teleconference behaviours and related CA characteristics as indicators of effective conversation and group interactivity.

Finally, Chapter 7 draws upon and provides a summary of all the above chapters and gives conclusions made, taking into account the whole process and results collected.

Chapter 2

2. Background

The background research for this study can be divided into three main topics; firstly characteristics of multi-party teleconferencing and the associated known issues surrounding group interaction through *mediated* channels. Secondly audio signals and research into the basics of audio, this includes how humans hear and perceive sounds. Lastly Conversation Analysis and the various parameters and speech attributes used within conversational speech. Additionally, this work will research standards used for subjective testing and objective measures. In this thesis *mediated* will be used to describe a conversation that is carried out through a synthetic transmission channel, e.g. a telemeeting or conferencing system. The term *unmediated* will be used to refer to a conversation in the normal sense, e.g. audio between talkers is transmitted through the air without a synthetic transmission system; however, *unmediated* could refer to a system where visual cues are not available as with some of the systems used in this thesis.

A notable point within this research area is the lack of related work on the assessment of delay within multi-party teleconferencing. This is a current topic with much discussion amongst ITU members to understand the impact of transmission delay within the multi-party teleconference environment and how best to measure this. The ITU recommendation, ITU P.1305 “Effect of delays on telemeeting quality” [6] discusses the factors effecting communication via telemeetings and recommendations for teleconferencing assessment. This recommendation is an extension upon the previous P.1301 standard [7].The proposal for using CA metrics for the subjective testing within P.1305 came directly from early work for this thesis, as previously mentioned in Chapter 1, Section 1.3. The testing methodology proposed in Annex A of this standard came directly from the proposals in this thesis work (see Chapter 3) and the problem statement regarding delays and “False Starts” as shown in Chapter 4 also came from this thesis work. These contributions were presented to the ITU by the author’s industrial supervisor Peter Hughes. However, P.1305 did not consider the evidence to support the use of CA in this context and the work of later chapters in this thesis is the first work to provide this evidence.

The methodology for this research is supported by some substantial pieces of work, by O’Conaill, Whittaker and Wilbur [8]. This research effectively examines conversation parameters between different forms of communication (e.g. face-to-face, through an ISDN connection and a LIVE-NET connection). Analysis of the results gathered uses objective measures of auditory/visual attributes and characteristics to determine the best method of mediated communication. Although the work done in [8] details audio-visual communication, e.g. video conferencing, many of the analytical tools used and associated

conversation parameters and behaviours are also valid for speech only conversation, which may be held over teleconferencing systems. Whittaker [9] also reviews various forms of mediated communication in comparison to unmediated face-to-face communication. The research pinpoints key differences between the two forms of communication, which assists in selecting the most appropriate communication channel to use in particular conditions. Also discussed are issues of interactions and behaviours lost over mediated communication; they are examined to better understand the effects these could have on conversation. Many elements of the work in [9] support the approach used in this thesis, as key areas associated with speech analysis and human auditory perception are referred to throughout. Elements of the research by both by O'Conaill, Whittaker and Wilbur [8] and Whittaker [9] provide information that assisted with forming the basis of the subjective testing phases of this thesis. The above-mentioned works both show specific fundamentals of speech, found within both mediated and unmediated communication channels, which provide an outline of what will be required to be examined from the corpus of data gathered during the testing stages of this work. Conversation attributes discussed in [8] such as turn taking, double talk and backchannels begin to describe the very basics of CA and highlight some of the initial measures to be considered during the analysis stage of the study from the acquired audio recordings.

The study by Kitawaki and Itoh [10] focuses on the effects of pure delay in telecommunications and looks at communication difficulties at delay levels around the 500ms mark. The work mainly looks at two party communication and how this is impacted; this thesis aims to examine multi-party four-way conversations and the effects of transmission delay on group interactions. Schoenenberg [11, 12, 13, 14] shows interactions and behaviours associated with conversations under transmission delay, but again focuses on two interlocutors over telephone conversations with models presented for speaker alternation for mainly two-way conversation, while task selection also focused on two party communication. Takahashi [15] and Raake et al. [16] addresses issues with using VoIP networks for telecommunications and the associated factors of transmission delay, the works purpose objective quality assessment methods that expands upon the current G.107 E-model [17]. Transmission delay is a key focus of this thesis with current recommendations by the ITU in G.114 [18] that one-way transmission delay should not exceed 400ms for general VoIP networks, but notes that highly interactive conversational tasks, such as telemeetings, may be susceptible to negative effects at lower levels of delay; acceptable amounts of delay for these highly interactive tasks are not fully quantified but are only predicted through use of the E-model [17] which begins to estimate the effects on conversational speech at delay levels below 500ms. This study plans to expand upon this by gaining exact metrics for highly interactive conversation-based tasks in the presence of varying levels of transmission delay, with the addition to different presentations of audio qualities. These recommendations do however assist with the decisions made concerning which levels of transmission delay to test throughout both of the subjective testing phases of this study.

Historical work carried out on the organisation and various components of turn taking within conversations by Sacks, Schegloff and Jefferson in [1] presents the importance and fundamental use of turn taking within unmediated conversation. This seminal paper discusses the commonly observed attributes in the construction of turn taking and outlines the techniques and behaviours associated with the organisation of turn taking within conversation. The notion is introduced of mostly "one-at-a-time" in terms of conversation structure, along with discussion on the use of *no-gap-no-overlap* as a common transition feature from talker to talker. However, Sacks et al. [1] does acknowledge that brief periods of overlap are relatively frequent in their occurrence too. A quote from the work defines these concepts: ''Transitions from (one turn to a next) with no gap and no overlap are common. Together with transitions characterised by slight gap or slight overlap, they make up the vast majority of transitions.'' [1]. Clift [19] also defines the role of CA, explaining further how CA provides insights into the mechanisms of *talking* and how a wide range of actions are accomplished through talk. Elements of both of these works provide the foundation of information required to approach the task of examining speech parameters from the acquired subjective testing phases and subsequent recordings for this thesis. Another study examined by Schegloff [20] described the organisation of turn taking within conversations and gives an account of speech overlaps within a conversational context. The paper observes practices employed to deal with simultaneous overlapping talk and also when overlap is delivered in a problematic fashion. Schegloff expanded further on the no-gap-no-overlap feature noted by Sacks [1] and characterises no-gap-no-overlap as "just a bit of space'' roughly equating one syllable, which corresponds to a silent interval in between turns of around 150–250ms. This has also been referred to by Walker and Trimboli [21] as smooth transitions with the belief that the gap of silence in between turns is closer to 200ms. The factors explored during the works [1, 20, 21] offer fundamental insights into overlapping speech and techniques deployed to deal with overlap, disruptive or otherwise; the works recognise the basic design feature of conversation being the "one-at-a-time" concept which is majority of the time achieved in face-to-face (unmediated) conversations. The organization of speech and overlapping speech is key when considering the introduction of end-to-end transmission delay planned for the conversation-based subjective testing phases of this study. Further studies to support this from the field of CA which were looked at to gain a greater understanding of highly interactive conversations and the processes involved included work by Stivers et al. [22] and Levinson [23]. Stivers and Levinson identify the complex underlying psychological processes involved in turn taking within conversation from one speaker to another normally occurs in rapid succession; gaps in between turns were discovered to be as short as 200ms. Although studies on language production by Indefrey and Levelt [24] explain how production of a single word can, on average, take 600ms. This suggests conversations require pre planning of produced utterances, meaning that planning by another speaker is already taking place whilst the initial speaker is still talking. Heldner and

Edlund [25] discusses how the previous CA works that define metrics for conversational features may not be as precise as originally thought, given the complex nature of turn organisation; this supports the basis of this thesis when monitoring how conversational features and the turn taking process has the potential to be disrupted under transmission delay impairments. This complicated set of subliminal processes further highlights how transmission delay could potentially disrupt this organised turn taking process and all its nuances, but for the purpose of this thesis work the findings from works described [1, 20, 21, 22, 24, 25] further assists with categorising delay levels of interest to be explored during the subjective testing stages of this work.

Recommendations from the ITU P.805 [26] and P.1301 [7] covers all aspects of subjective quality assessment and test set up which shall be implemented throughout the testing phases of this body of work. The recommendations in P.805 specify procedures and methods used when conducting conversation-based tests to assess the quality of the communication. It provides information detailing specifics relating to recommendations on the test facilities used, test and task design for various scenarios, impairments and degradations, subject recruitment, in addition to rating scales and procedures to estimate the subjective quality of telecommunication systems. P.805 was adhered to in all the tests carried out in this thesis. P.1301 gives guidelines for multi-party interactive teleconferencing services involving more than two participants and discusses all relevant aspects of multi-party communication to be considered when evaluating subjective quality of these systems. Guidance is given on accepted and recommended tasks for multi-party conversations with details on subject training prior to testing and objective measurements that can be expected to be achieved; this thesis work aims to take this guidance and extend further to incorporate CA techniques and measurements to improve upon current methods. Additionally, as mentioned previously the ITU standard P.1305 [6] (which extends from P.1301) details the proposal for using CA metrics for subjective testing measures which came directly from early work of this thesis. The methodologies disclosed in P.805, P.1301 and P.1305 will be followed when conducting both subjective testing blocks for this work, with reference to the implementation of these recommendations, along with expansions and improvements, set out in Chapter 3.

Some subjective testing methods that this thesis work plans to use, investigate and improve upon are standards defined by the ITU. A main ITU recommendation is the P.800 standard [27]. This standard refers to the 5-point MOS scale; this scale is used in subjective testing for the participants to evaluate the quality of the voice/speech as they perceive it. The scale defines 1 as the worst quality and 5 as excellent. This system of rating voice quality is good for significant impairments that are clearly audible. However, it does not perform as well for higher-quality system comparisons, where impairments may not be so apparent to an untrained listener. Another subjective test recommendation by the ITU is the BS.1534

standard [28]. The BS.1534 standard incorporates MUSHRA testing, which is used for intermediate quality audio systems. This has advantages over MOS testing, but still does not work as well for high-quality audio as it is not designed to deal with small audio impairments found with high-quality systems. Additionally, as MUSHRA testing uses repeated A-B testing it is not straightforward for measuring interactivity of live communication systems. A subjective test standard that is more suited to high-quality systems and audio is the BS.1116-1 ITU standard [29]. This standard specifies in much more detail the requirements for the testing room, equipment used, background noise level, etc. It is much more suited to high-quality audio, in which there are minor impairments only. The ITU standards [27, 28, 29] (mentioned above) have all be considered for use during the subjective testing stages of this project, with the MOS scale deemed as most suitable for this body of work.

Another form of testing used for audio quality is objective testing. Objective testing usually refers to some form of objective measurement from a test, rather than a subjective opinion. The ITU also has standards which specify objective assessment methods; one example is the P.862 standard [30]. PESQ is used in this standard and operates by predicting the outcome of a subjective test by using a perceptual model. This model works by taking the original audio signal and comparing it with the degraded signal. The degraded signal refers to how the original signal would be heard over the telephone network. A PESQ assessment comes up with a comparable result to that of the MOS scale, as seen in the P.800 standard. This means that the outcome would be the same every time with an identical input, which differs greatly from subjective testing where the outcome is based on the individual's opinions. PESQ is only beneficial for obtaining objective measures from narrowband speech codec tests. This objective quality assessment method has useful elements but may not be as suitable for identifying impairments in wideband or higher quality telephony. The ITU standard P.563 [31] again has the same function of PESQ, by determining an equivalent MOS score by prediction. P.563 models the human vocal tract and also how humans perceive distortion or impairments in speech signals. It uses this information to process the signal and outputs a score relating to the quality of the signal, this score is comparable to the MOS scale. This objective testing method also is only recommended to be used with narrowband speech codecs, so may not perform properly when used in conjunction with higher quality audio, such as planned for the spatialised conditions of the testing phases of this work. An appropriate objective model used for high-quality telephony is the PAMS model [32]. PAMS uses the same techniques as PESQ and P.563, by predicting the perceived quality of a signal and mapping its outcome to an equivalent MOS scale. After careful consideration, it was found that most of these existing standards concentrate only on a point-to-point transmission channel and offer no system for measuring multi-party channels and the quality of the system for allowing normal conversation flow. Thus, for the purpose of this thesis the only suitable standard to compare with was the ITU-T P.800 using the MOS scale as its quality rating system is not

limited in its usage scenario. Another reason for its use is because there is an existing understanding and use of this scoring scale within the industrial context of this project. MOS is also the most widely used rating system for traditional telephony-based systems [33].

Both objective and subjective assessment methods will be implemented within this study for the experimental phase, as they provide the appropriate approach in which to achieve a reliable set of results from which to draw conclusions from. Results from both the participant's perceived opinions and feedback, along with objective measures of data gathered during these experiments will provide this. Objective CA measures from the experiments will be provided by gathering data from the audio recordings made during both testing phases utilising current and developed speech analysis tools. These tools will assist in collating results from various speech parameters monitored throughout the tests. This will allow for both of these elements to be combined together, to give a much more conclusive description of the differences between the varying levels of transmission delay and presentation of different audio conditions used during the testing phases.

Chapter 3

3. Multi-Party Teleconferencing Environment: Monitoring the Effects of Transmission Delay - Experimental Design

3.1 Introduction

This chapter aims to outline one of the main areas of study untaken, to form the basis of this work and underpin a reliable and rigorous design for the testing phase of this project, which will then be expanded upon in further testing, as detailed in Chapters 5 and 6. Specifically, the purpose of this work is to provide a testing methodology for the main hypothesis stated in the Introduction: by measuring high-level conversation features (not the full content of speech) the quality of multi-party teleconferencing can be discerned. We begin by giving a broad overview of the initial subjective test design; this shows the planned approach to carry out the first round of testing, which focuses purely on monitoring the effects of transmission delay within a multi-party teleconferencing environment. We then continue by discussing the three-task design model and the process of task selection in order to determine if the task itself has implications on group interaction/behaviours and if it shows any effects on the conversation as a result of induced transmission delay. Also, the different levels of transmission delays are discussed in detail and the reasoning behind the levels under test and their association with known acceptable or tolerated levels of delay.

The creation of the three-task design model has been developed to combine different meeting or conference styles into a task driven format. Each one of the three tasks was tailored in an attempt to replicate a scenario similar to that of which it is trying to mimic, being one of three different meeting styles which have been identified through discussion with BT about common meeting styles and typical categories which conferencing calls generally fall under. The first meeting style considered would be classed as an informal meeting, consisting of general chat, with ease of conversation and flow as the main objectives; this is to be considered a natural interactive style of meeting. The second would take the form of a formal meeting, where the interaction is highly structured with clear objectives throughout the course of the conversation. This style of meeting is often chaired by one member of the conversation who oversees the objectives of the conversation; this would be deemed as a highly typical meeting style within a business setting and one that most people within the business sector are best accustomed with. The final meeting style would be one that is goal-orientated or dependant. The premise of this being group effort, focused to complete an interactive task which calls on all members to take part. The basis of the three-task design model was to simulate three different style tasks in order to see how transmission delay

is handled within varying situations and if any particular scenarios are more prone to delay disruption than others; this allows us to fully explore and understand how transmission delay effects the interaction between the groups and the overall call quality. The three tasks within the design model will be described in more detail in Section 3.3 and each of the tasks are entitled: *informal meeting task, formal meeting task* and *goal-oriented meeting task*.

Within the concluding sections we refer to commonly used subjective testing methods for audio quality assessment, and how they are integrated into both the initial and further testing phases. The final section draws upon all of the above-mentioned aspects of the study and evaluates the experiment design methods put in place to ensure fair and accurate testing across the varying tasks and delay levels incorporated in both testing phases.

3.2 Subjective Test Design

An essential part to the successful design of the testing phase included the accurate set up of the test environment along with clear objectives as to how both subjective and objective data were to be obtained. This section aims to provide a complete high-level overview detailing the nature of the tests including equipment used and the overall setup of the testing phase which was carried out at BT Technology Research Labs in Ipswich, Adastral Park.

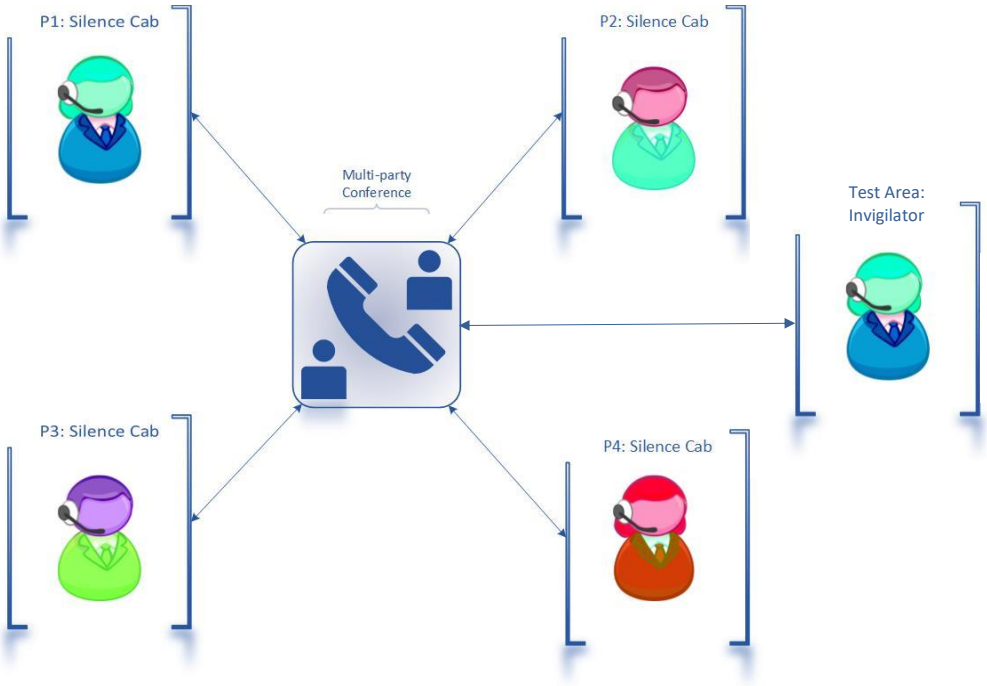


Figure 3.1: Multi-party Teleconferencing Subjective Test Set-Up

Figure 3.1 illustrates the physical testing phase setup and displays how participant communication was enabled between silence cabinets and the test invigilator. The selection process for the number of subjects per group was determined by two factors: firstly, BT have experience with running subjective tests and commented that they often had difficulty with recruiting even two people for tests at the same time; secondly, the author had a personal discussion with a vendor of large scale (global) teleconferencing systems who reported that mean conference sizes in Europe were between 3-4, whereas in the US it was 4-5. It is difficult to obtain publishable data on this as companies strive to maintain confidentiality about their platforms; the author is grateful to this source for revealing information that was widely known within the industry. Consequently, it was decided to use a group size of four; three is clearly the minimum for group communication and larger groups would have been logistically very difficult.

The format of the initial subjective tests consisted of participants being assembled into their groups of four. For the purpose of the test they communicated via headset with a microphone in a four-way audio only conversation. Each individual participant was allocated their own silence cabinet, effectively soundproof rooms which allow for no external noise to affect the conversational tests and whereby their only means of communication is via headset; there was no visual stream or communication present throughout the duration of the initial round of testing whilst carrying out the three-task design model. These soundproof rooms meet the recommendations for subjective testing as set in the ITU standard P.800 [27] and are housed within a professional grade and unique subjective testing facility at BT Adastral Park, Ipswich. A fifth person, the test invigilator, (also the author) was present via audio stream throughout the duration of the test as an invigilator to assist the group and help with any technical problems or questions which arose; also, to control the varying conditions of the test and provide feedback to the group members. The audio quality provided throughout the duration of the initial testing remained the same, as delay levels were the focus in the initial testing phase. The audio was provided by custom systems contained within the silence cabinets with pulse code modulated (PCM) mono at 16 kHz, 16-bit sampling to provide good quality audio which matched the bandwidth of wideband speech systems, but without codec degradations. The communication was between all four cabinets and the test invigilator located in the control area of the facility.

The conversations that participants had during the test whilst performing the three-task design model was recorded for later analysis. Each of the tests consisted of three conversation-based tasks, one task for each of the three transmission delays conditions being 0ms, 400ms and 800ms (with rational given in Section 3.4). This permitted each group to experience all three delay levels over a series of three varying tasks. The transmission delay level was different for each task. The participants were not made aware that there was a transmission delay or that there was any difference in the audio configuration for different

tasks, also they were not informed about any difference in transmission delay levels over the conference prior to the test. Participants were only given instructions regarding the format of the test by means of a presentation outside of the silence cabinets, whereby they were shown examples of the tasks and given the opportunity to ask any questions. They were told that the test consisted of three different types of task for them to carry out collectively as a group; the details of each of the tasks will be illustrated further in Section 3.3. They were told that each task was different and had varying lengths which will be timed by the test invigilator, one being ten minutes *(informal meeting task*), one at fifteen minutes (*formal meeting task*) and another at twenty minutes (*goal-oriented task*). It was stressed to the participants that it was not themselves and their own abilities that were under test, but the conference system itself. They were told during the informative presentation that after they had finished the tasks they would be asked to fill out a conversational experience survey which would ask them for their opinions and preferences with regards to all three tasks and audio conditions; this provided the means for subjective data collection. A copy of the all surveys and participant consent forms can be found in Appendix A1, A2, A3 and A4.

3.2.1 Subjective Test Recruitment

It was found that subject recruitment was a difficult task and using subjects from both BT and the University of Essex would not be enough, further incentives would be needed, e.g. subjects would need to be paid. Thus, to broaden the search, people will be employed from a local company based in Ipswich called “Find a Job” [34]. The reasoning behind recruiting both students and locals to Ipswich would be to gain a far greater spread in terms of age range and also to balance the tests with both male and female participants. Prerequisites to this included individuals having to the best of their knowledge good hearing in both ears and to be native English speakers, this will be discussed further in Section 3.5.3, along with not being part of any subjective testing of this nature within the past 6 months [24, 27].

3.3 Task Design and Specification

As previously mentioned in Section 3.2, the tasks have been designed to each reproduce circumstances similar to that of common meeting styles, which are often used within a business conferencing context. This notion originated from considering the end goal and where teleconferencing is most notably used; the business environment highly relies upon frequent group conference calls. Many studies have been carried out for two party conversation with varying task selection [13, 14, 35]. The author worked on and developed the three-task design model in the course of preparing this work after numerous testing stages to decide upon effective larger group tasks with a business lead focus. As a result of this work carried out by the author of this thesis, Peter Hughes (Industrial Supervisor) from BT Research (Adastral Park) presented these for discussion in ITU-T Study Group 12, Question 10. The study group accepted the recommendations and subsequently published these in Annex A of the ITU-T standard P.1305 [6]. This

section will highlight the details of each task in depth. A copy of the forms for each of the tasks described below can be found listed in Appendix A1, A2, A3 and A4. The duration of the tasks, described below, were determined after pretesting at BT as a compromise between obtaining enough conversation events and fatigue in the subjects had to be taken into account. In practice it was found that approximately five, formal, pre-tests using the developed protocol are required to ensure all the logistical details were corrected to provide a reliable test. Additionally, a significant number of informal tests were carried out to reach the final developed protocol. In practice, this informal pre-testing phase can be lengthy and should be continued until the various test parameters (e.g. format, length of task etc) are converging on usable parameters. The informal pre-tests were recorded and used as the baseline from which to modify the subsequent formal pre-tests. This preliminary stage provided seemly small, but significant findings upon which to improve the overall testing phase and allowed for the elimination of unwanted factors which can impede the test itself or tasks. An example being details down to which pens to use so as participants did not click the pen nib up and down thereby causing interference with the recorded audio; this would prove significant later on when attempting to automatically detect speaking and silent segments per participant. Thereby it is highly recommended that both informal and formal pre-testing takes place allowing as many nuisance variables as possible to be eradicated prior to the beginning of the official testing phase and formal data collection.

3.3.1 Informal Meeting Task

The aim of the informal meeting task was to reproduce as close to natural conversation as possible which would be very loosely structured whilst still allowing for adequate interaction between all four members of the group. The main criteria of the task would be to have all members engaged in flowing conversation without being too highly scripted or predetermined. It was decided that the task would need to have some element of structure, but loosely based in order to promote enough conversation. After considering conversation starters which everyone could engage in, it was decided that the most feasible option would be a task where the group collectively discuss holidays and travel; the idea being that most people would have had some form of holiday and/or travel experience, be it at home or abroad. If someone had had no holiday or travel experience it was felt that this topic would encourage conversation about places which they would like to visit in the future. The topic had been chosen to allow for general conversation flow with the aim to produce a general chat type environment, in which all participants feel that they could voluntarily take part. The task had been loosely structured to facilitate a natural conversation flow, whilst still providing some of form of guidance and a discussion topic to keep the task focused. The format of the task was broken down into two notable sections, a discussion section and a list compiling section. The duration of the task was set at ten minutes, which was divided into two five minute intervals for discussion and list compilation. The participants would have the first five minutes to have a conversation about

favourite holidays and travel experiences, within the remaining five minutes they were asked to list ten things that make a good holiday of which they collectively have to agree upon; this was set to encourage all members to participate who perhaps may not have been as active during the discussion phase. This section required the participants to come together as a group and was intended to permit people to share and discuss opinions. However, this does rely on participants being willing to discuss personal experiences with people they may not have meet before. Also, the element of collectively compiling a list ensured that the conversation would not be too general, with no overall measurable end goal and provided clear objective data as to whether they would manage to finish the list within the allocated time and if it was accurate amongst all four members.

3.3.2 Formal Meeting Task

The aim of the formal meeting task was to reproduce a formal conference style call which is highly structured and is chaired by one member of the team. This task contrasted the previous informal holiday task with the criteria set to meet very different needs. The key objectives for this task were to have each member of the group allocated a role which they must carry out for the duration of the task. This approach aimed to highly structure the conference by giving the group a set narrative to adhere to and each person within the group a predefined role. This method was developed to have a likeness to a formal business type conference call and to highlight any features of this style of meeting which may be affected by transmission delay. The task was set to be based around a card game call "You be the Judge" by Paul Lamond games [36] which features real life court cases for players to decide what was the verdict of the case. The format this task took was that participants were presented with two real life court case scenarios; each case was allotted a seven and a half minute time frame for group discussion and debate. The overall task duration was set at fifteen minutes to allow for the two cases to be presented and debated. The task was rigorously structured and consist of the following roles:

- **Chairperson**: The chair will lead the meeting by reading aloud the case and verdict cards and making sure the agenda is followed to time for both case cards. This role shall be carried out by the person responsible of running and coordinating the tests; this will ensure a precise and accurate description of the cases will be given each time. This will allow for the task to remain precise and structured whilst keeping the group within the set time frame. The chair shall give instructions, whilst making clear the details of each case, but will allow for the floor management of the conversation to be decided amongst the four participants.
- **Counsel For**: Two members of the team will be allocated the role of presenting reasons and opinions why the court would rule **for/in favour** of each case. Their task shall be to lead the debate **for** each of the cases. Both members will be notified that they have been allocated the role of

counsel for, but they will not be given any time to confer or discuss between them the argument **for** the cases. Floor management and presentation of ideas will be determined between them and the rest of the group during the task.

- **Counsel Against**: Two members of the team will be allocated the role of presenting reasons/opinions why the court would rule against/not in favour of each case. Their task shall be to lead the debate **against** each of the cases. Both members will be notified that they have been allocated the role of council **against**, but they will not be given any time to confer or discuss between them the argument **against** the cases. Floor management and presentation of ideas will be determined between them and the rest of the group during the task.

After each case has been presented via the chairperson, debate from both sides will take place. This will then be followed by the team having to agree or come to a conclusion on what they collectively feel the actual verdict for each case was. Participants are told that even though they have predetermined roles of counsel for or against, they can come to any conclusion they feel best suits the case. The chairperson will then read the final verdict aloud. The formal meeting style task had been designed to replicate that of a formal business conference call; the task consists of predefined typical roles which are usually found within a conferencing scenario. The roles of counsel for and counsel against aim to reproduce the occurrence of differing opinions between various members within a conference, with the end objective similar to that of a real-life call where the calls participants would normally come to some form of conclusion or agreement about the discussed topic.

3.3.3 Goal-Oriented Task

The idea of a task specifically designed to draw out the features of a conference call that may be affected by transmission delay is the basis for the third task within the three-task design model. The aim was to reproduce a task driven and dependent conference style call which is highly interactive. The task itself would need to actively encourage all members to participate whist eliminating the possibility of inactive group members, which is perhaps a weakness of the previous two tasks discussed, although attempts had been made to counteract this by means of list compilation and role assigning. The task would also need to accommodate for a measurable result to see how accurately the groups were able to complete, or how close they got to completing the task within the set time. The task which it was thought would best suit all of the above needs was called a *goal-oriented task* which is defined as below.

The goal-oriented task shall consist of all four members focusing together as a team to help each other complete a route around a map. A predefined path will have been determined for the map, but participants will be unaware of the correct and complete path; they shall only be presented with a quarter of the path. The map shall be divided into four sections, with each participant only having part of the

route/path marked on their own map. Through discussion they must help each other navigate the entire map from the start to the end point and draw on their individual copies of the map. The map task has been designed so that each one of the four participants gets the chance to direct the other three members of the group. This is done by each participant having a different section of the map marked with an incomplete path; they will each have 5 minutes to direct the team around their section of the map. This will ensure that each person gets to take on the navigator/instructor and listener role. Also, the maps will vary slightly between the four participants; some may have landmarks or labels missing from their map, which may be present on others. This should add to the complexity of the task. The differences between the maps should cause confusion as the delay within the call increases; this is because participants may have to backtrack and repeat themselves or interrupt each other in order to make sure that they are at the same point on the map as each other throughout. The test shall be done under timed conditions again, with participants given 20 minutes to successfully complete the map task. A copy of an example map used can be found in Appendix A5. Chapter 5 of this work provides further details on the goal-orientated tasks expansion for the subsequent testing phase of this study with the extended map design and overall improved test design; it illustrates the findings collated from the original map design and how it can be further refined to evoke maximum participant interactions and conversation.

3.4 Transmission Delay Levels

With the three-task design model complete, the next stage was to decide upon which induced transmission delay levels would be used. As each test itself has a total run time of 45 minutes, without taking into account the initial presentation and breaks, it was decided that each task would only be able to accommodate one delay level due to time constraints and how many conditions would be suitable to not stress or fatigue the participants. Although varying levels of delay were used for the tasks, the presentation order of the delay levels were randomised, along with the order in which the tasks are presented to the participants. A full illustration of the testing order used for the first round of subjective tests can be found in Appendix A8. The randomisation of the delay levels, conditions and presentation order will be discussed in further detail in Section 3.6. The delay values that were used are listed as follows:

- Low delay (virtually no delay) – 0ms[1]
- Medium delay – 400ms
- High delay – 800ms

[1] It should be noted that there was a notional system delay of 2ms which should be added to the values above. The 2ms delay value was determined by BT through acoustic click delay from microphone to headphone and measured on an oscilloscope.

These values where decided upon to permit for a full test of low, medium and high delay to see how transmission delay at such varying levels can affect a conference call and what effects these may have on the four-way group dynamics of a conference call. The low delay level was included to examine no to low delay over a conference style call with the upper values being selected to see what may happen when the acceptable delay limits are pushed; with the knowledge that normal conversation is understood to degrade and show signs of degradation within interactivity after or around the 200ms mark [22]. In addition to the ITU recommendations from the G.107 E-model [17] and G.114 [18] which acknowledges that one-way transmission delay should not exceed 400ms for general VoIP networks, but notes that highly interactive conversational tasks, such as teleconferencing, may be susceptible to negative effects at lower levels of delay.

3.5 Subjective Testing Methods

3.5.1 Mean Opinion Score (MOS)

The subjective testing method used for rating the audio quality during the tests is the P.800 standard [27], which uses a Mean Opinion Score (MOS) based system. This standard recommends a listening-quality scale whereby participants can rate the quality of speech with a score from 1 – 5. The scale used gives the participants the following options from which to choose from after the test is complete. A rating of 1 equates to bad, 2 equals poor, 3 equals fair, 4 equals good and lastly 5 equates to excellent. This MOS system will be used throughout the tests to identify the subjective feedback from participants based on each task and their perceived measure of audio quality. This technique of determining audio quality and perceived speech was chosen as participants that would be recruited to do the tests will not have been trained in performing conversation-based tests (e.g. they were non-expert subjects). One advantage of the MOS scale is that it provides a fairly easily interpreted system on which to score audio quality, especially for previously untrained participants. Also, this method is most commonly used for rating telephony systems and many other objective testing methods such as PESQ and PAMS map their scores to a MOS scale [30,31]. Other commonly used subjective testing mechanisms are not appropriate e.g. MUSHRA [28] among others, as they generally involve switching between degradation conditions which would be impossible during a task orientated test as this work sets out to achieve.

3.5.2 Subjective Testing Forms

As previously discussed in Section 3.3 of this chapter, in an overview of the task design and specification, a series of forms are presented to the participants on arrival to the test. Each participant will be handed a folder marked with their silence cabinet number on; this will remain their designated silence cabinet throughout and the folder will house all of the relevant documents they will need for the testing phase. These forms provide the means of gathering all the required data from the tests, all of which will be

analysed and reported on in Chapter 4. In total five forms were formulated for the testing phase; the decision was taken that these forms should be hard copies as there could be several issues with computerised, on screen forms. The first issue would be that participants would have to tick boxes and fill out information on screen; the data from this would then need to be submitted to some form of database in order to record the information. This posed a problem as the tests consisted of three tasks, each of which consisted of a different style of task and different levels of transmission delay associated with them. If forms where not correctly filled out or submitted after each task during the test, the risk of data loss of the participant's responses from each task was high. Also, the implementation of such a structure during the test would have taken a considerable amount of time to design and get to function as desired. The second issue was that this method may have been more confusing to participants as paper copies of the forms would be easier to fill out whilst doing the test and could be previewed before the test began, especially for the map task which would have required drawing onto a computerised version. This also provided the participants with the chance to ask any questions regarding how to complete the forms before the test began. Overall, hard paper copies of the forms seemed to be a much safer and reliable option. Now we will continue by discussing each of the forms in detail. It should be noted that the forms where submitted to the Faculty Ethics Committee and the University's Ethics Committee to which they were approved. Copies of each form can be found in the Appendix A1, A2, A3 and A4 at the end of this thesis.

3.5.3 Experiment Consent Form

The experiment consent form is a mandatory form that must be given to each participant on arrival to the test. This form outlines the format of the test, explains all the factors involved in the test and what participants will be required to do. It informs the participants that their conversations during the tests will be recorded for the purpose of speech analysis afterwards, to which they have to agree before the test can begin. The form also includes detailed safety procedures during the test and who to contact with regards to any further questions that they may have to ask. All information in the form will be explained to the participants before the test begins, giving them an overview of what the test involves and to reiterate the fact that conversations will be recorded. The form requires their signature to show that they agree to the terms and conditions that have been outlined and that they agree to participate. This also gives the participants a chance to ask any questions that they may have regarding the test, before they begin. Also, the consent form had a number of questions in order to provide some important information, they are presented in the form of tick boxes or blank sections for them to detail answers further. The first questions they are asked are for their name, gender and age. This data will be recorded along with how many of the other participants they knew prior to the test. The second question asks them which, if any, videoconferencing or audio-conferencing applications they have used before; they are given the option

to tick as many as apply and also room to list any others which are not already included. It should be noted that there was also a prerequisite of the testing phase that all individuals had to the best of their knowledge good hearing in both ears, English as their first spoken language and not have taken part in subjective tests of this nature within the past 6 months [24, 27]. These requirements where essential as the tests relied on participants being able to hear any audio differences between conditions and also being native English speakers due to the social and cultural differences in the nature of conversation structure and speech patterns found in other languages. In addition to not having taken part in similar subjective testing within the stipulated time frame so as to ensure that they were not predisposed to what the full purpose of the test may be or be accustom to listening out for impairments which they may have become familiar with in previous tests, thereby they may bias the results.

It is recommended that these prerequisites should apply to all conversation-based type subjective testing to ensure an accurate and cohesive data set is obtained. Extensive pre-testing was a crucial element of the test design, even down to each of the forms and surveys which participants were required to fill out; it is important that vital information is captured on these forms during the tests. It should be noted that the execution of the forms, surveys and data collection should be planned down to the smallest of details to ensure the most effective subjective testing environment possible, with a minimum of five trials as a recommendation. Minor details, including pen selection and presentation of paperwork, may appear to be insignificant or trivial, but when capturing audio, it is sensible to eliminate as much interference and background noise or pollution as feasibly possible. One example of this being pens with a nib that click up and down when intended for use. This would impede the audio captured in the silence cabinets as clicking of pen nibs would cause issues when processing the audio, even for simple VAD on/off speech recognition the additional noise created would throw off even simple mark-up of audio. Without the preliminary tests this discovery, which upon reflection may seem obvious, would have resulted in detected “speech” when in fact it was a nuisance variable which could have been removed through pre-trialling of the test environment.

3.5.4 Conversational Experience Survey

The conversation experience survey is a form given to all participants at the end of each of the three tasks during the experiment. These surveys require them to answer questions about the three tasks and conditions which they have just been subject to. These forms are clearly marked to correspond to each task and given to the participants in order in their folders, so each conversational experience survey for each task directly follows after they have completed that task. The survey consists of five MOS scale style questions, with the final survey after the third task asking participants to elaborate on any differences that they may have noticed between the tasks. Firstly, the form displays an evaluation scale from 1-5 as

follows, (5) Excellent, (4) Good, (3) Fair, (2) Poor and (1) Bad. This scale relates back to the MOS scale previously discussed in Section 3.5.1 and allows the participants to rate the audio and speech quality from each of the three tasks. The first question asks, “How would you rate the audio quality for this task?” This is followed by the numbers 1-5; they must circle their answer. This will provide the MOS score from the participants for each delay level they have experienced. This is then preceded by a series of questions which have possible answers and statements attached to them; the subjects must simply circle the statement that they feel is most suitable to their thoughts about each of the different tasks. An example of the type of questions asked refers to how well they felt they were understood, did they have any problems talking or hearing over the connection, how easy did they find it to communicate over the connection and how would they rate the ease of the conversation flow. The content of the questions in the survey were discussed and approved with BT, as it was felt that they would be appropriate to ask for later analysis. Details of all questions included in this form shall be referred to in Appendix A4. The series of questions asked are to gain subjective opinions from each participant about their thoughts and feelings towards each audio condition and their associated levels of transmission delay. All the answers to these questions for each audio condition will shed light on the subject’s feelings and thoughts on the conditions of each task and provide additional information other than the MOS result, although the answers to these questions could be mapped to that of a MOS score. The last question after the third task asks the participants to consider all three of the tasks and to answer either “yes” or “no” as to whether they found any differences in the conversational difficulty or their ability to communicate. After this question they are provided space to elaborate if they answered yes, with reference to specific tasks with what they thought the differences were or anything which caused them problems in communicating.

3.6 Experiment Design Methods

This section aims to highlight and underline experiment design methods that have been selected for use during the testing phase of this study. It will begin with a discussion on Graeco-Latin square design, which is a popular method employed for ensuring suitable randomisation within tests and their designs. Also, the topic of presentation order will be discussed, along with how to avoid or minimise possible learning effects that could be carried over from task to task.

3.6.1 Graeco-Latin Square Design, Randomisation, Presentation Order and Learning Effects

Graeco-Latin square design is a commonly employed method when it comes to designing experiments. Graeco-Latin squares allow nuisance factors within experiments to be blocked to allow for the effects of the primary factor of interest to be studied; this is often referred to as the treatment factor [37, 47]. Graeco-Latin squares are formed by using two Latin square designs merged together; this is required within tests which contain three or more nuisance variables. The advantages of this type of design is that

it can allow an experiment to be run only a small number of times, whilst providing a truly random order allowing the effect of each audio condition to effectively be monitored. This principle was useful when designing the subjective tests for this work. It allowed the primary factor and interest in this case to be the main focus of the tests, rather than the other elements that were involved. With specific relation to the subjective tests for this body of work, it highlighted a method which could be used to allow the overall experience encountered with varying levels of delay to affect the participant's opinions, rather than the presentation order of the delay values or order in which they undertook the different tasks. This method of randomisation was essential as the order in which the delay levels and tasks were presented to the subjects could affect their opinions. It is common for the last condition heard to be remembered the best, therefore the participants may believe it was better than the previous conditions purely because it was the most recent one. In an attempt to counteract this, randomising the order in which the delay levels and tasks were being presented to subjects was vital. A short survey was given to participants at the end of each task to capture their thoughts as quickly and as accurately as possible after a task was complete. A copy of this has been included in the Appendix A4 section of this report, along with other supporting documents used for the subjective testing phase. A total of three different permutations from a 3x3 Graeco-Latin square were employed over the course of the subjective testing phase; this totalled nine separate tests which were carried out. One test consisted of four participants, all of which would undergo all three of the tasks together as a team, whilst being unaware that each of the tasks had a different transmission delay levels associated with it.

The experimental design also had to consider the effects the independent variables in the test would have on the dependent variables obtained during the test. The independent variables are the transmission delay levels and the different tasks; these would be controlled and manipulated as discussed above throughout the tests. The dependent variables obtained from the tests are the variables which we measure during the test, such as duration of each task, the average number of turns per minute, the number of overlapping talk instances per minute, and also the overlapping talk ratio. The presentation order of both the delay levels and the tasks had to be manipulated in order to avoid any confounding, uncontrolled variables arising in the tests. The selected method used for the subjective testing in this project was to use related samples rather than unrelated samples. Related samples refer to the data being collected from the same group of people, which entails each group performing all three of the conditions in the test. Unrelated samples refer to the data being gathered from different groups for each condition in the test, this would result in three groups being required in order for one test to be completed. Unrelated samples were not chosen as this method would require a much larger number of participants; this would equate to three times as many than would be needed if related samples were used. Also, the main drawback is that major differences between individual's ability to accomplish the experimental tasks

can have an influence on the results gained from the tasks. Related samples provided a reduction in background variation by elimination of any individual differences between groups for each condition as the same group is used for all three conditions. Also, fewer participants are needed to conduct the tests, which had to be taken into account due to time and financial constraints of this project. The disadvantage of this method is that it introduces order effects or learning effects into the design. Learning effects can be carried over from task to task, with the assumption that this could lead to an improvement in participants performances as they become more practised and familiar with the tasks. This could be due to an increasing understanding of the equipment used in the tests and also a greater awareness of the demands of the task. It could also have the opposite effect and lead to performance deterioration throughout the duration of the tasks. Deterioration in their ability to perform the tasks can be associated with fatigue or boredom as the test goes on, or even a loss or lack of concentration. Both deterioration and improvement can transpire within participants at the same time during a test, so a suitable solution to this problem had to be implemented in order to offset this [37]. The second phase of subjective testing, as discussed in Chapter 5, aims to improve further upon these type of learning effects by incorporating an orientation round and also a face-to-face condition. The orientation rounds intended purpose is to familiarise the participants with the test set up, expose them unknowingly to each condition prior to data capture, whilst giving sufficient time for the learning effects to take place and their behaviours to potentially adapt. This ensures that any possible improvements or alterations in their approach are not associated with the conditions under test, thereby giving unfair advantages dependent upon the order in which the conditions are presented to them. A face-to-face condition was also implemented to reference all mediated conditions with; this condition gave a “best case” scenario where it is thought that participants would perform at their best and most natural due to the nature of the condition, e.g. all participants in the same room without the use of any equipment to converse and without impairments effecting interactions, such a transmission delay and audio quality. This condition would aim to replicate as close as possible normal face-to-face speech whilst blocking visual cues with the use of professional grade sound isolation baffles. It is essential that no visual cues were involved during the face-to-face condition so as not to inadvertently give an advantage to this condition as no visual channel would be available throughout the testing phase over any of the mediated audio conditions.

A further method employed to stabilize and control the order and learning effects within this study was to counterbalance the experiments by allowing a randomised combination and sequence of the tasks to be carried out once to ensure a fair trial. This is where elements of Graeco-Latin square design and counterbalancing come together to create a viable solution of how the tests would be carried out. This method would allow for fewer participants, whilst still providing a reliable and accurate data set as an end result. The approach taken to formulate the finalised number of tests that would be run was to

firstly account for each condition and variable within the tests. Figure 3.2 on the following page depicts how initially two Latin squares were designed and then combined together to form a Graeco-Latin square. This provided the foundations to discover possible combinations and sequences in which the tests could be run by combining the audio delay conditions with the different tasks themselves.

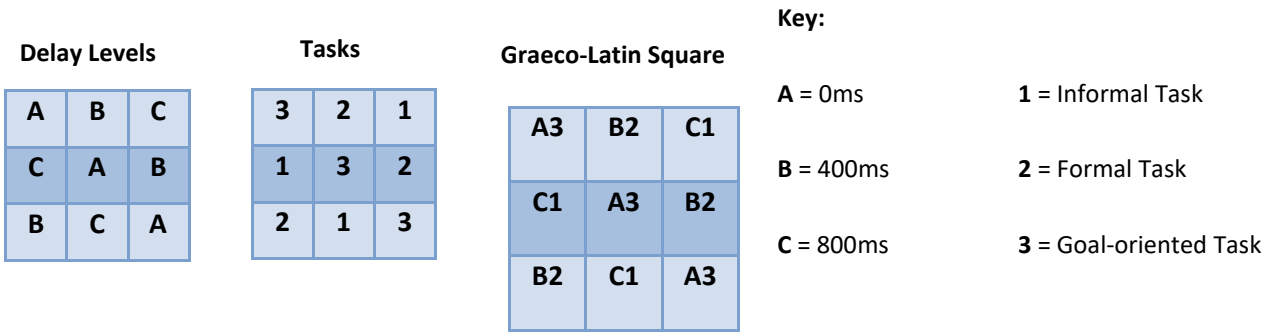


Figure 3.2: Two Latin Squares combined to make a Graeco-Latin Square Design

As can been seen from Figure 3.2, each of the letters and numbers appear once per row and once per column; when the two Latin squares are joint this provides a randomised design in the form of a Graeco-Latin square. The Graeco-Latin square provides the test run order for nine tests, this could be run numerous times using different permutations of the square to create various data sets. This also provided a starting point for listing all the possible combinations in which the tests could be run. It was found that 36 combinations of delay level presentation orders and task presentation orders could be made from the Graeco-Latin square. This number would not be achievable as it would require 36 groups of participants to provide a reliable data set in which each combination would be used once. This would require a large cohort of 144 participants which is a very high number of people to recruit, also this would create a very substantial data set to analyse. The data collection would have to rely on a much more feasible option of employing a Graeco-Latin square to provide the randomised design for a combination of nine tests to be run using one Graeco-Latin square, which would require a more manageable number of 36 participants. By using the Graeco-Latin square design it would allow for nuisance factors to be taken into account and aid with the prevention of any confounding variables. In addition, it gives the security that by means of randomisation the effects of presentation order of the tasks and the transmission delay levels could be mitigated as much as possible without running many more tests or introducing more conditions [37].

3.7 Summary

In summary, this chapter explores the intricate design and set up of the initial subjective testing phase of this body of work. The three-task design model was evolved to accommodate for multi-party conversation-based tasks that would pinpoint specific areas of interest within the field of CA whilst

highlighting their potential sensitivity to transmission delay. The aim of the three-task design being to mimic varying meeting and teleconferencing styles to see how they would withstand fluctuating induced transmission delay levels. Through industrial lead discussions with BT and widely recognised meeting formats the meeting styles chosen to replicate were informal, formal and goal-oriented meetings as they were thought to encompass the vast majority of uses of a teleconferencing system. As discussed, task selection is key when it comes to fully testing a teleconferencing type system or a subjective test platform of any nature; it can demonstrate how inherent impairments associated with mediated forms of communication can impact on conversation and group interaction. A focus on pre-testing of all aspects of a subjective test design proves that preliminary trials of tests are highly recommended and a crucial part to the success or failure of the overall test and task design. It should be emphasised that the design and testing phase of this study proved to be a demanding task, with subject recruitment taking careful consideration along with the logistical demands related with employing the number of people needed to carry out the tests. Other aspects of the design included the selection of the transmission delay levels to put under test, again the final figures for the delay levels came as a result of close discussion with BT based upon industrial known acceptable measures of delay and through the various pre-tests carried out.

Chapter 4

4. Induced Transmission Delay: Subjective Test Results

4.1 Introduction

This chapter aims to detail the results gathered during the first phase of subjective testing and define the CA parameters that will be used. It will present both subjective and objective results obtained, whilst discussing the various stages of audio data and CA parameter extraction, along with the methods used to employ CA techniques across this body of work. A crude measurement of around 38,000 data points from the audio gathered were manually reviewed during the first round of subjective tests. Initially, we begin by highlighting the multiple phases of processing under which the audio collected from the testing phase was subject to; this gives an insight into the inherent difficulties associated with VAD and accurate mark-up. We then continue by addressing the issue of how simple measurements from the audio do not provide enough information from which to identify how well a conversation or interaction between groups went. This is because the context of a conversation, not just timings of speech, plays a major role in determining the level of interactivity between the group members and CA methods can assist in this area. Finally, we conclude with a discussion of the results from the first subjective testing phase, where both subjective and objective results will be shown and presented in further detail.

4.2 Utterance-by-Utterance Adaptive VAD MATLAB Code, Praat: Speech Analysis Software and R Code

The first step in determining CA features was to look at gross measures of information such as simple talking/not-talking from the conversations recorded. Open source MATLAB utterance-by-utterance adaptive VAD code was used to process the audio gathered from each silence cabinet during the testing phase. This automated process utilised MATLAB code developed by Tomi Kinnunen and Padmanabhan Rajan [38, 39] from the University of Eastern Finland (UEF). This provided the initial round of processing that the raw audio files went through in order to detect silent and sounding portions of the conversational speech; this converted the VAD outcome into files known as textgrids. All the information contained in the textgrids was in a simple text file format. These textgrid files in their basic form contained all the data relating to the recordings; it listed each silent or sounding (talking/not-talking) section as intervals with each showing a start and end time for that interval, along with a label being either 0 (silence/not talking) or 1 (sounding/talking). However, due to the frequent inaccurate labelling of the recordings processed by the VAD code, manual adjustments were needed. The adaptive VAD code did allow for adjustments to be made within relation to the thresholds set for the speech detection, but still this did not provide an

accurate enough data set. Adjustment of the VAD pre-processed textgrids was vital for all essential conversation attributes to be detected and included within the sounding potions of the audio files. The corpus of audio files and generated textgrids went through manual adjustments via Praat phonetics software.

Praat is a free speech analysis package that was developed by Paul Boersma and David Weenink from the University of Amsterdam (UvA) [40]. This software offered a semi-automated approach to processing conversational recordings. It allowed for audio files and prior created textgrids to be manually adjusted. This speech analysis tool provided the necessary means for adjustments to be made across all textgrids to ensure an accurate data set. The nature of this task proved to be highly repetitive and vastly time consuming, but essential to ensure that the most accurate data set possible was obtained; VAD alone is simply not accurate enough to capture each segment of audio and identify it correctly. Figure 4.1 is an example of a textgrid display from Praat which shows a small portion of a recording on a spectrogram.

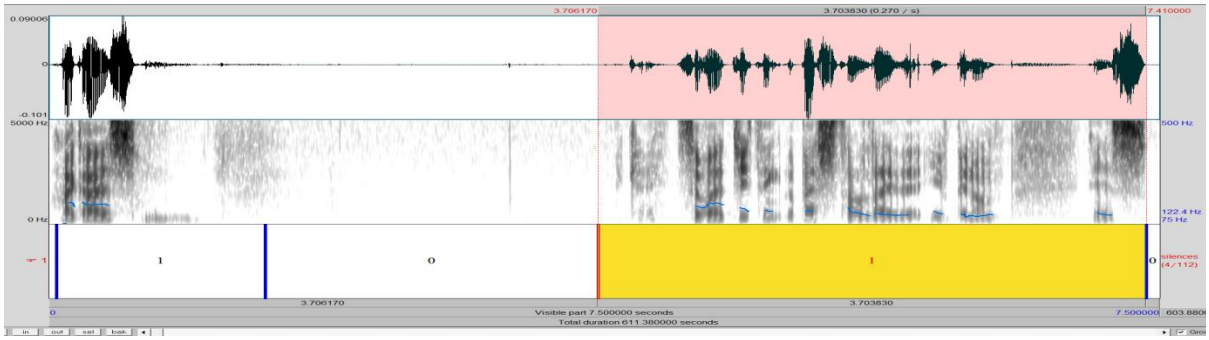


Figure 4.1: Example of Praat Textgrid with Spectrogram

Once accurate manual adjustments were made the next stage was for the audio files to be further processed through some code to help detect overlapping sections of speech. R code [41] which had been utilised within previous works [1] had to be significantly modified to accommodate a four-way conversation, as previous studies had only focused on two-way communication [13, 14, 35]. The R code had been modified to record in list form when an overlap in speech occurs between any of the four participants of a recording. The code would list when a double, triple or quad talk occurred within a recording, whilst also noting which participants were involved in the overlapping talk. The code quantises the timeline and checks which participants are talking at 100ms intervals; 100ms precision was deemed accurate enough, but this could be adjusted to ensure further accuracy if more granularity would be needed in the future. This formed the foundation from which to begin CA parameter detection from the creation of a basic, but accurate textgrids coupled with the utilisation of R code for overlapping speech detection.

4.3 Conversation Analysis Parameters & False Start Analysis

The audible effects of increasing transmission delay over the teleconferencing system deployed was quite evident on the conversations and interactions when manually checking textgrid mark-up. The main issue which was apparent as the delay increased was the confusion when participants were trying to claim the floor of the conversation or trying to establish who had the floor. This key area of interest is termed here as a *false start* situation and shall be explored further. The challenge with automatic false start detection is apparent in the numerous forms and circumstances in which false starts can occur.

To highlight a few of the most common false start situations Figure 4.2 illustrates a very small number of false start possibilities and how they can be produced or occur; this figure was later used as part of a contribution to an ITU standard P.1305: Effect of delay on telemeeting quality [6]. It should be noted that the task of identifying false starts from simply pattern matching and generating code which will look for these patterns is not trivial. This is due to a profoundly high number of ways and situations in which a false start can be produced. This work aims to go beyond any machine recognition as the resources are not yet currently available to automatically detect these kinds of complex and inherent interactional behaviours which to date are manually processed. Figure 4.2 merely displays the most common examples of conversational floor confusion, whilst also assisting in highlighting the nature of a false start.

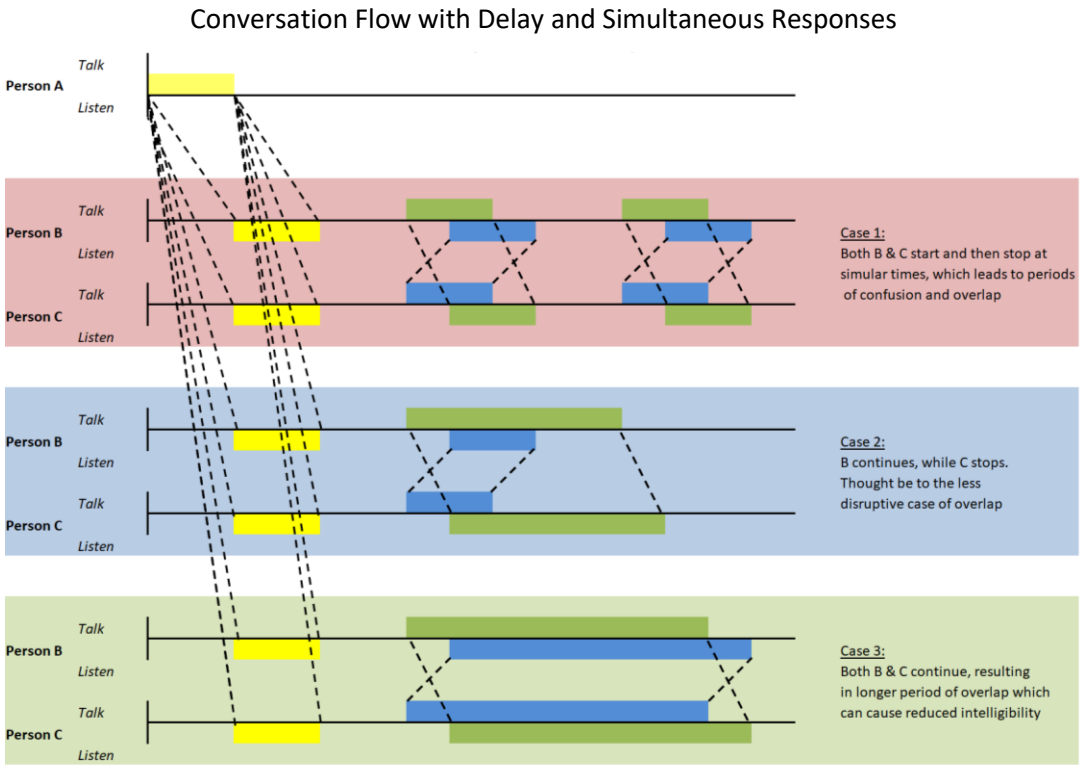


Figure 4.2: False Start Case Figure with Examples of False Start Scenarios

False starts can to some extent be rather crudely monitored by looking for examples of case 1 from the Figure 4.2. As shown in Figure 4.3 a simple graph showing the number of speakers along a timeline of a recording shows a sudden peak in the number of speakers from 0 – 3 or even 4 people, which could indicate a false start. This displays a firm example of case 1, shown previously in Figure 4.2.

Figure 4.3 depicts an example from a recording of an informal meeting task where the delay was set to the highest level, 800ms. The graph indicates that a number of false starts have taken place and that there is somewhat of confusion as to who should speak next to claim the floor of the conversation. The possible false starts have been highlighted with a red dashed line. Importantly as can be seen from the graph, these false start occurrences happen right at the beginning of the conversation when participants first begin to communicate and are unaware of any delay involved in the conference call. This is a typical example of a number of false starts which can possibly be linked to the delay within the call.

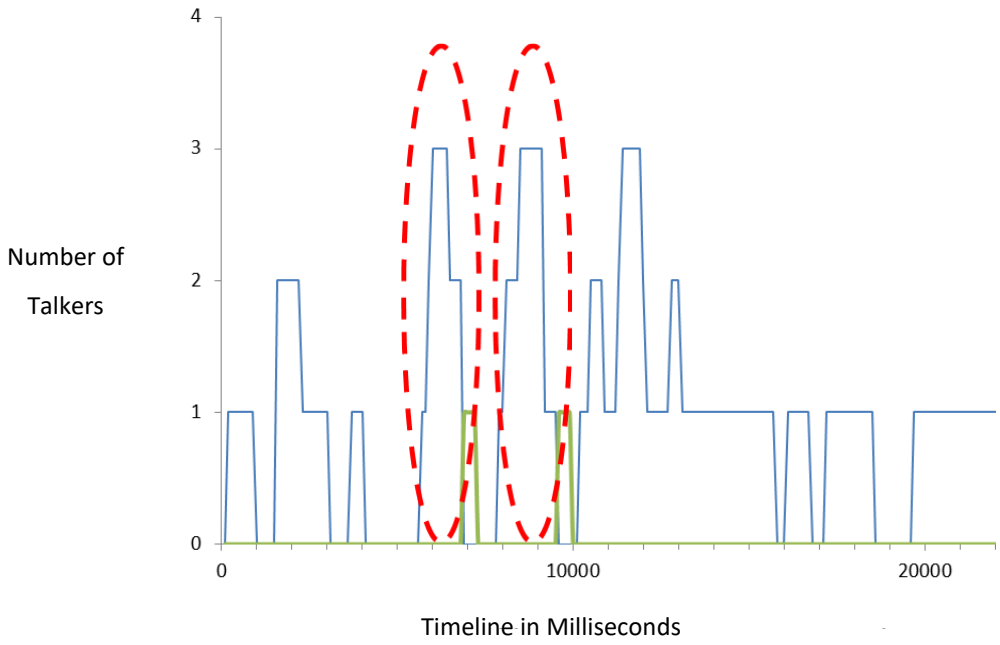


Figure 4.3: Examples of False Starts Relating to Case 1 of Figure 4.2

Figure 4.3 may show a simplistic method of false start identification, but this method could only be used to highlight the most basic of false start scenarios. Another issue which should be taken into account is that the graph cannot display the interactional events or what is being said at that particular point in the conversation. Even this method would require some manual processing, for example, pattern matching this peak in speaker activity would not be able to differentiate between a false start or perhaps a choral co-production activity such as laughter, or one speaker awaiting a short response or continuer

(e.g. yes or no) from multiple participants at the same time. This is where the manual element to this study focusing on conversation context cannot be simply processed by a piece of code to solely identify what is happening at any one point in a conversation. It is unlikely that a distinction between different speak features can be detected reliably at the current time, hence the motivation for the development of a mark-up system that will be described shortly. The R code does however help to identify when multiple participants are speaking at once and assists with a manual mark-up at these specific time periods. The following stage of this research aims to work at scripting as much as possible to make the manual mark-up of each recording as easy and quick as possible.

The scripting identifies through the R tool when a double, triple or quadruple overlap of speech happens within a recording and automatically opens and plays all four audio streams from that conversation. The script jumps to the time frame when the overlap in talk occurred and gives the user the option to manually code what is happening in the conversation. An important factor is that all four streams are aligned and played out together to allow for the whole conversation at that time from all four participants to be heard. This allows for an informed decision to be made about what is happening in the conversation at that particular time. Once a decision has been made it can be manually coded and then the script continues onto the next instance of overlap in the conversation until each overlap has been manually coded. Once all instances of overlap are coded for each conversation, a master textgrid is produced combining all four audio streams together into one textgrid. The way by which the R tool functions is by playing a small segment of the conversation both before and after the overlapping speech has occurred/ been detected. The tolerance for this can be easily adjusted, e.g. plays 2.5 seconds prior to overlap and 2.5 seconds after. This will then be followed by a playout of the exact overlap in speech with no time buffer added. This allows for the user of the tool to gain some of the context of the conversation to mark the speech as accurately as possible, followed by just the overlap portion which is being marked. This feature proves to be very useful as in some scenarios the overlaps occur in rapid succession with fractional gaps in between, as highlighted in Figure 4.2, false start case 1, and Figure 4.3. This makes them incredibly difficult to differentiate between and not so obvious which one has been identified for marking. The implementation and design of the R tool allows for the context to be gathered, followed by playout of just the overlap itself; this provides a far better solution when attempting to identify how to manually code each overlap in a conversation. A conversation analysis approach would facilitate the need to acquire quantifiable data from the testing phase, alongside the subjective based MOS results to enable further rigorous, in depth comparison and analysis. The R tool was used to code the following conversational parameters/behaviours which were identified as key parameters to monitor under varying levels of induced transmission delay. The following sections will present some widely recognised CA terms along with some extended CA parameters, which became apparent under the synthesised communication

conditions available throughout the testing phase. Many of these terms were introduced and discussed in Chapter 2, but they are repeated here to give a definitive list of terms, e.g. conversational parameters/behaviours, used in the processing stage of this study. The CA approach was adopted for conversation monitoring whilst identifying interactions and effectiveness of communication through use of manual mark-up of each CA parameter; below in the following section each CA term is expanded upon.

4.4 Explored Conversation Analysis Parameters

Detailed in this section are officially recognised and previously well-defined and documented CA terms/features. An example of each CA parameter measured throughout the course of this study are listed here in detail with examples and clear definitions of each CA parameter of interest.

- *Anticipatory Completion* – Defined as when a speaker or speakers go to complete another speaker's sentence or turn prior to the original speaker finishing.
- *Choral Co-production* – Multiple speakers or group simultaneous production of a turn, utterance or word.
- *Choral Laughter* – Multiple speakers or group simultaneous production of laughter.
- *Continuer* – A sign of acknowledgement and to continue to the current speaker, examples of this being: *yeah, mm, uh-huh mm-hm,* etc.
- *Inbreath* – An intake of breath can signal an attempt to break into a turn or as a signal that they are about to start speaking.
- *Laughter* – Single person laughter, not chorally produced.
- *No-Gap-No-Overlap* – A minimal or negligible gap between the finish of a turn from one speaker and the transition to another. The change over from speaker to speaker, turn to turn is very smooth and seamless.

4.5 Extended Conversation Analysis Parameters

As previously mentioned, some conversation parameters of interest were developed by extending beyond existing *official* CA features. The development of the parameters listed here came from listening to the recordings gathered during the testing phase and noting the effects that transmission delay had on the behaviours and conversational attributes of the participants interactions. It was apparent upon listening that delay itself lead to behaviours that may differ from that of a tradition face-to-face conversation with no mediated impairments. This prompted the development and expansion of CA parameters which previously may have not been considered or deemed appropriate within traditional CA practice. However, this work aims to expand upon recognised CA features to discover how conversations and interactions

adapt within synthesised environments, which in themselves inherently come with associated limitations and issues e.g. quality issues, communication degradations, transmission delay etc.

- *Continuer with a change of floor* – A continuer to a series of continuers produced as a way to break into a turn and take the floor of the conversation.
- *False Start* – Two or more (multiple) speakers start to talk at once; confusion can occur as to who has the floor of the conversation, who should continue to speak and who should stop? False starts usually occur in multiples and quick succession when delay is introduced, please refer to Figure 4.2 for common examples. This parameter is also recognised within the teleconferencing industry as a known factor in telemeetings.
- *Constructive Overlap* – Any form of overlap which occurs at: the end of Turn Constructional Units (TCU's), speaker transitions or Transition Relevance Places (TRP's) are deemed as "non-violative". The overlap is acknowledged at the appropriate time by the speaker.
- *Detrimental Overlap* – Any form of overlap which does **not** occur at acceptable periods, e.g. end of TCU's, speaker transitions or TRP's are deemed as "violative". This overlap can cause confusion within the conversation as it is mistimed due to delay.
- *Successful Interruption* – An Interruption is deemed successful when a speaker has managed to interrupt another and hold the floor of the conversation.
- *Unsuccessful Interruption* – An interruption is deemed unsuccessful when a speaker does not manage to break into another's turn and is then silenced, without holding the conversation floor.
- *Grey area/undefined* – Used to mark behaviours, overlaps or periods of conversation which are unclear/undefined

4.6 Conversational Experience MOS Feedback Data

As previously discussed in Chapters 2 and 3, gathering subjective data from the testing phase was a key element of this body of work as participants perception would be important to monitor to understand how impairments within teleconferences are perceived by users. The method used to gather subjective feedback was the widely recommended 5-point MOS scale as described with the ITU recommendation, P.800 standard [27]. This is a method traditionally used to collect subjective data whilst rating telephony-based systems to determine the perceived speech and audio quality and has the benefit of an easily interpreted scale on which previously untrained users can rate audio quality among other perception-based measures. Figure 4.4 presents the MOS results captured from the conversational experience surveys given to each participant during the subjective tests; the MOS results shown are from the corpus of data from all nine tests, inclusive of all three tasks from the three-task design model and they aim to

look at the observed effects of transmission delay on multi-party conversations over a teleconferencing system.

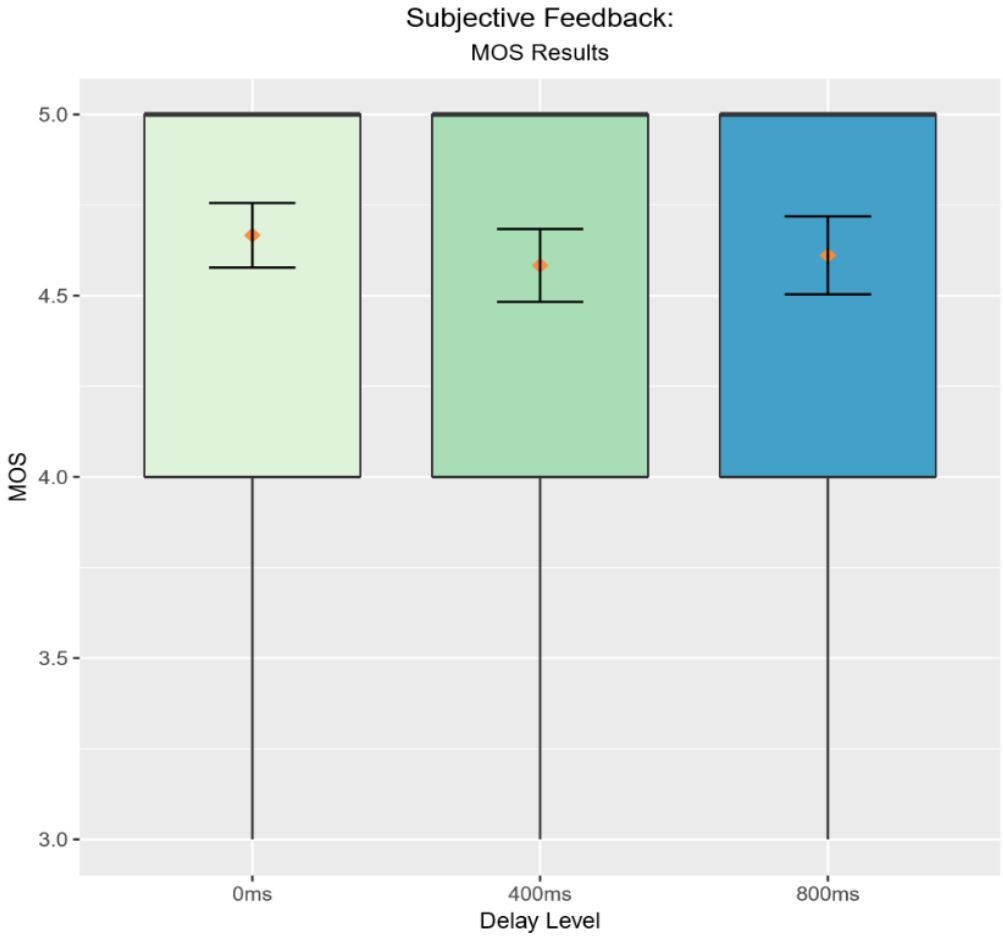


Figure 4.4: Audio Quality MOS Results Box Plot (Grouped by Delay)

As can be seen from the above figure, Figure 4.4, the MOS results show very little variation between all three levels of induced transmission delay that the participants were subject to. The MOS results clearly show the perceived audio quality from one delay level to another does not translate well through a 5-point based scale. It would be inappropriate to assume that participants could not tell the difference between the transmission delay levels, but the MOS results simply demonstrate the complications when testing system impairments paired with variation between task styles and using a basic scale to rate higher quality systems. It would also be inaccurate to presume that each of the transmission delay levels are the same given the MOS result or that they are perceived as the same by participants and that delay is having no effect on the conversation or quality of the teleconference. A factor involved in influencing this is the MOS scale itself and how the users of the system felt that they could rate the audio quality through varying tasks. Many participants throughout the initial testing phase and further developed testing phase, as to be discussed in Chapter 5, expressed the difficulty in which

they had rating the audio purely on a scale of 1-5. This was further confounded within the second subjective testing phase as subjects were unsure how they could compare the different communication modalities e.g. unmediated (face-to-face) and mediated (mono and spatial audio) on this scale alone. This further adds to the shortcomings of the MOS scale for this kind of testing platform and adds additional justification to look closer at more in depth objective measurements, which shall be discussed in the following Section, 4.7.

4.7 Conversation Analysis Objective Results

As previously shown in Section 4.6, MOS results and subjective feedback data alone are not detailed enough to explore the full effects of induced transmission delay on conversations, interactions and the change in participant behaviours in mediated teleconferencing environments. Subjective results and traditional relied upon methods of quality assessment, such as the widely used and recognised MOS scale [27], do not provide an accurate enough depiction of the overall effectiveness of the communication; they solely rely and fundamentally focus on participants thoughts and opinions, which do not always accurately reflect the communication quality itself. The current MOS system takes measures based on the perceived experience of the user of the system who may focus or base their feedback on a number of elements associated with the teleconference. This section sets out to display accurate representations of objective data gathered from the initial testing phase to help support and illustrate actual events which occurred within those conversations to better understand what impact and implications differing levels of transmission delay really had. We begin by presenting the collated CA objective results in box plot form to clearly highlight in the form of visual representations the differences across all delay levels used in the testing stage. This then leads onto the introduction of the preliminary and simple conversations metrics which lead to the further investigation into CA and how this provided a better understanding of conversations. These simplistic and crude measures demonstrate how simple measurements alone would not pinpoint any detail as to how well or not communication and interactions were, or how this was affected based on the different delay levels used. We conclude with the expansion of the key findings from the corpus of data and which CA parameters were to be discovered further in order to facilitate a second round of the testing as detailed in Chapters 5 and 6.

4.7.1 Box Plot Representation of CA Objective Results – Strongly Correlated Results

The full set of results from the induced transmission delay testing phase, comprising of the entire population of nine groups, detailing the three tasks undertaken by each group will be documented here. This subsection also approaches the topic of CA and extended speech parameters and discusses the comprehensive CA features that were monitored and analysed for the purpose of this work. Strongly correlated results shall be presented here, with discussion on remaining results shown in Section 4.7.3. The box plot shown in Figure 4.5 illustrates the effects of induced transmission delay on the conversation parameter no-gap-no-overlap at the varying delay levels of 0ms, 400ms and 800ms. The box plot displays the results obtained from the corpus of data gathered from the nine groups.

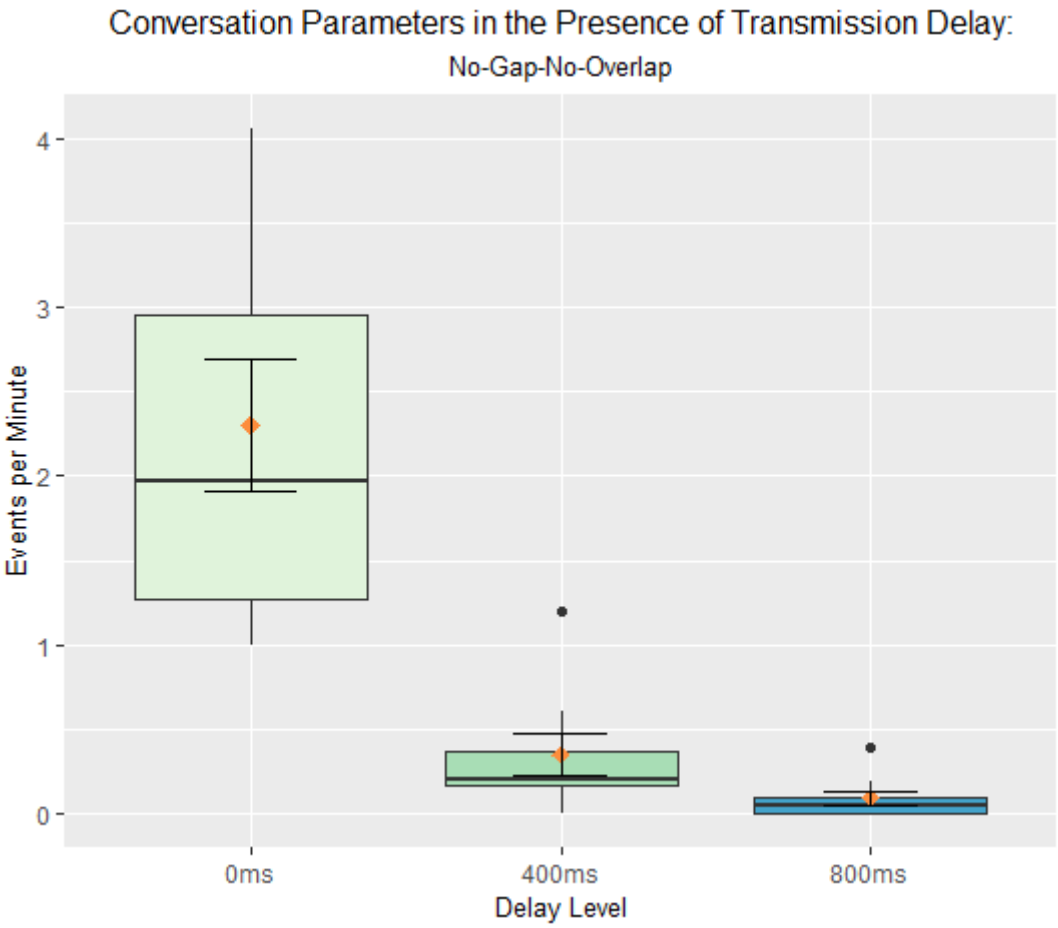


Figure 4.5: No-Gap-No-Overlap Conversation Parameter Box Plot (Grouped by Delay)

Figure 4.5, along with all proceeding figures in this chapter, presents the results in box-plot form and shows the mean (orange diamond) together with the Standard Error of mean in the form of error bars. The box plots present in the normal manner the median, the inter-quartile range and whiskers that extend to either the highest (lowest) value or 1.5 times the inter-quartile range whichever is the least [42,

43]. The boxplot shows the positively skewed distribution of data and a stark contrast is highlighted between the no delay scenario and the delay induced scenarios. In order to interpret the results a classification of the CA speech parameter in question must first be defined. No-gap-no-overlap is a common occurrence and behaviour within normal face-to-face speech, also being a well-recognised CA term, it can be defined as minimal or negligible gap between the finish of a turn from one speaker and the transition to another. The change over from speaker to speaker, turn to turn is very smooth and seamless [21]. As can been seen in Figure 4.5, a lower frequency of no-gap-no-overlap turn transitions occur with higher delay, this limits the amount of smooth transitions in talker turn alternation compared with that of a conversation with no delay or low delay. With higher transmission delay a less seamless communication at TRPs is apparent. This may seem an obvious finding in retrospect but can be a highly effective indicator of how well a conversation is flowing and can also be a good indicator of how much transmission delay is being experienced. The nature of no-gap-no-overlap does lend itself to potentially being automatically detectable as it is a relatively simple conversation attribute; this is due to it being based around timing within a conversation and not relying on the context of the speech itself. This is one aspect that sets no-gap-no-overlap apart from most of the other speech parameters investigated throughout this study, as it relates to *when* something is being said, rather than *what* is being said. This concept shall be explored further is Chapter 6 under future work.

The box plot shown in Figure 4.6, presents the data captured from the conversation attribute false starts, again at the varying levels of transmission delay under test in the initial round of subjective testing, being 0ms, 400ms and 800ms. As can quite clearly be seen from the visual representation shown in Figure 4.6, the stark contrast between transmission delay levels again is undoubtedly evident. The identification of false starts became apparent upon listening to recorded conversation with induced transmission delay as an impairment, where they distinctly present as a negative conversation attribute. This particular parameter became one of many extended CA features. This arose from a combination of observing the interactional behaviours associated with official CA features already widely acknowledged within the field of CA and discovering how these are adapted and modified to fit the associated delay level being experienced. Delay is not found within the usual unmediated face-to-face conversation environment which CA is accustom too, so therefore when attempting to recreate a teleconferencing environment, a highly synthesised representation of as close to normal speech as possible it would be sensible to consider how users may adapt to these types of impairments. This prompted the acknowledgment and need for extended CA parameters to be incorporated to fully appreciate the impact of these impairments on the conversations. The lack of studies into this area of multi-party teleconferencing interactions created a gap to accommodate new classifications of behaviours associated with transmission delay and not necessarily

found in everyday conversations. This formed an interdisciplinary approach to merging the two components, teleconferencing and CA.

False starts as an extended CA parameter are defined as two or more (multiple) speakers who start to talk at once; confusion can occur as to who has the floor of the conversation, who should continue to speak and who should stop? False starts usually occur in multiples and quick succession when delay is introduced and can be deemed as a form of violative interjection (please refer to Figure 4.2 for the most common examples) and as Figure 4.6 demonstrates the increase in transmission delay also promotes the significant increase in false start occurrences. False starts are thought to have a detrimental impact on the overall conversation performance and teleconferencing efficiently; they introduce an element of confusion within the group dynamic with multiple unintended overlaps in speech that, upon hearing, causes an audible negative influence on the conversation and break down within the conversational flow.

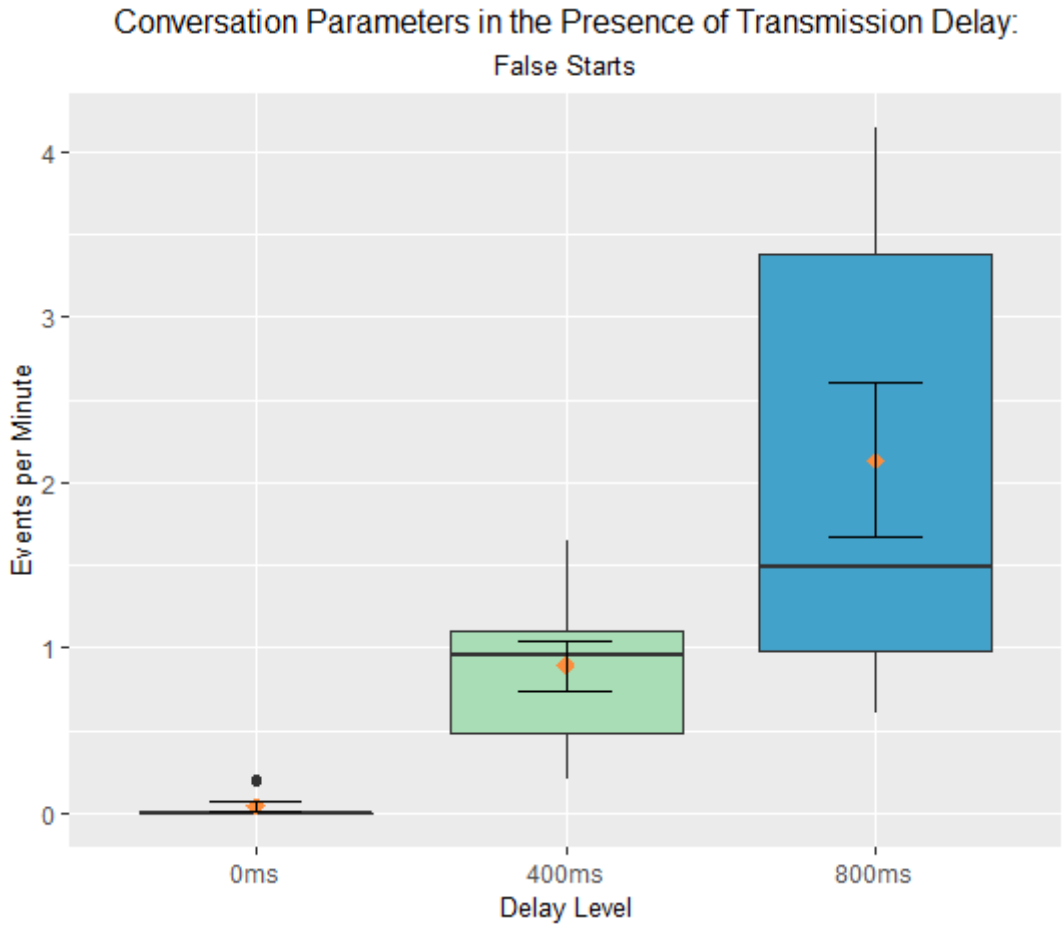


Figure 4.6: False Starts Conversation Parameter Box Plot (Grouped by Delay)

Another conversation parameter of interest to monitor under varying levels of delay were successful interruptions, the results of which can be seen in Figure 4.7. The findings from this extended CA attribute clearly highlights the negative effect that increasing levels of transmission delay are having

on the participants ability to successfully interrupt or interject in a suitable manner. By definition a successful interruption within this study is deemed as successful when a speaker has managed to interrupt another speaker's turn or utterance and continues to hold the floor of the conversation. For this type of exchange to effectively happen, the interruption would need to be timed at an acceptable period within the conversation, e.g. at a relevant speaker transition point such as a TCU or TRP. As the box plot shows the execution of this parameter drops in frequency considerably as higher levels of delay are experienced. This means that conversations and group interaction can suffer as a result, e.g. if effective floor exchange from speaker to speaker is not feasible through the use of successful interruptions.

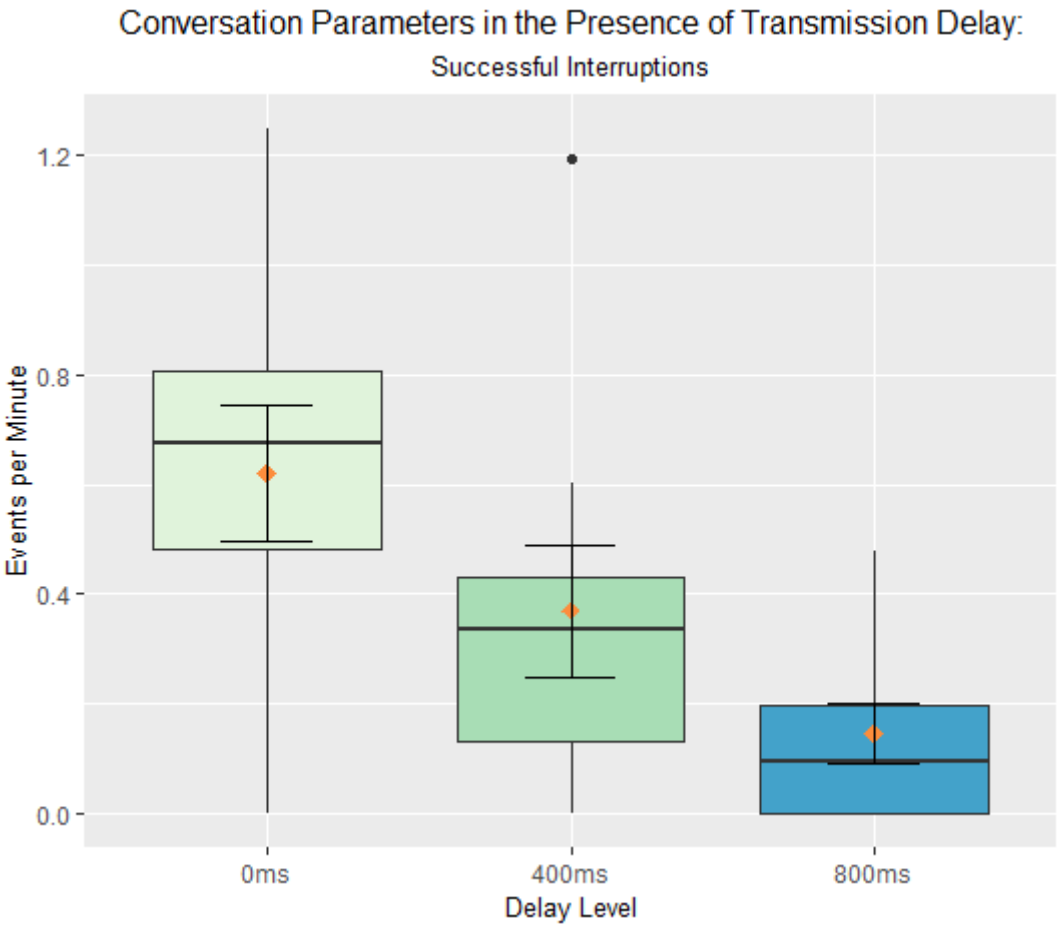


Figure 4.7: Successful Interruptions Conversation Parameter Box Plot (Grouped by Delay)

Figure 4.8 details the collated results for detrimental overlap which is another extended CA speech parameter where the focus of the overlap is based around violative interactions. A detrimental overlap is classified as any form of overlap which does **not** occur at acceptable periods, e.g. end of a TCU, speaker transitions or TRP and thus is deemed as "violative". This type of overlap can cause confusion within the conversation as it is mistimed due to the induced transmission delay impacting on the delivery of the overlap. The number of occurrences of events per minute drastically increases again with the rise

of induced transmission delay values. Detrimental overlap can be categorised as a negative conversation attribute and the presence of this parameter is shown to worsen, becoming more frequent as higher levels of delay are experienced.

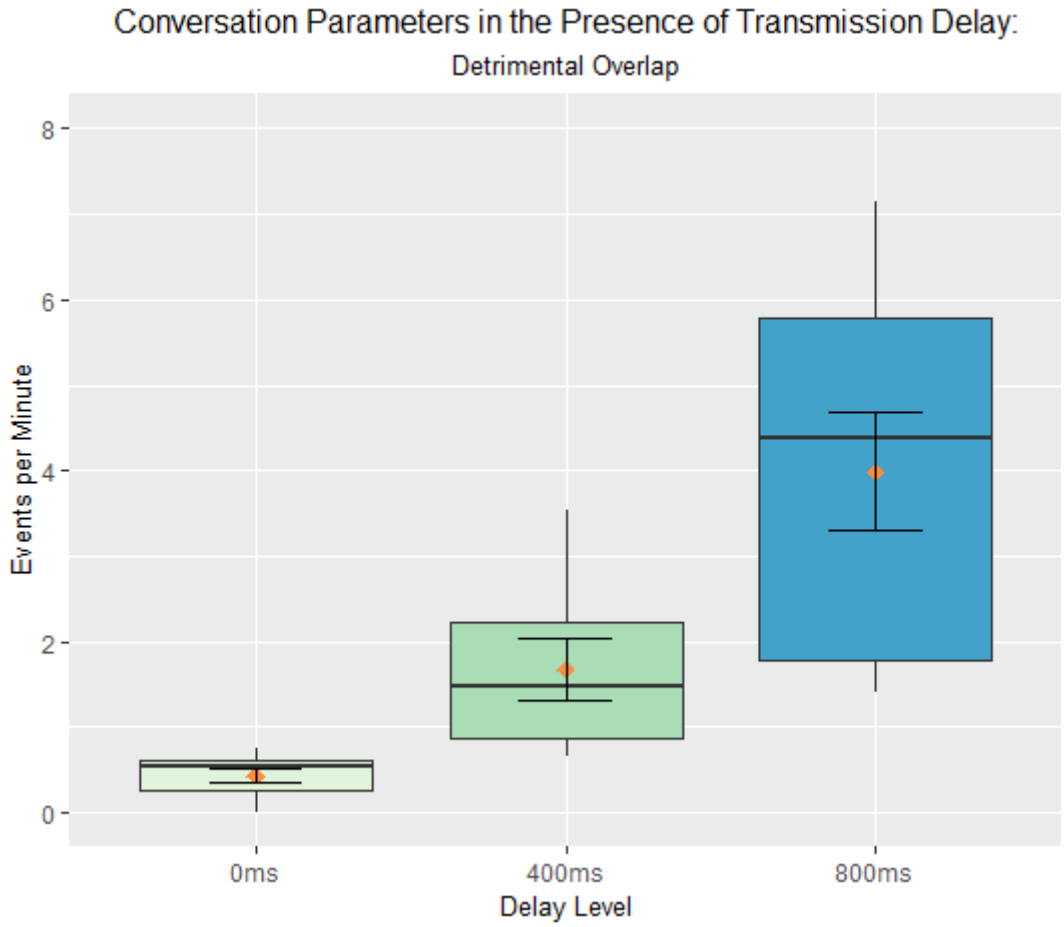


Figure 4.8: Detrimental Overlap Conversation Parameter Box Plot (Grouped by Delay)

In contrast, Figure 4.9 details the results obtained for constructive overlap which can be described as the opposite of the detrimental overlap parameter. This conversation attribute can be considered as closely aligned to behaviours commonly characterised as part of face-to-face natural conversations. A prominent feature of this parameter is that the presence of the overlap does not interrupt the flow of the conversation; the overlap is acknowledged by the other talkers and produced in a non-violative manner. The overlap takes place at constructive and sensible points within a conversation such as at speaker transitions, the end of TCU or TRP where it would be expected that some form of overlap may occur. As Figure 4.9 demonstrates the escalation in delay shows a drop in the number of events of constructive overlap. The difference between both 400ms and 800ms is minimal, which highlights how even a 400ms delay can have an adverse effect on the production of this speech parameter as it relies upon precise timing in order to line up an overlap with the end or pause in another's turn. Delay exacerbates the

problem and the potential for constructive overlap to occur in adequate timing; precision timing is required to ensure the overlap is heard and acknowledged but does not cause any form of disturbance or negative interruption in the conversation.

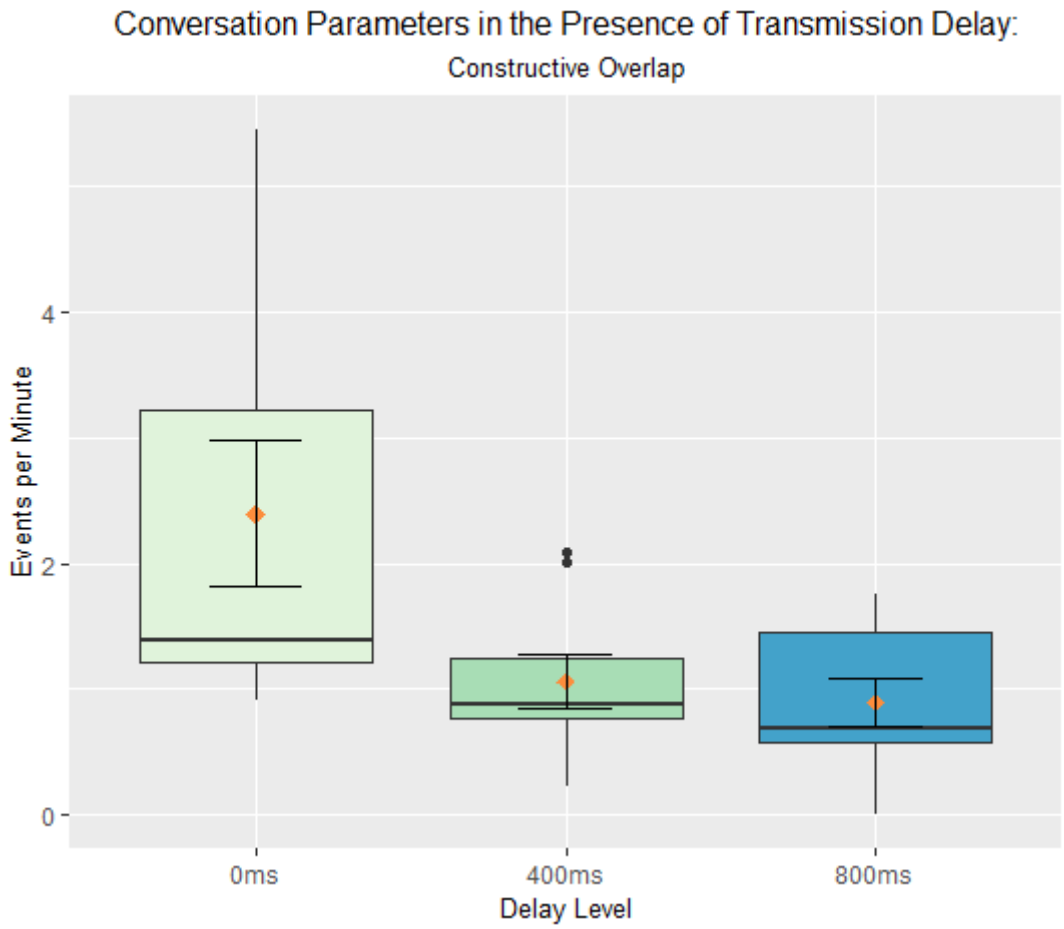


Figure 4.9: Constructive Overlap Conversation Parameter Box Plot (Grouped by Delay)

Figure 4.10 sets out to combine all conversation parameters that have a disruptive effect on the group conversation and interactions. The box plot consolidates three parameters together to show a collective result for false starts, detrimental overlap and unsuccessful interruptions. It illustrates the overall impact that increasing delay levels have with the rise in number of occurrences. The widest spread of data is noted for 800ms delay, whilst still showing a raised number of events compared with the lower/no delay levels. The task is thought to be having an influence, in addition to the levels of delay, as with all CA parameters monitored and discussed here in Chapter 4. The three-task design model encompasses three very different styles of task along with differing objectives e.g. free conversation, highly scripted and goal-oriented. A key objective within this study was not only to monitor the effects of transmission delay on multi-party teleconferences but through the process an emphasis was placed on the importance of task selection and design used for the conversation-based tests. Task selection was just

as crucial as which delay levels to test, as the task itself was found to have a bearing on how well users were able to engage with the system and whether or not the task prompted enough conversation from which to obtain sufficient amounts of data.

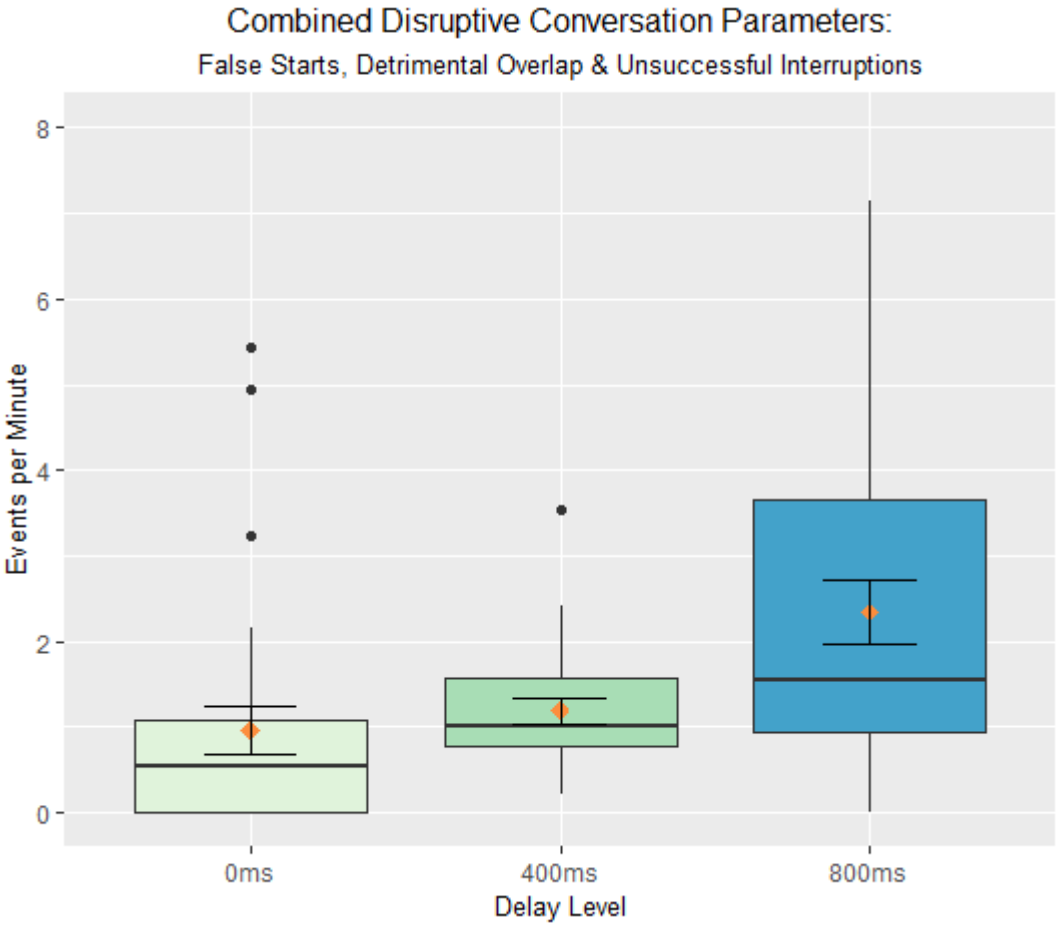


Figure 4.10: Combined Disruptive Conversation Parameters Box Plot (Grouped by Delay)

An example of the implications of task selection on the conversations are shown in Figure 4.11, which presents the results gathered for number of turns and false starts, grouped by task. The figure shows a greater correlation between the informal and goal-oriented tasks, rather than the informal and formal tasks. Both the informal and goal-oriented tasks display higher instances of number of turns and false starts, implying higher levels of interaction and exchange than compared with the formal task. This is encouraged to promote conversation, allowing for greater amounts of CA data to be captured and displays how the goal-oriented task can reproduce certain parameters similar to that of a free-flowing conversation (which the informal task sets out to replicate). These findings aid with the process of expanding upon the subjective test design for further investigations as discussed in Chapter 5.

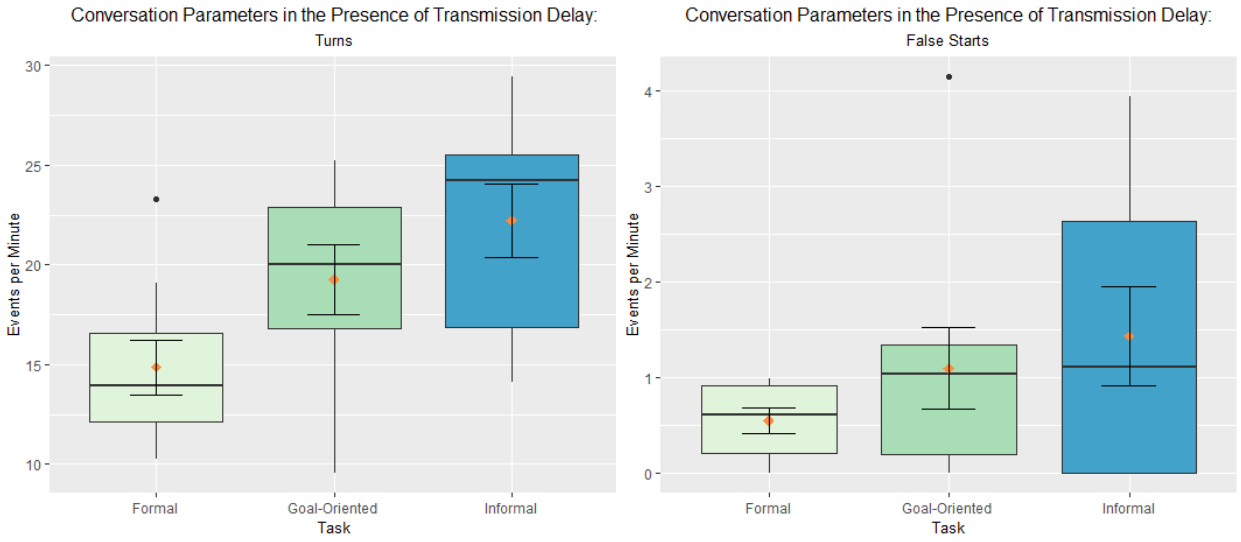


Figure 4.11: Joint Turns and False Starts Box Plot (Grouped by Task)

4.7.2 Discussion on Poorly Correlated Initial Results

This subsection looks at the initial and simplistic results gathered from the initial testing phase. Basic and rudimental measures of the mean instances of double talk, mean number of turns and mean turn duration in Tables 4.2, 4.3 and 4.4 respectively, shows that the variation in these results between the delay levels is small and does not highlight any significant or notable findings. It was initially presumed that these basic speech parameters would show statistically significant variation between the delay levels; however, the initial results were not solid enough to indicate that the increase in delay was having a detrimental effect on the conversations and the interactions between the participants.

The results which have been obtained from the conversational recordings have been selected based upon their appropriate use towards a typical industrially focused teleconference. The conversational parameters were chosen in order to determine whether transmission delay within the system had an impact upon basic conversation attributes and behaviours; the attributes selected for initial analysis were the following:

- Number of instances of double/overlapping talk
- Number of turns taken
- Mean turn duration

With the above conversational attributes in mind, they were selected as some of the most basic and easily obtainable attributes to measure. The end goal after having achieved metrics for overlapping talk, turns taken and turn duration would be that these parameters would lead towards further investigation

of more complex interactional behaviours which have been discussed in Section 4.7.1. The R code generated to obtain these results had been modified further to display when and just how many participants were talking in overlap, rather than purely the mean number of times any form of overlap happened. The results shown in Table 4.1 displays each of the above-mentioned conversational parameters. It should be noted that the following work shows a negative result; from this further understanding of how to detect overlapping talk and ways in which to measure this has been gained and have be shown and discussed prior in Section 4.7.1.

Test	Delay (ms)	Order	Length	Turns	Overlap	MeanTurnDur
1	0	1	611.3	9.8	4.1	6.1
1	400	2	816.5	6.7	5.2	8.3
1	800	3	1251.4	10.0	6.7	5.4
2	0	3	872.7	5.8	2.1	9.9
2	400	1	1157.3	13.9	3.3	3.7
2	800	2	615.9	19.0	7.8	2.5
3	0	2	1244.5	16.6	5.0	3.0
3	400	3	596.1	9.0	3.6	6.3
3	800	1	903.6	11.6	6.5	4.4
4	0	3	594.3	17.8	10.0	3.3
4	400	2	1252.9	12.4	7.9	4.5
4	800	1	968.7	8.9	6.0	5.8
5	0	1	1115.9	15.4	7.0	3.7
5	400	3	895.4	12.5	7.5	4.6
5	800	2	621.3	15.5	10.0	3.5
6	0	2	887.5	15.2	9.5	3.9
6	400	1	618.6	20.3	13.5	2.8
6	800	3	1059.4	15.5	12.1	3.5
7	0	1	561.6	19.2	10.3	2.7
7	400	2	1103.1	16.4	7.6	3.2
7	800	3	797.2	15.3	9.8	3.4
8	0	3	1271.3	16.0	7.8	3.6
8	400	1	903.6	9.6	7.4	5.9
8	800	2	625.8	14.8	8.3	3.7
9	0	2	740.1	13.0	6.3	3.9
9	400	3	601.3	17.9	10.2	2.9
9	800	1	1244.6	15.5	8.6	3.6

Table 4.1: Initial Results showing Double/Overlapping Talk, Turns and Mean Turn Duration

Table 4.2 shows the averages calculated for each transmission delay level in relation to the mean number of instances of double/overlapping talk per minute. On first appearance the mean number of overlaps seems to show a corroboration between increasing levels of delay showing increasing levels of

overlap. It may initially appear that the delay is having a detrimental effect, given the results, as the overlap could be seen as having negative connotations and a disruptive impact on the conversation. Without further investigation into other speech parameters this should not be assumed; as previously mentioned not all overlaps are negative, overlap may be produced in the form of constructive overlap or continuers to name a few examples which have a positive effect on the conversations and participants interactive behaviour.

Delay Level	Mean number of instances of double/overlapping talk
0ms	6.90
400ms	7.36
800ms	8.43

Table 4.2: Initial Results showing the Mean Number of Instances of Double/Overlapping Talk per Minute

Table 4.3 below shows the averages calculated for each transmission delay level in relation to the mean number of turns taken per minute. Taking a closer look at the number of turns, the results are unclear and does not display a link between the amount of turn taking decreasing or increasing with delay, as one might initially hypothesise.

Delay Level	Mean number of turns taken
0ms	14.32
400ms	13.16
8000ms	14.02

Table 4.3: Initial Results showing the Mean Number of Turns Taken per Minute

Table 4.4 below shows the averages calculated for each transmission delay level in relation to the mean turn duration. This further compounds the notion that these rudimental measures cannot correctly give a definitive answer whether the conversation quality is good or poor.

Delay Level	Mean turn duration
0ms	4.45
400ms	4.68
800ms	3.98

Table 4.4: Initial Results showing the Mean Turn Duration

As is evident from these basic and easily captured metrics, the need for greater investigation and in-depth study into the complex nature of speech and turn organisation in an interdisciplinary approach was

fundamental to the study. A soft-focus CA approach to the work helped in forming a greater understanding of multi-party teleconferencing and how induced transmission delay can impact on group conversation; this is where CA methods assisted with this body of work and allowed for the development and expansion of existing CA metrics.

4.7.3 Box Plot Representation of CA Objective Results – Poorly Correlated Results

This subsection presents some of the poorly correlated results gathered from the testing phase and gives a brief overview for each feature. It is important that poorly correlated results are considered as they can have the potential to assist further with the expanded subjective test design. They can shed light on how to adapt the tasks to pick up on more sensitive CA features that from the initial round of testing are not well understood and therefore need greater investigation.

The box plot in Figure 4.12 shows the results obtained for the official CA speech parameter anticipatory completion.

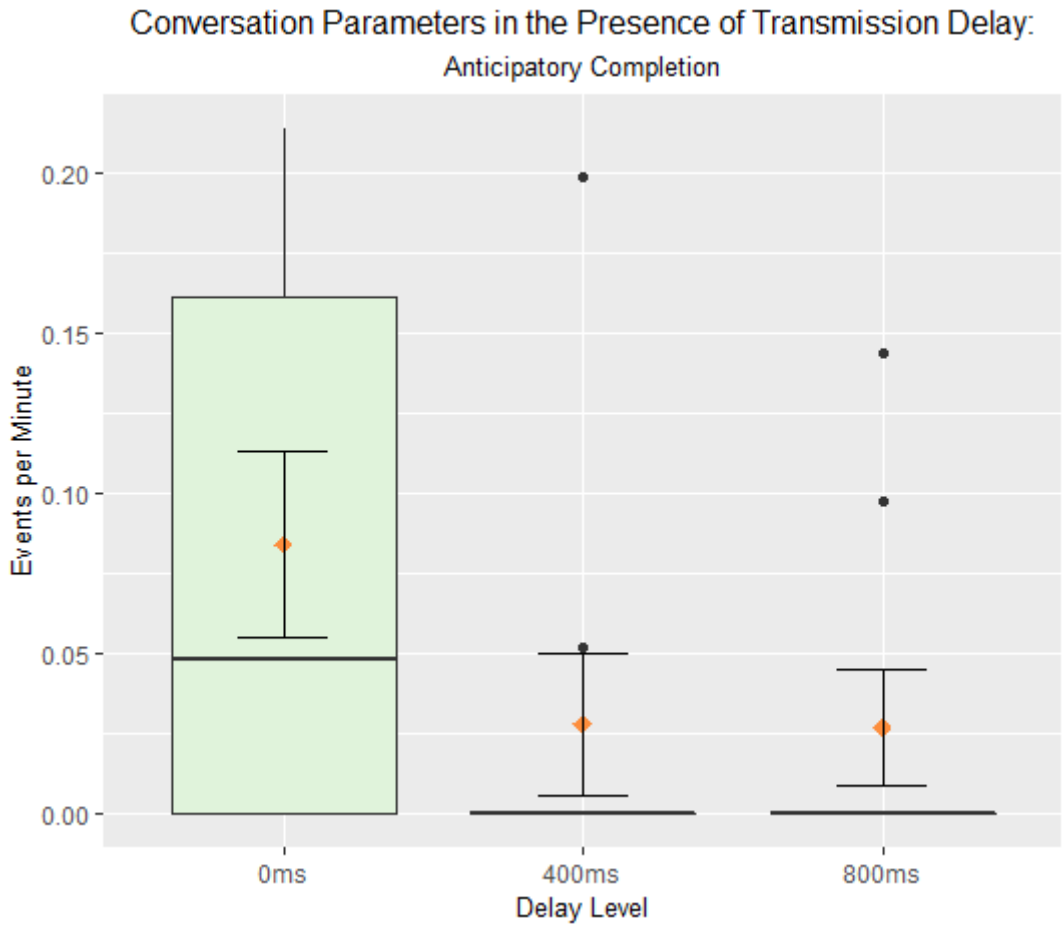


Figure 4.12: Anticipatory Completion Conversation Parameter Box Plot (Grouped by Delay)

The classification of an anticipatory completion is defined as a speaker or speakers attempt to complete another speaker's sentence or turn prior to the original speaker finishing. As can be seen from the box plot the number of instances of anticipatory completion is very low, therefore making it difficult to draw any conclusions due to the lack of information. The task selection could be a factor as well, as each of the tasks may not have brought out this particular conversation feature; again, a medium or high level of delay has an impact on its production with events falling further.

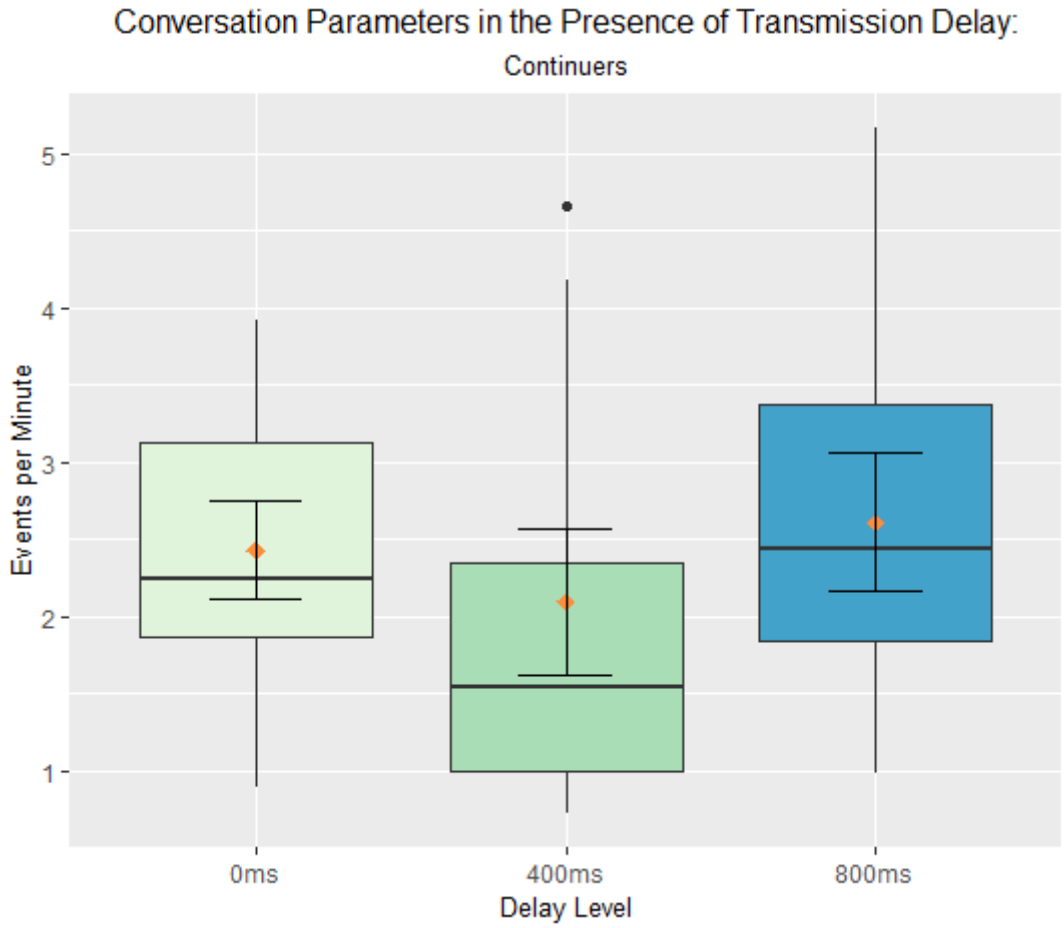


Figure 4.13: Continuers Conversation Parameter Box Plot (Grouped by Delay)

An area of interest in this study was the monitoring of the CA attribute continuers; this speech parameter is well documented in the field of CA and presents itself in normal face-to-face speech. Continuers can be categorised as a sign of acknowledgement and indication to continue to the current speaker, examples of this being short utterance such as: *yeah, mm, uh-huh mm-hm,* etc. These types of interactions would be expected to happen at relevant places in a conversation, such as speaker transitions, the end of a TCU or TRP. The hypothesis on continuer production in the presence of transmission delay was that as delay increases the frequency of continuer occurrences should subside

(also, as referred to in Chapter 6, a decrease in the volume at which continuers are produced should be apparent). This hypothesis was formulated upon listening to all the recorded audio from the testing phase and observing how the varying delay levels effected the behaviour of participants; one notable question was the impact on continuer use. The results shown in Figure 4.13 from the first round of testing does not clearly illustrate that continuers were correlated with the change in delay. This is thought to be due to a number of elements involved in the testing phase which would have an impact on continuer delivery. Continuers are thought to be highly individual in their delivery, with some groups more inclined than others to utilise them; also, continuers can be highly influenced by the style of task, with certain tasks requiring more frequent use of them than others. This shall be investigated and discussed further in Chapter 5.

Continuers with change of floor became an extended CA attribute that is closely linked with continuers. Figure 4.14 again shows poor correlation between the levels of transmission delay and the production of this attribute; however, as with continuers it is felt that with further investigation a greater understanding of the role of continuers with change of floor within teleconferencing can be discovered.

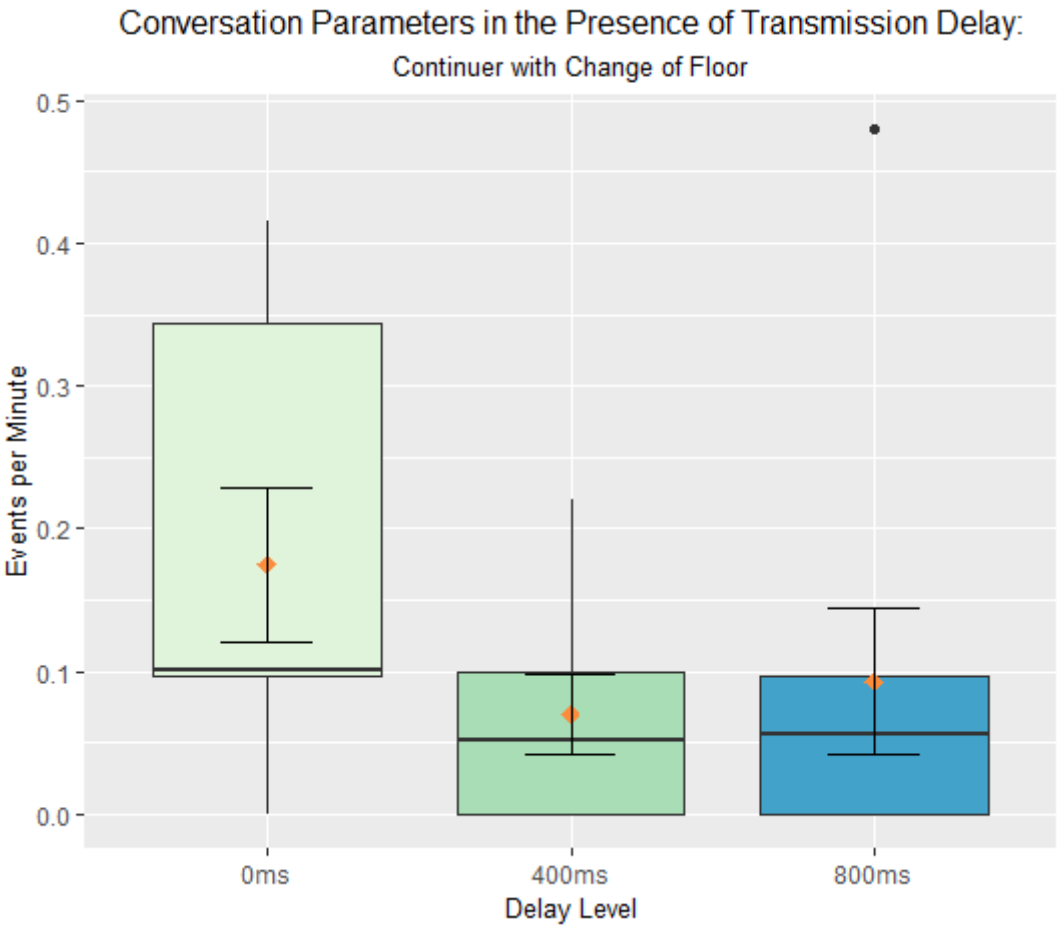


Figure 4.14: Continuer with Change of Floor Conversation Parameter Box Plot (Grouped by Delay)

Choral co-production results are exhibited in Figure 4.1.5 and shows again weak correlation between transmission delay levels. It would be presumed that choral co-production would be more frequent within low to no delay situations, when in fact the opposite is shown in the results. Choral co-production does rely on timing of delivery in order to make the execution of the utterance choral, so higher levels of delay should have an impact on this CA feature. This parameter could be affected by the task variation in the three-task design model, resulting in an unclear outcome. This conversation feature has the potential to be discovered further in the expansion of the subjective tests, with the intention to discover if refinement of the tasks can aid with understanding choral co-production behaviours in delay induced environments.

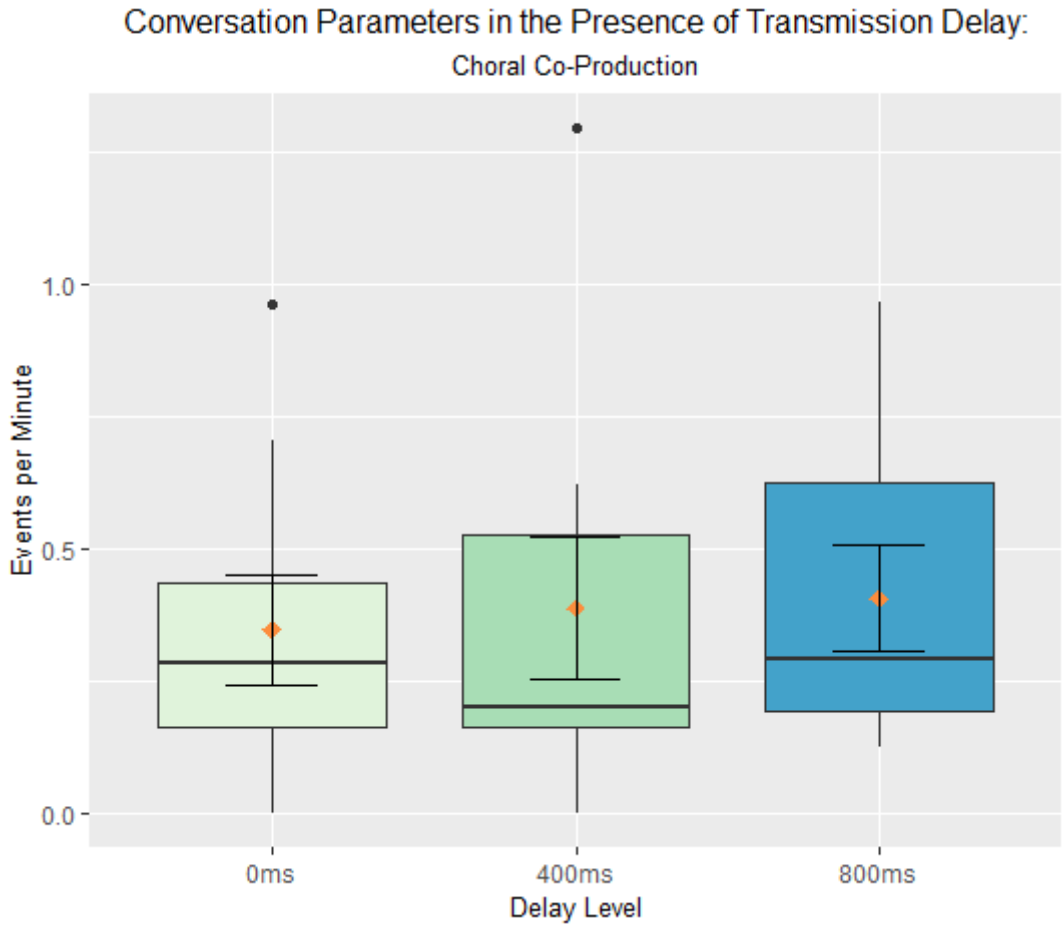


Figure 4.15: Choral Co-Production Conversation Parameter Box Plot (Grouped by Delay)

4.8 Discussion

This section aims to discuss the findings gathered from all the results mentioned in this chapter. It will summarise and review the presented results along with areas for expansion for the further subjective testing (which shall be discussed further in Chapter 5).

The initial round of subjective testing gave a promising set of results; existing and the newly extended CA measurements gave significant information on the interactions and behaviours displayed by participants in the conversation-based tests and how each of these CA features responded to the introduction of transmission delay. The results highlighted and confirmed many of the preconceived notions of issues associated with delay in teleconferencing by using CA methods that enabled the extraction of key conversation attributes. The results captured from some of the key attributes of interest, such as false starts and no-gap-no-overlap, are clearly in line with what would be expected when subject to delay impairments and displayed the degree of impact that higher levels of delay had on multi-party interactions. The results also prompted the need to understand further how a better referencing system for rating conversation-based tests could potentially be developed. The MOS results clearly demonstrated how subjective feedback alone was not enough to confidently state how well the system performed under varying levels of delay. With the help of CA methods, the second round of subjective tests can explore further how focusing on key conversation features can assist with a more accurate platform on which to rate conversation-based tests. The poorly correlated results also gave additional and invaluable information regarding the processing of conversation-based tests by highlighting some issues with the test and task design and provided sufficient data from which to expand on in future tests.

Some of the key findings and questions raised from the first round of testing provide the scope for expansion for further testing, as detailed in Chapter 5. Notable results included no-gap-no-overlap: this conversation parameter highlighted how a lower frequency of no-gap-no-overlap turn transitions occur with higher delay, this limits the amount of smooth transitions in talker turn alternation compared with that of a conversation with no delay or low delay. The lack of no-gap-no-overlap at the higher levels of delay meant that the communication was not as seamless at expected points such as TRP's. This may seem an obvious finding in hindsight but is a highly effective indicator of how well a conversation is flowing and also an indicator of how much delay is being experienced.

An unexpected result was attained for the continuers speech parameter with unclear evidence on how they were affected by transmission delay. With continuers being part of normal face-to-face speech, it was important to monitor their delivery in the presence of transmission delay as it was felt that they could potentially be an indicator of good, or poor quality, conversation. The nature of continuers sets them apart from most of the other conversation features under investigation and makes them a candidate for automatic detection if proven to show suitable benefits to conversation quality. The current results held for continuers does not follow the pattern that would have been expected or anticipated. The initial hypothesis was that as delay increases the frequency of their occurrence should subside (also as referred to in Chapter 6, a decrease in the volume at which continuers are produced should be apparent). The

results when sub-divided by group, delay level and task, still does not illustrate the initial concept. This is believed to be due to a number of key elements involved which would have an impact on continuer delivery:

- Highly group/individual dependent – Some groups are naturally more inclined to produce continuers than some other groups which are not as interactive with each other. Also, with the loss of visual cues, as all tests took place under audio only conditions, some groups or individuals feel the need to produce more continuers to reassure the other members of their active participation, whereas others may not compensate as much.
- Highly task dependent – Some tasks have been found to draw out the use of continuers to a greater or lesser extent than others. An example of this being the goal-oriented task, which is highly interactive, plus has the added competitive component of participants being asked to finish a map within a set time period. The task requires significantly more feedback from all group members than usual in order to effectively carry out and complete the task within the given time frame. The informal task, on the other hand, had a totally diverse approach as a freer flowing and relaxed scenario in order to create a natural conversation type environment. Although the task was still timed, the nature of the task does not prompt a competitive element like the goal-oriented task. This task also allows for some members of the group of remain inactive or less active in the conversation than others.

It should also be noted that tasks were of differing lengths with the informal task taking 10 minutes, the formal task taking 15 minutes and finally the goal-oriented task taking 20 minutes; each group only experienced each task with one delay level due to time constraints. This is an area for improvement for future testing, considering the time taken for each task to allow for all delay levels to be experienced per task, whilst also allowing for ample data collection. Also, the longer tasks did prove to be extremely difficult to process with the manual adjustments and coding of CA parameters taking a significant amount of time. In addition, the need for a face-to-face condition along with an orientation or familiarisation round was recognised**.** These requirements going forward with future testing would be recommended as there was no face-to-face condition to compare group performance against and to provide a “best case” scenario to help interpret results better. Also, there was no orientation or familiarisation round to allow groups to become accustom to the test environment, conditions or each other; this in turn can have an impact on their performance which has no bearing on the system under test itself and therefore should be strongly considered. The second round of subjective testing, as detailed in the following chapter, Chapter 5, aims to eradicate the above mentioned issued by way of modified test and task design.

In addition, delay levels considered for the testing phases were trialled extensively before commencing both rounds of official tests. Widely accepted recommendations such as the G.107 E-Model [17] and G.114 [18] standards acknowledge that one-way transmission delay should not exceed 400ms; this assisted in defining a reasonable upper delay limit at 400ms (used in both testing phases as the medium delay level). In order to fully test the system and push the boundaries on known acceptable delay levels it was decided that a larger delay should be incorporated to gauge the impact this would have on participants conversational flow and therefore how it would effect specific CA parameters that were monitored throughout. The highest delay level was set at 800ms (used in the initial set of tests as the high delay level) to observe if there were any major differences between the known upper accepted level of 400ms and that level doubled at 800ms. The low delay level was set at 0ms[2] to simulate a setting as close as possible to a face-to-face environment with no delay. This condition was also used as a benchmark as the "best case" scenario, without the inclusion of an actual face-to-face round, to give participants the ideal synthesised conditions in which to perform as they would in natural conversation. The value of 0ms was used in both testing phases as the low/virtually no delay level. A low/medium delay level was used during the second phase of tests and was set at 200ms, this value was chosen in order to test the theory that normal conversation is understood to degrade and show signs of deterioration in terms of interaction on or around 200ms [22]. Additional, 200ms was half the previously used and known accepted delay level of 400ms and was selected with the intention to observe any differences between ITU recommended delay levels and CA acknowledged findings. Non variable delay levels were used throughout and were chosen to give an accurate measure of user performance and interaction at each exact and specific metric. Variable delay between participants may be considered more akin to a real life setting over audio conferencing. However, in practice jitter buffers tend to keep the overall delay near constant, except in extreme error conditions. Additionally, using variable delay would have proven impractical when attempting to quantify and calculate the effects that specific delay values had on conversations over mediated channels.

[2] It should be noted that there was a notional system delay of 2ms which should be added to the values above. The 2ms delay value was determined by BT through acoustic click delay from microphone to headphone and measured on an oscilloscope.

Chapter 5

5. Multi-Party Teleconferencing: Monitoring Effects of Transmission Delay in the Presence of Spatial Audio - Experimental Design

5.1 Introduction

This chapter aims to outline the second phase of subjective testing untaken, which further expands upon and refines the initial test design as discussed in Chapters 3 and 4. The second phase of subjective testing explores some of the issues encountered within the initial test design and seeks to improve upon CA parameter extraction and refine task design to allow for maximum group interaction. The variation in tasks within the first round of testing was an issue when it came to collating the results, as all three tasks where very distinct from each other; also not all three of the tasks experienced all three of the delay levels due to the time constraints of the tests. The need for refinement was apparent with emphasis on narrowing down the task selection to just one task which would encompass all the positive elements of the three previous tasks into one. This chapter discusses the findings of extensive pretesting carried out and displays how the lessons learnt from these preliminary tests assisted in the improved test and task design incorporated in this second phase. A crude measurement of around 65,000 data points from the audio gathered were manually reviewed during the second round of subjective tests; this highlights the significant rise in data points to almost double the amount achieved from the initial testing phase (at around 38,000 data points). This confirms that participant interaction increased substantially and further endorses the refined goal-oriented task as having particular value for conversation-based testing to promote a highly interactive environment from which to capture sizable amounts of conversational data.

Within this chapter we also explore the added element of spatial audio in the presence of transmission delay with the end goal to understand further the role that audio quality and spatialisation plays paired with end-to-end delay. Spatialisation is starting to be included in telemeetings and other multi-party audio communication systems, such as online-games. Spatialisation aims to synthetically reproduce the effect of sound being localised; there are many spatialisation techniques which can be used [44]. In this work we are focussing on telemeetings which are likely to involve headphone presentation, therefore the technique chosen was similar to that used by an emerging BT product [45]. Specifically, the spatialisation was created using a head-related transfer function (HRTF) [44] which had been previously determined by previous work at BT and was incorporated into the *PAW* boxes used for the testing. In each listener's station the three other talkers were filtered using these HRTFs to give the audio illusion of one

talker in the middle and the other two approximately 30 degrees each side of the central talker in the horizontal plane.

With this in mind, a particular interest of this study targets if spatial audio can assist with intelligibly and interactivity when transmission delay is present, in addition to comparing with any benefits over lower quality mono audio under the same level of delay. We begin by giving a broad overview of the further refined subjective test design; this highlights the approach taken for the second round of testing. The added feature of the test comprising of differing audio quality will be detailed along with conclusions made from pretesting and information gathered from the initial test design. We draw upon task selection and lessons learnt from the three-task design model to further enhance and streamline the testing process to ensure an accurate and useful data set from which CA parameters of interest can be extracted. In addition to conclusions made from the results from the initial testing phase we discuss the varying levels of delay selected for further investigation to conclude if intermediate levels of delay, more commonly deemed as acceptable, have the same or lesser impact on conversation quality than higher levels which can be associated to conversation impairment or degradation.

We go on to discuss the expanded test design and how improvements to the overall design and format of the testing phase aim to accommodate for better extraction of previously identify CA areas of interest. The key objective being to identify what role transmission delay plays within a teleconferencing environment and how this effects group interaction and communication. In an environment where transmission delay is inevitable and can only be scaled down so far, how well can additional features such as spatial audio help to counteract some of the inherent disadvantages associated with delay. Finally, we conclude with the full and comprehensive set of objective results collated from the second subjective testing phase, where key results will be presented in further detail.

5.2 Spatial Audio Subjective Testing Experimental Design

The test design follows the previous format of the first testing phase as documented in Chapter 3 with all variables remaining the same including test location, number of participants, equipment used and all test pre-conditions; any supplementary features or equipment to the previous test and design are documented here in this chapter. Appendix A6 shows the modified conversational experience survey used for the second phase of testing. The test setup and layout also follow the same format as previously shown in Figure 3.1, with all four participants divided into their own silence cabinets for the duration of the test. The only exception to this comes in the form of an addition of a face-to-face condition, whereby all four participants along with the test coordinator were present in the same silence cabinet in an attempt to replicate a normal face-to-face scenario. The face-to-face condition comprised of the use of a professional grade sound isolation baffle being used as a means of separation between microphones for each

individuals audio stream, in addition to blocking visual cues so as not to give an unfair advantage to the face-to-face condition over the over conditions experienced by the users.

With further expansion on the previous test design, a face-to-face condition was incorporated into the design as a means of benchmarking the group and participants overall performance compared to other mediated forms of communication. Elements such as group performance measures, consisting of conversational effectiveness, levels of interaction and ability of complete tasks over varying mediated and unmediated forms were considered. The concept being that with no synthesised audio stream, the participant's communication and ability to carry out the tasks involved in the test should not be disadvantaged by any form of mediated channel and should allow for as natural as possible conversation flow. Although with the use of a baffle no visual cues were permitted, this condition was implemented to replicate as close as possible to everyday normal face-to-face communication from which the following tests using mediated channels of communication could be compared to. In addition, an orientation round was added to the test design, as previously mentioned in Chapter 3, as a supplementary measure to help with counteracting any leaning effects that could possibly be carried over and effect the results. The orientation round consisted of a shortened version of the main task and the intended purpose was to familiarise the participants with the test set up, whilst exposing them unknowingly to each condition prior to data capture. This would also ensure adequate time was allowed for any learning effects to take place and their behaviours to potentially adapt. This ensured that any possible improvements or alterations in their approach are not associated with the conditions under test, thereby giving unfair advantages dependent upon the order in which the conditions are presented to them. It should be noted that data from the orientation round was not processed or included in the results; the data is disregarded and purely for the purposes of familiarisation and the mitigation of learning effects.

In addition to the introduction of an unmediated channel of communication and an orientation round, refinement of task selection was prioritised in order to target some key CA parameters and behaviours identified from previous testing phase and results gathered, which are discussed in Chapter 4. The refinement of the tasks was key as the previous three-task design model attempted to replicate meeting styles, which had the disadvantage of introducing too many variables which made classification of results difficult. Preliminary trials were again carried out to allow for modifications to tasks to be comprehensively tested before deciding upon the best task or tasks to take forward. The findings from pretesting are presented here to demonstrate the process by which the refined task was chosen.

Pretesting initially focused on the test setup at BT Adastral park, with adjustments being made to accommodate for the additional condition to the testing phase, being the face-to-face condition. Trials were undertaken with and without the use of a sound isolation baffle when participants where in the

same silence cabinet; this tested if the baffle made a suitable difference with microphone separation when capturing each individual's audio stream and also if it suitably blocked any visual cues. The trials were conducted using both the informal and goal-oriented tasks and it was identified that the sound isolation baffle did assist with the separation of audio streams, whilst also adequately blocking visual cues between participants. With the test facility adequately adjusted for the addition of a face-to-face condition and the focus on task refinement, pretesting findings shall be presented here.

The goal-oriented task showed potential to fulfil the desired criteria of the testing phase by generating enough conversation and interaction to provide worthwhile CA analysis. In addition to this, the map navigation was found to actively encourages all group members to participate and interact, while a free conversation task such as the informal meeting task would allow for some members to remain inactive throughout the task and provide very little input to the conversation. The use of maps promoted participant engagement whilst providing a fun and different exercise each time to ensure participants did not lose interest mid-way through a test. This style of task did also present with similar features to that of the informal meeting task and therefore is thought to have the potential to encourage enough interaction as would be found in a freer flowing style of task, whilst having the added benefit of engaging all group members to provide a true multi-party situation. The initial subjective tests however highlighted that the task did require modification to prompt more frequent participant interaction and conversation between all group members. Careful consideration was needed to be given to the design of the maps used for the goal-oriented task with some alternative iterations of the map designs that were pretested being detailed here, along with modifications to the existing goal-oriented task also being discussed.

Map Types (Edraw – 3D directional map or Google API – akin to standard 2D grid formation map)

- 3D tilted directional style map seemed to have the benefit of allowing the navigation of the map to be less obvious; simple left, right, up and downs were eliminated purely based on the map design. This created an element of difficultly which required more interaction and conversation between participants in order to successfully navigate the map. A drawback of this particular design was that design and production process took longer as each map is "hand drawn" using Edraw software.
- Google API standard 2D style maps allowed for much quicker map production as pre-existing maps were utilised, with little work being required to remove existing labels and landmarks. However, map selection took some time as appropriate maps needed to be found to then be built upon. The disadvantages of this style of map were found to be that simple navigational instructions could be given (left, right, the first junction, the T junction etc.) which did not lend themselves well to group interaction or feedback between group members.

Path Length, Difficulty and Map Completion Timing

- Path length and difficulty should be controlled and kept as close as possible between all maps used during the testing phase. The path length was reduced prior to the trials taking place, this may have impacted on the interaction and group communication as the path were shortened to allow for map completion within the 2.5 minute time frame. One option was to increase the length and difficulty of the route so that map completion would not be feasible within the 2.5 minutes. This should prompt further interaction as more feedback from the group would be necessary. This however would need to be adjusted carefully so as not to make the route too difficult to navigate; this would in turn stress the participants and impact of the overall test experience.

  A reduction to a shortened time frame allowance per map was trialled by reducing the time given to 1.5 minutes per map in which to complete the path. The notion behind this was that with the reduction in time allowed, but while still keeping the path length and difficultly the same as previously trialled would make map completion not feasible. This added the need for extra communication due to the length of time given to complete to task, but without lengthening the path therefore making for a more difficult and less stressful task for participants.

Missing Landmark Designation

- Missing landmark designation also played a key role as it had the ability to impact on the group's ability to effectively and efficiently navigate the map. It was found that if the missing landmarks were too clearly illustrated on the non-leading team members maps then the task of identifying missing landmarks could easily be completed with little to no need to confer with the map leader. One issue that was addressed was the way in which missing landmarks were displayed on the maps to prompt further conversation between the map leader and the other group members. A modification was made to the map design whereby missing landmarks were implemented on all versions of the map, including the lead map which had the completed path from start to end. This was found to encourage communication and interaction between all group members as the participant with the lead map would still need to confer with the other team members in order to complete the task. Previous trials without this added feature seemed to prompt single talker conversation and much less interaction which was not the desired outcome and one of the shortcomings of the initial map design used for the first round of subjective tests. The refinement of the missing landmark designation was critical to prevent one sided, single talker conversations with simple responses.

After gathering all the required information from the pretesting, the final task design was selected to be a single task that was a modified version of the original goal-oriented task; this would allow for a streamlined testing platform, as this task fulfilled all the desired criteria to extract key CA features of interest. Also, by use of a single shorter task, this would accommodate all participants experiencing all levels of delay and audio quality without having interfering nuisance variables such as differing tasks to impact on the results.

5.2.1 Refined Goal-Oriented Task

The refined goal-oriented task consisted of all four members focusing together as a team to help each other complete a route around a series of maps. A predefined path was predetermined for each of the maps, but not all participants were aware of the correct and complete path. One participant at a time had what is known as the “master map”. The master map contained the full and complete path, from a start to an end point. The three other members of the group had basic maps with no route shown and no start or end point. Through discussion the aim was for them to help each other navigate the entire map, working as a team from the start to the end point and draw on their own individual copies of the map. All maps had various landmarks removed, including the master map; this was to allow for conversation between all participants to form in order to complete the task and to deter from simple yes/no, turn left/right/straight ahead etc. When a landmark was removed, to make it easier and quicker to identify a dotted line and question mark was inserted in its place to make it clear which parts needed to be enquired about in order to fill in and complete the map. This method also helped with making sure all missing landmarks were captured and also for easier marking of maps for accuracy measures after. The maps were road style maps, an improvement upon the design in Chapter 4, as the pre-tests showed it was easier for participants to follow road outlines rather than freehand drawn paths. The refined map task was also broken up into shorter more manageable maps than previously used in the first round of testing, this meant that each person was provided the chance to direct the other three members of the group around their complete path on a map. The maps varied slightly between the four participants; e.g. some may have landmarks missing from their map, which may be present on others. This increased the complexity of the task. The differences between the maps were designed to cause confusion as the delay within the call increases; this was because participants may have to backtrack and repeat themselves or interrupt each other in order to make sure that they are at the same point of the map as each other throughout. The test was carried out under timed conditions, with each participant given 1.5 minutes to successfully complete each map. An example copy of the refined goal-oriented map task can be found in Appendix A7.

5.3 Transmission Delay Levels and Audio Conditions

An important factor in the second round of subjecting testing was that each task involved in the testing phase utilised each transmission delay level under test; this was previously unattainable in the initial testing stage due to the length of the tasks involved in the three-task design model. Additionally, with the inclusion of a face-to-face round and varying qualities of audio being provided, a new test design had to be implemented to facilitate an accurate testing phase from which a reliable data set could be achieved. The transmission delay values that where decided upon came from observations and results from the first round of tests, combined with an interest in lower levels of delay to see how they would compare to the higher rates previously tested. Also, paired with the knowledge that normal conversation is understood to degrade and show signs of degradation within interactivity after or around the 200ms mark [22], 200ms was identified as a key delay level of interest to monitor both with and without the presence of spatial audio. The values employed during the second subjective testing phase are as follows:

- Low delay (virtually no delay) – 0ms[3]
- Low/Medium delay – 200ms
- Medium delay – 400ms

These values where decided upon to permit for a full test of low, low/medium and medium delay to see how transmission delay at varying levels can affect the four-way group dynamics of a conference call. These values were also of interest when taking into account the use of diverse levels of audio quality as conditions to also be explored during the testing phase. The three audio conditions that were utilised during the second subjective testing phase are as follows:

- Face-to-face
- Mono
- Spatial (as described earlier in Section 5.1)

Given the three levels of delay and three audio conditions under test it was preferable to allow for each group to experience each delay level with each audio condition, with the exception of the face-to-face condition which naturally would be an unmediated condition with no delay. This would also entail each participant having their own individual master map on which to lead with for every audio condition and delay level. This ensured that as many confounding variables as possible were removed from the testing phase and improved upon the previous test designs limitations. A breakdown of the test design is shown

[3] It should be noted that there was a notional system delay of 2ms which should be added to the values above. The 2ms delay value was determined by BT through acoustic click delay from microphone to headphone and measured on an oscilloscope.

below as an example of a typical format of which each of the subjective tests would follow. A full illustration of the testing order used is shown in Appendix A9. To permit a full test of all conditions, 12 groups in total were employed for this experiment; this allowed for every combination of audio condition and delay level to be experienced without having to solely rely on Graeco-Latin square design to build a randomised task block. It should be noted that the second round of testing involved considerably more conditions than the previous early tests; each group was subject to 14 conditions (7 from the orientation round and 7 from the main testing block). This resulted in a substantially larger corpus of data with 12 groups in total, consisting of 84 recordings (168 including the orientation round). Additionally, the orientation round, as can be seen from the example below, was intended to be a quick familiarisation round, so only 1 map was used per condition. The main test block however employed 4 maps per condition to ensure each participant had the role of leader of a map for each combined audio and delay condition.

Block 1 Orientation round: (around 10-15 minutes) – shown without any randomisation.

* As orientation round, only 1 map per condition – quick familiarisation round

Face-to-Face (1 map, 1.5 mins)

Mono 0ms (1 map, 1.5 mins) **Mono 200ms** (1 map, 1.5 mins) **Mono 400ms** (1 map, 1.5 mins)

Spatial 0ms (1 map, 1.5 mins) **Spatial 200ms** (1 map, 1.5 mins) **Spatial 400ms** (1 map, 1.5 mins)

Block 2: Main test block (around 45-50 minutes) – shown with examples of randomised ordering

Face-to-Face (4 maps, 6 mins)

Mono 400ms (4 maps, 6 mins) **Spatial 200ms** (4 maps, 6 mins) **Mono 0ms** (4 maps, 6 mins)

Spatial 0ms (4 maps, 6 mins) **Mono 200ms** (4 maps, 6 mins) **Spatial 400ms** (4 maps, 6 mins)

5.4 Spatial Audio Conversation Analysis Objective Results

As previously discussed in Chapter 4, the MOS scale is traditionally used to rate audio systems, but these type of measurements and subjective feedback data alone are not detailed enough to provide an accurate assessment of the system or the conversation quality. The MOS scale was used again in the second round of subjective tests, but as was found in the initial tests this did little to demonstrate the impact that delay or audio quality was having on the conversations. This section sets out to present precise depictions of objective data gathered from the refined second testing phase to help support and demonstrate actual events which occurred within the recorded conversations; this will assist with a better understanding of

what role differing levels of transmission delay coupled with added diverse audio conditions played. We begin by presenting the captured CA objective results in box plot form to clearly display, in the form of visual representations, the differences across all delay levels and audio conditions utilised during the testing stage. We conclude with a discussion of the CA measures gathered and draw upon the results to conclude on the effects of both delay and spatialisation within the context of teleconferencing systems.

5.4.1 Box Plot Representation of CA Objective Results – Strongly Correlated Results

The full set of results from the induced transmission delay in the presence of spatial audio testing phase, comprising of the entire population of twelve groups, detailing the various conditions undertaken by each group will be documented here. This subsection also approaches the topic of CA and extended CA speech parameters. It discusses the comprehensive CA features that were continually monitored, as within the initial tests, with a focus on some previous explored speech parameters with an interest to how they have performed under different delay levels and categories of audio.

The box plots shown in Figure 5.1, illustrates the effects of induced transmission delay and varying audio conditions on the conversation parameter no-gap-no-overlap at the varying delay levels of 0ms, 200ms and 400ms, along with the audio conditions consisting of face-to-face, mono and spatial. The series of box plots shown in the figure have subdivided the results into categories grouped by all delay levels with purely mono audio, all delay levels with purely spatial audio, all audio conditions including all delay levels and all audio conditions with 400ms delay only. The box plots in Figure 5.1 clearly display both the impact of delay on no-gap-no-overlaps, but also the benefit that spatial audio brings to this speech parameter, even in the presence of the highest delay value of 400ms. The spatial audio condition also shows improvements over standard mono audio at lower delay values of 200ms and 0ms.

The figure also highlights the correlation between the face-to-face condition and the spatial condition, with the spatial condition performing with not too dissimilar values as face-to-face with the range of data for the number of events per minute. The results present evidence that spatial audio is assisting with smoother talker transitions and alternation than compared to mono audio; this is more akin to a normal face-to-face condition thus demonstrating the benefits of spatial audio in the presence of end-to-end transmission delay.

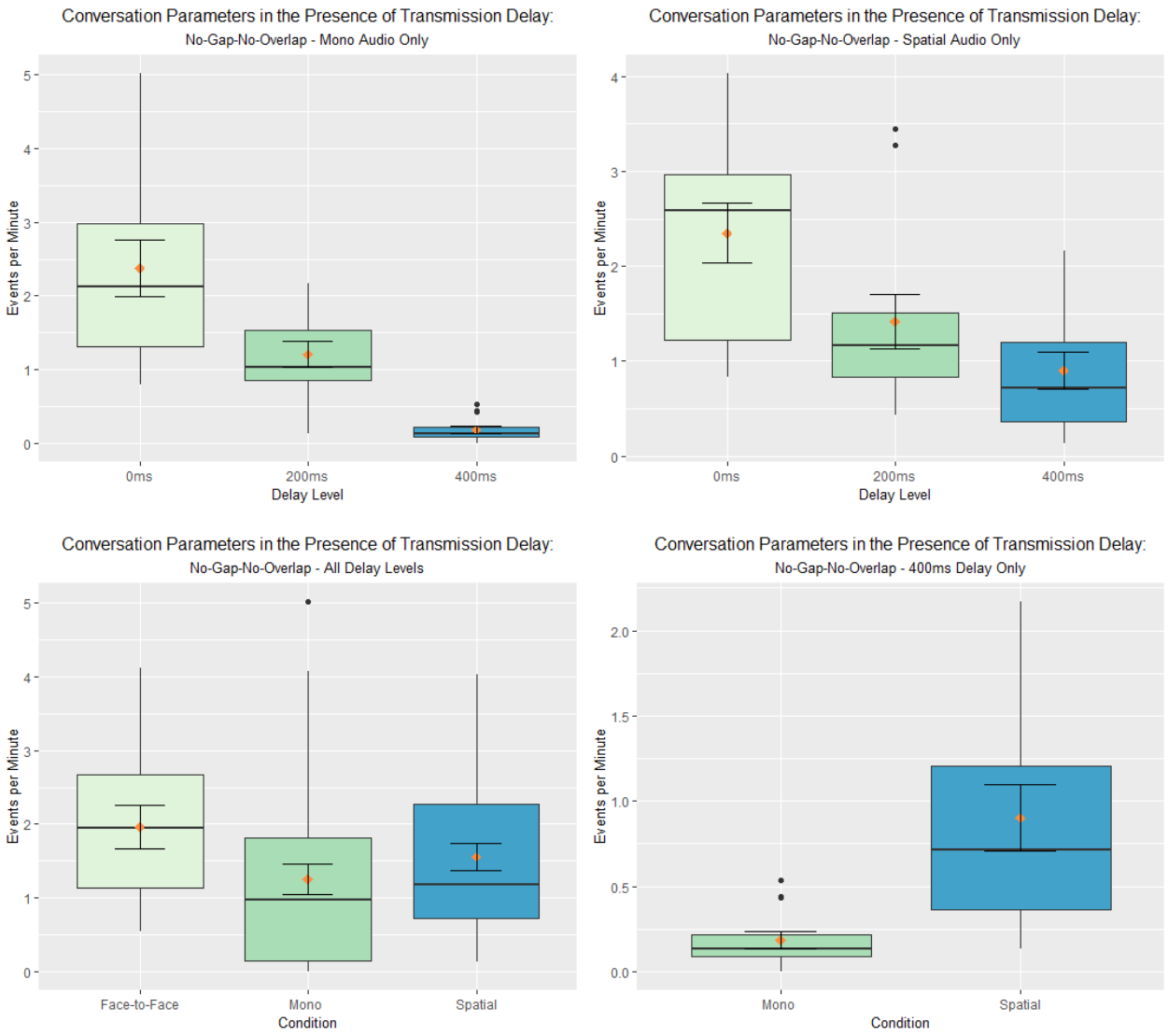


Figure 5.1: No-Gap-No-Overlap Box Plots subdivided by Delay Levels and Audio Conditions

In Figure 5.2 the box plots presents the effects of induced transmission delay and varying audio conditions on the conversation attribute false starts at the varying delay levels of 0ms, 200ms and 400ms, along with the audio conditions consisting of face-to-face, mono and spatial. The series of box plots shown in the figure have subdivided the results into categories grouped by all delay levels with purely mono audio, all delay levels with purely spatial audio, all audio conditions including all delay levels and all audio conditions with 400ms delay only. Figure 5.2 shows a strong correlation between the rise in the number of events of false starts with the introduction of 400ms end-to-end delay in conjunction with the use of lower quality audio in the mono condition. Interestingly, the low/medium delay level of 200ms displays considerably less sensitivity to false starts than the higher value tested of 400ms; in addition, this is true even for the lower grade mono audio condition. However, as transmission delay is increased to 400ms

the spatial condition shows a marked advantage over mono audio, clearly highlighting the benefits of spatially separating users of a teleconferencing system in the presence of delay.

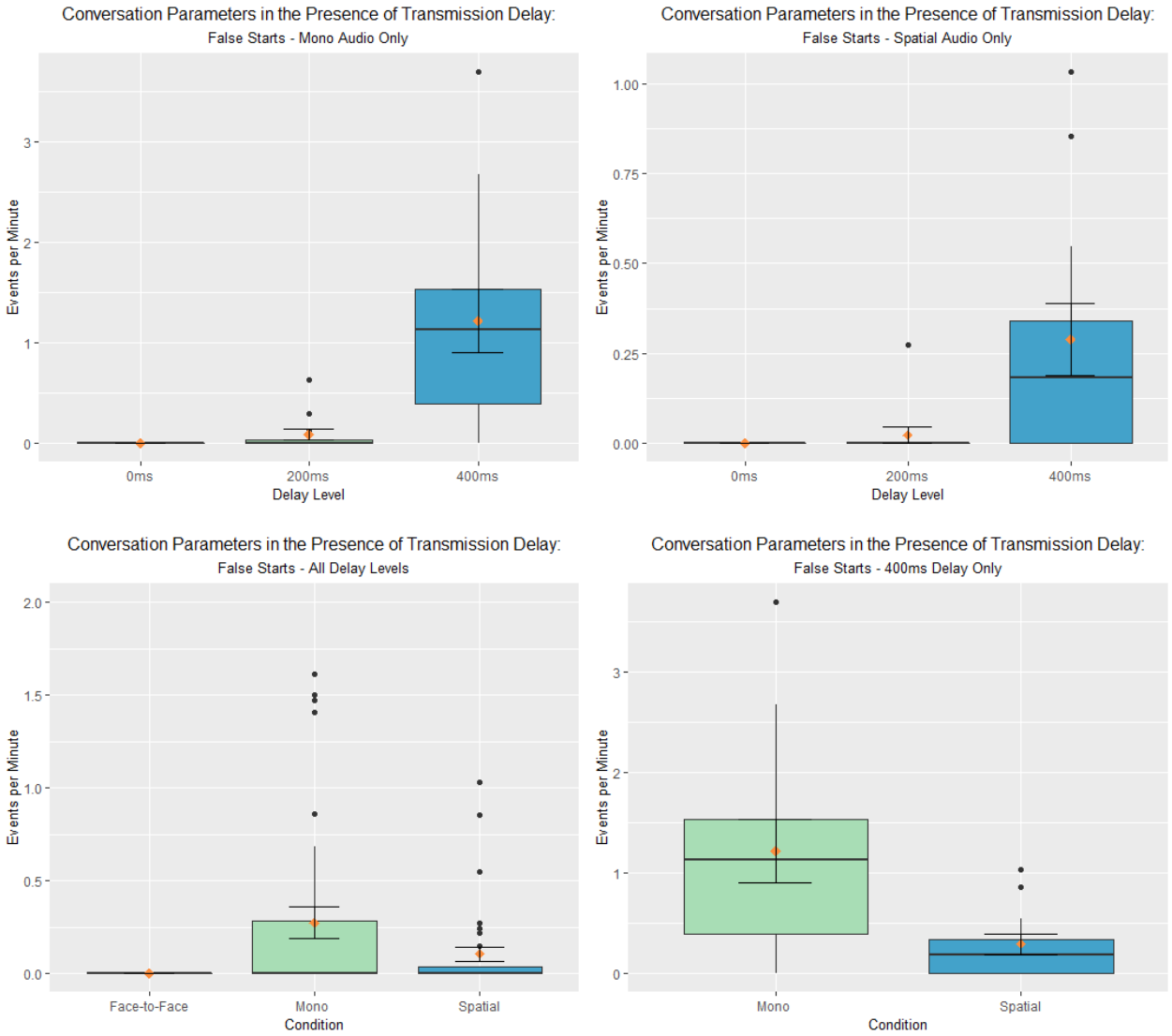


Figure 5.2: False Starts Box Plots subdivided by Delay Levels and Audio Conditions

The box plots shown in Figure 5.3 displays the effects of induced transmission delay and varying audio conditions on the speech parameter successful interruptions at the varying delay levels of 0ms, 200ms and 400ms, along with the audio conditions consisting of face-to-face, mono and spatial. The series of box plots shown in the figure have subdivided the results into categories grouped by all delay levels with purely mono audio, all delay levels with purely spatial audio, all audio conditions including all delay levels and all audio conditions with 400ms delay only. The stark difference between the behaviours seen with the mono and spatial audio conditions can easily be interpreted from the gradual decline in the number of successful interruptions as the delay increases in the mono audio only plot; spatial on the other

hand manages to maintain a roughly even range of successful interruptions across each of the delay values. Also, at 400ms of end-to-end delay, spatial demonstrates a better performance, in terms of a higher number of events per minute than mono. The face-to-face scenario naturally shows the optimum level of successful interruptions being achieved during the tasks; although spatial audio does not match this it is consistent in showing an improvement over the mono audio condition. This result again confirms the additional benefits of utilising spatial audio in a teleconferencing type setting; it demonstrates that by providing spatial separation it provides the users with the environment in which to successfully interrupt as and when necessary.

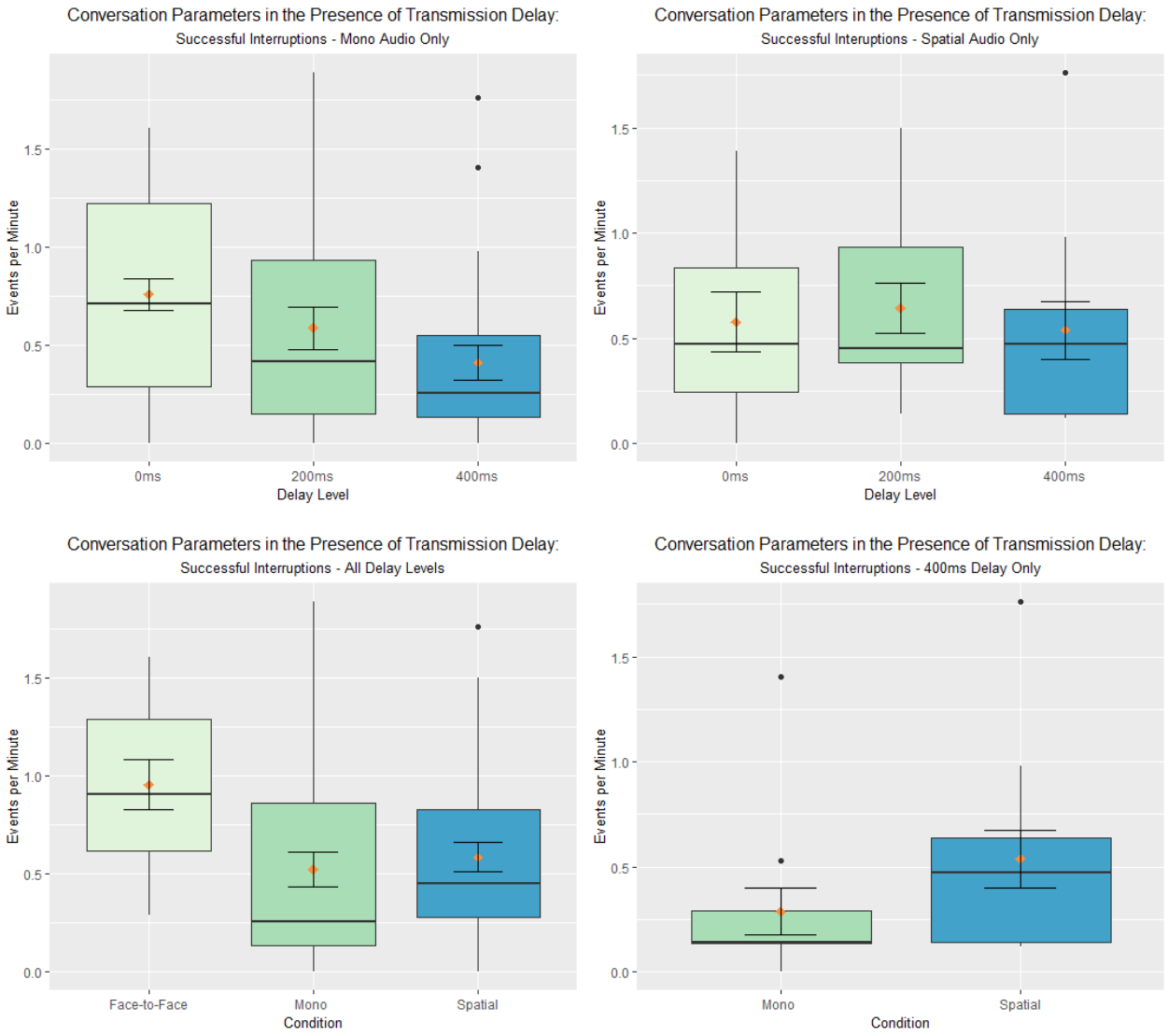


Figure 5.3: Successful Interruptions Box Plots subdivided by Delay Levels and Audio Conditions

Another CA attribute monitored was the extended CA parameter known as detrimental overlap. The box plots in Figure 5.4 visually represent the effects of induced transmission delay and varying audio conditions at delay levels of 0ms, 200ms and 400ms, along with the audio conditions consisting of face-

to-face, mono and spatial. The series of box plots shown in the figure have subdivided the results into categories grouped by all delay levels with purely mono audio, all delay levels with purely spatial audio, all audio conditions including all delay levels and all audio conditions with 400ms delay only. The number of incidences of detrimental overlap is shown to rise within the mono only audio conditions when compared with spatial audio. The number of events for this speech parameter at both 200ms and 400ms of transmission delay are shown to be lower when spatial audio was in use, which again highlights one of the benefit of spatial audio to help with offsetting negative conversation attributes in an inherent delay prone environment such as teleconferencing. In addition, the spatial condition can be seen to have improved performance over mono audio at the higher value of 400ms, which gives evidence that even in higher delay scenarios spatial separation of individuals helps to counteract the associated negative overlaps that can potentially disrupt conversations, task proficiency and group interactions.

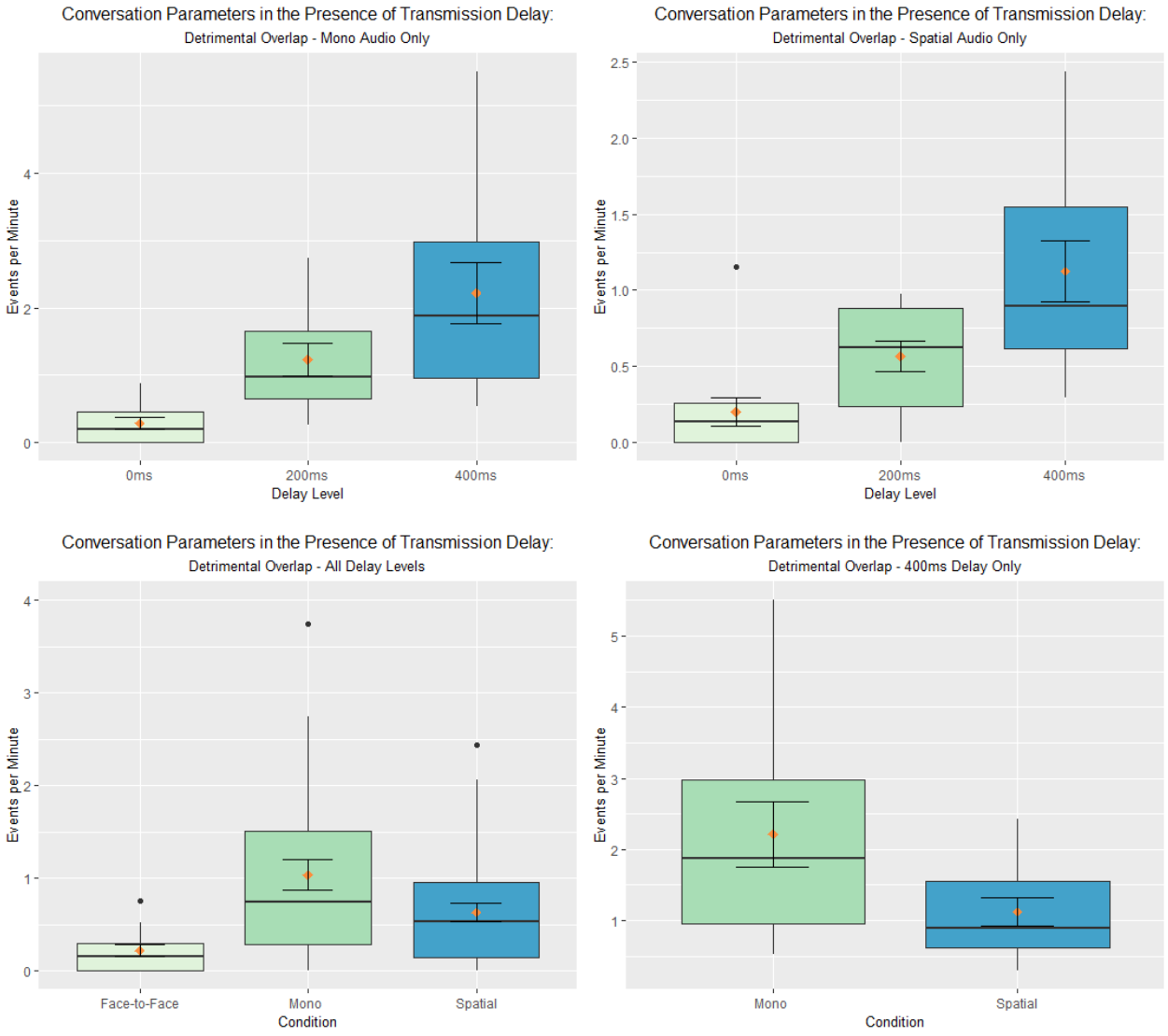


Figure 5.4: Detrimental Overlap Box Plots subdivided by Delay Levels and Audio Conditions

An additional CA parameter of interest was the extended CA term of constructive overlap, the results from which can be seen in Figure 5.5. The box plots visually embody the effects of induced transmission delay and varying audio conditions at delay levels of 0ms, 200ms and 400ms, along with the audio conditions consisting of face-to-face, mono and spatial. The series of box plots shown in the figure have subdivided the results into categories grouped by all delay levels with purely mono audio, all delay levels with purely spatial audio, all audio conditions including all delay levels and all audio conditions with 400ms delay only.

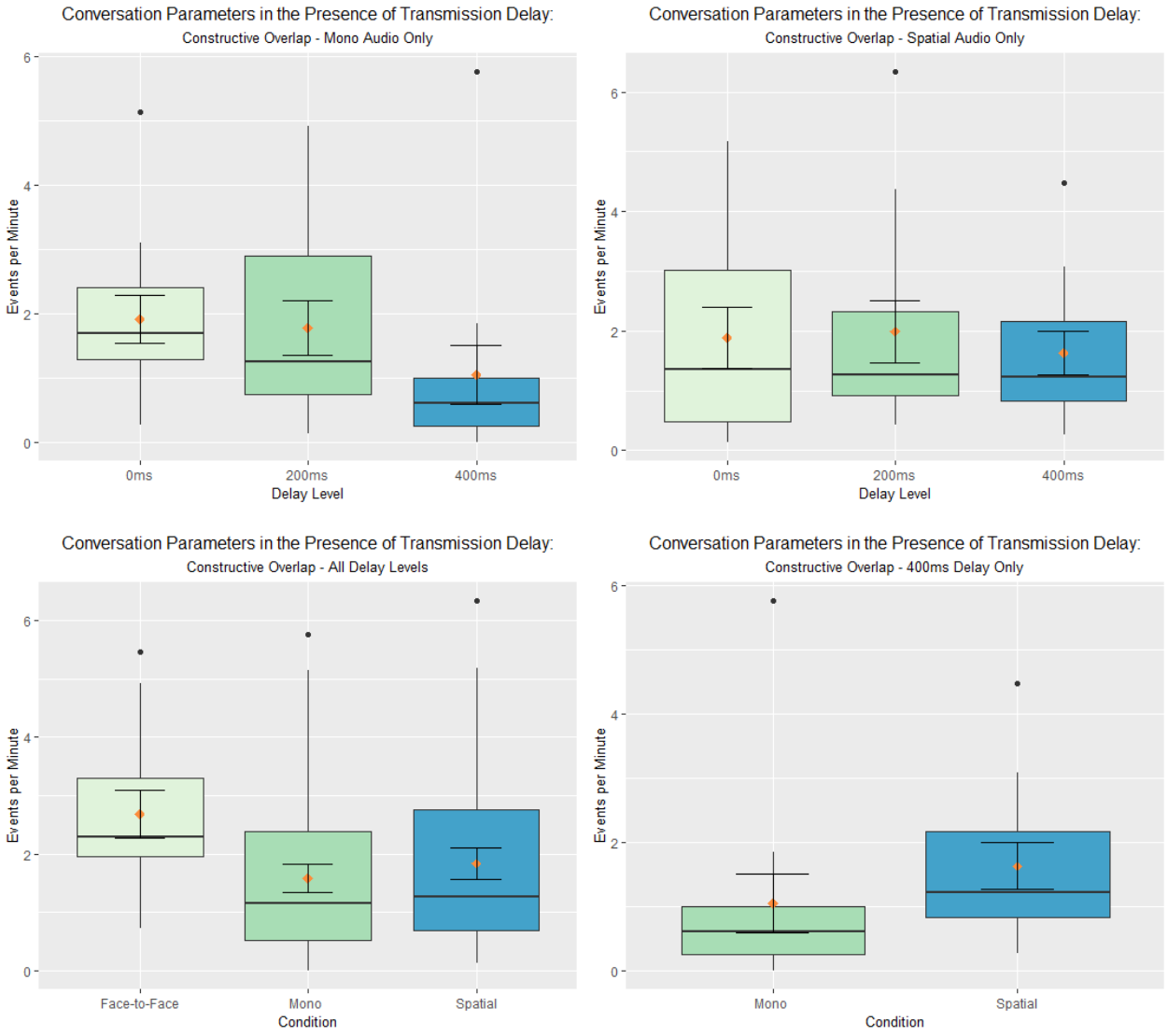


Figure 5.5: Constructive Overlap Box Plots subdivided by Delay Levels and Audio Conditions

As opposed to the previous results for detrimental overlap in Figure 5.4, constructive overlap demonstrates the reduction in the number of events as the transmission delay increases with mono audio, however as the number of events may be expected to reduce with delay, spatial audio has a positive effect on their production even at the highest delay of 400ms. A delay of 400ms demonstrates a more consistent

result similar to what spatial achieves at 200ms. Also, the spatial condition again demonstrates improved performance over mono audio at the higher value of 400ms with results close to double the rate at which constructive overlaps are achieved under mono only conditions. This highlights how positive speech parameters can be provided with a suitable environment in which to occur, even in the presence of delay, when paired with higher quality audio systems.

In Figure 5.6 the box plots showcase the effects of induced transmission delay and varying audio conditions on the conversation attribute continuers at the varying delay levels of 0ms, 200ms and 400ms, along with the audio conditions consisting of face-to-face, mono and spatial. The series of box plots shown in the figure have subdivided the results into categories grouped by all delay levels with purely mono audio, all delay levels with purely spatial audio, all audio conditions including all delay levels and all audio conditions with 400ms delay only.

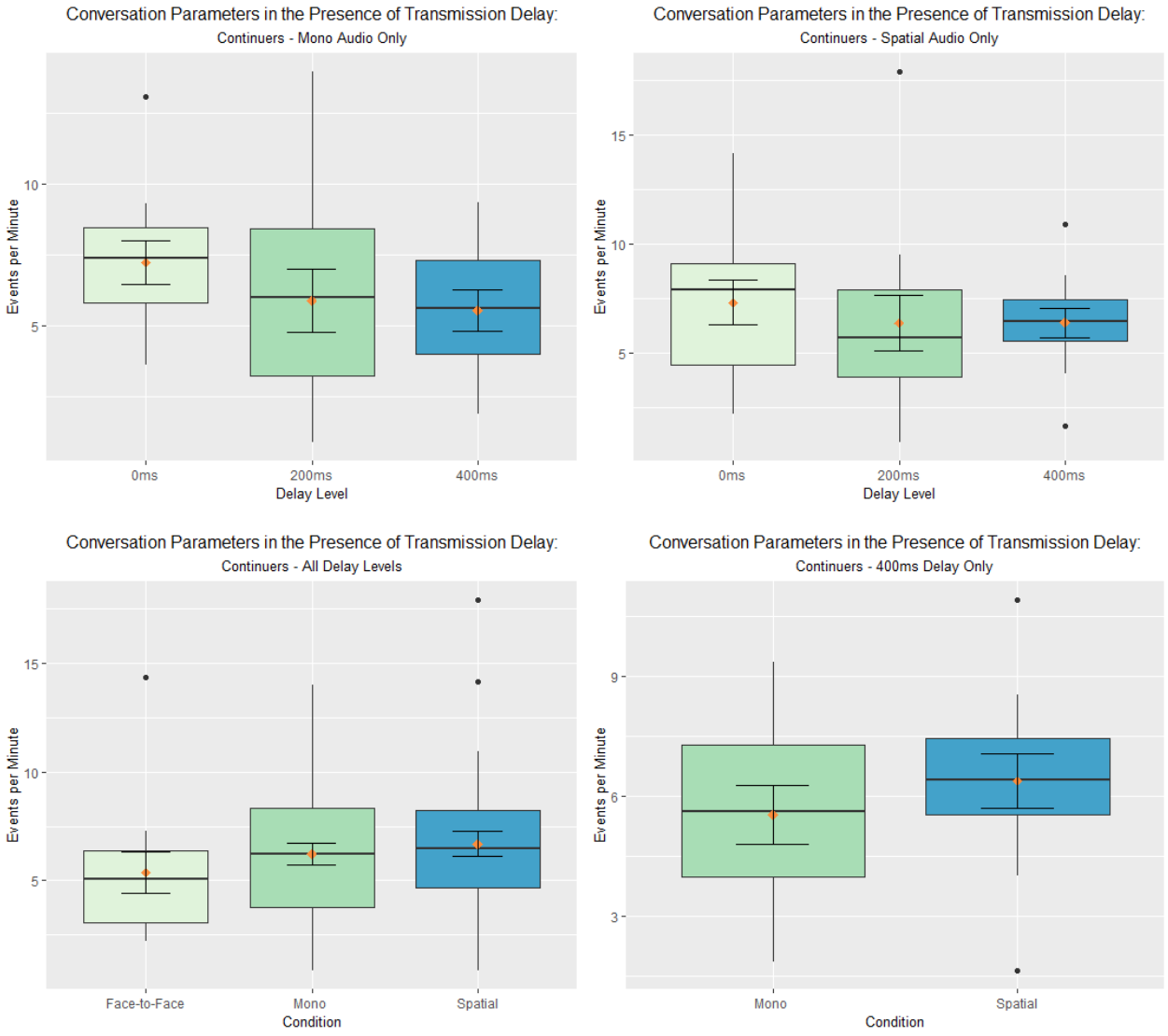


Figure 5.6: Continuers Box Plots subdivided by Delay Levels and Audio Conditions

Continuers were a key speech parameter of interest as a focus of this study hypothesised on the production of continuers and how they are affected in delay induced scenarios, as detailed further in Chapter 6. There is a slight reduction shown in the number of continuers produced within increasing delay, again with spatial providing some relief giving a small increase in their production at 400ms when compared with that of mono at the same delay value. Continuer production was thought to naturally decrease with the increase in delay, originally hypothesised and discussed in Chapter 6 and on reflection from the results gathered from both testing phases, a reduction in their creation may not necessarily be the case, but rather a reduction in the volume at which they are produced by participants. It should be noted that from the results perhaps more frequent continuer occurrences are linked to the use of mediated forms of communication, as in the face-to-face condition we see less activity. This would seem feasible given that continuers are signs of acknowledgement to the talker that their presence would be of more valued in a teleconferencing setting.

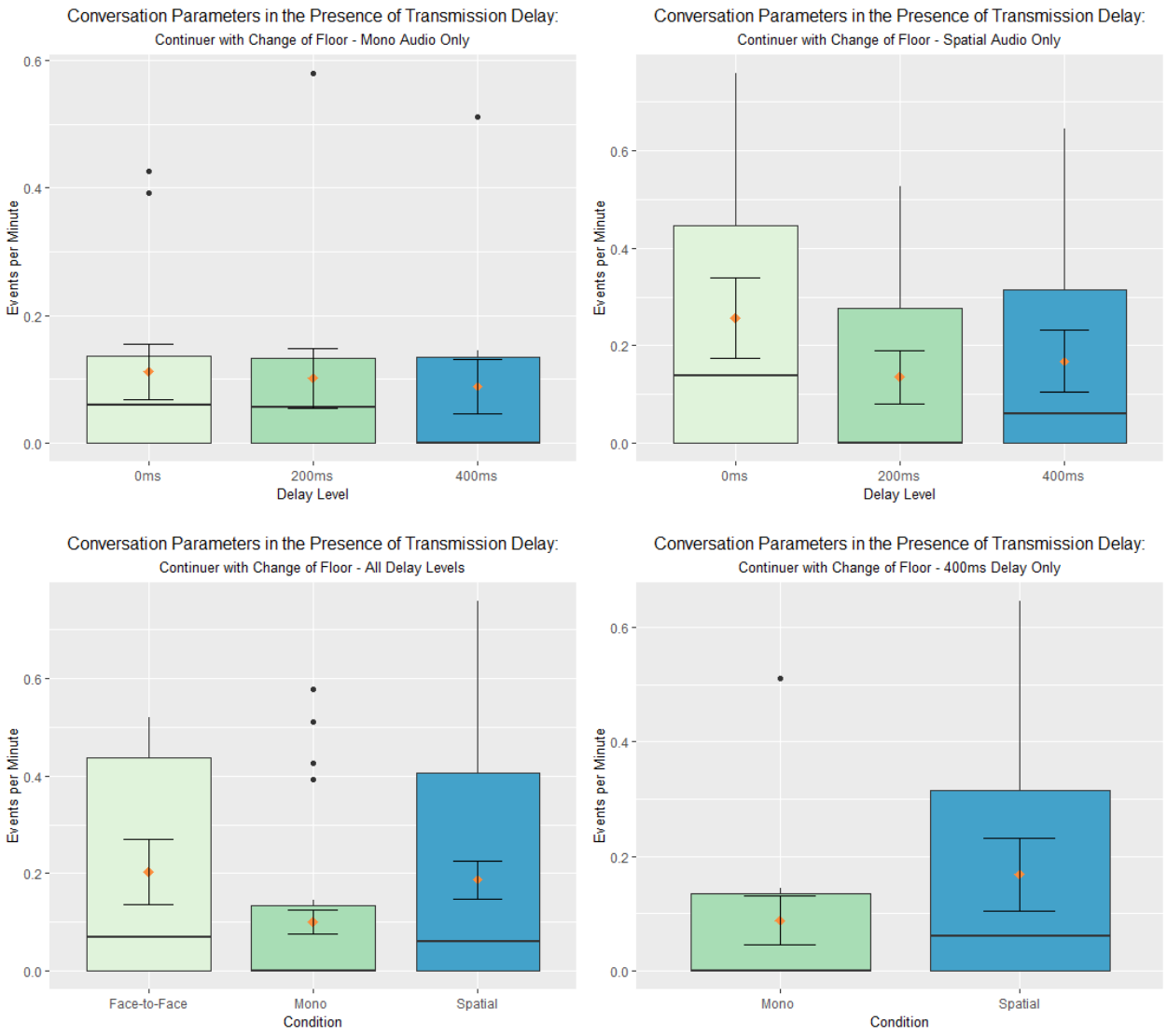


Figure 5.7: Continuers with Change of Floor Box Plots subdivided by Delay Levels and Audio Conditions

Another CA parameter monitored was the extended CA parameter defined as continuers with change of floor. The box plots in Figure 5.7 illustrates the effects of induced transmission delay and varying audio conditions at delay levels of 0ms, 200ms and 400ms, along with the audio conditions consisting of face-to-face, mono and spatial. The series of box plots shown in the figure have subdivided the results into categories grouped by all delay levels with purely mono audio, all delay levels with purely spatial audio, all audio conditions including all delay levels and all audio conditions with 400ms delay only. Intriguingly, continuers with change of floor for the spatial audio condition shows improvements across all levels of delay induction, including 0ms when compared with the mono equivalent. The spatial audio condition also performs in a more consistent manner as is seen with the face-to-face condition, along with again demonstrating better performance than mono audio.

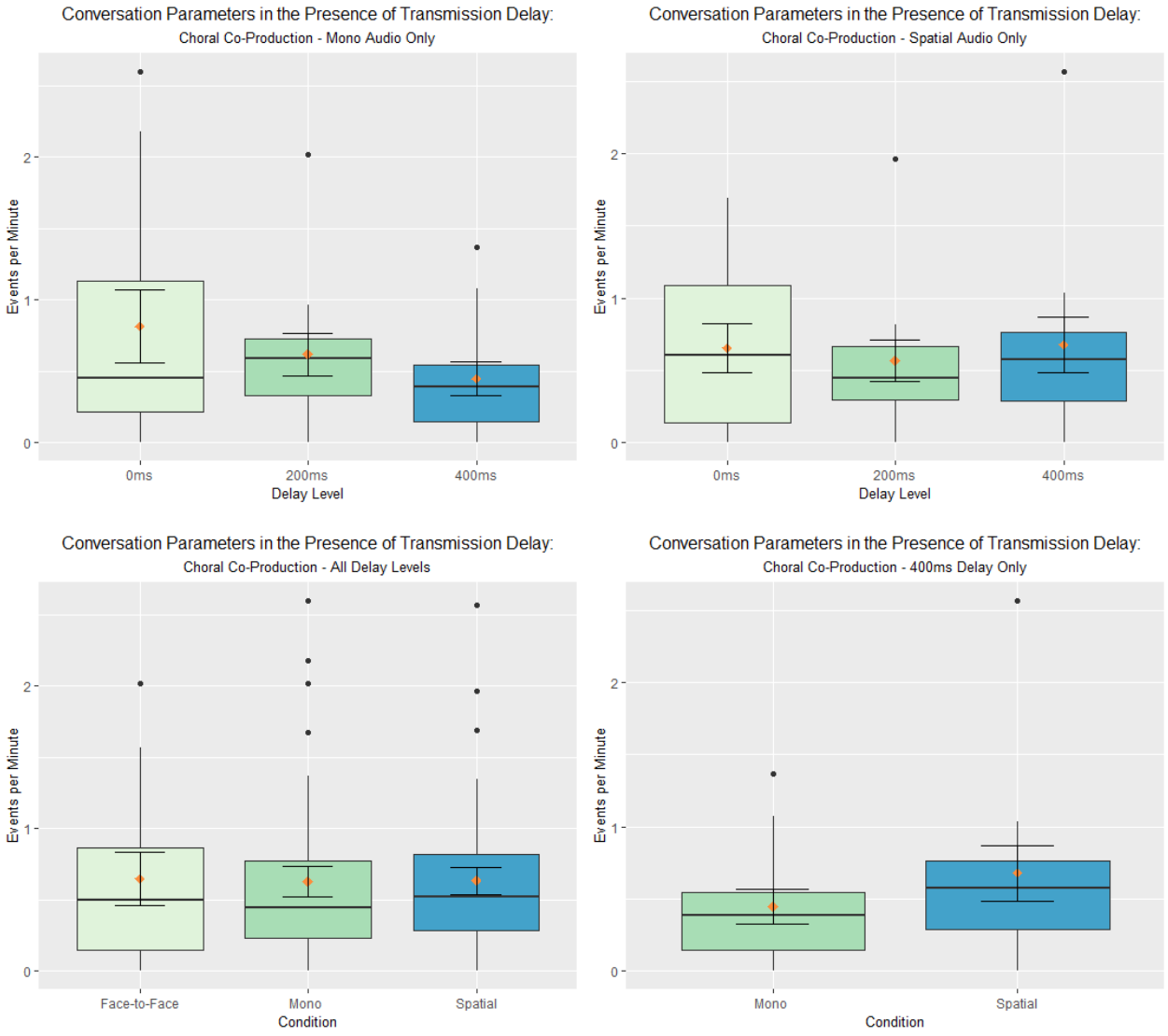


Figure 5.8: Choral Co-Production Box Plots subdivided by Delay Levels and Audio Conditions

The box plots in Figure 5.8 showcase the effects of induced transmission delay and varying audio conditions on the conversation parameter choral co-production at the varying delay levels of 0ms, 200ms and 400ms, along with the audio conditions consisting of face-to-face, mono and spatial. The series of box plots shown in the figure have subdivided the results into categories grouped by all delay levels with purely mono audio, all delay levels with purely spatial audio, all audio conditions including all delay levels and all audio conditions with 400ms delay only.

The results for choral co-production show more subtle effects of spatial audio over mono, with a slight increase in the number of events present for the spatial audio at 400ms of delay in comparison with mono audio at the same delay level. Also, spatial audio during the testing phase performed not too dissimilar to the face-to-face condition; again, this is subtle as mono's performance is close to both in comparison, so any benefits gained by having a spatial effect for this speech parameter are only slight.

The poorly correlated results from the second phase of subjective tests are presented here in list form:

- Anticipatory Completion
- Laughter

5.5 Summary

In summary, this chapter explores the design and results gathered from the second subjective testing phase of this body of work. The test and task design refinements accommodated for an improved version of the initial subjective test; with the information gathered from the results of the first round of test this provided the background knowledge needed to progress onto further testing. Preliminary pretesting were carefully employed before the final test design were completed, which assisted with the final task selection of a refined goal-oriented task. The focus on pretesting was vital to eradicate any issues from previous tasks and it gave the opportunity to attempt to pinpoint, through task selection, some key CA features of interest. Another element that was considered in the second round of tests was the introduction of different quality audio conditions in conjunction with varying levels of transmission delay. A main focus was to determine the effects of spatial audio in the presence of transmission delay and indeed if spatial separation of participants was of any benefit in these delay inherent environments, when compared with the lower quality of audio that the mono condition provided. Also, the delay values where adjusted to lower levels to detect if the sensitivity to certain CA parameters had such an impact at these lower values at around 200ms.

The results obtained from the second subjective testing phase gave very clear and consistent outcomes, which indeed highlighted the benefit of spatial audio in teleconferencing, even in the presence of higher delay situations. Some existing and newly extended CA measurements gave significant

information on the interactions and behaviours displayed by participants in the conversation-based tests and how each of these CA features responded to the introduction of transmission delay and differing levels of audio quality. The introduction of a face-to-face condition in this round of testing also gave additional information as to how well the other audio conditions performed in comparison to as realistic as possible "natural" condition. The results highlighted and confirmed many of the preconceived issues associated with delay in teleconferencing by using CA methods that enabled the extraction of key conversation attributes in order to give an overall objective measure and metric of each conversation feature. The results captured from some of the key attributes of interest, such as false starts and no-gap-no-overlap, are clearly in line with what would be expected when subject to delay impairments and they display the degree of impact that higher levels of delay had on multi-party interactions. The results also prompted the need to understand further how a better referencing system for rating conversation-based tests could potentially be developed, with automatic detection of some the CA parameters possible in near future work, especially for the conversation parameters which should be detectable in a non-contextual sense.

Chapter 6

6. Discussion and Future Work

This chapter will aim to look at comparing the first and second round of objective results gathered, as discussed previously in Chapters 4 and 5. The objective CA data held from both testing phases shows a reassuring consistency across results gathered from the two very different tests; this allows conclusions to be drawn on the impact of induced transmission delay on multi-party interaction in teleconferencing scenarios. In addition to transmission delay, audio quality was under test to explore any potential benefits of spatial audio over standard mono audio in a teleconferencing context. The consistency of results across the two testing phases shows the merit of CA metrics being utilised to measure conversation quality, in conjunction with further understanding group conversation over mediated communication channels with impairments and degradations. In addition, the number of data points manually reviewed for the first initial phase of subjective tests (at around 38,000) shows how the refinement and expansion of the work for the second phase achieved a more reliable test and task design which prompted even greater participant interaction with data points almost doubling (at around 65,000 data points). This was a key finding as it demonstrates more interaction between participants, therefore allowing us to conclude that the refined goal-oriented task was a better designed task to promote conversation.

This chapter, in addition, aims to discuss the role of CA and the value of the objective measures gained from its deployment, along with the expected impact of the findings of this research by highlighting future work areas; for example, the potential for machine learning exists now a data set has been obtained and this shall be explored along with ideas for a new measurement system for multi-party teleconferencing. Also explored in this chapter are various hypothesis developed from experience from the testing phases; future work could investigate further these notions to aid with even better identification of good or poor quality telemeetings. The chapter concludes with recommendations going forward for CA based referencing systems and how automatic detection of CA measures can be implemented.

The CA speech parameters clearly pinpoint the many features of a conversation that are impacted upon by the introduction of transmission delay and also showcased the effects of spatial separation and its ability to counteract some detrimental parameters at higher levels of delay. These results are valuable as they demonstrate how improvements in audio quality can help in an environment with an inherent susceptibility to transmission delay. The benefits of these findings are substantial as end-to-end transmission delay will always be a factor, with reductions only feasible to a certain extent given the

systems used, network considerations and also the distance involved in the teleconferences. In addition, by its very nature the number of participants active in a multi-party conference can stress a system. This is especially true as the number of participants increases causing further delay, paired with the apparent need for an environment where all the intricate nuances of group conversation can be accommodated delay will always need to be a consideration. The use of CA techniques to extract previously unknown and unmeasured factors from multi-party mediated conversations has the potential for future work to expand upon these findings and improve upon the way the systems are rated. The MOS scale traditionally used has obvious flaws when used to evaluate a high-quality system and also systems where deliberate impairments and degradations are employed to test the systems capabilities and how well it performs and can withstand these elements.

6.1 Discussion of CA and Objective Measures

The multidisciplinary approach taken by this study combining CA with engineering metrics and statistical analysis provided quantifiable metrics by which to evaluate human communication in the presence of transmission delay over mediated channels; this methodology was unlike any previous works that have been explored. The synthesis of the two distinct fields provided many advantages over current methods that are widely utilised, such as the E-Model [17] and/or the MOS scale [27]. Current standards used to determine transmission and audio quality focus on subjective data and the perceived or approximated quality from the end user. The novel methodology outlined in this thesis extended upon this further to incorporate objective measures to deliver accurate representations of the quality of communication experienced with statistical findings.

The benefits of using CA as the theoretical approach to this work are clear from the results from both testing phases, revealing previously undefined objective CA measures from teleconferences which traditionally would exclusively rely on participants perceptions of quality or estimated effects on conversational speech. A theoretical CA approach was chosen as a viable method to analyse human communication over telemeetings in the presence of transmission delay as CA itself analyses the organisation and various components of turn taking in the usual face-to-face environment. CA theory provided an in-depth understanding into normal human communication skills and interactions, thus providing a benchmark of common features associated with normal speech. CA therefore supported the definition of good and poor-quality teleconferencing communication based upon traditional CA measures on the construction of conversation. Through the use of CA, new insights were unveiled into a communication phenomena found with the introduction of transmission delay, commonly associated with teleconferencing systems widely in use today. The phenomena prompted the development of extended CA parameters beyond the scope of official CA measures. Examples of behavioural changes

compounded by the introduction of delay include false starts and detrimental overlap, which previously have been undiscovered or documented in an objective fashion as within this study. A focus of this thesis was to interpret the role of transmission delay in a teleconferencing environment; CA provided methodologies to understand the construction of normal face-to-face conversation and common conventions associated with group conversations, therefore helping in ascertaining the impact of delay on mediated communication. Induced transmission delay revealed a phenomena of conversational and behavioural changes which became apparent when comparing with CA theories surrounding normal conversation structure, such as the “one-at-a-time” and “no-gap-no-overlap” concepts [1]. Delay did provide some influence behind the selection of CA as the theoretical basis of this study; in order to quantify the impact that delay had on mediated conversation firstly a greater understanding of face-to-face conversation was needed to fully assess the differences between the two modalities.

The theoretical CA approach to this work created new objective measurables to ascertain the impact of delay on telemeetings with the expansion of official CA parameters, as discussed in Chapter 4. Issues did arise when attempting to combine previously explored conversation attributes with novel interpretations of behavioural changes. Some widely accepted CA parameters, such as continuers, were hypothesised to have a greater impact on the conversations than were documented from the results of both testing phases. Initial hypothesis anticipated there to be far fewer occurrence of continuers in the presence of transmission delay and that the frequency of continuers could be an indicator of good quality communication due to their regular occurrence in face-to-face conversations. This remains a partly unresolved hypothesis from the current set of results held, but with further expansion (as discussed later in Section 6.2.2) continuer volume may hold the key to continuer production within synthesised environments. Without the fusion of both areas in an interdisciplinary manner the data gained from the subjective testing stages would not have been as broadly understood as it is now. The research community can take encouragement in the knowledge that the combination of two distinct traditions helped to lead the way to greater understanding of communications technology and how we as individuals interact differently over them. This can assist in further developing how we use communication systems, which in turn can impact on how we design them to suit our needs and provide greater usability. This thesis touches on just part of what can be discovered with close cooperation between the social sciences and computer and engineering sciences.

A degree of integration between both CA and purely technology based objective measures was essential, neither alone could satisfactorily provide a methodology suitable to gain an understanding into the effects of transmission delay over audio only teleconferencing. A balance between the two fields was necessary as not all aspects of CA can currently provide useful data or automatically detectable

parameters without an element of context being involved. Contextual based CA parameters are currently beyond the automatic recognition capabilities of technology today; future work could exploit these findings when the ability for a machine to understand fully what is being said, not just when something is being said is possible. However, some solutions to minimise the effects and impact of delay on mediated conversations have been discovered using some non-contextual based CA parameters which rely on essentially pattern matching to identify interactions defined by certain characteristics. An example of this is the patent derived from this work to simulate no-gap-no-overlap situations by pre-empting gaps caused by delays and then inserting audio snippets into a listeners system ahead of a talkers stream so that negative situations, such as false starts, are less likely to occur. Also, the use of spatial audio has been shown to provide better conditions in which to communicate in the presence of delay due to the spatial separation of participants. In situations where overlap may occur due to delay, spatial separation helps listeners to better classify who is talking and hear multiple talkers at the same time.

It should be noted that this work intentionally did not incorporate the use of video streams over the controlled lab setting conferencing system and focused solely on audio only. Video conferencing was considered; however, the use of video would have introduced additional variables to account for such as time alignment between the audio and visual streams, the quality of video vs audio and the impact of delay on both streams. Also, in addition to a host of other elements to define and measure such as the use of visual cues in place of/alongside audible cues. The use of video was deemed out of the scope of this study without achieving the findings now held for audio only conferencing; future work may wish to take these findings into account when assessing video conferencing systems.

6.2 Future Work – Expected Impact of Research

6.2.1 Exploitation of the Work

The work of this thesis resulted in a number of discussions on how the work can directly impact operational systems in use by companies such as BT. The author of this thesis was the primary inventor on the patent filed by BT (in co-inventorship with both industrial and academic supervisors) [5] titled “Managing Streamed Audio Communication Sessions.” This patent arose from the observation in the work of the no-gap-no-overlap and false start CA parameters. The invention is designed to simulate no-gap-no-overlap situations by pre-empting gaps caused by delays in centralised telemeeting bridges and inserting audio snippets into a listeners system ahead of a talkers stream so that false starts are less likely. At this stage the invention is only a concept, developing it into a working system, and proving that it works will require future work.

Further exploitation of the work would be to incorporate the findings into a real life setting to provide some useful benefits to currently available teleconferencing systems, given some of the knowledge gained from the CA findings. Additional work is exploring how the corpus of CA marked audio can be used to train a machine learning system that will detect certain CA parameters (such as no-gap-no-overlap and continuers) so that the quality of a telemeeting can be determined in real-time. This could be used to automatically insert interventions such as moving conference bridges to locations that reduce delay or turning on the no-gap-no-overlap feature of the previously described invention. A networks manager could be provided with real time stats (or for assessment after a call has ended) relating to the quality being experienced on a teleconference; this would allow for accurate monitoring of conferencing quality and facilitate any changes to improve the overall experience. This work is in the early stages of invention drafting for a patent filing with BT.

6.2.2 Conversation Analysis Hypothesis and Machine Learning

Throughout the duration of this study the lack of readily available tools to reliably measure conversation quality and interactions prompted the need for many manual processes in order to examine accurately the behaviours of participants in a multi-party teleconferencing setting. As discussed in Chapter 3, VAD code [38, 39] was used to semi-automate some of the mark-up of the audio recordings. Due to inaccuracies found with the identification of silent or sounding (listening or talking) sections, each recorded audio stream from each participant required manual adjustments through the use of Praat software [40]. This proved time consuming but necessary to facilitate the next stage of marking any overlap in speech with the corresponding CA parameter value. Upon listening to each group conversation multiple times to enable these manual adjustments it became clear that the delivery of the CA attribute continuers was adapting given the delay induced circumstances. It was noted that the change in continuers deployment was linked to the delay induced conditions only, with the 0ms condition presenting normal continuer patterns. Multi-party behaviours hypothesised to change with delay shall be the topic of this subsection, detailing each with specific examples to support the initial observations. It should be noted that with regards to the continuers hypothesis, very early stages of coding have been started in an attempt to evidence the observations.

The use of continuers in multi-party conversations were of interest due to the significance of their occurrence within natural conversations. Continuers are used by participants of a conversation as a sign of acknowledgement and to signal to the talker to continue with their utterance. They can also be used as a sign of engaged and supportive listenership within a conversation. Examples of continuers are utterances which come between sentences and are found at TRP's which include the use of *yeah, mm, uh-huh mm-hm,* etc. TRP's indicate where a turn or floor exchange can take place between speakers.

Continuers can also be deployed in overlap with another speaker, so pose a challenge when attempting to automatically detect false starts as a simple continuer could easily be detected as interruption or false start. The continuers hypothesis that was developed assumed that with increased delay continuers are not only just less frequently produced, but also there is a tendency for the volume at which they are delivered to be significantly lower. This could be due to participants fearing that they may interrupt, distract the talker, be too late in their delivery or will simply not be heard/acknowledged. Also, a point to note is that continuer length is considerably shorter on average with higher delay; mainly consisting of *mm, uh-huh, mm-hm* rather than long quasi-turns such as *yeah, ok then, right,* etc.

In order to test this hypothesis, future work could assess if there is any decrease in volume for each individual participant when they produce a continuer as opposed to the volume at which they produce the rest of their speech within the task conversations. As it has been observed that continuers are less frequently delivered and the volume of continuers tends to be lower when a significant delay is introduced; this is believed to have a disruptive effect on the overall interaction between the groups and their ability to successfully complete the set out tasks. A measure of the volume of continuers is of interest as the general role of a continuer is to supply feedback to the talker from other group members (listeners), to provide a reassurance of acknowledgment and to continue on with their utterance. This can become even more important and continuers can be heavily relied upon in situations where a time constrained, highly interactive task is conducted and an aural only channel is the only means of communication. Without the manifestation of this key speech parameter, a reluctance to produce them or producing them at lower volume levels, the talker has little to no sign of group participation or interaction, which is fundamental to the successful completion of the tasks. The lack of visual cues hinders the process as an aural only channel is available and participants are unable to see any visual continuers, e.g. nod of the head, gestures, eye contact, body language, etc. Upon listening to the group conversations where a delay has been introduced, the change in dynamic within the group can be associated with participants being aware of a disruptive element (delay) in the conversation but are unsure what is causing it. Their behaviour adapts to accommodate to the set of circumstances, whereby the start of the conversation can be in great contrast to the middle and end of the conversation; this also differs greatly to face-to-face communication where an external component such as audio quality or delay level does not exist. With this in mind, a measure of the overall volume of a participant's speech and the volume of all continuer only speech will not be of much benefit, therefore it is suggested that the analysis could segment the continuers produced along a task's timeline into quarters so the volume of continuers at the beginning of a task can be compared to the volume of continuers produced near to the end of the task. Signal processing techniques would be needed to ensure an accurate representation of volume is achieved, for example by the use of an A-Weighting filter to account for the perceived relative loudness of a signal by

the human ear [46]. There are several factors to consider as the task also has a direct impact on the number of continuers produced. The first round of subjective tests mixes three different styles of task therefore providing results for three very different styles of conversation, informal, formal and goal-oriented. The second round of tests aimed to eliminate any impacting factors from different tasks and instead focused on the use of just one task with the intention to draw out conversation parameters, such as continuers, which are heavily relied upon in multi-party collaborative timed tasks. Future work could identify individual's continuer volume with consideration as to which group member produces or is party to the continuer delivery to allow for accurate volume metrics to prove or disprove this hypothesis.

Another observation taken from the in-depth study of the audio recordings from both testing phases is related to the CA parameter no-gap-no-overlap. It was noted that a lower frequency of no-gap-no-overlap turn transitions with higher delay levels; this limits the amount of smooth transitions in talker turn alternation compared with that of a conversation with no delay or low delay. Delay induced conditions suffered from less seamless communication at TRP's, which on reflection would seem obvious given that as delay increases the opportunity for no-gap-no-overlap to occur is narrow. However, due to the nature of no-gap-no-overlap being one of few CA features that this study explored that required no contextual evidence to identify, as it is purely associated with the timing in which is it delivered, this sets it apart and allows for the early stages of automatic detection. Automatic detection of this speech parameter could be considered as part of a new system to determine conversation quality and teleconferencing effectiveness, or to simply show a good or poor-quality conversation. As previously discussed, in the near future this could be achievable with the deployment of machine learning from the extensive data set acquired during this body of work.

6.2.3 Standardisation of Conversational Test Method and Analysis

Another option could be for this work to be embodied into an ITU standard relating to acceptable delay levels within multi-party teleconferencing and conversational testing strategies. This type of standard could also document the conversational testing methodologies used throughout this work; this could specify the type of tasks used to highlight certain conversational attributes and the overall test design.

6.3 Summary

In summary, this chapter explores future work and also presents a discussion of the results from both the first and second round of subjective testing. As discovered as a key finding is the consistency between the two testing phases; the CA parameters extracted from both very different tests validates the use of CA metrics in this area in providing reliable and consistent results across different testing platforms, whilst clearly having benefits for understanding better how to process conversation-based testing. The evidence gathered from this study gives the foundations to explore future opportunities, such as to utilise machine

learning with the potential (as described under Section 6.2.1) to determine the quality of a telemeeting in real-time. Also, the continuers hypothesis was explored in detail with suggestions as to how to detect any reduction in the volume of continuer production in the presence of delay. In addition, the no-gap-no-overlap observations assisted in the development of a patent and provides grounds for further investigation for machine learning. This was due to the nature of no-gap-no-overlap being non-contextual and again being a CA parameter with the possibility to be detected in real-time across telemeetings. This work is in the early stages of invention drafting for a patent filing with BT.

Chapter 7

7. Conclusions

The work described and discussed within this thesis has presented how induced transmission delay with and without the presence of spatial audio in teleconferencing effects the interactions and behaviours of the users of the system. With a soft-focus on an interdisciplinary approach of combining CA metrics with objective engineering measurements to fully assess multi-party teleconferencing quality. The conversation-based tests consisting of a mono and in some cases spatial audio condition, trialled alongside a novel representation of a "normal" face-to-face speech condition. The testing platform was employed to carry out a series of subjective tests of which the results have been discussed in detail from the corpus of data attained during the testing phases. This data did not only provide subjective measures from the tests, but also provided objective measures from analysis of the conversations themselves. The CA extraction design has been accomplished through uses of various software packages and tools used to process the data from the experiments, along with a significant amount of modification to existing tools and programs. We begin by reviewing the chapters and contributions made throughout this work.

Chapter 3 gave a comprehensive overview of the experimental methodology and background for the initial round of testing, whilst detailing the specification of the developed subjective test design with the aid of a diagram that demonstrated the set-up of the testing facility and test platform. Also presented was the subjective testing method employed (MOS scale), all of which had been specified with the help of background research and industrially supervision from BT. The 5-point MOS scale for evaluating perceived speech and audio quality was selected as participants that were recruited were previously untrained and not familiar with performing conversation-based tests. The advantage that the MOS scale had was that it provided a reasonably easily interpreted system on which to rate the speech quality, especially for previously untrained participants. Also, MOS is widely recognised and commonly used system for rating telephony systems. The chapter also discusses the important of use subjective testing forms, highlighting the essential subjective data gathered from each participant who took part in the subjective testing phase. Chapter 3 also documented the subject recruitment process and described in detail the three-task design model with the specifics of each chosen task and why they were selected. Finally, Chapter 3 concludes with a full description of the experiment design method used to ensure a fair and effective test design was achieved. The method used incorporated elements of Graeco-Latin square design to provide a form of randomisation to the tests, in an attempt to eliminate the risks of learning effects impacting on the tasks. In addition, the importance of presentation order was acknowledged for both the levels of transmission delay and the selection of tasks; this is where elements of Graeco-Latin

square design provided a solution to this issue. The concept of trialling each possible combination of test order, referring to delay levels and task presentation order, was also detailed with reasons as to why this was not feasible due to constraints on the length of the tests.

Chapter 4 presents the implementation of the first subjective testing phase and introduced the explored CA parameters and extended CA parameters monitored during the tests; discussion of both strongly and poorly correlated results captured from the initial subjective testing phase were all presented. Each tool, piece of software, developed code and program used in the creation of the testing phase was listed and then discussed in depth. The modification of many of the components involved in the design was a crucial element in accomplishing the required functionality. The chapter also discussed the CA aspect of this work and presented how the recordings were processed to gain objective results after the testing stage was complete. The use of Praat, speech analysis software and specialist utterance-by-utterance adaptive VAD MATLAB code are referred to as tools that assisted to achieve some form of semi-automated results for the speech analysis. Also, the inclusion of R code to process the recordings and extract the raw objective data was an achievement which will prove useful to further adapt and develop for use on future studies of this nature. The results gathered during the first initial round of tests were promising with significant findings in relation to key CA attributes, such as false starts and no-gap-no-overlap in conjunction with increasing levels of transmission delay. The chapter also highlighted areas for improvement with regards to the planned second subjective testing phase. The need for task refinement was apparent with the belief that the variation in task selection made the interpretation of some the CA metrics and results difficult to quantify or inconclusive.

Chapter 5 presents the CA objective results obtained from the second expanded testing phase of this work, in addition to the refinement of the task design and details all modifications made to the system. The chapter addressed the added elements to the second phase of conversation-based tests by discussing the inclusion of varying audio conditions, in addition to refined transmission delay levels. An emphasis was made on pretesting of all refined tasks with discussion as to how the refined goal-oriented task was eventually selected. Clearly laid out in this chapter was also the improvements made to the test design from the previous round, with an emphasis on the inclusion of an orientation round and the addition of a face-to-face condition. The decision to incorporate these elements into the test design proved valuable in ensuring the most reliable and consistent set of results were achieved. The CA section of Chapter 5 presented the results of the CA parameters that were monitored and that showed some significant evidence on the benefits of spatial audio over standard mono audio. Also demonstrated was the similarity between the face-to-face and spatial conditions. This has great value when considering

spatial audio for use within a teleconferencing context, as it provided a closer, more efficient form of communication to that of “normal” speech when compared with mono.

Finally, Chapter 6 draws on the progression of possible future work from this study. It details the work achieved so far regarding a patent and potential future patents, whilst discussing machine learning from the corpus of data from both sets of subjective conversation-based tests, which provided a substantial data set which can be processed in the future. Also presented are the metrics which in some cases are more easily measured than the MOS 5-point scale. The recommendations from this body of work are that CA extraction techniques employed did in fact provide more accurate representations of the quality of teleconferences than the subjective MOS results provided. CA parameter monitoring to determine conversation quality in these type of scenarios provided a measurable benefit over the traditional MOS scale with high-quality systems under tests; MOS lacks the granularly to differentiate between audio conditions, especially with the use of higher quality conditions such as spatial audio and also with the introduction of impairments and degradations such as transmission delay. The advantages offered by CA in this type of interdisciplinary study can clearly be seen from both the official and extended CA parameters documented, which enabled the accurate depiction and objective measures of audio quality experienced by participants during the testing phases to be analysed. The addition of gathering quantifiable metrics based on actual events within the recorded conversations provided a much more reliable set of results than would have been feasible with subjective based assessments only. CA played a major role in the embodiment of this work and gave an insight into some conversation behaviours and non-context based accepted patterns associated with transmission delay. In the near future, with the data sets achieved and with the assistance of machine learning, CA parameters could give the potential for automatic measures to be taken from teleconferences to give a definitive metric on conversation and telemeeting quality. The inclusion of two distinct areas, both the study of human communication from the field of CA paired with engineering metrics and statistical analysis, has resulted in new and innovative solutions to objectively measure conversation quality over mediated channels of communication. Throughout the course of this study, with the introduction of transmission delay into teleconferencing systems, a communication phenomena has been identified through the deployment of a new and novel methodology. This has provided new and interesting insights into the problems associated with such impairments in synthesised environments, along with possible solutions to help counteract the impact of transmission delay.

Bibliography

[1] H. Sacks*, et al.*, "A simplest systematics for the organization of turn-taking for conversation," *Language,* pp. 696-735, 1974.

[2] Hughes, P J. *ITU-T COM12-C142-E. Measuring the performance of wideband communications.* Geneva : ITU, December 2013.

[3] Hughes, P J., Bailey, K L. *ITU-T COM12-C278-E. Results from a 4-way audio conference test using free conversations with delay.* Geneva : ITU, May 2015.

[4] K.L Bailey, "Comparing Narrowband and Wideband Speech Communication," MSc Dissertation, University of Essex, December 2013.

[5] I. Kegel, K.Bailey, M.Reed, P.Hughes, "Managing Streamed Audio Communication Sessions," Worldwide Patent WO2019122343, June 2019.

[6] ITU-T Rec. P.1305, *"Effect of delays on telemeeting quality"*, July 2016.

[7] ITU-T Rec. P.1301, *"Subjective quality evaluation of audio and audiovisual multiparty telemeetings"*, July 2012.

[8] B. O'Conaill, S. Whittaker and S. Wilbur, "Conversations over Video Conferences: An Evaluation of the Spoken Aspects of Video-Mediated Communication," in *Human-Computer Interaction,* Vol. 8, Lawrence Erlbaum Associates, 1993, pp. 389-428.

[9] S. Whittaker, "Theories and Methods in Mediated Communication," AT&T Res. Labs, New Jersey, NY, 2003.

[10] N. Kitawaki and K. Itoh, "Pure delay effects on speech quality in telecommunications," *Selected Areas in Communications, IEEE Journal on,* vol. 9, pp. 586-593, 1991.

[11] K. Schoenenberg, A.Raake, S.Egger and R. Schatz, "On interaction behaviour in telephone conversations under transmission delay", Speech Communication. 63–64, (Sep. 2014), 1–14.

[12] A. Raake, J. Skowronek and K. Schoenenberg. "*Dependence of speech quality on delay and interactivity, and implications for the E-model."* Geneva : s.n., 2013. ITU-T COM12-C79.

[13] K. Schoenenberg, A. Raake, and J. Koeppe. "Why are you so slow? – Misattribution of transmission delay to attributes of the conversation partner at the far-end." *International Journal of Human-Computer Studies*72.5, pp.477-487, 2014.

[14] K.Schoenenberg, "The quality of mediated-conversations under transmission delay." 2016.

[15] A. Takahashi*, et al.*, "Objective assessment methodology for estimating conversational quality in VoIP," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 14, pp. 1984-1993, 2006.

[16] A. Raake*, et al.*, "Predicting speech quality based on interactivity and delay," in *INTERSPEECH*, pp. 1384-1388, 2013.

[17] ITU-T Rec. G.107, "The E-model: A computational model for use in transmission planning", Geneva, April 2009.

[18] ITU-T Rec. G.114 , *"One-Way Transmission Time"*, May 2003.

[19] R. Clift, "4. Conversation analysis." *Pragmatics of Discourse* 3 pp. 97-113, 2014.

[20] E.A. Schegloff, "Overlapping talk and the organization of turn-taking for conversation", *Language in Society* 29: 2000 1–63.

[21] M.B. Walker and C.Trimboli, "Smooth transitions in conversational interactions." The Journal of Social Psychology, 117, 305–306, 1982.

[22] T. Stivers, N.J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, … S.C. Levinson, "Universals and cultural variation in turn-taking in conversation." *Proceedings of the National Academy of Sciences*, 2009.

[23] S. C. Levinson, "Action formation and ascription," *The handbook of conversation analysis,* pp. 101-130, 2013.

[24] P.Indefrey and W.J Levelt, "The spatial and temporal signatures of word production components." *Cognition*, *92*(1-2), pp.101-144, 2004.

[25] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics,* vol. 38, pp. 555-568, 2010.

[26] ITU-T Rec. P.805 , *"Subjective evaluation of conversational quality"*, April 2007.

[27] ITU-T Rec. P.800, "*Methods for Subjective Determination of Transmission Quality*," 1996.

[28] ITU-R, BS.1534, "*Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems,*" 2003.

[29] ITU-R, BS. 1116-1, "*Methods for the Subjective Assessment of Small Impairments in Audio Systems including Multichannel Sound Systems,"* 1997.

[30] ITU-T Rec. P. 862, *"Perceptual Evaluation of Speech Quality,"* 2001.

[31] ITU-T Rec. P. 563, *"Single-ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications,"* 2004.

[32] A. Rix and M. Hollier, "Perceptual Speech Quality Assessment from Narrowband Telephony to Wideband Audio," *AES 107th Convention,* New York, 24th – 27th September 1999, pp.1 – 11.

[33] T. A. Hall, "Objective speech quality measures for Internet telephony," in *Voice over IP (VoIP) Technology, Proceedings of SPIE*, 2001, pp. 128-136.

[34] Find a Job Website [Online]. Available: https://findajobea.co.uk/

[35] Issing J, Nikolaus N, Conversational quality as a function of delay and interactivity. In: International Conference on Software, Telecommunications and Computer Networks, IEEE, pp. 1–5, 2012

[36] You be the Judge by Paul Lamond Games. [Online]. Available: https://paul-lamond.com/

[37] P. Harris, *Designing and reporting experiments in psychology*: Open University Press, 2008.

[38] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data", *Proc. Int. Conf. on Acoustics, Speech and Signal Processing* (ICASSP 2013), pp. 7229--7233, Vancouver, Canada, May 2013.

[39] T. Kinnunen. (May 13). *Utterance-by-utterance adaptive voice activity detector code* [Online]. Available: http://cs.joensuu.fi/pages/tkinnu/webpage/

[40] Praat.org. *Praat Speech Analysis Tool* [Online]. Available: http://www.praat.org

[41] R Core Team. (2012). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-02012* [Online]. Available: http://www.R-project.org/

[42] S.Wickham, L. Hadley, "40 years of boxplots" (PDF). November 2011.

[43] H. Wickham, "ggplot2: Elegant Graphics for Data Analysis." Springer-Verlag New York, 2016.

[44] M. Naef, O. Staadt, M. Gross, "Spatialized audio rendering for immersive virtual environments." In processings of the ACM symposium on Virtual reality software and technology (VRST'02). Association of Computing Machinery, New York, 65-72, 2002. [Online]. Available: https://doi.org/10.1145/585740.585752

[45] BT MeetMe with Dolby Voice data sheet. [Online]. Available: https://www.btconferencing.com/downloads-library/datasheets/en/meetme-with-dolby-voice.pdf

[46] IEC 61672-1:2013 Electroacoustics - Sound level meters - Part 1: Specifications

[47] S.M. Ross, “Testing Statistical Hypotheses,” and “Hypotheses Tests Concerning Two Populations,” in *Introductory Statistics*, 2nd ed. Morgan Kaufmann Publishers, 2005, pp.387-497.

Appendix

A1. Sample Copy of Holiday Task Form

Name:…….

Holiday Task:

Through discussion with the group, please collectively agree and list 10 things which you feel makes a good holiday. You have 5 minutes to discuss your travel experiences/favourite holidays and 5 minutes to agree upon 10 things which make a good holiday.

1. …….
2. …….
3. …….
4. …….
5. …….
6. …….
7. …….
8. …….
9. …….
10. …….

A2. Copy of You be the Judge Task Form

Name:…….

You Be The Judge Task:

This task will involve the discussion of 2 case and verdict cards related to various real life court cases. As a group you must discuss and debate the 2 cases presented within the time frame of 15 minutes.

The role you have been allocated is:

- **Counsel Against**:

  Your role as Counsel Against involves presenting reasons and opinions why the court would rule **against/ not in favour** of each of the 2 cases. You shall be responsible with your other team member for leading the debates **against** each of the cases.

- **Case 1 : Cereal Killer**

  In Australia in 1911, a driver left his wagon, loaded with bags of wheat at the side of the road, guarded by his dog. Two horses came along and, undaunted by the dog's barking, tore open the bags and ate so much that they died soon after from the effects of over-eating. The horses' owner sued for the loss of his horses, and the wagon owner sued for the loss of his wheat.

  Did the court rule *FOR or AGAINST* the **owner of the horses**?

- **Case 2 : Crash Landing**

  A small plane was coming in to land at a major airport when it became apparent that it was in the flight path of a much larger jumbo jet. The air traffic control tower radioed the pilot of the small plane and instructed him to abort the landing and fly to one side to make way for the jet. The plane followed the instructions, but was caught in the air turbulence from the jet and crashed. The pilot sued the airport for not warning him of the possible turbulence.

  Did the court rule *FOR* or *AGAINST* the **pilot**?

A3. Copy of Consent Form

Name: ……

Gender: Male / Female

Age: ……..

- How many of the other participants do you know? Please tick one box.

  ☐ No-one

  ☐ Everyone

  ☐ Only the following people (please list their names below):

  ……

  ……

  ……

  ……

  ……

- Which of the following videoconferencing or audio conferencing applications do you use? Please tick as many as apply.

  ☐ Skype

  ☐ Google+ Hangouts

☐ Apple FaceTime

☐ Telephony-based conferencing systems

☐ Other – please give name(s): ……

☐ None

Your Consent

I give my consent for:

- Audio recordings to be made during the group conversation sessions in the silence cabinets
- Monitoring via webcam whilst in the silence cabinets for safety purposes
- Written notes to be taken throughout the experiment

These will be recorded by representatives of BT plc and the University of Essex and will be used for research purposes.

You can withdraw your consent at any time by contacting Karis Bailey at either karis.bailey@bt.com or kbailej@essex.ac.uk.

Name: ……

Signature: ……

Date: ……..

A4. Copy of Conversational Experience Survey for First Phase of Subjective Tests

Name:…….

Conversational Experience Survey

Please take a moment to rate your experience throughout the experiment. Please circle the most appropriate number or statement for each question that relates to your experience. Numbered questions correspond with the evaluation scale below. Thank you for taking part and providing us with your feedback.

Evaluation Scale: (5) Excellent (4) Good (3) Fair (2) Poor (1) Bad

1st Task:

How would you rate the audio quality for this task?	**5**	**4**	**3**	**2**	**1**
Did you or your partners have any difficulty in talking or hearing over the connection?			**Yes**	**No**	
How well do you feel you were understood?	**Very well**	**Well**	**Fairly well**	**Not well**	**Not at all**
How easy did you find it to understand the other participants?	**Very easy**	**Easy**	**Fairly easy**	**Difficult**	**Very difficult**
How would you rate the ease of conversation flow within this task?	**Very easy**	**Easy**	**Fairly easy**	**Difficult**	**Very difficult**

Name:…….

Conversational Experience Survey

Please take a moment to rate your experience throughout the experiment. Please circle the most appropriate number or statement for each question that relates to your experience. Numbered questions correspond withthe evaluation scale below. Thank you for taking part and providing us with your feedback.

Evaluation Scale: (5) Excellent (4) Good (3) Fair (2) Poor (1) Bad

2nd Task:

How would you rate the audio quality for this task?	**5**	**4**	**3**	**2**	**1**
Did you or your partners have any difficulty in talking or hearing over the connection?			**Yes**	**No**	
How well do you feel you were understood?	**Very well**	**Well**	**Fairly well**	**Not well**	**Not at all**
How easy did you find it to understand the other participants?	**Very easy**	**Easy**	**Fairly easy**	**Difficult**	**Very difficult**
How would you rate the ease of conversation flow within this task?	**Very easy**	**Easy**	**Fairly easy**	**Difficult**	**Very difficult**

Name:…….

Conversational Experience Survey

Please take a moment to rate your experience throughout the experiment. Please circle the most appropriate number or statement for each question that relates to your experience. Numbered questions correspond with the evaluation scale below. Thank you for taking part and providing us with your feedback.

Evaluation Scale: (5) Excellent (4) Good (3) Fair (2) Poor (1) Bad

3rd Task:

How would you rate the audio quality for this task?	**5**	**4**	**3**	**2**	**1**
Did you or your partners have any difficulty in talking or hearing over the connection?			**Yes**	**No**	
How well do you feel you were understood?	**Very well**	**Well**	**Fairly well**	**Not well**	**Not at all**
How easy did you find it to understand the other participants?	**Very easy**	**Easy**	**Fairly easy**	**Difficult**	**Very difficult**
How would you rate the ease of conversation flow within this task?	**Very easy**	**Easy**	**Fairly easy**	**Difficult**	**Very difficult**

Considering all three tasks, did you find any differences in conversational difficulty or your ability to communicate? **Yes** **No**

If **yes**, please elaborate below with reference to specific tasks:

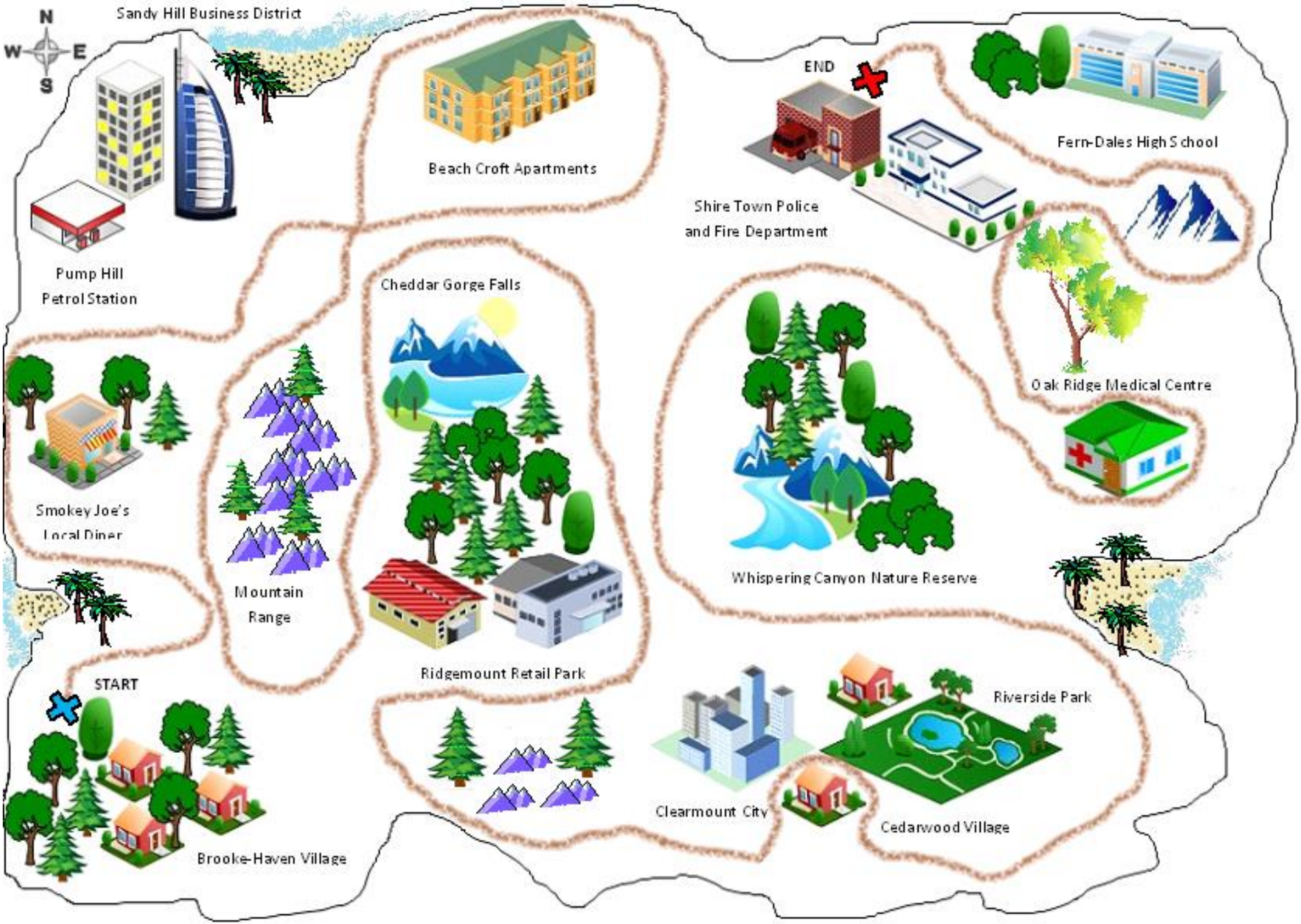
.....

.....

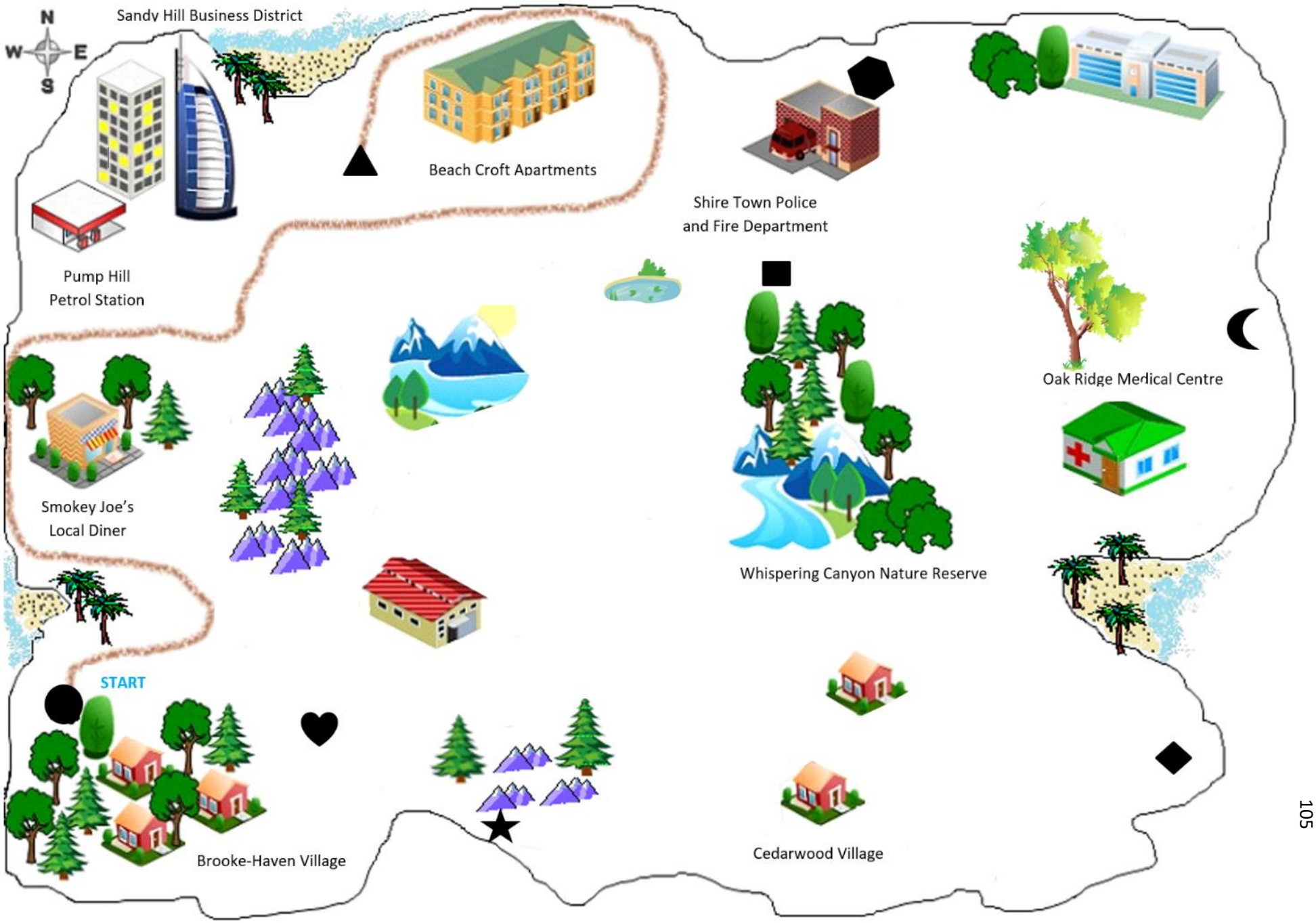
.....

.....

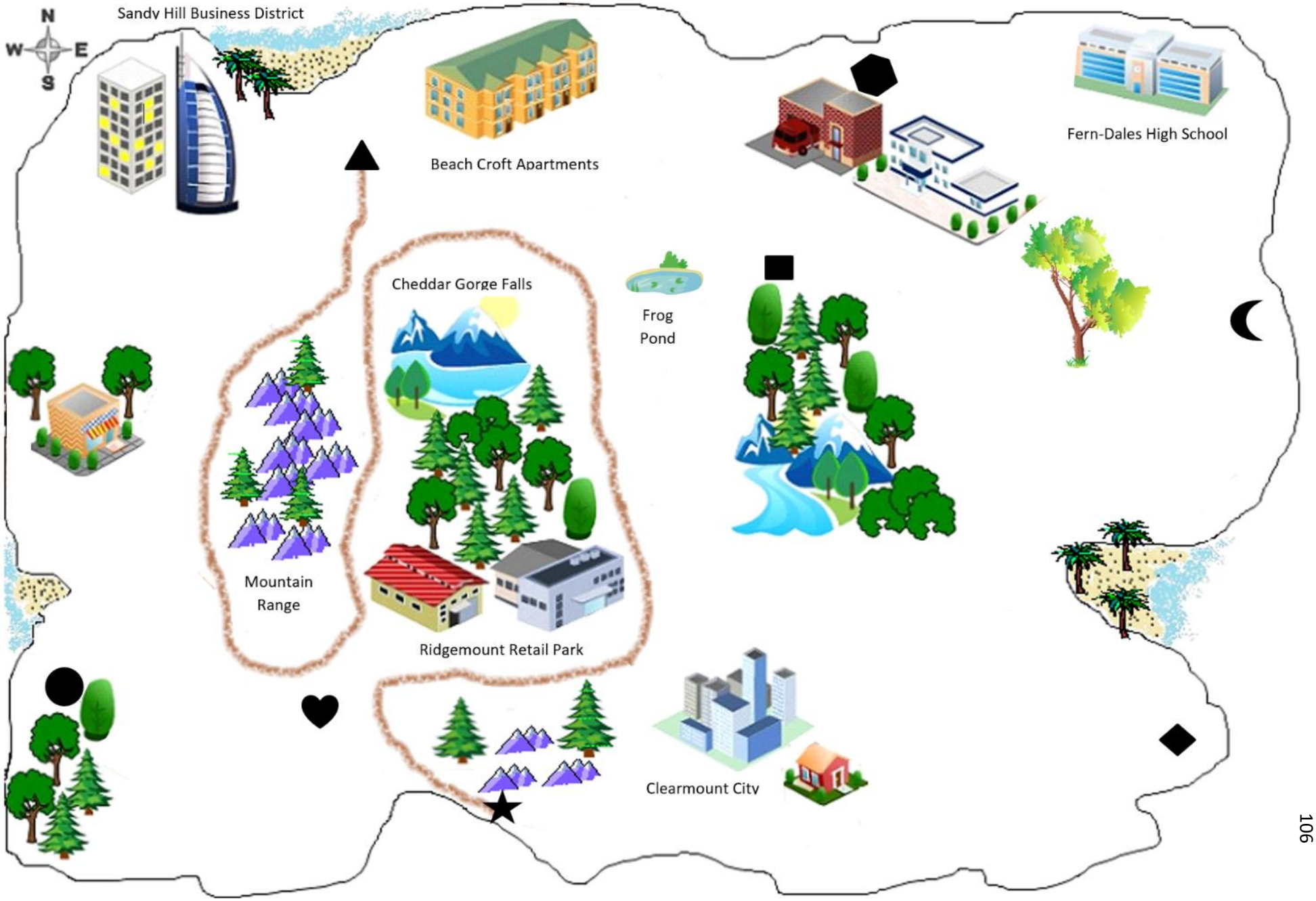
A5. Copy of Example Goal-Oriented Task Map for First Phase of Subjective Tests (Master Map)



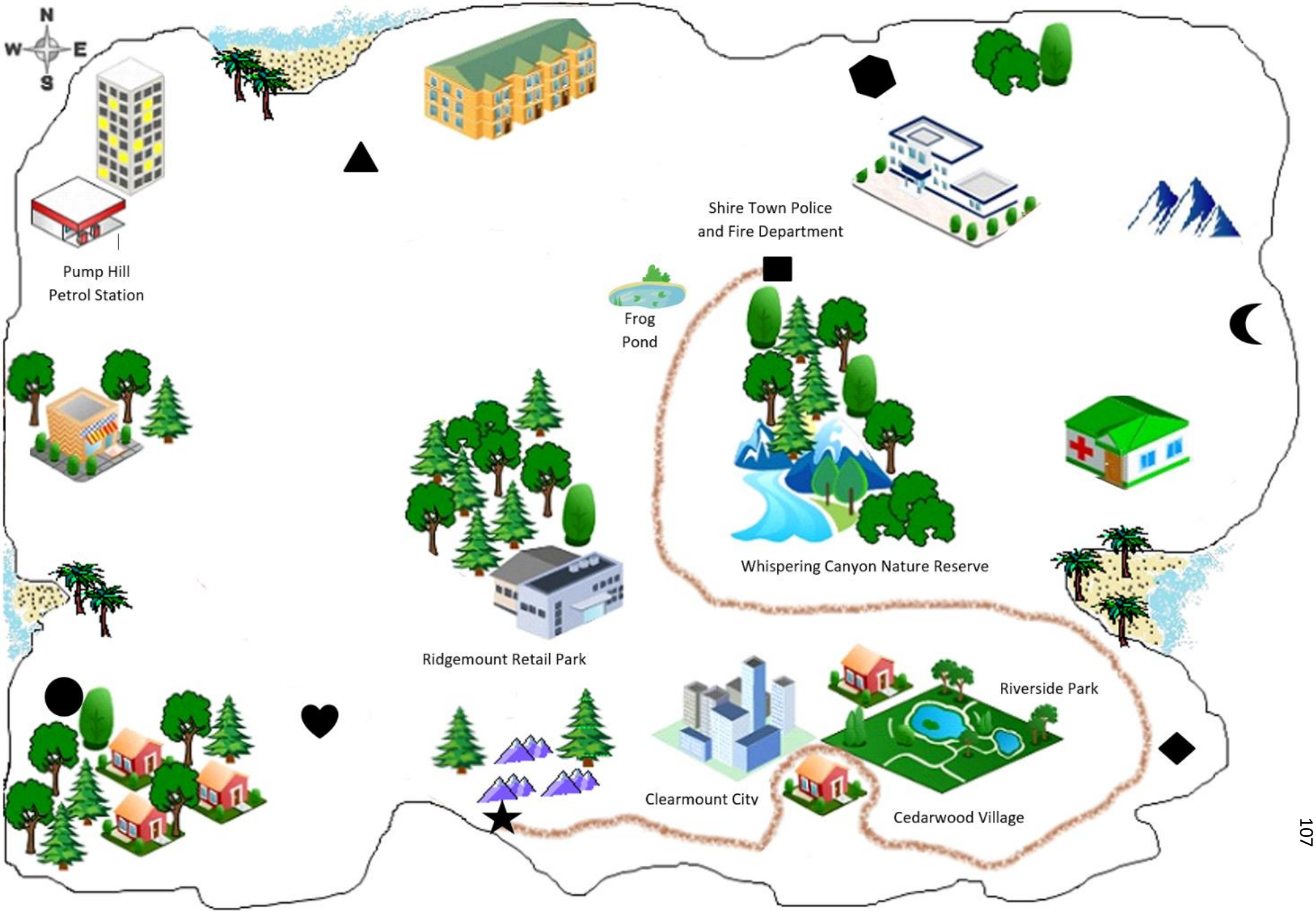
A5. Copy of Example Goal-Oriented Task Map for First Phase of Subjective Tests (Cabinet 3)



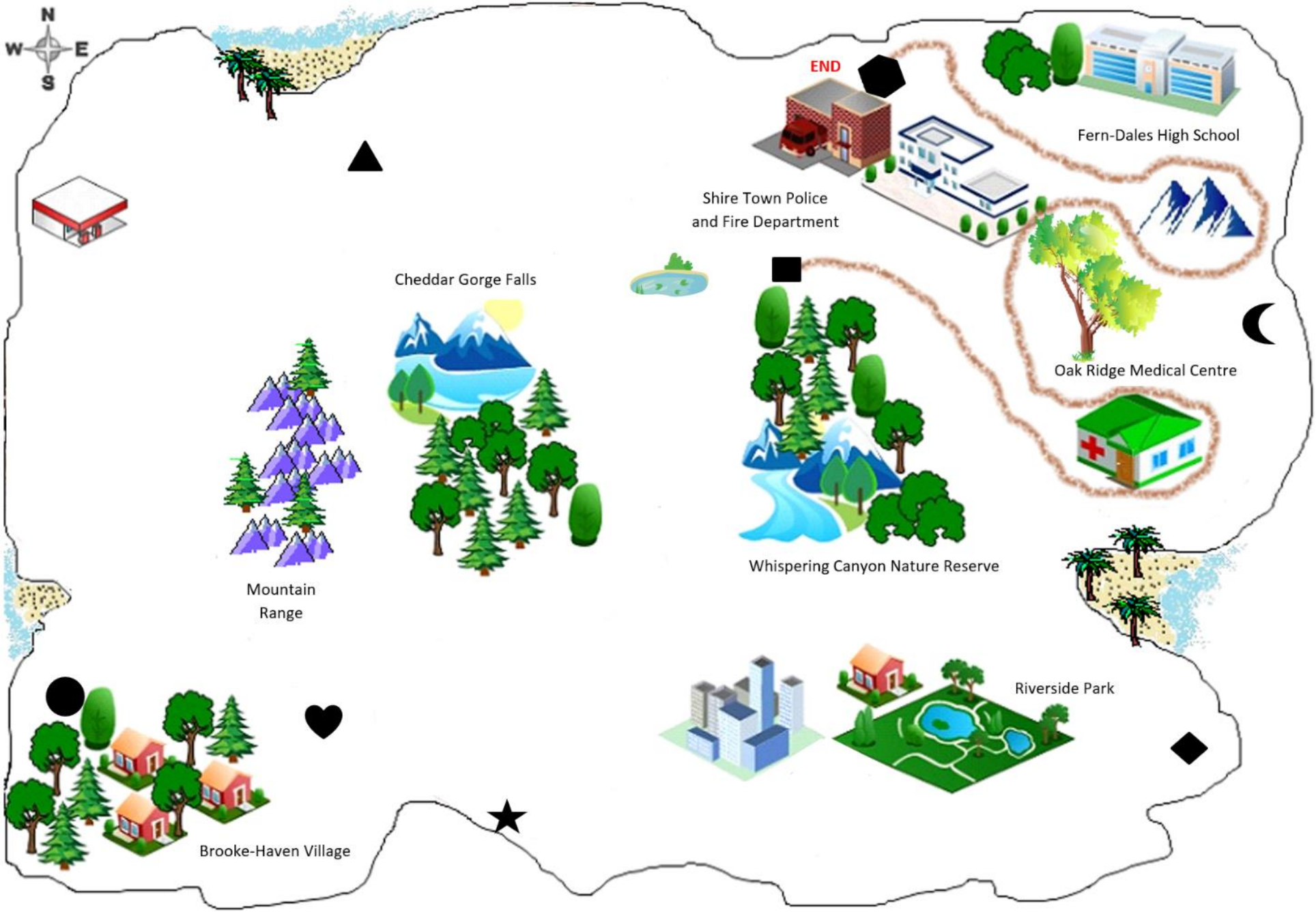
A5. Copy of Example Goal-Oriented Task Map for First Phase of Subjective Tests (Cabinet 4)



A5. Copy of Example Goal-Oriented Task Map for First Phase of Subjective Tests (Cabinet 5)



A5. Copy of Example Goal-Oriented Task Map for First Phase of Subjective Tests (Cabinet 6)



A6. Copy of Conversation Experience Survey for Second phase of Subjective Tests

Name:…….

Conversational Experience Survey

Please take a moment to rate your experience throughout the experiment. Please circle the most appropriate number or statement for each question that relates to your experience. Numbered questions correspond with the evaluation scale below. Thank you for taking part and providing us with your feedback.

Evaluation Scale: (5) Excellent (4) Good (3) Fair (2) Poor (1) Bad

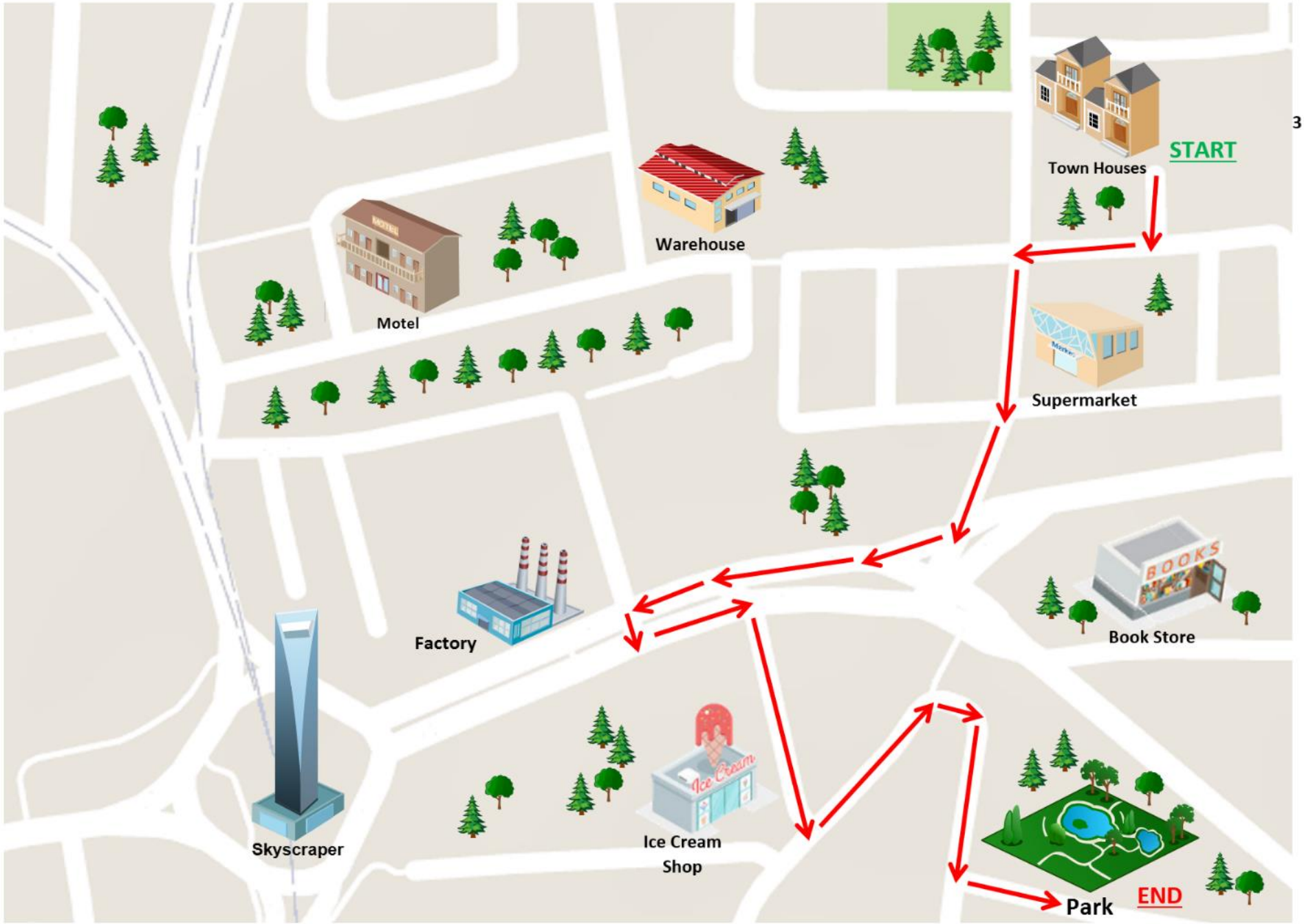
Maps 1 - 7

How would you rate the audio quality for this task? **5 4 3 2 1**

Did you or your partners experience any difficulty in communicating or hearing over the connection? **Yes No**

How attentive were your conversation partners during the call?	**Very attentive**	**Attentive**	**Fairly attentive**	**Inattentive**	**Very inattentive**
How would you judge the effort required to interrupt the other participants?	**Very easy**	**Easy**	**Fairly easy**	**Difficult**	**Very difficult**
How easy did you find it to communicate using the system?	**Very easy**	**Easy**	**Fairly easy**	**Difficult**	**Very difficult**
How easy did you find it to understand the other participants?	**Very easy**	**Easy**	**Fairly easy**	**Difficult**	**Very difficult**
How would you rate the ease of conversation flow within this task?	**Very easy**	**Easy**	**Fairly easy**	**Difficult**	**Very difficult**

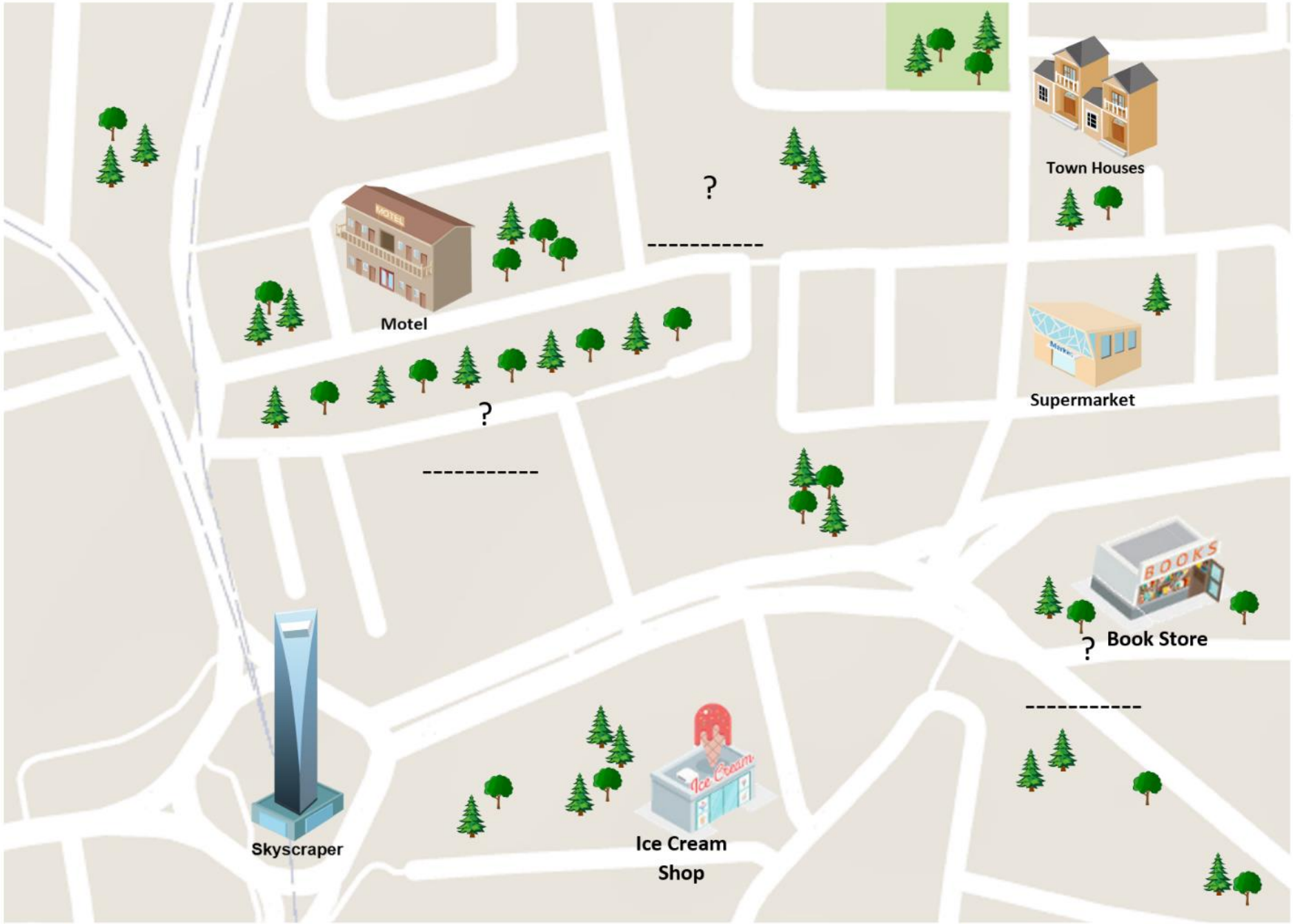
A7. Copy of Example Goal-Oriented Task Map for Second Phase of Subjective Tests (Master Map)



A7. Copy of Example Goal-Oriented Task Map for Second Phase of Subjective Tests (Cabinet 4)

A7. Copy of Example Goal-Oriented Task Map for Second Phase of Subjective Tests (Cabinet 5)

A7. Copy of Example Goal-Oriented Task Map for Second Phase of Subjective Tests (Cabinet 6)



A8. BT First Subjective Testing Phase: Test Order

Test 1 – Friday 23rd May – Morning Group

1) Informal Meeting Task 0ms
2) Formal Meeting Task 400ms
3) Goal-Oriented Task 800ms

Test 2 – Friday 23rd May – Afternoon Group

1) Goal-Oriented Task 400ms
2) Informal Meeting Task 800ms
3) Formal Meeting Task 0ms

Test 3 – Friday 23rd May – Late Afternoon Group

1) Formal Meeting Task 800ms
2) Goal-Oriented Task 0ms
3) Informal Meeting Task 400ms

Test 4 – Thursday 5th June – Afternoon Group

1) Formal Meeting Task 800ms
2) Goal-Oriented Task 400ms
3) Informal Meeting Task 0ms

Test 5 – Friday 6th June – Morning Group

1) Goal-Oriented Task 0ms
2) Informal Meeting Task 800ms
3) Formal Meeting Task 400ms

Test 6 – Friday 6th June – Afternoon Group

1) Informal Meeting Task 400ms
2) Formal Meeting Task 0ms
3) Goal-Oriented Task 800ms

Test 7 – Friday 6th June – Late Afternoon Group

1) Informal Meeting Task 0ms
2) Goal-Oriented Task 400ms
3) Formal Meeting Task 800ms

Test 8 – Monday 9th June – Morning Group

1) Formal Meeting Task 400ms
2) Informal Meeting Task 800ms
3) Goal-Oriented Task 0ms

Test 9 – Monday 9th June – Afternoon Group

1) Goal-Oriented Task 800ms
2) Formal Meeting Task 0ms
3) Informal Meeting Task 400ms

A9. BT Second Subjective Testing Phase: Test Order

Test 1 - Monday 27th June – Morning Group

Orientation round (Face-to-Face, Mono 0ms, Mono 200ms, Mono 400ms, Spatial 0ms, Spatial 200ms, Spatial 400ms)

1) Face-to-Face
2) Mono 0ms
3) Mono 200ms
4) Spatial 400ms
5) Mono 400ms
6) Spatial 200ms
7) Spatial 0ms

Test 2 – Friday 1st July – Morning Group

Orientation round (Face-to-Face, Mono 0ms, Mono 200ms, Mono 400ms, Spatial 0ms, Spatial 200ms, Spatial 400ms)

1) Face-to-Face
2) Mono 200ms
3) Mono 400ms
4) Mono 0ms
5) Spatial 0ms
6) Spatial 400ms
7) Spatial 200ms

Test 3 – Friday 1st July – Afternoon Group

Orientation round (Face-to-Face, Mono 0ms, Mono 200ms, Mono 400ms, Spatial 0ms, Spatial 200ms, Spatial 400ms)

1) Face-to-Face
2) Mono 400ms
3) Spatial 0ms
4) Mono 200ms
5) Spatial 200ms
6) Mono 0ms
7) Spatial 400ms

Test 4 – Thursday 7th July – Morning Group

Orientation round (Face-to-Face, Mono 0ms, Mono 200ms, Mono 400ms, Spatial 0ms, Spatial 200ms, Spatial 400ms)

1) Face-to-Face
2) Spatial 0ms
3) Spatial 200ms
4) Mono 400ms
5) Spatial 400ms
6) Mono 200ms
7) Mono 0ms

Test 5 – Monday 25th July – Morning Group

Orientation round (Face-to-Face, Mono 0ms, Mono 200ms, Mono 400ms, Spatial 0ms, Spatial 200ms, Spatial 400ms)

1) Face-to-Face
2) Spatial 200ms
3) Spatial 400ms
4) Spatial 0ms
5) Mono 0ms
6) Mono 400ms
7) Mono 200ms

Test 6 – Monday 25th July – Afternoon Group

Orientation round (Face-to-Face, Mono 0ms, Mono 200ms, Mono 400ms, Spatial 0ms, Spatial 200ms, Spatial 400ms)

1) Face-to-Face
2) Spatial 400ms
3) Mono 0ms
4) Spatial 200ms
5) Mono 200ms
6) Spatial 0ms
7) Mono 400ms

Test 7 – Tuesday 26th July – Morning Group

Orientation round (Face-to-Face, Mono 0ms, Mono 200ms, Mono 400ms, Spatial 0ms, Spatial 200ms, Spatial 400ms)

1) Face-to-Face
2) Mono 0ms
3) Mono 200ms
4) Mono 400ms
5) Spatial 0ms
6) Spatial 200ms
7) Spatial 400ms

Test 8 – Wednesday 27th July – Morning Group

Orientation round (Face-to-Face, Mono 0ms, Mono 200ms, Mono 400ms, Spatial 0ms, Spatial 200ms, Spatial 400ms)

1) Face-to-Face
2) Mono 200ms
3) Mono 0ms
4) Spatial 400ms
5) Spatial 200ms
6) Mono 400ms
7) Spatial 0ms

Test 9 – Thursday 28th July – Morning Group

Orientation round (Face-to-Face, Mono 0ms, Mono 200ms, Mono 400ms, Spatial 0ms, Spatial 200ms, Spatial 400ms)

1) Face-to-Face
2) Mono 400ms
3) Spatial 400ms
4) Mono 200ms
5) Mono 0ms
6) Spatial 0ms
7) Spatial 200ms

Test 10 – Wednesday 10th August – Morning Group

Orientation round (Face-to-Face, Mono 0ms, Mono 200ms, Mono 400ms, Spatial 0ms, Spatial 200ms, Spatial 400ms)

1) Face-to-Face
2) Spatial 0ms
3) Mono 400ms
4) Spatial 200ms
5) Mono 200ms
6) Spatial 400ms
7) Mono 0ms

Test 11 – Wednesday 10th August – Afternoon Group

Orientation round (Face-to-Face, Mono 0ms, Mono 200ms, Mono 400ms, Spatial 0ms, Spatial 200ms, Spatial 400ms)

1) Face-to-Face
2) Spatial 200ms
3) Spatial 0ms
4) Mono 0ms
5) Spatial 400ms
6) Mono 200ms
7) Mono 400ms

Test 12 – Tuesday 16th August – Evening Group

Orientation round (Face-to-Face, Mono 0ms, Mono 200ms, Mono 400ms, Spatial 0ms, Spatial 200ms, Spatial 400ms)

1) Face-to-Face
2) Spatial 400ms
3) Spatial 200ms
4) Spatial 0ms
5) Mono 400ms
6) Mono 0ms
7) Mono 200ms