Do We Hyperarticulate on Zoom?

Sam O'Connor Russell, Ayushi Pandey, Naomi Harte

Sigmedia Lab, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland
{russelsa,nharte}@tcd.ie

Abstract

Hyperarticulation is a form of speech which helps overcome multimodal impediments to communication. However, it can degrade the performance of automatic speech recognition (ASR). Videoconferencing is in widespread use and is often supported by ASR for captioning and diarisation. Hence, there is a need to understand the nature of speech production in videoconferencing. We ask whether 'Zoom speech' - characterised by increased pauses and formality - is hyperarticulation. We conduct a comparative study of in-person and Zoom conversational interactions. We find some but not all features of classic hyperarticulation in Zoom interactions. Consistent with hyperarticulation we find more pauses, longer vowels, and an increased F0. Changes to the articulation rate, F0 range and vowel space are not consistent with hyperarticulated speech. To the best of our knowledge, our work is the first to assess videoconferencing speech for the presence of hyperarticulation. We discuss whether videoconferencing merely disrupts interaction, or induces an atypical form of multimodal hyperarticulation.

Index Terms: human-computer interaction, Lombard speech, hyperspeech, videoconferencing

1. Introduction

Speakers hyperarticulate to overcome impediments to communication, e.g. when addressing a hearing-impaired listener [1]. This involves changes to speech [2, 3] and nonverbal communication [4] which increase intelligibility [5, 3]. Speech-based changes result in hyperspeech (also called clear speech [2]), characterised by a higher F0, slower speaking rate and an expanded vowel space [1, 6, 2]. Visually, mouth and head movements become intensified [7, 8]. Hyperarticulation benefits listeners and is equivalent to a 3-6 dB reduction in noise [5].

Automatic speech recognition (ASR) systems are generally not robust to changes in human speech production. They are typically trained on speech recorded in noise-free conditions with noise mixed into the signal in post-processing [9]. Thus, hyperspeech is not encountered during training [9, 10, 11] and performance is degraded when hyperspeech is encountered by the trained model [9, 11]. Including hyperspeech during training can improve performance [9, 10]. Hence, the detection of hyperarticulation is an active area of research [11]. While having a detrimental effect on ASR, visual speech recognition experiences a performance boost in the presence of hyperarticulation [10, 9].

Lindblom proposed the hyper/hypo (H&H) theory explaining the mechanism behind hyperarticulation [12]. Hyperarticulation is an effortful form of communication. As such, it is only produced when necessary. Speakers continually monitor how well they are understood, making dynamic adjustments to their articulation as required [12]. A well-known type of hyperspeech is Lombard speech which occurs in the presence of background noise [13]. However, hyperspeech is more general. It can occur whenever a speaker perceives an issue with communication, e.g. when noise is experienced by the listener only [2]. Recent work by Trujillo et al. demonstrated a visual trigger of acoustic hyperspeech. Trujillo et al. used screens and cameras to connect interlocutors [4]. Interlocutors produced hyperspeech when visibility decreased by way of an artificial visual blur and audio quality remained optimal [4]. Thus, hyperspeech is triggered by both visual and acoustic disturbances.

Could hyperarticulation occur on Zoom? Meetings are increasingly occurring remotely, often supported by ASR for record-keeping and live captioning. However, modern videoconferencing (VC) platforms including Zoom (which has become synonymous with VC) and Microsoft Teams present a number of challenges to interaction [14, 15, 16]. VC interaction is characterised by heightened formality, lack of spontaneity and interlocutor fatigue [14, 15, 17, 18]. The human-computer interaction literature considers this the result of the intrinsic constraints of VC [14, 15]. Nonverbal communication is limited to yes/no nods and shakes of the head [16]. Mutual gaze cannot be achieved [19, 20] despite being critical to fluid interaction [21]. Zoom has an inherent time-varying latency, disrupting the rhythm of conversation [22]. However, to the best of our knowledge, no study has considered the possibility that 'Zoom speech' is hyperarticulation. Therefore, it is unknown if the deployment of state-of-the-art ASR (the performance of which is typically *not* robust to hyperarticulation) in VC-mediated interactions is inherently problematic.

In this paper, we present a novel perspective on VC-mediated interaction. We ask: *do we hyperarticulate on Zoom*? Our argument is as follows. Hyperspeech is triggered when an interlocutor perceives multimodal disruptions to communication. VC presents a number of multimodal disruptions to human interaction. Thus, consistent with H&H theory [12], we argue that a VC speaker responds by producing hyperspeech. We examine speech produced both in-person and in VC for 5 classic features of acoustic hyperspeech. We use two publicly-available corpora of VC and in-person conversational interactions: the RoomReader [23] and Multisimo corpora [24]. The RoomReader corpus was recorded over Zoom whilst the Multisimo corpus was recorded in-person. The corpora share the same task which prompts spontaneous interaction, making the dialogues similar. We demonstrate that VC speech is partially consistent with hyperspeech. We find an increased articulation rate in the VC dialogues; the opposite of what is typically observed in hyperspeech. We find more frequent pauses, longer vowels and an elevated F0; which are all characteristics of hyperspeech [3, 5]. We discuss whether these transformations are an atypical form of multimodal hyperarticulation, concluding that the answer to the question posed in this paper is nuanced.

2. Eliciting and Quantifying Hyperspeech

The audio features of hyperspeech have been consistently reported across studies [2]. Hazan and Baker [2] note that compared with normal speech, hyperspeech is characterised by: a higher median F0 and a higher F0 range [1, 6, 2]; a reduced speaking rate [6]; more frequent and longer pauses [1, 6]; an expanded vowel space [6]; longer vowels [25]; and higher energy in the 1-3kHz range [26, 2].

A number of experimental approaches to elicit hyperspeech have been reported. Picheny et al. asked 3 speakers of English to read sentences normally and then as if they were addressing a hearing-impaired individual [1]. They found hyperspeech features in the latter, including a slower speaking rate and longer pauses [1]. In a separate study, they found that this resulted in increased intelligibility [3]. Hyperspeech has also been found in spontaneous communication [2]. Hazan and Baker recruited 20 male and 20 female normal-hearing speakers of English who conducted a spot-the-difference task under two conditions: one noise-free and another in a setting where the listener experienced auditory noise [27]. This induced hyperspeech in the speaker's voice, characterised by an increased median F0, increased energy and slower speaking rate [2]. Hyperspeech exhibits invariance to a speaker's native (L1) / non-native (L2) status. Smiljanić and Bradlow compared English and Croatian hyperspeech with 20 L1 English and 30 L1 Croatian speakers [6]. They found hyperspeech in both languages is very similar [6]. Granlund et al. compared the characteristics of English language hyperspeech produced by L1 and L2 speakers [28]. Twelve native Finnish speakers were recruited who spoke English with Finnish accents. The native Finnish speakers produced English hyperspeech with similar characteristics to L1 English speakers [28]. For speech rate, Duran found that fluent non-native speakers' pause behaviour reflects their speaking style in their native language and articulation rate does not indicate fluency [29].

Hyperspeech has been recently studied in a multimodal setting. Trujillo et al. asked a pair of Dutch-speaking participants standing opposite one another to communicate an action in any way they wished, for example tracing an object [8]. 20 male and 32 female participants were recruited. Noise emitted through the headphones increased hand and mouth movements and the intensity of speech [8]. A second study by Trujillo et al. (described in Section 1) induced a visual blur finding similar changes to hand gesture, body movement and speech [4]. To the best of our knowledge, the characteristics of speech produced in conversational VC interaction have not been explored. VC was used in a recent hyperspeech study by Pham and Karuza [30]. However, this experimental design is not applicable to our work. In their study, VC was used as a means to elicit hyperspeech by exposing participants to acoustic noise [30].

3. Corpora

There are few suitable publicly available corpora to aid this study. Corpora in the hyperspeech literature involve an experimental design that introduces a disruption to communication; e.g. the LUCID corpus released by Hazan and Baker induces acoustic noise [27]. Such corpora are unsuitable for our analysis; we do not wish to study hyperspeech produced in response to noise, a well-established phenomenon. Instead, we require everyday VC mediated and in-person interactions to conduct a comparative analysis. We choose to make use of publicly available datasets. The **RoomReader (RR)** [23] corpus is one of only publicly available corpus of conversational VC interaction, recorded over Zoom [31]. The corpus follows a scenario that is led by a tutor which encourages spontaneous, unscripted collaboration and interaction. The **Multisimo (MM)** [24] corpus uses an identical scenario and was recorded in-person, enabling us to conduct a comparison of speech production in the two corpora.

The RR [23] and MM [24] corpora each consist of n=30 and n=18 English language, small-group, interactions. In the MM corpus, participants sat around a table and audio and video recordings were captured. In the RR corpus, participants joined a Zoom [31] call using a personal device from a quiet location of their choosing. Audio and video recordings were captured. The mean session duration in the RR corpus is 17 minutes and 9.1 minutes in the MM corpus. Both corpora are provided with transcriptions and diarised audio. There are 4-5 (mode=4) participants in each session of the RR corpus and 3 participants in each session of the MM corpus. A **scenario prompts spontaneous conversational interaction**. 1 participant assumes the role of a tutor (T) who asks the student (S) participants to name and rank by their popularity the top three most popular answers to a question previously asked to 100 people. An example is ***”Name an instrument in a symphony orchestra”***. In total, there are 115 participants in the RR corpus (n=50 male, n=65 female) and 56 participants in the MM corpus (n=20 male, n=34 female). All participants are fluent speakers of English [24, 23] of which n=89 (n=37 male, n=52 female) are native speakers in the RR corpus and n=12 (n=8 male, n=4 female) in the MM corpus. Irish-English[1] is the dominant dialect of native speakers in both corpora n=63 (n=24 male, n=39 female) participants in the RR corpus and n=9 (n=5 male, n=4 female) participants in the MM corpus. There are 2 T participants in the RR corpus (n=1 male, n=1 female) both of whom are native Irish-English speakers and each partake in 15/30 of the sessions. There are 3 T participants in the RR corpus (n=3 female) all of whom are native Greek speakers fluent in English. Each tutor partakes in 7, 9 or 2 of the 18 sessions. The mean speaker age is 23 (std 5.2) in the RR corpus and 30.5 (std 6.8) in the MM corpus.

4. Methodology

We select 5 audio features of hyperspeech to explore. Due to differences in recording conditions between the RoomReader (RR) and Multisimo (MM) corpora, we exclude energy-based metrics that use the amplitude of the speech waveform. We use five phoneme- and frequency-based metrics: **1) vowel duration, 2) articulation rate, 3) pause frequency and duration, 4) F0 median and range, and 5) vowel space**.

We use the Montreal Forced Aligner (MFA) [32] to obtain phoneme boundaries. An inspection revealed accurate and precise boundaries. We identify instances of primary and secondary stress of the 8 English monophthongs /æ, a, ɔ, ɛ, i, ʊ, u/ (i.e the vowels in *had, hard, hod, head, hid, heed, Hudd and who'd* respectively) [33]. We collapse each form of stress into a single category of **vowel**. We compare the median **vowel duration** of the two corpora [25]. We restrict our analysis to student (S) participants, as tutor (T) participants appear in multiple sessions. We consider gender-based differences as Alghamdi et al. reported larger vowel lengthening in female hyperspeech [25].

We define the **articulation rate** as a participant syllable count ($nsyll$) divided by speaking time in seconds ($lenS$): $articulation\_rate = nsyll/lenS$. We use the automated method of de Jong et al. to estimate syllable locations [34]. We define a **pause** as the silence between adjacent phonemes from the same speaker. We use end and start times to compute **pause**

[1] Differences in demographic data mean we identify speakers with Irish-English dialects as native speakers who are also Irish nationals in MM and native speakers who have lived in Ireland for 5 years in RR

length. We set a minimum duration of a pause to 200 milliseconds [35]. We define **pause rate** as the number of pauses in a speaker dialogue (n_pause) divided by total speech time in minutes ($lenM$): $pause\_per\_min = n\_pause/lenM$.

We use Praat [36] to compute fundamental frequency **F0** at the midpoint of each vowel [6], reducing contextual effects. We use a 10 millisecond width, a 70 Hz pitch floor and a 500 Hz ceiling. We compute the **median F0** for each speaker: $medianF0 = median(F0_i)_{i=1}^{N}$ for all the N measurements of F0 from a speaker. We define the **F0 range** as the interquartile range (IQR) of all F0 values of a speaker: $range(F0) = IQR(F0_i)_{i=1}^{N}$ [2]. Each T participant occurs in multiple sessions. This leads to far more data for T participants than for S participants. Hence, any difference between the median F0 could originate from differences between the tutors' F0 rather than a hyperspeech effect. To avoid this, we restrict our analysis of F0 to S participants. To investigate the **vowel space**, we compute F1 and F2 for each vowel at the midpoint using Praat (Burg estimation). We use a 25 millisecond window, a pre-emphasis filter with a +3dB point at 50 Hz, and maximum formant ceilings of 5000 Hz for males and 5500 Hz for females [37]. Thus far, we have not made a distinction between native and non-native speakers due to the similarities of L1 and L2 hyperspeech (Section 2). For the formant analysis we select only groups of male and female Irish-accented speakers, as accent impacts formant values [33]. Each vowel provides an estimate of F2 and F1 which we analyse as a **vowel space** (i.e. a 2D Cartesian plane) [6, 1]. In hyperspeec,h the vowel space is expanded which Smiljanić and Bradlow quantify in terms of the *vowel space area* and *vowel space dispersion* statistics [6]. We define the *vowel space area* of a speaker as the area in Hz^2 of the polygon formed by the median F2 and F1 values [6]. We define the *vowel space dispersion* of a speaker as the average of all Euclidean distances between each F2 F1 estimate pair and the centroid of the polygon formed by all F2 and F1 values [6].

We remove statistical outliers from the MM and RR corpora using the definition $outlier \notin [-1.5Q_1, 1.5Q_3]$ where Q_1 and Q_3 are the first and third quartiles. Upon analysis, features (e.g. pause duration) exhibit non-normality as evidenced by the Shapiro-Wilk test. Thus, we use the Mann-Whitney U (MWU) non-parametric statistical significance test to compare differences in the median of the two groups. This test is appropriate for both normal and non-normally distributed statistics. We present the median statistic in the form $median(Q_1, Q_3)$.

5. Results

5.1. Vowel duration

The median vowel duration in the videoconferencing (VC) mediated RoomReader (RR) corpus is longer than in the in-person Multisimo (MM) corpus. There are 5421 vowels in the RR corpus and 3156 in the MM corpus produced by student (S) participants. The median duration of a vowel is 70 (40, 110) and 60 (40, 100) milliseconds in the RR and MM corpora (Table 1). There is a significant increase of 10 milliseconds in median vowel duration in the RR corpus (MWU $40 \times 10^6 p < 0.001$). Males have a median vowel duration of 70 (40,100) milliseconds in the RR corpus and 60 (40,100) in the MM corpus which is a significant difference (MWU $10 \times 10^6 p < 0.001$). Females exhibit a larger increase in median vowel duration in the RR corpus 80 (50,120) milliseconds than in the MM corpus at 60 (40,109) milliseconds (MWU $97 \times 10^5 p < 0.001$). Thus we find evidence of vowel elongation in the RR corpus **consistent with the presence of hyperspeech**.

Table 1: *Median, first and third quartiles of vowel duration for student participants in the RoomReader and Multisimo corpora*

	Vowel duration [milliseconds]					
	RoomReader			**Multisimo**		
Participants	**Median**	**Q₁**	**Q₃**	**Median**	**Q₁**	**Q₃**
All S	70	40	110	60	40	100
Male S	70	40	100	60	40	100
Female S	80	50	120	60	40	109

5.2. Articulation rate

The articulation rate is faster for S participants in the RR corpus at 4.62 syllables per second in the RR corpus compared with 4.24 syllables per second in the MM corpus (MWU $U = 20 \times 10^5, p < 0.001$, Table 2) equivalent to a 7% increase. We find the same trend for tutor (T) participants. The median articulation rate for the T participants is 4.93 syllables per second compared with 3.97 syllables per second in the MM corpus (MWU $U = 22 \times 10^5, p < 0.001$, Table 2). This is equivalent to a 24% increase in the articulation rate. The increased articulation rate in the RR corpus is an unexpected finding as it is **inconsistent with classic hyperspeech**.

5.3. Pause frequency and duration

Pauses are more frequent in the RR corpus for S participants at 11.52 and 7.02 pauses per minute in the RR and MM corpora respectively (Table 2, MWU $U = 20. * 10^5, p < 0.001$). This is equivalent to a 64% increase. We find a more dramatic increase in the median pause rate for T participants. Their median pause rate is 18.21 pauses per minute in the RR corpus which is over double the median 7.98 pauses per minute in the MM corpus (Table 2, MWU $U = 22. \times 10^5, p < 0.001$). The increased frequency of pauses is **consistent with the presence of hyperspeech** in the RR corpus dialogues. We find differing trends in the median pause duration for S and T participants in the two corpora. We find that S participants in the RR corpus have a median pause duration which is 40 milliseconds less than S participants in the MM corpus (Table 2, MWU $U = 20 \times 10^5, p < 0.001$). This is inconsistent with the presence of hyperspeech. However, we find the reverse trend for T participants: in the RR corpus the median pause duration is 160 milliseconds greater than in the MM corpus (Table 2, MWU $U = 20.2 \times 10^5, p < 0.001$), a finding consistent with the presence of hyperspeech.

Table 2: *Median, first and third quartiles of the articulation rate, pause rate and duration in the RoomReader (RR) and Multisimo (MM) corpora for tutors (T) and students (S).*

	RoomReader			**Multisimo**		
Statistic	**Median**	**Q₁**	**Q₃**	**Median**	**Q₁**	**Q₃**
Articulation rate [nsyll/s] - S	4.62	4.46	4.86	4.24	3.80	4.66
Articulation rate [nsyll/s] - T	4.93	4.47	4.99	3.97	3.83	4.10
Pause rate [npause/min] - S	11.52	8.85	15.73	7.02	4.98	12.30
Pause rate [npause/min] - T	18.21	17.16	21.14	7.98	7.37	10.30
Pause duration [ms] - S	480	310	880	520	320	940
Pause duration [ms] - T	600	370	1060	440	310	750

5.4. F0 median and range

The median F0 is higher for both males and females in the RR corpus. The median F0 for male S participants in the RR corpus is 124 Hz and 127 Hz in the MM corpus (MWU $U = 11 \times 10^5, p < 0.001$). For female S participants, the median F0 is higher in the RR corpus than in the MM corpus at 189 Hz and 198 Hz respectively (MWU $U = 14 \times 10^6, p < 0.001$). This finding is **consistent with the presence of hyperspeech**. We compare the range of F0 values, finding no evidence of a statistically significant difference in the median F0 range for male (118.7 vs 129 Hz in MM and RR; MWU $U = 460, p = 0.97$) and female (189.2 and 287 Hz in MM and RR; MWU $U = 528, p = 0.4$) S participants in the two corpora. This finding is **inconsistent with the presence of hyperspeech**.

5.5. Vowel space

The median number of vowels (and hence estimates of F1 and F2) of each type produced by S participants is 11 (6, 20) and 22 (13, 31) in the MM and RR corpora. We find no evidence of a change in the median vowel space area of male speakers: 198×10^3 (130×10^3, 318×10^3) Hz^2 in RR vs. 270×10^3 (230×10^3, 359×10^3) Hz^2 in MM (MWU $U = 30.5, p = 0.264$). We also find no evidence of a change in the vowel space dispersion: 409 (349, 465) Hz in RR and 408 (318, 604) Hz in MM (MWU $97.5, p = 0.167$). We also find no evidence of a difference for female speakers. This is **inconsistent with the presence of hyperspeech**. Male and female vowel space plots are omitted due to space limitations.

6. Discussion

In our comparative analysis of Zoom and in-person interactions, we found an increased pause rate, vowel length and median F0 - all consistent with hyperspeech - in videoconferencing (VC) interactions. However, we found no change in the F0 range nor the vowel space; usually apparent in hyperspeech. Moreover, our finding of an increased articulation rate is the opposite of what is typically observed in hyperspeech.

Have we failed to identify hyperspeech on Zoom? Firstly, we must consider the body of work establishing the characteristic features of hyperspeech. These are studies which considered acoustic disruptions to communication (Section 2), whereas VC presents a range of auditory and visual disruptions. Some disruptions are static e.g. the absence of mutual gaze [19, 20] and others vary e.g. the duration of latency [22]. Given this complexity, we begin to understand that the answer to the question posed in this paper is not clear-cut. We have clearly demonstrated alterations to speech production in VC. According to the Human-Computer Interaction (HCI) literature, these transformations merely reflect the disruptive effects of VC [14, 15, 16]. From this perspective, increased pauses can be explained by the poor turn-taking and hesitancy to speak which plague VC-mediated interaction [14, 15, 18]. However, we have found ample evidence that the pause frequency, pause duration (for the tutors) and vowel length are *all* altered over VC in a manner consistent with hyperspeech [3, 5]. Thus, jointly considering the nature of VC interaction and recent work establishing the multimodal nature of hyperarticulation [4], we now wonder whether VC induces a unique form hyperarticulation only partially consistent with the literature.

What do our findings mean for automatic speech recognition (ASR)? Any deployment of ASR in VC must take into account the potential impact of changes in speech production on ASR performance as clearly, VC impacts speech production. However, it is less clear if 'Zoom speech' fits into the well-understood category of hyperspeech. Our work establishes a basis for future studies developing our understanding of multimodal interaction in VC. In particular, our finding of an increased speech rate demands further investigation. The optimal speech rate for VC may be the opposite of what is usually observed in hyperspeech. Another future study could compare head and lip movements in across VC and in-person settings as they are intensified during hyperarticulation [4, 8].

We chose to use publicly available corpora of natural unconstrained human interaction (Section 3), motivated by our study of hyperspeech in natural communication. However, this imposes limitations on our findings. Firstly, there are limited numbers of speakers to compare (Section 3); in particular for the vowel space where only native English speakers were used (Section 5). Secondly, prior studies compare the same speakers in two communication settings, for example, noise-free and with background noise (e.g. [2]). In our study, the corpora contain different sets of participants. Hence, there is a possibility that natural variation in the speech production of participants influences the result, and observed differences are not due to the medium of VC. However, we note that our study contains similar numbers of participants to prior work (Section 2) and that the Zoom interactions exhibit an increase in the frequency of pauses consistent with prior observations in the HCI literature. Other studies have considered carefully controlled settings, removing these confounding factors [2]. Such designs could be considered in future experiments comparing in-person and VC communication, which would necessitate data collection. However, there is an inevitable trade-off between constraining the interaction (or the recording setup) and obtaining natural spontaneous speech. Nevertheless, our work forms an initial exploration of speech production in VC, leveraging public data of unconstrained everyday conversational interactions.

7. Conclusion

Online meetings are now an everyday occurrence. However, the automatic speech recognition (ASR) systems which often support them are not robust to changes in speech production. Hence there is a need to understand the nature of speech production in videoconferencing (VC). We conducted a comparative study of comparable in-person and VC mediated conversational interactions. To the best of our knowledge, our study is the first to consider if the changes to interaction induced by VC are hyperspeech. We found that VC speech is different to in-person speech. It exhibits many of the characteristic features of classical hyperspeech such as more frequent pauses. However, our finding of an increased articulation is the opposite of what is typically observed in hyperspeech. Hence, our work raises important questions about the nature of VC speech. Given the lack of robustness of modern ASR to changes in speech production, our work highlights the need to further understand and categorise 'Zoom speech'. Considering the emerging understanding of hyperarticulation as a multimodal phenomenon with visual triggers, we argue that an atypical form of hyperspeech may be produced by interlocutors in VC-mediated interactions.

8. Acknowledgements

This research was conducted with the financial support of Science Foundation Ireland at ADAPT, the SFI Research Centre for AI at Trinity College Dublin [13/RC/2106_P2].

9. References

[1] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing ii: Acoustic characteristics of clear and conversational speech," *Journal of Speech, Language, and Hearing Research*, vol. 29, no. 4, pp. 434–446, 1986.

[2] V. Hazan and R. Baker, "Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. 2139–2152, 2011.

[3] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing i: Intelligibility differences between clear and conversational speech," *Journal of Speech, Language, and Hearing Research*, vol. 28, no. 1, pp. 96–103, 1985.

[4] J. P. Trujillo, S. C. Levinson, and J. Holler, "A multi-scale investigation of the human communication system's response to visual disruption," *Royal Society Open Science*, vol. 9, no. 4, p. 211489, 2022.

[5] S. Liu and F.-G. Zeng, "Temporal properties in clear speech perception," *The Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 424–432, 2006.

[6] R. Smiljanić and A. R. Bradlow, "Production and perception of clear speech in croatian and english," *The Journal of the Acoustical Society of America*, vol. 118, no. 3, pp. 1677–1688, 2005.

[7] C. Davis, J. Kim, K. Grauwinkel, and H. Mixdorff, "Lombard speech: Auditory (a), visual (v) and av effects," in *Proceedings of the third international conference on speech prosody*. Citeseer, 2006, pp. 248–252.

[8] J. Trujillo, A. Özyürek, J. Holler, and L. Drijvers, "Speakers exhibit a multimodal lombard effect in noise," *Scientific Reports*, vol. 11, no. 1, p. 16721, 2021.

[9] R. Marxer, J. Barker, N. Alghamdi, and S. Maddock, "The impact of the lombard effect on audio and visual speech recognition systems," *Speech communication*, vol. 100, pp. 58–68, 2018.

[10] P. Ma, S. Petridis, and M. Pantic, "Investigating the Lombard Effect Influence on End-to-End Audio-Visual Speech Recognition," in *Interspeech 2019*. ISCA, Sep. 2019, pp. 4090–4094.

[11] K. Kakol, G. Korvel, G. Tamulevičius, and B. Kostek, "Detecting lombard speech using deep learning approach," *Sensors*, vol. 23, no. 1, p. 315, 2023.

[12] B. Lindblom, "Explaining phonetic variation: A sketch of the h&h theory," *Speech production and speech modelling*, pp. 403–439, 1990.

[13] O. Tuomainen, L. Taschenberger, S. Rosen, and V. Hazan, "Speech modifications in interactive speech: effects of age, sex and noise type," *Philosophical Transactions of the Royal Society B*, vol. 377, no. 1841, p. 20200398, 2022.

[14] B. O'Conaill, S. Whittaker, and S. Wilbur, "Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication," *Human-computer interaction*, vol. 8, no. 4, pp. 389–428, 1993.

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

[16] E. A. Isaacs and J. C. Tang, "What video can and can't do for collaboration: a case study," in *Proceedings of the first ACM International Conference on Multimedia*, 1993, pp. 199–206.

[17] J. N. Bailenson, "Nonverbal overload: A theoretical argument for the causes of zoom fatigue," 2021.

[18] A. Bleakley, D. Rough, J. Edwards, P. Doyle, O. Dumbleton, L. Clark, S. Rintel, V. Wade, and B. R. Cowan, "Bridging social distance during social distancing: exploring social talk and remote collegiality in video conferencing," *Human–Computer Interaction*, vol. 37, no. 5, pp. 404–432, 2022.

[19] B. J. Kushner, "Eccentric gaze as a possible cause of "zoom fatigue"," *Journal of Binocular Vision and Ocular Motility*, vol. 71, no. 4, pp. 175–180, 2021.

[20] S. A. Moubayed, J. Edlund, and J. Beskow, "Taming mona lisa: communicating gaze faithfully in 2d and 3d facial projections," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 1, no. 2, pp. 1–25, 2012.

[21] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta psychologica*, vol. 26, pp. 22–63, 1967.

[22] J. E. Boland, P. Fonseca, I. Mermelstein, and M. Williamson, "Zoom disrupts the rhythm of conversation." *Journal of Experimental Psychology: General*, vol. 151, no. 6, p. 1272, 2022.

[23] J. Reverdy, S. O. Russell, L. Duquenne, D. Garaialde, B. R. Cowan, and N. Harte, "Roomreader: A multimodal corpus of online multiparty conversational interactions," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 2517–2527.

[24] M. Koutsombogera and C. Vogel, "Modeling collaborative multimodal behavior in group dialogues: The multisimo corpus," in *Proceedings of the Eleventh Language Resources and Evaluation Conference*, 2018.

[25] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown, "A corpus of audio-visual lombard speech with frontal and profile views," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. EL523–EL529, 2018.

[26] J. C. Krause and L. D. Braida, "Acoustic properties of naturally produced clear speech at normal speaking rates," *The Journal of the Acoustical Society of America*, vol. 115, no. 1, pp. 362–378, 2004.

[27] R. Baker and V. Hazan, "Lucid: a corpus of spontaneous and read clear speech in british english," in *DiSS-LPSS Joint Workshop 2010*, 2010.

[28] S. Granlund, V. Hazan, and R. Baker, "An acoustic–phonetic comparison of the clear speaking styles of finnish–english late bilinguals," *Journal of Phonetics*, vol. 40, no. 3, pp. 509–520, 2012.

[29] Z. Duran-Karaoz and P. Tavakoli, "Predicting l2 fluency from l1 fluency behavior: The case of l1 turkish and l2 english speakers," *Studies in Second Language Acquisition*, vol. 42, no. 4, pp. 671–695, 2020.

[30] C. T. Pham and E. A. Karuza, "Noise-induced differences in the complexity of spoken language," *Quarterly Journal of Experimental Psychology*, p. 17470218221124869, 2022.

[31] Zoom Video Communications Inc., "Zoom meetings & Chat." Retrieved from https://zoom.us/meetings, 2019.

[32] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[33] E. Ferragne and F. Pellegrino, "Formant frequencies of vowels in 13 accents of the british isles," *Journal of the International Phonetic Association*, vol. 40, no. 1, pp. 1–34, 2010.

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

[35] M. B. Walker and C. Trimboli, "Smooth transitions in conversational interactions," *The Journal of Social Psychology*, vol. 117, no. 2, pp. 305–306, 1982.

[36] P. Boersma and V. Van Heuven, "Speak and unspeak with praat," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.

[37] K. Pisanski, E. C. Mora, A. Pisanski, D. Reby, P. Sorokowski, T. Frackowiak, and D. R. Feinberg, "Volitional exaggeration of body size through fundamental and formant frequency modulation in humans," *Scientific reports*, vol. 6, no. 1, p. 34389, 2016.