On the Dynamics of Overlap in Multi-Party Conversation

Kornel Laskowski [1], *Mattias Heldner* [2], *Jens Edlund* [3]

[1] Carnegie Mellon University, Pittsburgh PA, USA
[2] Department of Linguistics, Stockholm University, Stockholm, Sweden
[3] KTH Speech, Music and Hearing, Stockholm, Sweden
kornel@cs.cmu.edu, heldner@ling.su.se, edlund@speech.kth.se

Abstract

Overlap, although short in duration, occurs frequently in multi-party conversation. We show that its duration is approximately log-normal, and inversely proportional to the number of simultaneously speaking parties. Using a simple model, we demonstrate that simultaneous talk tends to end simultaneously less frequently than in begins simultaneously, leading to an arrow of time in chronograms constructed from speech activity alone. The asymmetry is significant and discriminative. It appears to be due to dialog acts which do not carry propositional content, and those which are not brought to completion.

Index Terms: multi-party conversation, overlap, turn-taking.

1. Introduction

At first sight, the simultaneous production of speech by parties to a conversation seems like an exotic event. After all, speech produced in this way is more difficult to hear, and if the speakers present different lines of thought, listeners are likely to be distracted. The popular description of conversational conduct as *a taking of turns* [1] virtually guarantees that each of us believe that simultaneous speech, or *overlap*, is rare.

In the last decade, the collection of large corpora of naturally occurring spontaneous speech, and the availability of both algorithm implementations and powerful computers, has demonstrated that overlap is frequent [2, 3], albeit short-lived. This departure from our intuition suggests that humans model the occurrence of overlap well enough to not be surprised by it too often. If that is true, then conversational machines should be designed not to ignore overlap — on the grounds that it is allegedly rare — but to anticipate it [2].

At the current time, there is considerable knowledge about overlap in conversation, almost exclusively in the form of *time-independent prior probabilities* of its occurrence. In some cases (e.g. [3]), those probabilities have been conditioned on conversation type, participant group, degree of involvement, and other quantities which are either constant or near-constant throughout an interval of conversation. The time independence of this knowledge puts conversational machines at the mercy of chance: they can anticipate that overlap will happen at some point, but are unable to anticipate when.

In this work, we look at the *dynamics* of overlap to begin to rectify that blind spot. We observe that overlap duration is inversely proportional to the number of simultaneously speaking participants. We then employ a very simple model, a first-order Markov state machine over binary speech activity states, to demonstrate that it is possible — with high accuracy — to infer the direction of time flow [4]. The observed temporal asymmetry indicates that entrance into and egress out of overlap differ systematically, even when the scene is so dramatically stripped of linguistic information. Finally, we explore which specific types of speech are responsible for the asymmetry. Our experiments show that it is the dialog acts (DAs) with low propositional content, as well as those not brought to completion, whose temporal deployment is least symmetric.

Our findings have an important impact on spoken dialogue systems which must contend in real-time with speakers who are better at conversational conduct than they are. Overlap which becomes predictable for systems has likely always been easily predictable for humans, and should not be construed as "misconduct" by either party. As such, recovery from it should receive less attention from system designers than from overlap which cannot be successfully predicted.

2. Data

Analysis and experiments are performed using the ICSI Meeting Corpus [5, 6]. The corpus consists of 75 meetings which would have occurred even if they had not been recorded. Each meeting was attended by 3 to 9 participants[1]. The total meeting time in the corpus, excluding speaker calibration intervals, and initial and terminal silence, is 67 hours.

The corpus has been orthographically transcribed [5]; to explore the occurrence of overlap, we use only the start and end times of words, made available as part of the DA annotation in [6]. In the last part of the current work, we also use the DA annotation itself. This allows us to evaluate the observed overlap as a function of specific DA types, such as statements, questions, backchannels (e.g. "uh-huh"), floor hold and grab mechanisms, etc. A detailed account of the DAs occurring in the corpus can be obtained from its annotation manual [7].

3. Interval Duration

The distribution of the durations of intervals of overlap in the ICSI corpus is shown in Figure 1. Each curve in the plot can be seen to be unimodal, and indicates an approximate log-normal distribution. This makes Markov modeling a suitable paradigm for our proposed studies. As can also be observed, the duration of contiguous intervals in which at least n parties talk simultaneously is approximately inversely proportional with n.

4. Degree-of-Overlap Dynamics

As in much of our previous modeling work, we begin with a binary speech ■ versus non-speech □ *chronogram* **Q** of each

[1] A tenth participant in one of the meetings was not wearing a microphone and a forced-alignment segmentation is not available for her/him.

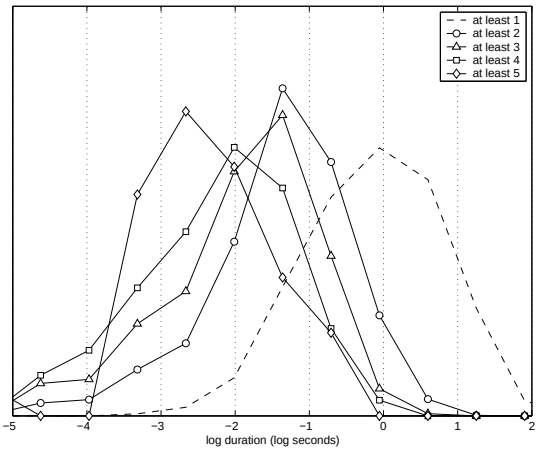

Figure 1: Log_e durations of contiguous intervals of talk by at least n parties, for $n \in \{1, 2, 3, 4, 5\}$. Probabilities along the y-axis normalized to yield unity area under each curve.

conversation [8, 9]. This is a matrix of T columns $\mathbf{q}_t$ of non-overlapping 100-ms frames, $1 \leq t \leq T$; columns contain as many entries K as there are parties to the conversation. Each cell has a value $\mathbf{q}_t\,[k] \in \{\square, \blacksquare\}$, $1 \leq k \leq K$.

4.1. A 1st-Order Markov Model

Instead of modeling the chronogram in its entirety, we model the number c_t of parties talking simultaneously at instant t,

$$c_t \quad = \quad \sum_{k=1}^{K} \delta\left(\mathbf{q}_t\,[k]\,, \blacksquare\right) \quad . \tag{1}$$

where $\delta\,(\cdot)$ is the Kronecker delta. We refer to c_t as the *degree of overlap* (strictly speaking, overlap occurs only when $c_t > 1$). The sequence $\{c_t\}$ indicates *how many* parties were speaking, not *which ones*.

Informed by Figure 1, we treat c_t as a Markov process. Its time-independent transition probabilities, from degree n_i to n_j,

$$a_{ij}^{+} \quad = \quad P\left(c_t = n_j | c_{t-1} = n_i\right) \quad , \tag{2}$$

are inferred by counting the observed transitions in our data. The maximal degree of overlap is 6, meaning that the entire model $A^{+} = (a_{ij}^{+})$ is a 7×7 matrix, each entry of which is bounded in $[0, 1]$. The superscript $^{+}$ indicates a forward direction of time, which we contrast with the backward direction denoted by $^{-}$, as explained below.

When trained on all of the ICSI meetings, the probabilities have the values shown in the upper half of each cell in Table 1 (additionally annotated with an arrow $\rightarrow$ for clarity). Unigram probabilities are shown in the right-most column. The table shows that the most likely value of c_t is unity (i.e., one speaker at a time), with a time-independent prior probability of 66.9%. When in this state, the conversation is likely to remain in it (with a transition probability of 90.5%): this is observed by looking at the alternatives in the row labeled "1". The next-most likely next state is silence (cf. the column labeled "0" in that row), with a transitional probability of 7.7%. True overlap ($c_t > 1$) arises out of the $n_i = 1$ state with only a 1.8% probability.

The table also shows the time-reversed model,

$$a_{ij}^{-} \quad = \quad P\left(c_{t-1} = n_j | c_t = n_i\right) \quad . \tag{3}$$

The values found from all of the ICSI Meeting Corpus are shown in the bottom half of each cell, annotated with a $\leftarrow$. It is apparent that in general neither the forward process nor the backward process observe detailed balance, $a_{ij}^{+} \neq a_{ji}^{+}$ and $a_{ij}^{-} \neq a_{ji}^{-}$: entering higher degrees of overlap is less likely than egressing from them, in either case. However, that $a_{ij}^{+} \neq a_{ij}^{-}$ indicates that chronograms are not symmetric in time.

The discrepancies appear to follow a pattern:

$$\text{when} \quad \begin{array}{lll} n_j < n_i - 1 \;\;, & a_{ij}^{+} < a_{ij}^{-} \\ n_j = n_i - 1 \;\;, & a_{ij}^{+} > a_{ij}^{-} \\ n_j = n_i \;\;, & a_{ij}^{+} = a_{ij}^{-} \\ n_j = n_i + 1 \;\;, & a_{ij}^{+} < a_{ij}^{-} \\ n_j > n_i + 1 \;\;, & a_{ij}^{+} > a_{ij}^{-} \end{array} \tag{4}$$

What this means, in words, is that an increase in the number of currently speaking participants, by one, is *less* likely than reversing the outcome in the same way (looking at the transition with time reversed). On the other hand, an increase in the number of currently speaking participants, by two or more, is *more* likely than reversing the outcome. This indicates that intervals of overlap, when they are not left-right-symmetric, are more likely to be created (or upgraded) two or more participants at a time than one participant at a time, but are more likely to be terminated one participant at a time than two or more at a time.

4.2. Time's Arrow

To determine whether the differences between A^{+} and A^{-} are significant, we ask the more stringent question of whether they are systematically *discriminative*. To answer, we propose to randomly decide whether or not to time-reverse chronograms $\mathbf{Q}$, and then attempt to classify the direction of time flow using only A^{+} and A^{-}, following application of Equation 1.

The experiment is carried out in a round-robin fashion: for each meeting in the ICSI corpus, we train a model A^{+} and a model A^{-} using the remaining 74 meetings. We then score the unreversed test meeting in each fold, and select either the forward model or the backward model using depending on which yields the highest log-likelihood. We expect the forward model to win in each case, if our chronograms are asymmetric and our simple models can capture the differences. The null hypothesis is that the forward model wins in 50% of the cases.

The results are that in 74 of the 75 cases, A^{+} yields a higher likelihood; we therefore unambiguously discard the null hypothesis. The asymmetry is due to overlap alone. To see this, consider a conversation with no overlap at all, where $c_t \in \{0, 1\}$ for all t. Model training material consisting exclusively of such conversations would necessarily entail $A^{+} \equiv A^{-}$.

5. Ablation of Speech by Dialog Act Type

Having shown that the chronograms are asymmetric in time, we would like to finger the culprit speech phenomena which make them so. Fortunately, our proposed framework makes this surprisingly easy, particularly because the ICSI Meeting Corpus is accompanied by DA annotation. We propose to eliminate from our chronograms that speech which implements specific DA types, and verify whether the direction of time can still be inferred. In each case, a new c_t is computed from the modified chronograms, and new models are trained, in round robin fashion as before. Our results are shown in Table 2.

The table is divided into 5 panels. In each, we removed speech labeled with DA tags of specific groups. The first panel

n_i, # of speakers in the "from" state	n_j, # of speakers in the "to" state: 0	1	2	3	4	5	6	unigram prob.
0	⇆ 81.1578	→ 18.3344 ← **18.5440**	→ **0.4970** ← 0.2950	→ **0.0104** ← 0.0032	→ *0.0003* ← *0.*	→ *0.* ← *0.*	→ *0.* ← *0.*	27.7883
1	→ **7.6993** ← 7.6123	⇆ 90.5279	→ 1.7364 ← **1.8289**	→ **0.0354** ← 0.0305	→ **0.0009** ← 0.0003	→ *0.0001* ← *0.0001*	→ *0.* ← *0.*	66.9285
2	→ 1.6624 ← **2.8006**	→ **24.8201** ← 23.5646	⇆ 71.4525	→ 2.0225 ← **2.1415**	→ **0.0425** ← 0.0399	→ *0.* ← *0.0008*	→ *0.* ← *0.*	4.9317
3	→ 0.2670 ← **0.8774**	→ 6.2055 ← **7.1974**	→ **32.0575** ← 30.2772	⇆ 59.2319	→ 2.0600 ← **2.3398**	→ **0.1780** ← 0.0763	→ *0.* ← *0.*	0.3294
4	→ *0.* ← *0.4255*	→ 0.8511 ← **2.9787**	→ 10.0000 ← **10.6383**	→ **39.1489** ← 34.4681	⇆ 46.3830	→ 3.6170 ← **4.8936**	→ *0.* ← *0.2128*	0.0197
5	→ *0.* ← *0.*	→ *1.8868* ← *1.8868*	→ *1.8868* ← *0.*	→ 11.3208 ← **26.4151**	→ **43.3962** ← 32.0755	⇆ 39.6226	→ *1.8868* ← *0.*	0.0022
6	→ *0.* ← *0.*	→ *0.* ← *0.*	→ *0.* ← *0.*	→ *0.* ← *0.*	→ *25.* ← *0.*	→ *0.* ← *25.*	⇆ 75.	0.0002

Table 1: First-order Markov transition probabilities for the forward-in-time $a^{+}_{n_i,n_j} = P\left(\|\mathbf{q}_t\| = n_j | \|\mathbf{q}_{t-1}\| = n_i\right)$ and for the time-reversed $a^{-}_{n_i,n_j} = P\left(\|\mathbf{q}_{t-1}\| = n_j | \|\mathbf{q}_t\| = n_i\right)$. Cells for which $a^{+}_{n_i,n_j}$ and/or $a^{-}_{n_i,n_j}$ are based on fewer than 3 occurrences are shown in italics, for completeness. In all other off-diagonal cells, the larger of $a^{+}_{n_i,n_j}$ (identified with "→") and $a^{-}_{n_i,n_j}$ (identified with "←") is shown in bold. Diagonal entries are identical for both models, i.e. $a^{+}_{n_i,n_j} \equiv a^{-}_{n_i,n_j},\ \forall n_i = n_j$, and are denoted with "⇆". "# of speakers" is the number of simultaneously speaking participants. The total number of events, at a frame step of 100 ms, from which these probabilities were computed is 2.4 million.

contains the result of the previous section, in which no speech is removed. CERR is the error encountered when attempting to classify the direction of time, across all 75 meetings. We observe, in the second panel, that although some speech was not labeled, was labeled as nonspeech, or was undecipherable, removal of these — collectively referred to as $\mathcal{X}$ — has no impact on our classification score.

In panel 3, it can be seen that the removal of interrupted and abandoned DAs increases our CERR slightly, and that removing both types of disruption forms $\mathcal{D}$ results in the same CERR as removing only interrupted DAs.

Panel 4 shows that the removal of acknowledgments, assessments, and backchannels increases CERR in each case, and that the effects are additive when all three $\mathcal{B}$ are removed together. Furthermore, removing both $\mathcal{D}$ and $\mathcal{B}$ is additive; when they are both removed, chronograms are 34.6%rel of the way to being symmetric under time reversal.

The last type of DA considered are floor mechanisms $\mathcal{F}$, of which there are three subtypes. Removing each increases CERR slightly, but the error incurred when removing all of them is no larger than that incurred by removing only floor grabbers. We note however that removing all of $\mathcal{X}$, $\mathcal{D}$, $\mathcal{B}$, and $\mathcal{F}$ yields a CERR of 34.7%, easily the majority (75.4%rel) of the way to symmetry. It should be noted that in this last case, as shown in the second column of Table 2, only 46.5 hours of speech remain, of the 66.6 hours in total.

DA Types Removed	Speech Left	CERR, %
—	66:34	1.3
$\mathcal{X}$, Unlabeled Phenomena (Groups 12 and 13)		
nonlabeled z	63:41	1.3
nonspeech x	64:22	1.3
indecipherable %	64:22	1.3
$\mathcal{X} \equiv$ z ∪ x ∪ %	63:37	1.3
$\mathcal{D}$, Disruption Forms (Group 12)		
$\mathcal{X}$ ∪ abandoned %- -	58:59	2.7
$\mathcal{X}$ ∪ interrupted %-	61:22	5.3
$\mathcal{X} \cup (\mathcal{D} \equiv$ %- ∪ %- -)	56:44	5.3
$\mathcal{B}$, Backchannels & Acknowledgments (Group 4)		
$\mathcal{X}$ ∪ acknowledgment bk	62:08	2.7
$\mathcal{X}$ ∪ assessment ba	62:37	3.0
$\mathcal{X}$ ∪ backchannel b	61:35	5.3
$\mathcal{X} \cup (\mathcal{B} \equiv$ b ∪ ba ∪ bk)	59:08	10.7
$\mathcal{X} \cup \mathcal{D} \cup \mathcal{B}$	52:22	17.3
$\mathcal{F}$, Floor Mechanisms (Group 3)		
$\mathcal{X}$ ∪ hold h	63:06	2.7
$\mathcal{X}$ ∪ floor holder fh	59:28	2.7
$\mathcal{X}$ ∪ floor grabber fg	61:43	5.3
$\mathcal{X} \cup (\mathcal{F} \equiv$ fg ∪ fh ∪ h)	57:03	5.3
$\mathcal{X} \cup \mathcal{D} \cup \mathcal{F}$	50:48	12.0
$\mathcal{X} \cup \mathcal{D} \cup \mathcal{B} \cup \mathcal{F}$	46:31	34.7

Table 2: Error rates (CERRs) for detecting the direction of the flow of time, as a function of the types of DAs removed; DA types grouped as in the annotation manual [7]. Remaining speaking time shown in hrs:min format.

6. Discussion

Table 1 indicates that for a snippet of conversation bounded by silence at both ends, containing no other silence, and containing exactly one interval of overlap $c_t = 2$, the most likely c_t sequence is $\{0, 1, 1, \ldots, 1, 1, 2, 2, \ldots, 2, 2, 1, 1, \ldots, 1, 1, 0\}$; namely, overlap is entered from $c_t = 1$ and egressed to $c_t = 1$, and its occurrence is symmetric in time. Somewhat less likely is $\{0, 2, 2, \ldots, 2, 2, 1, 1, \ldots, 1, 1, 0\}$, but $\{0, 1, 1, \ldots, 1, 1, 2, 2, \ldots, 2, 2, 0\}$ is much less likely. This difference, and others like it for higher degrees of overlap, allows for the inference of a time's arrow. Overlap is more likely to begin at the beginning of some other speaker's talk than to end at the end of it; alternately, overlap more frequently terminates

silence than initiates it.

Given our observations in this work, we revisit several seminal claims of conversation analysis, as summarized in [10].

6.1. Degree of overlap

[10] claims that

> "[...] it turns out with great regularity that, when more than one person is talking at a time, TWO persons are talking at a time, and not more; this appears to be invariant to the number of participants in the interaction" (page 7).

Although we do not contrast conversations of differing numbers of participants, the conversations under study here contain between 3 and 9 participants each, with a median of 6 participants. Across this large range, as evidenced in the right-most column of Table 1, $c_t = 3$ is 15 times less likely than $c_t = 2$, $c_t = 4$ is 30 times less likely than $c_t = 3$, etc. This progression has already been quantified in [3].

> "[...] the vast majority of instances of three talking at a time involve two speakers who simultaneously start next turns in terminal overlap with the incipient turn completion of a third [...]" (page 7).

Our observations do not agree with this conclusion of [10]. Table 1 indicates that the probability of entering $c_t = 3$ from $c_t = 1$ is 0.0354%, whereas that of entering it from $c_t = 2$ is 2.0225%, more than 50 times greater. Because the probability of being in $c_t = 1$ in the first place is only 13 times higher than being in $c_t = 2$, we conclude that it is not the simultaneous start of 2 participants during the turn completion of a third participants that accounts for the majority of time in $c_t = 3$. However, our condition of simultaneity is that two events occur within less that 100 ms of one another, which may be more precise than [10] assumed.

6.2. Reducing the degree of overlap

[10] also states that

> "Talk by MORE than two at a time seems to be reduced to two (or to one) even more effectively than talk by two is reduced to one" (page 7).

This is clearly in evidence in Table 1. The probability of a $c_{t-1} = 2$ to $c_t = 1$ transition is 24.8%, whereas the probabilities of $c_{t-1} = 3$ to $c_t \in \{1, 2\}$ and of $c_{t-1} = 4$ to $c_t \in \{1, 2, 3\}$ are $6.2 + 32.1 = 38.3\%$ and $0.9 + 10.0 + 39.1 = 50.0\%$, respectively.

6.3. The resulting duration of overlap

The distribution of overlap durations in Figure 1 supports the qualitative claim in [10] that

> "Most overlaps are over very quickly" (page 10).

Assuming that a "single beat" in [10] corresponds to a typical syllable duration of 200-300 ms allows us to check whether

> "Many overlaps are resolved after a single beat by the withdrawal of one or both parties at the first evidence that simultaneous talk is in progress" (page 22).

Indeed, the most frequent duration of contiguous intervals of talk by two or more participants ($n \geq 2$ in Figure 1) and by three or more participants ($n \geq 3$) is about 200-300 ms. Intervals of overlapping talk by at least four and at least five participants are usually even shorter. Furthermore, these overlap durations are substantially shorter than the intervals of talk by at least one participant, where the mode of the distribution os located near 1000 ms.

Similarly, Figure 1 indicates that most intervals of talk by two or more participants, and practically all intervals of overlap of higher numbers of participants are resolved to a lower degree of overlap after 1 second, verifying the claim that

> "[...] the vast majority of overlaps are resolved to a single speaker by the third beat" (page 24).

7. Conclusions

We have explored the dynamics of overlap in multi-party conversation. The durations of intervals of overlap were observed to be approximately log-normally distributed, with duration inversely proportional to degree. We showed that simultaneous talk tends to end simultaneously less frequently than it begins simultaneously. The difference is significant and discriminative. Additional experiments revealed that the asymmetry largely disappears if we remove those productions which are not syntactically complete, or those which implement non-propositional dialog act types, used to manage conversational flow. The findings, we have argued, have an impact on the design of dialogue systems, because they augment current knowledge that overlap will occur with knowledge of *when* it is likely to.

8. Acknowledgements

The work was supported in part by the Riksbankens Jubileumsfond (RJ) project Samtalets Prosodi.

9. References

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, pp. 696–735, 1974.

[2] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findings and implicatins for automatic processing of multi-party conversation," in *Proc. EUROSPEECH*, Aalborg, Denmark, 2001, vol. 2, pp. 1359–1362.

[3] Ö. Çetin and E. Shriberg, "Overlap in meetings: ASR effects and analysis by dialog factors, speakers, and collection site," in *Proc. MLMI*, Bethesda MD, USA, 2006, vol. 4299 of *Springer LNCS*, pp. 212–224.

[4] A. S. Eddington, *The Nature of the Physical World*, The University Press, Cambridge, UK, 1928.

[5] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI Meeting Corpus," in *Proc. ICASSP*, Hong Kong, China, April 2003, pp. 364–367.

[6] E. Shriberg, R. Dhillon, S. Bhagat, S. Ang, and H. Carvey, "The ICSI Meeting Recorder Dialog Act (MRDA) Corpus," in *Proc. SIGdial*, Cambridge MA, USA, April 2004, pp. 97–100.

[7] R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg, "Meeting recorder project: Dialog act labeling guide," Tech. Rep. TR-04-002, ICSI, Berkeley CA, USA, February 2004.

[8] K. Laskowski, "Modeling norms of turn-taking in multi-party conversation," in *Proc. ACL*, Uppsala, Sweden, July 2010, pp. 999–1008.

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

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