9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Silence and overlap in chat and chunk phases of multiparty casual conversation

Emer Gilmartin[1], *Maria O'Reilly*[1], *Christian Saam*[1], *Benjamin R. Cowan*[2], *Carl Vogel*[3], *Nick Campbell*[4], *Vincent Wade*[1]

[1]ADAPT Centre, School of Computer Science and Statistics, Trinity College Dublin, Ireland
[2]University College Dublin, Ireland
[3]Computational Linguistics Group, SCSS, Trinity College Dublin, Ireland
[4]Speech Communication Lab, Trinity College Dublin, Ireland

gilmare@tcd.ie, moreil12@tcd.ie, christian.saam@adaptcentre.ie, benjamin.cowan@ucd.ie, vogel@tcd.ie, nick@tcd.ie, vincent.wade@adaptcentre.ie

Abstract

Casual conversation, 'talk for the sake of talking', is often multiparty, with no clear practical goal, and can last up to several hours. Longer conversations proceed in phases of chat and chunk, where chat is highly interactive and chunks are dominated by one speaker. It is likely that prosodic features will vary between the two phases. Greater understanding of such casual conversation is vital to the design of human-like artificial dialogue, and the need for clearer modelling has prompted our explorations into silence and overlap in six manually segmented long (c. 1 hr) informal multiparty conversations. We test automatic segmentation on the data, and find manual segmentation is necessary to accurately capture speech activity. We analyse speech activity at the end of intervals where one participant speaks in the clear for a second or more, and categorise patterns of overlap and turn change or retention in chat and chunk phases. We also report on a study of a subset of our dataset, taken from a 5-party conversation, comprising over 200 manually annotated intonational phrases (IP) adjacent to silences and overlaps, analysing IP-final tunes with the IViE intonational transcription system, and measuring IP duration to investigate prosodic patterns in the different conditions.

Index Terms: multiparty dialogue, human-computer interaction, turntaking

1. Casual Conversation

Casual conversation, the unmarked or base case of conversation or 'talking just for the sake of talking'[1], is regarded as 'interactional' rather than 'instrumental' or task-based [2]. Described as an emergent behaviour of co-present humans [3], its function is considered to be the building and maintenance of social bonds [4] and avoidance of threatening silence [5, 6, 7]. Casual conversation is not monolithic but develops in stages or phases, with different arrangements of speaker participation and genres of talk appearing in different phases. Such subgenres include smalltalk, gossip, and conversational narrative. Casual conversation often involves multiple participants rather than the dyads normally found in instrumental interactions or examples from conversation analysis [8]. In addition, while task-based conversations are bound by task completion and tend to be short, casual conversation can go on indefinitely. For interactional talk, prosody, the 'how' of conversation, has been considered as important as the 'what' - the propositional content [9, 10].

There are several descriptions of sub-conversational phases – Ventola describes ritualised opening greetings, followed by approach segments of light uncontroversial small talk. Longer conversations proceed to more informative centre phases interleaved with further lighter approach stages, and then back to ritualised leavetakings [11]. Slade and Eggins describe alternating segments of 'chat' (interactive exchanges involving short turns by several participants) and 'chunks' (stretches dominated by a single speaker). Figure 1 shows examples from our data of chat and chunk phases. We are interested in speaker change patterns around gaps and overlaps in casual conversation in general, and in contrasts between such patterns in chat and chunk phases. Several researchers on casual conversation have noted that their analyses were limited as they were based on transcripts and thus lacked vital timing and multimodal information [7, 8, 6]. We hope that our analyses of multimodal recordings of long form casual talk will help to address this gap in understanding of this fundamental speech exchange system, and encourage the creation of further resources.

Very few multimodal data collections of multiparty casual conversations exist. Early analysis was often based on transcripts, or even memory, and timing information has not always been considered in great depth. Multimodal and audio corpora often comprise spoken tasks or interactions specific to particular domains such as meetings or telephone calls, or 'information gap' activities used to elicit conversation [12, 13, 14, 15, 16]. Findings from such data may not generalize to other genres of natural conversation [17]. Casual talk has been included in larger collections [18, 19, 20, 21], as fragments recorded in natural settings [22], or more recently in collections of short, often dyadic, casual talk or 'first encounters' and sometimes of dialogues between friends [23, 24, 25]. Such data provide valuable examples of the approach or chat stages of conversation but may not be long enough for full analysis of sequences of 'chat' and 'chunks'. For our studies, we have assembled a collection of multimodal recordings of six long multiparty casual conversations as described below.

2. Data and Annotation

Our dataset is a collection of conversations of around one hour each, drawn from the d64, DANS, and TableTalk corpora [26, 27, 28]. In each conversation participants were free to talk or not as the mood took them. Table 1 gives an overview of the conversations.

10.21437/SpeechProsody.2018-77

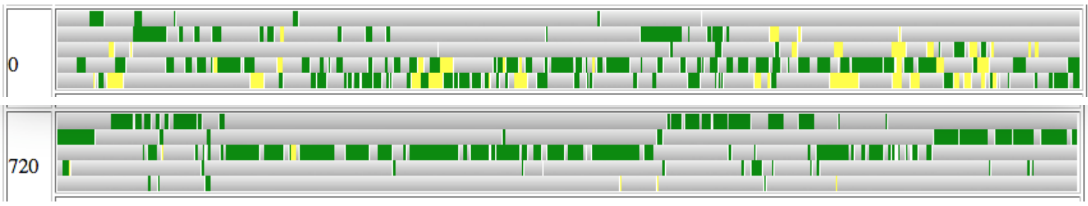

Figure 1: *Examples of chat (top) and chunk (bottom) phases in two-minute stretches from a 5-party conversation. Each row denotes the activity of one speaker across 120 seconds. Speech is green, and laughter is yellow on a grey background (silence). The chat frame, taken at the beginning of the conversation, can be seen to involve shorter contributions from all participants with frequent laughter. The chunk frame shows longer single speaker stretches later in the interaction.*

Conversation	Participants	Gender	Duration (s)
A	5	2F/3M	4164
B	3	1F/2M	4672
C	4	1F/3M	4378
D	3	2F/1M	3004
E	4	2F/2M	2072
F	5	3F/2M	4740

Table 1: *Conversations used in dataset.*

There has been much progress in the automatic analysis of spoken conversation [29, 30]. However, attempts to automatically segment our data were unsuccessful, so the data were segmented manually. There are valid concerns about manual segmentation, as humans tend to interpret what they hear and can miss or indeed imagine silences of short duration [31], or have difficulty recalling disfluencies [32]. However, these results were based on speakers timing pauses with a stopwatch in a single hearing. In the current work, using Praat [33] and Elan [34], speech could be slowed down and replayed, and annotators could see the speech waveform and spectrogram and check doubtful cases using video recordings. Therefore, it was hoped that problems due to annotators only picking up perceptually salient silences could be avoided. A subsequent analysis with the OpenSMILE toolkit[1] [35] implementing the VAD described in [36] confirmed that even under artificially good conditions the best average error per segment was around half the duration of the silences to be studied. False alarms were filtered under the premise that a beam-former could decide whether speech activity came from the targeted speaker or from cross talk. Short gaps under 60ms were bridged and a collar of 30ms was applied around the segment boundaries during scoring. The speech activity detection was scored with NIST SCTK 2.4.0 md-eval.pl version 22 using the guidelines of the Spring 2006 (RT-06S) Rich Transcription Meeting Recognition Evaluation Plan, Section 7 DIARIZATION — 'SPEECH ACTIVITY DETECTION' MDE [37] (n=3142 missed=279.34s falarm=1753.06s total=2032.4s avg. missed=0.089s/1 avg. falarm=0.558s/1 avg. total=0.647s/1).

Segmentation and transcription was at the intonational phrase level (IP), rather than a more theory dependent utterance or coarser inter-pausal unit (IPU) level. Labels covered speech, silence, coughs, breaths, and laughter. The speech label was applied to verbal and non-verbal vocal sounds (except laughter) and thus included contributions such as filled pauses, and short utterances such as 'oh' or 'mmhmm'. Laughter was annotated inline with speech. For this study, IPs were concatenated to IPUs, and annotated coughs, breaths, and laughter intervals were converted to silence. Chat and chunk phases were marked using an annotation scheme based on that used in Slade's treatment of chat and chunk phases in casual talk [38]. The 'chunks' were marked using the first, structural part of Slade and Eggins' definition - 'a segment where one speaker takes the floor and is allowed to dominate the conversation for an extended period' [1]. All other interaction was considered chat. A total of 213 chat and 358 chunk phases were identified across the conversations. A detailed description of the annotation process can be found in [39]. Annotation of intonation contours was carried out using the IViE system [40] on a subset of Conversation A as described in Section 4.

3. Speaker Change Activity across Dataset

We examine intervals where a particular speech/silence configuration holds, rather than more theory-dependent utterances or turns. As an example, an interval could consist of 342 milliseconds of Participants 2 and 3 speaking while Participants 1 and 4 are silent. The dataset contains 30688 changes in speech/silence configuration over a total of 23030 seconds, an average of 1.3 per second.

The most common conversational situation in terms of time was single participant speaking in the clear (68%), with global silence accounting for 23% of the conversational time. The remaining time (9%) is overlapping speech by two or more participants, with instances dropping sharply as the number of overlapping speakers increases. The vast bulk of overlap in all conversations involved two speakers. Figure 3 shows the distribution of the number of speakers in chat and chunk phases by time – there is significantly less overlap and more single party speech in chunk phases. The duration of intervals varies widely, and distributions are heavily right skewed. The log distributions more closely approximate Normal - the geometric mean duration of a single speaker interval is significantly higher in chunk than in chat phases (0.68s vs 0.46s). Overlap and silence interval mean durations are similar across both conditions - around 0.24 and 0.31s. However, when the distribution of silence is compared in chat and chunk, there is greater variability in the length of silences in chat, perhaps reflecting the fact that these phases involve more speaker change, while most silences in chunks are within speaker and may reflect a certain uniformity in duration.

In an n-party conversation where each participant may be speaking or silent at any moment, there are n^2 possible

[1] SMILExtract 2.3 with vad_opensource.conf, post-processed with a standard cRnnVad2 component configuration.

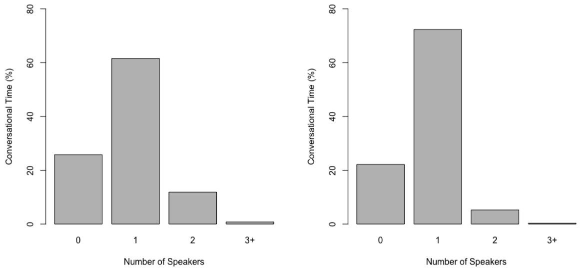

Figure 2: *Distribution of the floor in terms of % duration in chat (left) and in chunk (right) phases. X-axis shows number of speakers (0,1,2,3+) speaking concurrently.*

states for the dialogue at any time. To further explore gaps and overlap in such conversations, we consider the possibilities for a speaker, SpX, speaking in the clear. This single-party interval can conclude in silence, overlap, or a smooth switch to one or more speakers. The likelihood of two or more speakers starting at the same instant or one speaker starting immediately as another finishes is very small, given Praat's precision. In the dataset, there are 14932 intervals of single party speech in the clear; global silence follows 73% of these, while speech follows 27%. There are 35 intervals where there is simultaneous onset of speech after a silence, and 89 cases of smooth switching, accounting for less than 1% of the data. These 124 intervals were omitted from the data.

To further explore dynamics after a single speaker makes a contribution other than a backchannel or short utterance, we subset our data into situations where one speaker speaks alone for at least one second, followed by either global silence or overlap. This results in four conditions of interest, two around overlap and two around silence. We do not make any attempt to distinguish between backchannels or longer utterances from incoming speakers after the silence or overlap at this stage, although a finer analysis of the length of these contributions may help distinguish backchannelling from taking a turn. The conditions for SpX are:

WSS — Within Speaker Silence - SpX speaks on either side of a silence

BSS — Between Speaker Silence - SpX speaks before silence, SpX is not speaking after silence

Osame — Within Speaker Overlap - SpX is overlapped, SpX continues after the overlap

Odiff — Between Speaker Overlap - SpX is overlapped, SpX is not speaking after overlap

We first performed a general description and analysis of state changes around gap and overlap across the entire dataset. We then selected random samples from each of these four conditions for three speakers from conversation A, as the recordings were of highest quality, for prosodic analysis. We were interested in differences in the distributions of the four conditions in chat and chunk phases of conversation, and in any connection between prosodic tunes in SpX's original utterance and the following silence or overlap configuration.

3.1. Gaps and Overlap

In the dataset, looking at the situation to the right of a single speaker, there are a total of 14807 change points, comprising 6583 WSS, 4289 BSS, 1993 Osame, and 1943 Odiff. We focus on overlap and gaps to the right of a single speaker stretch of at least one second in duration, and impose a minimum silence threshold of 60ms to reduce the chance of counting stop occlusions as within speaker silences. This reduces the number of change points to 5358 – 3221 WSS, 1116 BSS, 597 Osame, and 424 Odiff.

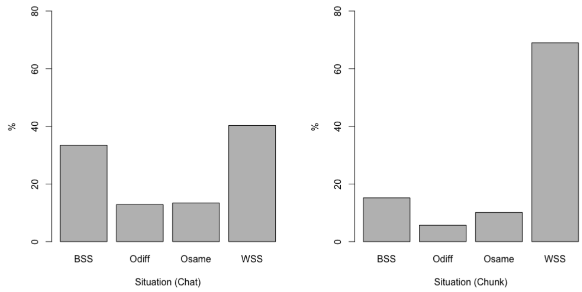

Figure 3: *Distribution of the four conditions of interest in chat (left) and chunk (right) phases.*

Figure 3 shows the distribution of the four conditions in chat and chunk segments for these 5358 cases. Across the entire conversation and particularly in chunks, WSS are the most common change points, reflecting the fact that chunks largely comprise one participant speaking. BSS are more common in chat than chunk phases. Osame and Odiff overlaps are more common in chat than chunk, and Osame overlap is more common than Odiff in chunk phases.

4. Prosodic Analysis of 5-party conversation

We analysed a subset of the data for intonation contours, and classified the nuclear (IP-final) tunes using the IViE transcription system [40]. This was to explore the distribution of tune types in the four conditions. We opted for analysis by a human labeller, as the quality of the recordings would not permit consistent automatic labelling. We drew random samples of 20 each of the four conditions of interest for each of the three male speakers in the conversation, a total of 240 samples. For each sample, a Praat Textgrid file was created with two seconds before the silence or overlap and two after, to provide context for the labeller to mark SpX's Intonational Phrase (IP) at onset of overlap or leading up to silence.

Figure 4 presents the distributions of nuclear tunes in the four conditions. The five main tunes are: fall (H*+L%), downstepped fall (!H*+L%), high plateau (H*%), low plateau (L*%) and fall-rise (H*+LH%); any other tune is subsumed under 'other'.

Falling nuclei dominate across the data, with some fall-rises and very infrequent other tunes. The H*+L% tune is the most common tune in each of the conditions, and overwhelmingly dominates in WSS. In the other three conditions (BSS, Osame, and Odiff), the downstepped variant, !H*+L%, occurs more frequently than it does in WSS. The fall-rise H*+LH% tune, while not used much overall, occurs twice as frequently (or more) in change of speaker (BSS and Odiff) compared to when the turn is maintained by the same speaker (WSS and Osame). No conspicuous trend is found for H*%, L*% or 'other'(these tunes are much less frequent).

We also examined the timing of overlap onset (Figure 5)

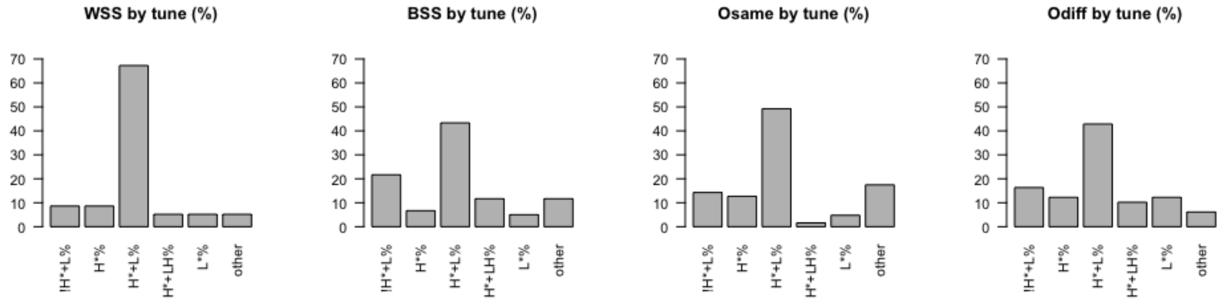

Figure 4: *Prosodic tunes observed in the 4 conditions of interest - WSS, BSS, Osame, Odiff.*

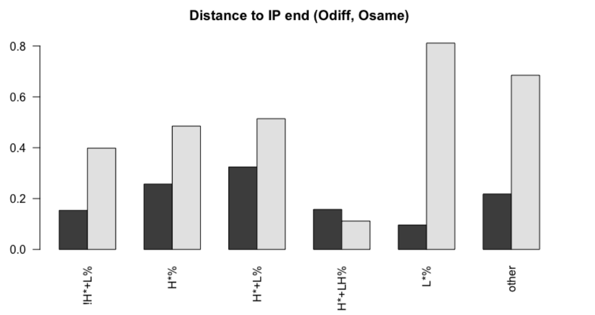

Figure 5: *Distance in seconds from onset of overlap to end of IP for between speaker overlap (Odiff - black) and within speaker overlap (Osame - grey).*

relative to IP end in the overlap data (Osame and Odiff). Overlap onset is consistently earlier in Osame than Odiff across all tunes; in H*+LH% however, overlap onset is early in both conditions. The early onset of overlap in all tunes except H*+LH% can be interpreted as a tendency for the interlocutor to wait until closer to the IP end when she/he intends to take over, and to speak earlier when simply back-channelling.

5. Discussion

Our analysis shows that more solo speaker utterances end in silence than overlap in both chat and chunk phases, although chat sequences, as expected, show more overlap and between-speaker activity. Our analysis takes the contribution of the first speaker after a gap or overlap as the defining factor in classifying the gap or overlap. However, in these conversations, it is not uncommon to see quite a lot of to and fro of short utterances around turn changes. Finer distinctions could be made by considering the length of the second speaker's contribution, thus further classifying BSS into before backchannel and before turn change silences. In future work, we will look at longer right-hand contexts to distinguish how long the incoming speaker persists. Within speaker silences are very common in the data, and some such silences may be a function of breathing. High quality annotated recordings are needed to further explore this question. The differences between silence and overlap distribution in chat and chunk phases add to earlier results showing differences in duration and laughter distribution.

For intonation, nuclear tunes do not appear to be uniquely tied to the conditions of interest. This is perhaps not surprising, since one tune can serve multiple functions in discourse. Nevertheless, the results help elucidate the role of intonation in speaker change conditions. For the two most frequent nuclear tunes we observe an interesting trend: the overwhelming use of H*+L% in WSS may indicate the speaker firmly holding the turn, while the down-stepped fall could indicate a lesser intent to keep the turn. Adding measurements such as f0 timing and scaling could reveal finer prosodic distinctions. For instance, H*+L in WSS would likely exhibit higher peaks (H*) than in BSS, the IP-final L would reach a lower level, etc. Larger datasets would allow for description of non-falling nuclear tunes, L*%, H*% and 'other', underrepresented in these data. In overlap, the timing of overlap onset appears more informative than intonation for turntaking – across practically all tunes the overlap starts closer to IP end in turn change than retention.

6. Conclusions

We have described our preliminary investigations into silence and overlap in multiparty casual talk, a speech genre fundamental to human social life. We compared and contrasted silence and overlap dynamics across a dataset of six long multiparty conversations. We analysed the distribution of within and between speaker silence and overlap following a single speaker speaking in the clear for at least one second. We identified differences in intonational and timing patterns in chat and chunk phases and in silence and overlap conditions. Further analysis in this area will depend on the availability of suitable datasets. Understanding the dynamics of task based interaction has been greatly aided by efforts to record high quality corpora, and it is hoped that our explorations will strengthen the case for the production of high quality multiparty casual conversation corpora.

7. Acknowledgements

This work is supported by the European Coordinated Research on Long-term Challenges in Information and Communication Sciences and Technologies ERA-NET (CHISTERA) JOKER project, JOKe and Empathy of a Robot/ECA: Towards social and affective relations with a robot, and by the ADAPT Centre for Digital Content Technology, which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

8. References

[1] S. Eggins and D. Slade, *Analysing casual conversation.* Equinox Publishing Ltd., 2004.

[2] G. Brown and G. Yule, *Teaching the spoken language.* Cambridge University Press, 1983, vol. 2.

[3] B. Malinowski, "The problem of meaning in primitive languages," *Supplementary in the Meaning of Meaning*, pp. 1–84, 1923.

[4] R. Dunbar, *Grooming, gossip, and the evolution of language.* Harvard Univ Press, 1998.

[5] R. Jakobson, "Closing statement: Linguistics and poetics," *Style in language*, vol. 350, p. 377, 1960.

[6] K. P. Schneider, *Small talk: Analysing phatic discourse.* Hitzeroth Marburg, 1988, vol. 1.

[7] J. Laver, "Communicative functions of phatic communion," *Organization of behavior in face-to-face interaction*, pp. 215–238, 1975.

[8] S. Thornbury and D. Slade, *Conversation: From description to pedagogy.* Cambridge University Press, 2006.

[9] D. Abercrombie, *Problems and principles: Studies in the Teaching of English as a Second Language.* Longmans, Green, 1956.

[10] S. I. Hayakawa, *Language in thought and action.* Houghton Mifflin Harcourt, 1990.

[11] E. Ventola, "The structure of casual conversation in English," *Journal of Pragmatics*, vol. 3, no. 3, pp. 267–298, 1979.

[12] A. Anderson, M. Bader, E. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller *et al.*, "The HCRC map task corpus," *Language and speech*, vol. 34, no. 4, pp. 351–366, 1991.

[13] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, and A. Stolcke, "The ICSI meeting corpus," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 1, 2003, pp. I–364.

[14] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, and V. Karaiskos, "The AMI meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005.

[15] G. Beattie, *Talk: An analysis of speech and non-verbal behaviour in conversation.* Open University Press, 1983.

[16] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, pp. 517–520.

[17] J. L. Lemke, "Analyzing verbal data: Principles, methods, and problems," in *Second International Handbook of Science Education.* Springer, 2012, pp. 1471–1484.

[18] S. Greenbaum, "ICE: The international corpus of English," *English Today*, vol. 28, no. 7.4, pp. 3–7, 1991.

[19] BNC-Consortium, "British national corpus," *URL http://www. hcu. ox. ac. uk/BNC*, 2000.

[20] D. Biber, S. Johansson, G. Leech, S. Conrad, E. Finegan, and R. Quirk, *Longman grammar of spoken and written English.* Longman London, 1999, vol. 2. [Online]. Available: http://www.tesl-ej.org/wordpress/issues/volume4/ej15/ej15r14/?wscr=

[21] J. Allwood, M. Björnberg, L. Grönqvist, E. Ahlsén, and C. Ottesjö, "The spoken language corpus at the department of linguistics, Göteborg University," in *FQS–Forum Qualitative Social Research*, vol. 1, 2000. [Online]. Available: http://www.ling.gu.se/ jens/publications/bfiles/B45.pdf

[22] J. W. DuBois, W. L. Chafe, C. Meyer, and S. A. Thompson, *Santa Barbara Corpus of Spoken American English. CD-ROM. Philadelphia: Linguistic Data Consortium*, 2000.

[23] J. Edlund, J. Beskow, K. Elenius, K. Hellmer, S. Strömbergsson, and D. House, "Spontal: A Swedish Spontaneous Dialogue Corpus of Audio, Video and Motion Capture." in *LREC*, 2010.

[24] P. Paggio, J. Allwood, E. Ahlsén, and K. Jokinen, "The NOMCO multimodal Nordic resource–goals and characteristics," 2010.

[25] A. J. Aubrey, D. Marshall, P. L. Rosin, J. Vandeventer, D. W. Cunningham, and C. Wallraven, "Cardiff Conversation Database (CCDb): A Database of Natural Dyadic Conversations," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, Jun. 2013, pp. 277–282.

[26] C. Oertel, F. Cummins, J. Edlund, P. Wagner, and N. Campbell, "D64: A corpus of richly recorded conversational interaction," *Journal on Multimodal User Interfaces*, pp. 1–10, 2010.

[27] S. Hennig, R. Chellali, and N. Campbell, "The D-ANS corpus: the Dublin-Autonomous Nervous System corpus of biosignal and multimodal recordings of conversational speech." Reykjavik, Iceland, 2014.

[28] N. Campbell, "Multimodal processing of discourse information; the effect of synchrony," in *Universal Communication, 2008. ISUC'08. Second International Symposium on*, 2008, pp. 12–15.

[29] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, Oct. 2010.

[30] K. Laskowski, "Predicting, detecting and explaining the occurrence of vocal activity in multi-party conversation," Ph.D. dissertation, Carnegie Mellon University, 2011.

[31] J. G. Martin, "On judging pauses in spontaneous speech," *Journal of Verbal Learning and Verbal Behavior*, vol. 9, no. 1, pp. 75–78, 1970.

[32] J. Deese, *Pauses, prosody, and the demands of production in language.* Mouton Publishers, 1980.

[33] P. Boersma and D. Weenink, *Praat: doing phonetics by computer [Computer program], Version 5.1. 44*, 2010.

[34] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: a professional framework for multimodality research," in *Proceedings of LREC*, vol. 2006, 2006.

[35] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: ACM, 2013, pp. 835–838, 00218. [Online]. Available: http://doi.acm.org/10.1145/2502081.2502224

[36] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on.* IEEE, 2013, pp. 483–487.

[37] J. G. Fiscus, J. Ajot, M. Michel, and J. S. Garofolo, "The rich transcription 2006 spring meeting recognition evaluation," in *International Workshop on Machine Learning for Multimodal Interaction.* Springer, 2006, pp. 309–322.

[38] D. Slade, *The texture of casual conversation: A multidimensional interpretation.* Equinox, 2007.

[39] E. Gilmartin and N. Campbell, "Capturing Chat: Annotation and Tools for Multiparty Casual Conversation." in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

[40] E. Grabe, "The IViE labelling guide," *Journal of the Acoustical Society of America*, vol. 101, pp. 3728–3740, 2001.