A COMPARISON OF ACOUSTIC-PROSODIC ENTRAINMENT IN FACE-TO-FACE AND REMOTE COLLABORATIVE LEARNING DIALOGUES

Nichola Lubold

CIDSE
Arizona State University
Tempe, AZ, USA
nlubold@asu.edu

Heather Pon-Barry

Department of Computer Science
Mount Holyoke College
South Hadley, MA, USA
ponbarry@mtholyoke.edu

ABSTRACT

Today, people are just as likely to have a business meeting remotely as they are face-to-face. Individuals obtain college degrees remotely and sick patients can visit the doctor from home. Especially important in light of this popularity, remote settings are posing communication challenges that are not present in face-to-face settings. Visual cues such as facial expressions and body language are either degraded or nonexistent. In this paper, we are interested in how remote settings affect spoken dialogue when compared to face-to-face settings. We focus on entrainment, a phenomenon of conversation where individuals adapt to each other during the interaction. Specifically, we investigate acoustic-prosodic entrainment, where individuals become more similar in their pitch, loudness, or speaking rate. We explore three different measures of acoustic-prosodic entrainment, comparing remote settings to face-to-face settings on a turn-by-turn basis. Our results indicate that the two settings do differ for different forms of entrainment, suggesting that the presence or absence of visual cues such as facial expressions and body language has an impact on the degree of entrainment.

1. INTRODUCTION

Face-to-face conversations are universal to all human societies, they require no special skills or training, and are the means by which children first acquire language [1]. The characteristics of face-to-face conversations enable participants to communicate not just through their speech but through numerous perceptual cues; speakers can see what the other participant is doing, how they are reacting and whether they are engaged in the conversation.

In the past, face-to-face settings were not only the most basic, fundamental setting of language use, they were the dominant form of communication. With the growth of technology, we are seeing a shift in how people communicate and live. Increasingly, individuals are just as likely to interact *remotely* over a distance as they are face-to-face. Work-at-home opportunities are thriving with co-workers collaborating over video conferencing with simple hand-held mobile devices. Massive open online courses are opening up educational opportunities to an unlimited number of participants through distance education. Unfortunately, with the advantages of remote interactions come the disadvantages of losing the context of face-to-face interaction. The perceptual cues when participants are physically present become stilted or completely unavailable in remote settings.

Given that non-verbal, perceptual cues such as facial expressions are stilted or even absent in many remote settings, we explore in this paper how speech differs from remote to face-to-face settings. We target one specific aspect of spoken dialogue for our comparison, a phenomenon called *entrainment*, also known as adaptation or accommodation. Entrainment, appearing in many dimensions of human-human interaction, occurs when participants adapt to each other during the course of a conversation and begin to act similarly. For instance, in speech, people may adapt their pitch, speaking rate, or loudness. This particular type of entrainment is called *acoustic-prosodic entrainment* and is the form of entrainment we investigate here. Previous works have shown the presence of entrainment in face-to-face settings leads to greater dialogue success and quality [2, 3]. In addition, acoustic-prosodic entrainment has been shown to be positively correlated with certain social behaviors and with learning [4, 5].

While there are many domains in which to analyze the effect of remote settings on entrainment in spoken dialogue, we are interested in supporting remote settings for collaborative distance learning. In this paper, we explore how spoken dialogue changes when people engage in collaborative learning remotely versus face-to-face. With the increasing popularity of online courses, understanding how spoken dialogue differs in remote settings can help guide future applications of spoken dialogue in remote learning scenarios as well as provide direction for supporting face-to-face settings. Using entrainment as a potential indicator of dialogue success, quality, and/or learning, we can design interfaces for both face-to-face and remote settings which utilize the presence or absence of entrainment to assess task success and learning, flag when an intervention may be needed, instigate a shift in the focus of the collaborators, or even suggest a change in partners.

The goal of this paper is to examine differences in *acoustic-prosodic entrainment*, how people adapt their pitch, intensity, or speaking rate, between remote settings and face-to-face settings when people are engaged in a collaborative learning task. We target two questions with this work: (1) we validate that acoustic-prosodic entrainment exists in both settings and (2) we examine how entrainment varies from remote to face-to-face settings, analyzing whether remote settings have more or less entrainment than face-to-face settings.

In Section 2 of this paper, we discuss related work on entrainment in conversational dialogue and comparisons of communication in remote and face-to-face settings. We describe our method for collecting math problem-solving dialogues from pairs of students in distance learning and face-to-face settings in Section 3, and our method for measuring entrainment using acoustic and prosodic features in Section 4. In Section 5, we present a comparison of entrainment in the two experimental conditions: (1) face-to-face interaction, and (2) remote interaction. We discuss our conclusions and future work in Section 6.

2. RELATED WORK

Previous work on how people become more similar to each other during a conversation spans a broad spectrum, including facial expression and gesture [6], text-based word similarity [4], and speaking style [2]. In all of these dimensions, the interactions which took place were face-to-face. To our knowledge, this is the first examination of acoustic-prosodic entrainment in remote versus face-to-face settings.

Approaches for measuring entrainment are varied. Bonin and de Looze [7], Levitan and Hirschberg [8], Thomason et al. [5], and Ward and Litman [9] introduce different methodologies. One parameter that varies is whether to analyze features at the turn-by-turn level or the conversational level (for instance comparing the first half of the conversation to the second half). Our analysis is at the *turn* level. While entrainment has been shown to occur at both the turn level and the conversation level, localized entrainment occurring near turn boundaries and backchannels has been found to be an important co-construction mechanism in social interaction [10, 11]. In addition, we adopt the methodology put forth by Levitan and Hirschberg that identifies three entrainment patterns: proximity, convergence, and synchrony. We consider all three patterns in our analysis of entrainment, and describe these in more detail in Section 4.

In comparing dialogue in remote to face-to-face settings, previous works have examined telephone conversations and video-mediated communication, focusing primarily on characterizing the differences in terms of the dialogue structure. For instance, O'Conaill et al. identified that in remote settings, listeners employed fewer backchannels and interruptions [12]; Clark [1] and Hopper [13] note the lexical, lyrical, and syntactic differences which arise in remote conversations. Several works have found that participants in the remote interaction employed more words and more turns to replace the lack of visual signals [14, 15]. Few works have investigated acoustic-prosodic features and measures of entrainment, which is our focus in this paper.

In hypothesizing how entrainment might differ in remote settings, we consider how previous research on entrainment relates to previous work on the dialogue structure of remote settings. Entrainment has been shown to be positively correlated with number of turns, negatively correlated with latency, positively related to interruptions, and present before backchannels [10, 11]. As remote settings have been shown to have more turns, more latency, fewer interruptions, and fewer backchannels, we might assume that individuals in the remote setting will exhibit less entrainment. Number of turns appears to be the only factor in which there is both a positive correlation with entrainment and support for finding that feature within the remote dialogues. Latency, interruptions, and backchannels all appear to present contrary evidence for entrainment in remote settings. We therefore hypothesize that we will see less entrainment in the remote setting.

3. GENERATING A CORPUS OF COLLABORATIVE PROBLEM-SOLVING DIALOGUES

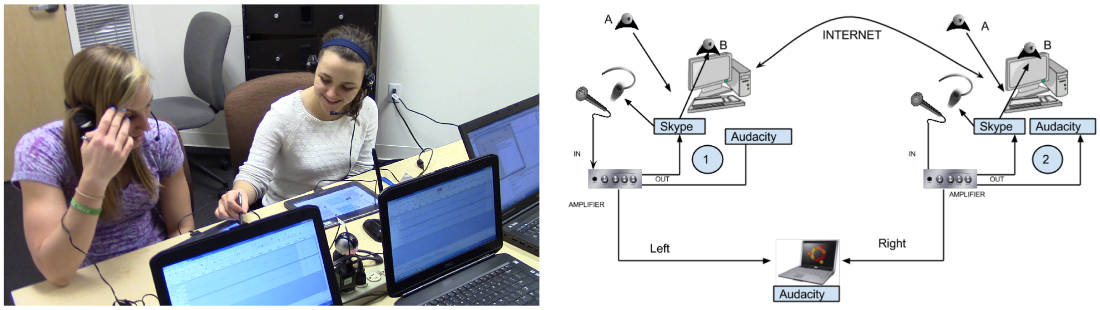
This section outlines our method for collecting a corpus of spoken dialogues in a collaborative, problem-solving setting. A key aspect of the data collection method is that students interact in pairs, in either a *face-to-face* setting or *remote* setting. Other than the interaction distance, all aspects of the experiment procedure are identical in the two settings.

We collect a corpus of 14 spoken dialogues. Each dialogue is between a pair of students working together to solve a mathematical-reasoning problem. Rather than using a pencil and paper, each student uses a tablet-based application called FACT (Formative Assessment with Computation Technologies[1]). The workspace within each tablet is shared to further encourage collaborative interaction and problem solving. The shared workspace enables both students to simultaneously write and see the other's changes, as if they were writing on the same paper at the same time.

The FACT application is designed to support and provide formative assessment for K-12 students solving mathematical problems. The math problems in the application are part of the Mathematics Assessment Project.[2] The math problems are designed with a goal to make knowledge and reasoning visible. They are well-suited for analysis of dialogue; the iterative refinement required to solve a problem generates conversation and drives collaboration. An example dialogue from our corpus is shown below.

[1] `http://fact.engineering.asu.edu/`

[2] `http://map.mathshell.org/materials/index.php`



(a) Face-to-Face Setup (b) Remote Setup

Fig. 1. *Experiment conditions: (a) face-to-face interaction, and (b) remote interaction.*

A: Ohhh . . . negative. Wait, this doesn't help anything
B: Well it's just a bad equation because it's a fraction
A: I clearly can't do this
B: No it's okay we can do it. So y equals 10 minus x . . . I mean negative x plus 10

The participants are undergraduate students with basic knowledge of algebra and geometry. They do not receive any mathematics-based training before the experiment. Descriptive statistics of the corpus are shown in Table 1. We find that the remote setting has on average a higher number of turns; however, the length of turns is on average shorter when compared to the face-to-face setting.

	Remote	**Face-to-Face**
Avg. session length (min)	18.5	16
Avg. number of turns	260.7	175.4
Avg. turn length (sec)	4.3	6.4
Number of dyads	6	8
Male-female dyads	2	4
Female-female dyads	3	1
Male-male dyads	1	3

Table 1. *Corpus of remote and face-to-face dialogues*

3.1. Procedure

Students begin with a 10-minute introductory exercise to ensure they are comfortable using the tablet-interface of the FACT application. In the body of the experiment, the dyads work together to solve two math problems (grade level 9 and above) using the tablets, lasting approximately 20 minutes.

3.2. Interaction Distance

The face-to-face and remote collaboration experiments are designed with the goal of minimizing exogenous factors as much as possible. As shown in Figure 1, in the face-to-face condition, students sit in the same room whereas in the remote condition, students sit in different rooms, connected through audio and video over Skype.

3.3. Audio Recording and Segmentation

We record high-quality audio data using unidirectional microphones with separate audio channels for each speaker. We manually label dialogue turns in the following manner. We identify the beginning of a turn as anytime a participant introduces some verbal articulation and the end of turn as either when the participant ceases that articulation or concludes the overall utterance. Laughter and filled-pauses are included in the turns. Overlapping speech results in overlapping turns.

We further segment each turn into inter-pausal units or IPUs. An IPU is a pause-free unit of speech separated from any other speech by at least 50ms (see [8]). Turns are composed of one or more IPUs. IPUs are the basic unit of prosodic analysis in our measures of entrainment.

4. ACOUSTIC-PROSODIC ENTRAINMENT

This section describes the acoustic-prosodic features utilized for the entrainment metrics and discusses our approach for measuring entrainment.

4.1. Acoustic-Prosodic Features

We examine five acoustic-prosodic features at the IPU level, when measuring entrainment: **intensity**, **pitch (F0)**, **jitter**, **shimmer**, and **speaking rate**.

Intensity is the normalized intensity, and pitch is the fundamental frequency F0. Jitter is defined as the varying pitch and is calculated as the period length deviations; shimmer consists of variations in loudness and is calculated as the amplitude deviations between pitch period lengths. The speaking rate is measured as estimated syllables per second.

For intensity, pitch, jitter and shimmer, we use OpenSmile [16] to obtain two functionals: **mean** and **standard deviation**.[3] For the speaking rate, we apply the approach from de Jong and Wempe, which utilizes peaks and dips in intensity to estimate syllables per second [17].

[3]openSMILE config file is located at `http://www.public.asu.edu/~nlubold/publications/entrainment_config.html`

When comparing speakers with different vocal tracts, we need to ensure that the features affected by the vocal tract lie in the same range. This is primarily an issue with gender so we normalize the female pitch mean and maximum by scaling them to lie in the same range as the male values; all other non-pitch features are raw.

4.2. Measuring Entrainment

In this paper, we focus on a form of local entrainment looking at how close speakers are to each other at the turn-level. We explore three different patterns of entrainment: *proximity*, *convergence*, and *synchrony*. Following Levitan and Hirschberg [8], we run a series of statistical tests to determine significant acoustic features for these measures.

Proximity is the similarity between two speakers' acoustic-prosodic features at each turn boundary. We calculate proximity by finding the difference between each speaker and their partner at each adjacent IPU, as in equation 1. We then compare how close the speakers are to each other at each turn as compared to the rest of the conversation, using equation 2 to calculate the average difference between each speaker and their partner at 10 other non-adjacent IPUs.

$$\Delta adjacent = |IPU_{a(i)} - IPU_{b(i)}| \quad (1)$$

$$\Delta other = \frac{\sum_{j=1, j \neq i}^{10} |IPU_{a(i)} - IPU_{b(j)}|}{10} \quad (2)$$

We identify whether speakers are entraining by proximity for a particular acoustic feature by comparing the similarities between partners at each turn, $\Delta adjacent$, to the similarities between partners at non-adjacent turns across the corpus, $\Delta other$. If these two similarities are significantly different, we consider there to be localized proximity on that feature.

Convergence is the degree to which speakers become more similar to each other over the course of the entire conversation. If convergence exists, the difference in the two speakers' acoustic-prosodic feature values will reduce over time. If there is divergence, the difference in the two speakers' feature values will increase. We consider there to be convergence on an acoustic-prosodic feature if we find a correlation between time and the absolute difference of a speaker and their partner at adjacent turns.

Synchrony is the quality of interaction which occurs when speakers stay "in sync" as they converse. Speakers may have distant acoustic-prosodic feature values, but as they converse, they modulate these values in tandem. To find synchrony, we look for correlations between the two speakers' raw feature values at adjacent turns, considering synchrony to be present if we find significant correlations.

For all three tests, we follow Levitan and Hirschberg [8] in considering results with $p < 0.01$ to be statistically significant and the results with $p < 0.05$ to approach significance.

5. RESULTS AND ANALYSIS

In this section we first verify there is acoustic-prosodic entrainment in both the remote and face-to-face dialogues using the three measures of entrainment discussed in 4.2. We then analyze the observed entrainment in terms of our hypothesis that dyads in the remote setting will entrain less than dyads in the face-to-face setting. We present results at the group level (Section 5.1) and at the dyad level (Section 5.2),

5.1. Entrainment at the Group Level for Remote versus Face-to-Face Settings

Looking across the entire set of dialogue for each setting, we examine each type of entrainment separately, exploring how each measure either supports or refutes our hypothesis that remote settings result in less entrainment. We find that for both proximity and synchrony, our hypothesis is not supported; the two settings are actually more similar than different in the amount of entrainment for these two measures. However, for convergence, there is a distinct difference between the two settings in support of the hypothesis. The results for those features which we found to be significant for each measure across both settings are depicted in Table 2.

For **proximity**, we see evidence of significant entrainment in both conditions for nearly all features. This means that on a local, turn-by-turn basis, speakers are staying relatively close to each other in both the face-to-face setting and the remote setting in their intensity, pitch, shimmer mean, and speaking rate. Based on this initial analysis, we do not find a salient difference in proximity between the two groups, contradicting our hypothesis that we would observe less entrainment in the remote setting.

In contrast, **convergence** is almost non-existent in the remote setting; intensity mean is the only feature for which convergence is present. The correlation is negative and relatively weak, with $r = -0.06$. In the face-to-face setting, there are several features for which we observe convergence, including pitch, jitter, and shimmer. Taking into account that remote settings generally have fewer backchannels and that speakers' converge on pitch and intensity before backchannels [11, 12], these results adhere to previous findings and imply that convergence in remote settings differs in comparison to face-to-face settings. Face-to-face settings may be more conducive to this type of entrainment which supports our hypothesis.

For **synchrony**, we see significant entrainment in both settings. The correlation coefficients in the face-to-face setting are slightly stronger for most features. In the remote setting the correlations for pitch, while weak, are negatively correlated. This suggests that instead of adapting pitch in the same direction as their partner, individuals may be moving in the opposite direction. These results do not allow us to accept our hypothesis. As with proximity entrainment, we find that, for synchronous entrainment, the face-to-face and remote settings are more similar than different.

	Proximity (t)		Convergence (r)		Synchrony (r)	
	Remote	*F2F*	*Remote*	*F2F*	*Remote*	*F2F*
Intensity Mean	10.23^*	12.09^*	-0.06	-	0.11	0.19^*
Intensity Std Dev	10.14^*	14.41^*	-	-	0.15	0.11^*
F0 Mean	4.81	5.28^*	-	0.11^*	-0.01^*	0.11^*
F0 Std Dev	12.47	2.54	-	0.09^*	-0.08^*	0.10^*
Jitter Mean	4.24	-	-	0.23^*	0.05	-
Jitter Std Dev	2.38	-	-	0.12^*	-	-
Shimmer Mean	9.35	10.50^*	-	-0.06	-	-
Shimmer Std Dev	-	2.89^*	-	-	-	-
Speaking Rate	9.35^*	10.25^*	-	-	-	-

Table 2. *Significant acoustic-prosodic features of three measures of entrainment for remote and face-to-face settings (F2F). Proximity is found using a paired t-test; Convergence and Synchrony are found Pearson's correlation coefficient in a two-tailed t-test. All values shown are significant at $p < 0.05$; values with $p < 0.01$ are marked with an asterisk (*).*

5.2. Entrainment at the Dyad Level

In Section 5.1, we examined entrainment at the group level. However, we believe that within groups, each dyad might exhibit different entrainment patterns for social and interpersonal reasons beyond our control. In this section, we report on the same three measures of entrainment at the *dyad* level.

In the remote setting, all dyads exhibit some degree of *convergence*. While convergence is less significant across the entire set of remote dialogues, it is actually a common form of entrainment within the dyads. In the face-to-face setting, only half of the dyads exhibit convergence. These dyads exhibit strong convergence, while the other half of the face-to-face dyads do not exhibit any convergence. These results do not support the hypothesis that there is less entrainment in remote settings, conflicting with our group-level findings.

To better understand the dyad-level results, we summarize in Table 3 the three most significant acoustic-prosodic features for the dyads in each setting, for each measure of entrainment. For proximity, entrainment in jitter and pitch is more frequent for remote dyads, whereas intensity entrainment is more frequent for face-to-face dyads. For convergence, the remote dyads exhibit the most entrainment for pitch and intensity whereas the face-to-face dyads exhibit the most entrainment for intensity and jitter. For synchrony, intensity entrainment is the most common for both groups. The implications of these observations are that we must consider variances between dyads if we are to develop applications that utilize entrainment to analyze dialogue.

Remote	**Face-to-Face**
Proximity (t)	
Jitter mean (7.03)	Intensity std. dev. (2.87)
Pitch mean (6.52)	Intensity mean (4.08)
Intensity mean (4.95)	Pitch mean (5.76)
Convergence (r)	
Pitch mean (-0.13)	Intensity std. dev. (-0.08)
Intensity mean (-0.21)	Jitter mean (-0.05)
Jitter mean (-0.01)	Shimmer mean (-0.09)
Synchrony (r)	
Intensity mean (0.20)	Intensity mean (0.20)
Intensity std. dev. (0.14)	Pitch std. dev (0.23)
Speaking Rate (-0.24)	Pitch mean (0.21)

Table 3. *Top three features per measure and the average measure across the dyads with that significant feature.*

6. CONCLUSIONS

The goal of this paper is to investigate how collaborative learning dialogues differ between remote and face-to-face settings using three measures of acoustic-prosodic entrainment. Targeting two questions, we (1) validate that acoustic-prosodic entrainment is present in both settings and (2) compare how acoustic-prosodic entrainment differs between the two settings for the three measures of entrainment. We assess how entrainment differs by validating the results against our hypothesis that entrainment is less present in remote settings.

Examining dialogues at the group level, we observe significant entrainment in both settings. We find that *proximity* and *synchrony* are present in both settings to a significant degree. In terms of convergence, only the face-to-face setting exhibits a significant amount of entrainment. Thus, our data supports the hypothesis that people will entrain less in the remote setting for *convergence* but not for proximity and synchrony. This suggests that convergence of acoustic-prosodic speaking styles may be harder to achieve in distance learning environments.

On the other hand, examining entrainment at the dyad level, our results indicate that dyads often entrain in different ways than the whole group. For convergence in particular, we feel that further analysis, ideally on a larger data set, is needed. To incorporate the variance of within- and between-dyad comparisons, future work will include performing a mixed model analysis to explore how the differences in setting affect acoustic-prosodic entrainment with the added complexity of the dyads. An alternative hypothesis to consider is that the absence of face-to-face body language signals might actually bolster acoustic-prosodic entrainment in remote interactions.

One of the challenges in conducting this work is how to maintain the quality of the audio in the remote setup considering issues such as packet loss, latency, and jitter over the network. The participants were connected over a gigabyte network and while perceptions of the quality were that it was acceptable, an area of future work may include testing for the effect of different quality of service and quality of experience measures on the amount or degree of entrainment.

7. ACKNOWLEDGMENTS

This work is supported in part by the Ira A. Fulton Schools of Engineering through a Dean's Fellowship. We thank Kurt VanLehn, Sree Aurovindh Viswanathan, Sujith Gowda and the rest of the FACT team, for their help and support in the data collection. We thank Arun Reddy Nelakurthi for guidance in the dialogue setup and preparation, and Erin Walker for providing insightful feedback in the writing process.

8. REFERENCES

[1] Herbert H Clark, *Using language*, Cambridge University Press, 1996.

[2] David Reitter and Johanna D Moore, "Predicting success in dialogue," in *Annual Meeting-Association for Computational Linguistics*, 2007, vol. 45, p. 808.

[3] Tanya L Chartrand and John A Bargh, "The chameleon effect: The perception–behavior link and social interaction.," *Journal of personality and social psychology*, vol. 76, no. 6, pp. 893, 1999.

[4] Heather Friedberg, Diane Litman, and Susannah BF Paletz, "Lexical entrainment and success in student engineering groups," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 404–409.

[5] Jesse Thomason, Huy V Nguyen, and Diane Litman, "Prosodic entrainment and tutoring dialogue success," in *Artificial Intelligence in Education*. Springer, 2013, pp. 750–753.

[6] Jessica L Lakin, Valerie E Jefferis, Clara Michelle Cheng, and Tanya L Chartrand, "The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry," *Journal of nonverbal behavior*, vol. 27, no. 3, pp. 145–162, 2003.

[7] Francesca Bonin, Céline De Looze, Sucheta Ghosh, Emer Gilmartin, Carl Vogel, Anna Polychroniou, Hugues Salamin, Alessandro Vinciarelli, and Nick Campbell, "Investigating fine temporal dynamics of prosodic and lexical accommodation," in *INTERSPEECH*, 2013.

[8] Rivka Levitan and Julia Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proceedings of Interspeech*, 2011.

[9] Arthur Ward and Diane J Litman, "Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora.," in *SLaTE*, 2007, pp. 57–60.

[10] Rivka Levitan, Agustín Gravano, Laura Willson, Stefan Benus, Julia Hirschberg, and Ani Nenkova, "Acoustic-prosodic entrainment and social behavior," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2012, pp. 11–19.

[11] Rivka Levitan, Agustín Gravano, and Julia Hirschberg, "Entrainment in speech preceding backchannels," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, 2011, pp. 113–117.

[12] Claire O'Malley, Steve Langton, Anne Anderson, Gwyneth Doherty-Sneddon, and Vicki Bruce, "Comparison of face-to-face and video-mediated interaction," *Interacting with computers*, vol. 8, no. 2, pp. 177–192, 1996.

[13] Robert Hopper, *Telephone conversation*, vol. 724, Indiana University Press, 1992.

[14] Abigail J Sellen, "Remote conversations: The effects of mediating talk with technology," *Human-computer interaction*, vol. 10, no. 4, pp. 401–444, 1995.

[15] Gwyneth Doherty-Sneddon, Anne Anderson, Claire O'Malley, Steve Langton, Simon Garrod, and Vicki Bruce, "Face-to-face and video-mediated communication: A comparison of dialogue structure and task performance.," *Journal of Experimental Psychology: Applied*, vol. 3, no. 2, pp. 105, 1997.

[16] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[17] Nivja H de Jong and Ton Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.