

Errata 1

This fascicle comprises the first errata (03 November 2010) to

K. LASKOWSKI (2010), “Modeling norms of turn-taking in multi-party conversation”, in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL2010)*, Uppsala, Sweden, 11–16 July, pages 999–1008.

SCOPE

The fascicle updates the numerical results in Tables 1 and 2, found on page 8 of the article (page 1006 as published by the ACL). The claim of a relative reduction in perplexity of 75% for all instants t , and of 78% for all those instants for which $\mathbf{q}_{t-1} \neq \mathbf{q}_t$, is modified to 50% and 54%, respectively. These numbers appeared on pages 1, 8, and 9 (pages 999, 1006, and 1007 as published by the ACL). Changes are identified with gray changebars and red lettering in the modified version of the article, which begins on page 3 of this fascicle.

Code replicating all numbers found in the corrected Tables 1 and 2 in this fascicle is available at the time of this writing at

http://www.cs.cmu.edu/~kornel/software_edo.html.

CORRECTIONS

This fascicle describes 4 corrections, one of which is major.

Correction 1 (minor): Table 1

The original article fails to mention that the models described by these results were smoothed by allocating a proportion of probability mass for events not seen during training. Smoothing was performed identically for all models; the held-out proportion of probability mass was 1×10^{-5} ; it was distributed uniformly over all unseen alternatives, separately for each conditioning context.

Sharing the held-out proportion of probability mass had been incorrectly implemented (it was done prior to ensuring that transition probabilities summed to unity). Correcting this error leads to updates in the mismatched conditions (B+A and D+C) for direct compositional models, and in all conditions for extended degree-of-overlap models. The corrections are identified in red lettering; their absolute values are no larger than 0.05%rel for direct compositional models for “all” instants, no larger than 0.7%rel for direct compositional models for “sub” instants, and no larger than 0.09%rel for extended degree-of-overlap models for both “all” and “sub” instants. These variations are negligible, and do not reverse the conclusions of the article as published.

The new results can be replicated using the MATLABTM script `ACL2010_Table1.m` (for direct compositional models) and the Bourne Again shell script `ACL2010_Table1.sh` (for extended degree-of-overlap models).

Correction 2 (major): Table 2

Extended degree-of-overlap model perplexities for unseen conversations are underestimated in the original article. The error involves the constant K_{max} , mentioned in Equations 18, 19, and 20. In the original application of the EDO model, in vocal activity detection, it was determined that improved results can be obtained by interdicting multi-participant vocal activity states with high degrees of overlap; this interdiction was achieved by eliminating states whose degree of overlap exceeded K_{max} . In the current application (in which vocal activity is observed rather than to-be-detected), all multi-participant vocal activity states should be allowed.

The modification requires two separate K_{max} thresholds, one governing the number of parameters in the EDO model (K_{max}^{model}), the other governing the licensed degree of overlap ($K_{max}^{topology}$) in the decoding

topology (for which transition probabilities are deterministically derived from the EDO model). In the current experiments, $K_{max}^{topology} \equiv K$, where K is the known number of participants in the test conversation. K_{max}^{model} governs the size complexity of the EDO model, trained without knowledge of the K of any test conversation to which it may be applied. (In contrast, for decoding, $K_{max}^{topology} = K_{max}^{model}$.)

The entries of Table 2, when corrected as explained in the preceding paragraph, are modified as shown in red lettering. Grossly, EDO perplexities do not approach the oracle Θ^{CD} perplexities to the same extent as appeared in the originally published article. The reduction of perplexity, relative to the Θ_{any}^{MI} model which ignores interaction, is 50% rather than 75%. Furthermore, the originally published article suggested that $K_{max}^{model} = 4$ was the optimal value for predicting vocal activity in the ICSI Meeting Corpus, at 100-ms frame steps. The new results suggest no ceiling on K_{max}^{model} , although perplexities with $K_{max}^{model} = 5$ are close to those with $K_{max}^{model} = 6$, beyond the precision reported in the table.

The relative improvement of 50%rel rather than 75%rel (and 54%rel rather than 78%rel for those instances for which $\mathbf{q}_{t-1} \neq \mathbf{q}_t$) appear in: (1) the last paragraph of the Introduction, on page 1 of the original article (page 999 as published by the ACL); the last sentence of the Conclusions, on page 9 of the original article (page 1007 as published by the ACL); and in Subsection 7.2 on page 8 of the original article (page 1006 as published by the ACL). All have been corrected, as shown with changebars and red lettering. The results of Table 2 can be replicated using the MATLABTMscripts `ACL2010_Table2_CD.m` and `ACL2010_Table2_UI.m`, and the Bourne Again script `ACL2010_Table2.sh`.

Correction 3 (typographic): Θ^{MI}

The model identifier Θ^{MI} on line 7 of Table 1 and on line 4 of Table 2 has been corrected to Θ_{any}^{MI} . The impact of this error is assumed to be of no consequence.

Correction 4 (typographic): Θ^{EDO}

The model mentioned on the last line of Subsection 7.2, Θ^{EDO} , has been corrected to Θ_{any}^{MI} . The impact of this error is assumed to be of no consequence.

Modeling Norms of Turn-Taking in Multi-Party Conversation

Kornel Laskowski

Carnegie Mellon University

Pittsburgh PA, USA

kornel@cs.cmu.edu

Abstract

Substantial research effort has been invested in recent decades into the computational study and automatic processing of multi-party conversation. While most aspects of conversational speech have benefited from a wide availability of analytic, computationally tractable techniques, only qualitative assessments are available for characterizing multi-party turn-taking. The current paper attempts to address this deficiency by first proposing a framework for computing turn-taking model perplexity, and then by evaluating several multi-participant modeling approaches. Experiments show that direct multi-participant models do not generalize to held out data, and likely never will, for practical reasons. In contrast, the Extended-Degree-of-Overlap model represents a suitable candidate for future work in this area, and is shown to successfully predict the distribution of speech in time and across participants in previously unseen conversations.

1 Introduction

Substantial research effort has been invested in recent decades into the computational study and automatic processing of multi-party conversation. Whereas sociolinguists might argue that multi-party settings provide for the most natural form of conversation, and that dialogue and monologue are merely degenerate cases (Jaffe and Feldstein, 1970), computational approaches have found it most expedient to leverage past successes; these often involved at most one speaker. Consequently, even in multi-party settings, automatic systems generally continue to treat participants independently, fusing information across participants relatively late in processing.

This state of affairs has resulted in the near-exclusion from computational consideration and from semantic analysis of a phenomenon which occurs at the lowest level of speech exchange, namely the relative timing of the deployment of speech in arbitrary multi-party groups. This phenomenon, the implicit taking of turns at talk (Sacks et al., 1974), is important because unless participants adhere to its general rules, a conversation would simply not take place. It is therefore somewhat surprising that while most other aspects of speech enjoy a large base of computational methodologies for their study, there are few quantitative techniques for assessing the flow of turn-taking in general multi-party conversation.

The current work attempts to address this problem by proposing a simple framework, which, at least conceptually, borrows quite heavily from the standard language modeling paradigm. First, it defines the perplexity of a vector-valued Markov process whose multi-participant states are a concatenation of the binary states of individual speakers. Second, it presents some obvious evidence regarding the unsuitability of models defined directly over this space, under various assumptions of independence, for the inference of conversation-independent norms of turn-taking. Finally, it demonstrates that the extended-degree-of-overlap model of (Laskowski and Schultz, 2007), which models participants in an alternate space, achieves by far the best likelihood estimates for previously unseen conversations. This appears to be because the model can learn *across* conversations, regardless of the number of their participants. Experimental results show that it yields relative perplexity reductions of approximately 50% when compared to the ubiquitous single-participant model which ignores interlocutors, indicating that it can learn and generalize aspects of interaction which direct multi-participant models, and merely single-participant models, cannot.

2 Data

Analysis and experiments are performed using the ICSI Meeting Corpus (Janin et al., 2003; Shriberg et al., 2004). The corpus consists of 75 meetings, held by various research groups at ICSI, which would have occurred even if they had not been recorded. This is important for studying naturally occurring interaction, since any form of intervention (including occurrence staging solely for the purpose of obtaining a record) may have an unknown but consistent impact on the emergence of turn-taking behaviors. Each meeting was attended by 3 to 9 participants, providing a wide variety of possible interaction types.

3 Conceptual Framework

3.1 Definitions

Turn-taking is a generally observed phenomenon in conversation (Sacks et al., 1974; Goodwin, 1981; Schegloff, 2007); one party talks while the others listen. Its description and analysis is an important problem, treated frequently as a sub-domain of linguistic pragmatics (Levinson, 1983). In spite of this, linguists tend to disagree about what precisely constitutes a *turn* (Sacks et al., 1974; Edelsky, 1981; Goodwin, 1981; Traum and Heeman, 1997), or even a turn boundary. For example, a “yeah” produced by a listener to indicate attentiveness, referred to as a *backchannel* (Yngve, 1970), is often considered to not implement a turn (nor to delineate an ongoing turn of an interlocutor), as it bears no propositional content and does not “take the floor” from the current speaker.

To avoid being tied to any particular sociolinguistic theory, the current work equates “turn” with any contiguous interval of speech uttered by the same participant. Such intervals are commonly referred to as *talk spurts* (Norwine and Murphy, 1938). Because Norwine and Murphy’s original definition is somewhat ambiguous and non-trivial to operationalize, this work relies on that proposed by (Shriberg et al., 2001), in which *spurts* are “defined as *speech regions uninterrupted by pauses longer than 500 ms*” (italics in the original). Here, a threshold of 300 ms is used instead, as recently proposed in NIST’s Rich Transcription Meeting Recognition evaluations (NIST, 2002). The resulting definition of talk spurt, it is important to note, is in quite common use but frequently under different names. An oft-cited example is the

inter-pausal unit of (Koiso et al., 1998)¹, where the threshold is 100 ms.

A consequence of this choice is that any *model of turn-taking behavior* inferred will effectively be a model of the distribution of speech, in time and across participants. If the parameters of such a model are maximum likelihood (ML) estimates, then that model will best account for what is most likely, or most “normal”; it will constitute a *norm*.

Finally, an important aspect of this work is that it analyzes turn-taking behavior as independent of the words spoken (and of the ways in which those words are spoken). As a result, strictly speaking, what is modeled is not the distribution of speech in time and across participants but of binary *speech activity* in time and across participants. Despite this seemingly dramatic simplification, it will be seen that important aspects of turn-taking are sufficiently rare to be problematic for modeling. Modeling them jointly alongside lexical information, in multi-party scenarios, is likely to remain intractable for the foreseeable future.

3.2 The Vocal Interaction Record \mathbf{Q}

The notation used here, as in (Laskowski and Schultz, 2007), is a trivial extension of that proposed in (Rabiner, 1989) to vector-valued Markov processes.

At any instant t , each of K participants to a conversation is in a state drawn from $\Psi \equiv \{S_0, S_1\} \equiv \{\square, \blacksquare\}$, where $S_1 \equiv \blacksquare$ indicates speech (or, more precisely, “*intra-talk-spurt instants*”) and $S_0 \equiv \square$ indicates non-speech (or “*inter-talk-spurt instants*”). The *joint* state of all participants at time t is described using the K -length column vector

$$\begin{aligned} \mathbf{q}_t \in \Psi^K &\equiv \Psi \times \Psi \times \dots \times \Psi \\ &\equiv \{S_0, S_1, \dots, S_{2^K-1}\}. \end{aligned} \quad (1)$$

An entire conversation, from the point of view of this work, can be represented as the matrix

$$\begin{aligned} \mathbf{Q} &\equiv [\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_T] \\ &\in \Psi^{K \times T}. \end{aligned} \quad (2)$$

\mathbf{Q} is known as the (discrete) vocal interaction (Dabbs and Ruback, 1987) record. T is the total number of frames in the conversation, sampled at $T_s = 100$ ms intervals. This is approximately the duration of the shortest lexical productions in the ICSI Meeting Corpus.

¹The inter-pausal unit differs from the *pause unit* of (Seligman et al., 1997) in that the latter is an intra-turn unit, requiring prior turn segmentation

3.3 Time-Independent First-Order Markov Modeling of \mathbf{Q}

Given this definition of \mathbf{Q} , a model Θ is sought to account for it. Only time-independent models, whose parameters do not change over the course of the conversation, are considered in this work.

For simplicity, the state $\mathbf{q}_0 = \mathbf{S}_0 = [\square, \square, \dots, \square]^*$, in which no participant is speaking (* indicates matrix transpose, to avoid confusion with conversation duration T) is first prepended to \mathbf{Q} . $P_0 = P(\mathbf{q}_0)$ therefore represents the unconditional probability of all participants being silent just prior to the start of any conversation². Then

$$\begin{aligned} P(\mathbf{Q}) &= P_0 \cdot \prod_{t=1}^T P(\mathbf{q}_t | \mathbf{q}_0, \mathbf{q}_1, \dots, \mathbf{q}_{t-1}) \\ &\doteq P_0 \cdot \prod_{t=1}^T P(\mathbf{q}_t | \mathbf{q}_{t-1}, \Theta), \end{aligned} \quad (3)$$

where in the second line the history is truncated to yield a standard first-order Markov form.

Each of the T factors in Equation 3 is independent of the instant t ,

$$\begin{aligned} P(\mathbf{q}_t | \mathbf{q}_{t-1}, \Theta) &= P(\mathbf{q}_t = \mathbf{S}_j | \mathbf{q}_{t-1} = \mathbf{S}_i, \Theta) \end{aligned} \quad (4)$$

$$\equiv a_{ij}, \quad (5)$$

as per the notation in (Rabiner, 1989). In particular, each factor is a function only of the state \mathbf{S}_i in which the conversation was at time $t - 1$ and the state \mathbf{S}_j in which the conversation is at time t , and not of the instants $t - 1$ or t . It may be expressed as the scalar a_{ij} which forms the i th row and j th column entry of the matrix $\{a_{ij}\} \equiv \Theta$.

3.4 Perplexity

In language modeling practice, one finds the likelihood $P(\mathbf{w} | \Theta)$, of a word sequence \mathbf{w} of length $\|\mathbf{w}\|$ under a model Θ , to be an inconvenient measure for comparison. Instead, the *negative log-likelihood* (NLL) and *perplexity* (PPL), defined as

$$\text{NLL} = -\frac{1}{\|\mathbf{w}\|} \log_e P(\mathbf{w} | \Theta) \quad (6)$$

$$\text{PPL} = 10^{\text{NLL}}, \quad (7)$$

²In reality, the instant $t = 0$ refers to the beginning of *the recording* of a conversation, rather than the beginning of the conversation itself; this detail is without consequence.

are often preferred (Jelinek, 1999). They are ubiquitously used to compare the complexity of different word sequences (or corpora) \mathbf{w} and \mathbf{w}' under the same model Θ , or the performance on a single word sequence (or corpus) \mathbf{w} under competing models Θ and Θ' .

Here, a similar metric is proposed, to be used for the same purposes, for the record \mathbf{Q} .

$$\text{NLL} = -\frac{1}{KT} \log_2 P(\mathbf{Q} | \Theta) \quad (8)$$

$$\begin{aligned} \text{PPL} &= 2^{\text{NLL}} \\ &= (P(\mathbf{Q} | \Theta))^{-1/KT} \end{aligned} \quad (9)$$

are defined as measures of *turn-taking perplexity*. As can be seen in Equation 8, the negative log-likelihood is normalized by the number K of participants and the number T of frames in \mathbf{Q} ; the latter renders the measure useful for making duration-independent comparisons. The normalization by K does not *per se* suggest that turn-taking in conversations with different K is necessarily similar; it merely provides similar bounds on the magnitudes of these metrics.

4 Direct Estimation of Θ

Direct application of bigram modeling techniques, defined over the states $\{\mathbf{S}\}$, is treated as a baseline.

4.1 The Case of $K = 2$ Participants

In contrast to multi-party conversation, dialogue has been extensively modeled in the ways described in this paper. Beginning with (Brady, 1969), Markov modeling techniques over the joint speech activity of two interlocutors have been explored by both the sociolinguist and the psycholinguist community (Jaffe and Feldstein, 1970; Dabbs and Ruback, 1987). The same models have also appeared in dialogue systems (Raux, 2008). Most recently, they have been augmented with duration models in a study of the Switchboard corpus (Grothendieck et al., 2009).

4.2 The Case of $K > 2$ Participants

In the general case beyond dialogue, such models have found less traction. This is partly due to the exponential growth in the number of states as K increases, and partly due to difficulties in interpretation. The only model for arbitrary K that the author is familiar with is the GroupTalk model (Dabbs and Ruback, 1987), which is unsuitable for the purposes here as it does not scale (with K ,

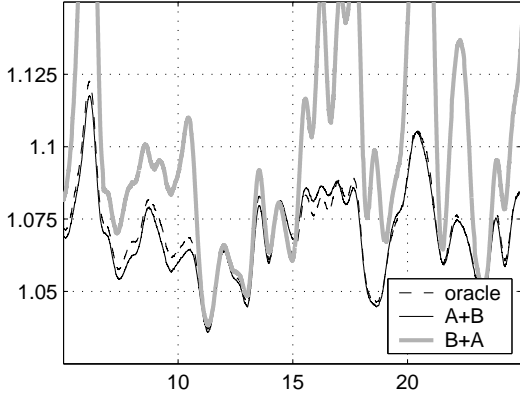


Figure 1: Perplexity (along y -axis) in time (along x -axis, in minutes) for meeting Bmr024 under a conditionally dependent global oracle model, two “matched-half” models (A+B), and two “mismatched-half” models (B+A).

the number of participants) without losing track of speakers when two or more participants speak simultaneously (known as *overlap*).

4.2.1 Conditionally Dependent Participants

In a particular conversation with K participants, the state space of an ergodic process contains 2^K states, and the number of free parameters in a model Θ which treats participant behavior as *conditionally dependent* (CD), henceforth Θ^{CD} , scales as $2^K \cdot (2^K - 1)$. It should be immediately obvious that many of the 2^K states are likely to not occur within a conversation of duration T , leading to misestimation of the desired probabilities.

To demonstrate this, three perplexity trajectories for a snippet of meeting Bmr024 are shown in Figure 1, in the interval beginning 5 minutes into the meeting and ending 20 minutes later. (The meeting is actually just over 50 minutes long but only a snippet is shown to better appreciate small time-scale variation.) The depicted perplexities are not unweighted averages over the whole meeting of duration T as in Equation 8, but over a 60-second Hamming window centered on each t .

The first trajectory, the dashed black line, is obtained when the entire meeting is used to estimate Θ^{CD} , and is then scored by that same model (an “oracle” condition). Significant perplexity variation is observed throughout the depicted snippet.

The second trajectory, the continuous black line, is that obtained when the meeting is split into two equal-duration halves, one consisting of all instants prior to the midpoint and the other of all

instants following it. These halves are hereafter referred to as A and B, respectively (the interval in Figure 1 falls entirely within the A half). Two separate models Θ_A^{CD} and Θ_B^{CD} are each trained on only one of the two halves, and then applied to those same halves. As can be seen at the scale employed, the matched A+B model, demonstrating the effect of training data ablation, deviates from the global oracle model only in the intervals [7, 11] seconds and [15, 18] seconds; otherwise it appears that more training data, from later in the conversation, does not affect model performance.

Finally, the third trajectory, the continuous gray line, is obtained when the two halves A and B of the meeting are scored using the mismatched models Θ_B^{CD} and Θ_A^{CD} , respectively (this condition is henceforth referred to as the B+A condition). It can be seen that even when probabilities are estimated from the same participants, in exactly the same conversation, a direct conditionally dependent model exposed to over 25 minutes of a conversation cannot predict the turn-taking patterns observed later.

4.2.2 Conditionally Independent Participants

A potential reason for the gross misestimation of Θ^{CD} under mismatched conditions is the size of the state space $\{\mathbf{S}\}$. The number of parameters in the model can be reduced by assuming that participants behave *independently* at instant t , but are conditioned on their *joint* behavior at $t - 1$. The likelihood of \mathbf{Q} under the resulting *conditionally independent model* Θ^{CI} has the form

$$P(\mathbf{Q}) \doteq P_0 \cdot \prod_{t=1}^T \prod_{k=1}^K P(\mathbf{q}_t[k] | \mathbf{q}_{t-1}, \Theta_k^{CI}), \quad (10)$$

where each factor is time-independent,

$$P(\mathbf{q}_t[k] | \mathbf{q}_{t-1}, \Theta_k^{CI}) = P(\mathbf{q}_t[k] = S_n | \mathbf{q}_{t-1} = \mathbf{S}_i, \Theta_k^{CI}) \quad (11)$$

$$\equiv a_{k,in}^{CI}, \quad (12)$$

with $0 \leq i < 2^K$ and $0 \leq n < 2$. The complete model $\{\Theta_k^{CI}\} \equiv \{\{a_{k,in}^{CI}\}\}$ consists of K matrices of size $2^K \times 2$ each. It therefore contains only $K \cdot 2^K$ free parameters, a significant reduction over the conditionally dependent model Θ^{CD} .

Panel (a) of Figure 2 shows the performance of this model on the same conversational snippet

as in Figure 1. The oracle, dashed black line of the latter is reproduced as a reference. The continuous black and gray lines show the smoothed perplexity for the matched (A+B) and the mismatched (B+A) conditions, respectively. In the matched condition, the CI model reproduces the oracle trajectory with relatively high fidelity, suggesting that participants’ behavior may in fact be assumed to be conditionally independent in the sense discussed. Furthermore, the failures of the CI model under mismatched conditions are less severe in magnitude than those of the CD model.

Panel (b) of Figure 2 demonstrates the trivial fact that a conditionally independent model Θ_{any}^{CI} , tying the statistics of all K participants into a single model, is useless. This is of course because it cannot predict the next state of a generic participant for which the index k in \mathbf{q}_{t-1} has been lost.

4.2.3 Mutually Independent Participants

A further reduction in the complexity of Θ can be achieved by assuming that participants are *mutually independent* (MI), leading to the participant-specific Θ_k^{MI} model:

$$P(\mathbf{Q}) \doteq P_0 \cdot \prod_{t=1}^T \prod_{k=1}^K P(\mathbf{q}_t[k] | \mathbf{q}_{t-1}[k], \Theta_k^{MI}). \quad (13)$$

The factors are time-independent,

$$P(\mathbf{q}_t[k] | \mathbf{q}_{t-1}[k], \Theta_k^{MI}) = P(\mathbf{q}_t[k] = S_n | \mathbf{q}_{t-1}[k] = S_m, \Theta_k^{MI}) \quad (14)$$

$$\equiv a_{k,mn}^{MI}, \quad (15)$$

where $0 \leq m < 2$ and $0 \leq n < 2$. This model $\{\Theta_k^{MI}\} \equiv \{a_{k,mn}^{MI}\}$ consists of K matrices of size 2×2 each, with only $K \cdot 2$ free parameters.

Panel (c) of Figure 2 shows that the MI model yields mismatched performance which is a much better approximation to its performance under matched conditions. However, its matched performance is worse than that of CD and CI models. When a single MI model Θ_{any}^{MI} is trained instead for all participants, as shown in panel (d), both of these effects are exaggerated. In fact, the performance of Θ_{any}^{MI} in matched and mismatched conditions is almost identical. The consistently higher perplexity is obtained, as mentioned, by smoothing over 60-second windows, and therefore underestimates poor performance at specific instants (which occur frequently).

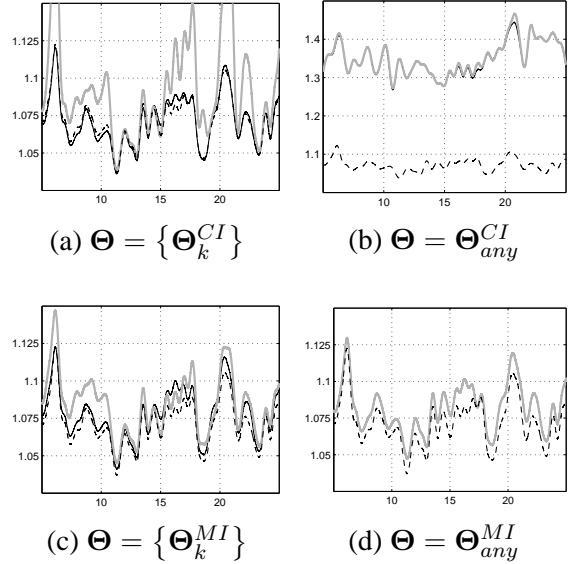


Figure 2: Perplexity (along y -axis) in time (along x -axis, in minutes) for meeting Bmr024 under a conditionally dependent global oracle model, and various matched (A+B) and mismatched (B+A) model pairs with relaxed dependence assumptions. Legend as in Figure 1.

5 Limitations and Desiderata

As the analyses in Section 4 reveal, direct estimation can be useful under oracle conditions, namely when all of a conversation has been observed and the task is to find intervals where multi-participant behavior deviates significantly from its *conversation-specific* norm. The assumption of conditional independence among participants was argued to lead to negligible degradation in the detectability of these intervals. However, the assumption of mutual independence consistently leads to higher surprise by the model.

5.1 Predicting the Future Within Conversations

In the more interesting setting in which only a part of a conversation has been seen and the task is to limit the perplexity of what is still to come, direct estimation exhibits relatively large failures under both conditionally dependent and conditionally independent participant assumptions. This appears to be due to the size of the state space, which scales as 2^K with the number K of participants. In the case of general K , more conversational data may be sought, from exactly the same group of participants, but that approach appears likely to be

insufficient, and, for practical reasons³, impossible. One would instead like to be able to use other conversations, also exhibiting participant interaction, to limit the perplexity of speech occurrence in the conversation under study.

Unfortunately, there are two reasons why direct estimation cannot be tractably deployed across conversations. The first is that the direct models considered here, with the exception of Θ_{any}^{MI} , are K -specific. In particular, the number and the identity of conditioning states are both functions of K , for Θ^{CD} and $\{\Theta_k^{CI}\}$; the models may also consist of K distinct submodels, as for $\{\Theta_k^{CI}\}$ and $\{\Theta_k^{MI}\}$. No techniques for computing the turn-taking perplexity in conversations with K participants, using models trained on conversations with $K' \neq K$, are currently available.

The second reason is that these models, again with the exception of Θ_{any}^{MI} , are \mathbf{R} -specific, independently of K -specificity. By this it is meant that the models are sensitive to participant index permutation. Had a participant at index k in \mathbf{Q} been assigned to another index $k' \neq k$, an alternate representation of the conversation, namely $\mathbf{Q}' = \mathbf{R}_{kk'} \cdot \mathbf{Q}$, would have been obtained. (Here, $\mathbf{R}_{kk'}$ is a matrix rotation operator obtained by exchanging columns k and k' of the $K \times K$ identity matrix \mathbf{I} .) Since index assignment is entirely arbitrary, useful direct models cannot be inferred from other conversations, even when their $K' = K$, unless K is small. The prospect of naively permuting every training conversation prior to parameter inference has complexity $K!$.

5.2 Comparing Perplexity Across Conversations

Until \mathbf{R} -specificity is comprehensively addressed, the only model from among those discussed so far, which exhibits no K -dependence, is Θ_{any}^{MI} , namely that which treats participants identically and independently. This model can be used to score the perplexity of *any* conversation, and facilitates the comparison of the distribution of speech activity *across* conversations.

Unfortunately, since the model captures only durational aspects of *one*-participant speech and non-speech intervals, it does not in any way encode a norm of turn-taking, an inherently interac-

tive and hence *multi*-participant phenomenon. It therefore cannot be said to rank conversations according to their deviation from turn-taking norms.

5.3 Theoretical Limitations

In addition to the concerns above, a fundamental limitation of the analyzed direct models, whether for conversation-specific or conversation-independent use, is that they are theoretically cumbersome if not vacuous. Given a solution to the problem of \mathbf{R} -specificity, the parameters $\{a_{ij}^{CD}\}$ may be robustly inferred, and the models may be applied to yield useful estimates of turn-taking perplexity. However, they cannot be said to directly validate or dispute the vast qualitative observations of sociolinguistics, and of conversation analysis in particular.

5.4 Prospects for Smoothing

To produce Figures 1 and 2, a small fraction of probability mass was reserved for unseen bigram transitions (as opposed to backing off to unigram probabilities). Furthermore, transitions into never-observed states were assigned uniform probabilities. This policy is simplistic, and there is significant scope for more detailed back-off and interpolation. However, such techniques infer values for under-estimated probabilities from *shorter truncations of the conditioning history*. As K -specificity and \mathbf{R} -specificity suggest, what appears to be needed here are back-off and interpolation *across states*. For example, in a conversation of $K = 5$ participants, estimates of the likelihood of the state $\mathbf{q}_t = [\square \blacksquare \blacksquare \blacksquare \square]^*$, which might have been unobserved in any training material, can be assumed to be related to those of $\mathbf{q}'_t = [\square \square \blacksquare \blacksquare \square]^*$ and $\mathbf{q}''_t = [\square \blacksquare \blacksquare \square \square]^*$, as well as those of $\mathbf{R}\mathbf{q}'_t$ and $\mathbf{R}\mathbf{q}''_t$, for arbitrary \mathbf{R} .

6 The Extended-Degree-of-Overlap Model

The limitations of direct models appear to be addressable by a form proposed by Laskowski and Schultz in (2006) and (2007). That form, the Extended-Degree-of-Overlap (EDO) model, was used to provide prior probabilities $P(\mathbf{Q} | \Theta)$ of the speech states of multiple meeting participants simultaneously, for use in speech activity detection. The model was trained on *utterances* (rather than talk spurts) from a different corpus than that

³This pertains to the practicalities of re-inviting, instrumenting, recording and transcribing the same groups of participants, with necessarily more conversations for large groups than for small ones.

used here, and the authors did not explore the turn-taking perplexities of their data sets.

Several of the equations in (Laskowski and Schultz, 2007) are reproduced here for comparison. The EDO model yields time-independent transition probabilities which assume conditional inter-participant dependence (cf. Equation 3),

$$P(\mathbf{q}_{t+1} = \mathbf{S}_j \mid \mathbf{q}_t = \mathbf{S}_i) = \alpha_{ij} \cdot \quad (16)$$

$$P(\|\mathbf{q}_{t+1}\| = n_j, \|\mathbf{q}_{t+1} \cdot \mathbf{q}_t\| = o_{ij} \mid \|\mathbf{q}_t\| = n_i),$$

where $n_i \equiv \|\mathbf{S}_i\|$ and $n_j \equiv \|\mathbf{S}_j\|$, with $\|\mathbf{S}\|$ yielding the number of participants in \blacksquare in the multi-participant state \mathbf{S} . In other words, n_i and n_j are the *numbers* of participants simultaneously speaking in states \mathbf{S}_i and \mathbf{S}_j , respectively. The elements of the binary product $\mathbf{S} = \mathbf{S}_1 \cdot \mathbf{S}_2$ are given by

$$\mathbf{S}[k] \equiv \begin{cases} \blacksquare, & \text{if } \mathbf{S}_1[k] = \mathbf{S}_2[k] = \blacksquare \\ \square, & \text{otherwise,} \end{cases} \quad (17)$$

and o_{ij} is therefore the number of same participants speaking in \mathbf{S}_i and \mathbf{S}_j . The discussion of the role of α_{ij} in Equation 16 is deferred to the end of this section.

The EDO model mitigates \mathbf{R} -specificity because it models each bigram $(\mathbf{q}_{t-1}, \mathbf{q}_t) = (\mathbf{S}_i, \mathbf{S}_j)$ as the modified bigram $(n_i, [o_{ij}, n_j])$, involving three scalars each of which is a *sum* — a commutative (and therefore rotation-invariant) operation. Because it sums across only those participants which are in the \blacksquare state, completely ignoring their \square -state interlocutors, it can also mitigate K -specificity if one additionally redefines

$$n_i = \min(\|\mathbf{S}_i\|, K_{max}) \quad (18)$$

$$n_j = \min(\|\mathbf{S}_j\|, K_{max}) \quad (19)$$

$$o_{ij} = \min(\|\mathbf{S}_i \cdot \mathbf{S}_j\|, n_i, n_j), \quad (20)$$

as in (Laskowski and Schultz, 2007). K_{max} represents the maximum model-licensed degree of overlap, or the maximum number of participants allowed to be simultaneously speaking. The EDO model therefore represents a viable conversation-independent, K -independent, and \mathbf{R} -independent model of turn-taking for the purposes in the current work⁴. The factor α_{ij}

⁴There exists some empirical evidence to suggest that conversations of K participants should not be used to train models for predicting turn-taking behavior in conversations of K' participants, for $K' \neq K$, because turn-taking is inherently K -dependent. For example, (Fay et al., 2000) found that qualitative differences in turn-taking patterns between

in Equation 16 provides a deterministic mapping from the conversation-independent space $(n_i, [o_{ij}, n_j])$ to the conversation-specific space $\{a_{ij}\}$. The mapping is deterministic because the model assumes that all participants are identical. This places the EDO model at a disadvantage with respect to the CD and CI models, as well as to $\{\Theta_k^{MI}\}$, which allow each participant to be modeled differently.

7 Experiments

This section describes the performance of the discussed models on the entire ICSI Meeting Corpus.

7.1 Conversation-Specific Modeling

First to be explored is the prediction of yet-unobserved behavior in conversation-specific settings. For each meeting, models are trained on portions of that meeting only, and then used to score other portions of the same meeting. This is repeated over all meetings, and comprises the mismatched condition of Section 4; for contrast, the matched condition is also evaluated.

Each meeting is divided into two halves, in two different ways. The first way is the A/B split of Section 4, representing the first and second halves of each meeting; as has been shown, turn-taking patterns may vary substantially from A to B. The second split (C/D) places every even-numbered frame in one set and every odd-numbered frame in the other. This yields a much easier setting, of two halves which are on average maximally similar but still temporally disjoint.

The perplexities (of Equation 9) in these experiments are shown in the second, fourth, sixth and eighth columns of Table 1, under “all”. In the matched A+B and C+D conditions, the conditionally dependent model Θ^{CD} provides topline ML performance. Perplexities decrease as model complexities fall for direct models, as expected. However, in the more interesting mismatched B+A condition, the EDO model performs the best. This shows that its ability to generalize to unseen data is higher than that of direct models. However, in the easier mismatched D+C condition, it is outperformed by the CI model due to behavior differences among participants, which the EDO model

small groups and large groups, represented in their study by $K = 5$ and $K = 10$, and noted that there is a smooth transition between the two extremes; this provides some scope for interpolating small- and large- group models, and the EDO framework makes this possible.

Model	Hard split A/B (first/second halves)				Easy split C/D (odd/even frames)			
	A+B		B+A		C+D		D+C	
	“all”	“sub”	“all”	“sub”	“all”	“sub”	“all”	“sub”
Θ^{CD}	1.0905	1.6444	1.1224	1.8276	1.0915	1.6555	1.0993	1.7413
$\{\Theta_k^{CI}\}$	1.0915	1.6576	1.1154	1.7749	1.0925	1.6695	1.0961	1.7086
$\{\Theta_k^{MI}\}$	1.0978	1.7236	1.1090	1.7829	1.0991	1.7381	1.0992	1.7398
Θ_{any}^{MI}	1.1046	1.8047	1.1047	1.8059	1.1046	1.8050	1.1046	1.8052
Θ^{EDO}	1.0976	1.7248	1.0984	1.7309	1.0977	1.7259	1.0981	1.7302

Table 1: Perplexities for conversation-specific turn-taking models on the entire ICSI Meeting Corpus. Both “all” frames and the subset (“sub”) for which $\mathbf{q}_{t-1} \neq \mathbf{q}_t$ are shown, for matched (A+B and C+D) and mismatched (B+A and D+C) conditions on splits A/B and C/D.

does not capture.

The numbers under the “all” columns in Table 1 were computed using all of each meeting’s frames. For contrast, in the “sub” columns, perplexities are computed over only those frames for which $\mathbf{q}_{t-1} \neq \mathbf{q}_t$. This is a useful subset because, for the majority of time in conversations, one person simply continues to talk while all others remain silent⁵. Excluding $\mathbf{q}_{t-1} = \mathbf{q}_t$ bigrams (leading to 0.32M frames from 2.39M frames in “all”) offers a glimpse of expected performance differences were duration modeling to be included in the models. Perplexities are much higher in these intervals, but the same general trend as for “all” is observed.

7.2 Conversation-Independent Modeling

The training of conversation-independent models, given a corpus of K -heterogeneous meetings, is achieved by iterating over all meetings and testing each using models trained on all of the other meetings. As discussed in the preceding section, Θ_{any}^{MI} is the only one among the direct models which can be used for this purpose. It also models exclusively single-participant behavior, ignoring the interactive setting provided by other participants. As shown in Table 2, when all time is scored the EDO model with $K_{max} = 6$ is the best model (in Section 7.1, $K_{max} = K$ since the model was trained on the same meeting to which it was applied). Its perplexity gap to the oracle model is only **half** of the gap exhibited by Θ_{any}^{MI} .

The relative performance of EDO models is even better when only those instants t are considered for which $\mathbf{q}_{t-1} \neq \mathbf{q}_t$. There, the perplexity gap to the oracle model is smaller than that of

⁵Retaining only $\mathbf{q}_{t-1} \neq \mathbf{q}_t$ also retains instants of transition into and out of intervals of silence.

Model	PPL		Δ PPL (%)	
	“all”	“sub”	“all”	“sub”
Θ^{CD}	1.0921	1.6616	—	—
Θ_{any}^{MI}	1.1051	1.8170	14.1	23.5
Θ^{EDO} (6)	1.0985	1.7329	7.0	10.8
Θ^{EDO} (5)	1.0985	1.7329	7.0	10.8
Θ^{EDO} (4)	1.0986	1.7330	7.1	10.8
Θ^{EDO} (3)	1.0997	1.7366	8.3	11.3

Table 2: Perplexities for conversation-independent turn-taking models on the entire ICSI Meeting Corpus; the oracle Θ^{CD} topline is included in the first row. Both “all” frames and the subset (“sub”) for which $\mathbf{q}_{t-1} \neq \mathbf{q}_t$ are shown; relative increases over the topline (less unity, representing no perplexity) are shown in columns 4 and 5. The value of K_{max} (cf. Equations 18, 19, and 20) is shown in parentheses in the first column.

Θ_{any}^{MI} by 54%.

8 Discussion

The model perplexities as reported above may be somewhat different if the “talk spurt” were replaced by a more sociolinguistically motivated definition of “turn”, but the ranking of models and their relative performance differences are likely to remain quite similar. On the one hand, many inter-talk-spurt gaps might find themselves to be within-turn, leading to more \blacksquare entries in the record \mathbf{Q} than observed in the current work. This would increase the apparent frequency and duration of intervals of overlap. On the other hand, alternative definitions of turn may exclude some speech activity, such as that implementing backchannels. Since backchannels are often produced in overlap

with the foreground speaker, their removal may eliminate some overlap from \mathbf{Q} . (However, as noted in (Shriberg et al., 2001), overlap rates in multi-party conversation remain high even after the exclusion of backchannels.) Both inter-talk-spurt gap inclusion and backchannel exclusion are likely to yield systematic differences, and therefore to be exploitable by the investigated models in similar ways.

The results presented may also be perturbed by modifying the way in which a (manually produced) talk spurt segmentation, with high-precision boundary time-stamps, is discretized to yield \mathbf{Q} . Two parameters have controlled the discretization in this work: (1) the frame step $T_s = 100$ ms; and (2) the proportion ρ of T_s for which a participant must be speaking within a frame in order for that frame to be considered \blacksquare rather than \square . $\rho = 0.5$ was chosen since this posits approximately as much more speech (than in the high-precision segmentation) as it eliminates. Higher values of ρ would lead to more \blacksquare , leading to more overlap than observed in this work. Meanwhile, at constant ρ , choosing a T_s value larger than 100 ms would occasionally miss the shortest talk spurts, but it would allow the models, which are all 1st-order Markovian, to learn temporally more distant dependencies. The trade-offs between these choices are currently under investigation.

From an operational, modeling perspective, it is important to recognize that the choices of the definition for “turn”, and of the way in which segmentations are discretized, are essentially arbitrary. The investigated modeling alternatives, and the EDO model in particular, require only that the multi-participant vocal interaction record \mathbf{Q} be *binary-valued*. This general applicability has been demonstrated in past work, in which the EDO model was trained on *utterances* for use in speech activity detection (Laskowski and Schultz, 2007), as well as in (Laskowski and Burger, 2007) where it was trained separately on talk spurts and *laugh bouts*, in the same data, to highlight the differences between speech and laughter deployment.

Finally, it should be remembered that the EDO model is both time-independent and participant-independent. This makes it suitable for comparison of conversational genres, in much the same way as are general language models of words. Accordingly, as for language models, density estimation in future turn-taking models may be im-

proved by considering variability across participants and in time. Participant dependence is likely to be related to speakers’ social characteristics and conversational roles, while time dependence may reflect opening and closing functions, topic boundaries, and periodic turn exchange failures. In the meantime, event types such as the latter may be detectable as EDO perplexity departures, potentially recommending the model’s use for localizing conversational “hot spots” (Wrede and Shriberg, 2003). The EDO model, and turn-taking models in general, may also find use in diagnosing turn-taking naturalness in spoken dialogue systems.

9 Conclusions

This paper has presented a framework for quantifying the turn-taking perplexity in multi-party conversations. To begin with, it explored the consequences of modeling participants jointly by concatenating their binary speech/non-speech states into a single multi-participant vector-valued state. Analysis revealed that such models are particularly poor at generalization, even to subsequent portions of the same conversation. This is due to the size of their state space, which is factorial in the number of participants. Furthermore, because such models are both specific to the number of participants and to the order in which participant states are concatenated together, it is generally intractable to train them on material from other conversations. The only such model which may be trained on other conversations is that which completely ignores interlocutor interaction.

In contrast, the Extended-Degree-of-Overlap (EDO) construction of (Laskowski and Schultz, 2007) may be trained on other conversations, regardless of their number of participants, and usefully applied to approximate the turn-taking perplexity of an oracle model. This is achieved because it models entry into and egress out of specific degrees of overlap, and completely ignores the number of participants actually present or their modeled arrangement. In this sense, the EDO model can be said to implement the qualitative findings of conversation analysis. In predicting the distribution of speech in time and across participants, it reduces the unseen data perplexity of a model which ignores interaction by 50% relative to an oracle model.

References

- Paul T. Brady. 1969. A model for generating on-off patterns in two-way conversation. *Bell Systems Technical Journal*, 48(9):2445–2472.
- James M. Dabbs and R. Barry Ruback. 1987. Dimensions of group process: Amount and structure of vocal interaction. *Advances in Experimental Social Psychology*, 20:123–169.
- Carole Edelsky. 1981. Who’s got the floor? *Language in Society*, 10:383–421.
- Nicolas Fay, Simon Garrod, and Jean Carletta. 2000. Group discussion as interactive dialogue or as serial monologue: The influence of group size. *Psychological Science*, 11(6):487–492.
- Charles Goodwin. 1981. *Conversational Organization: Interaction Between Speakers and Hearers*. Academic Press, New York NY, USA.
- John Grothendieck, Allen Gorin, and Nash Borges. 2009. Social correlates of turn-taking behavior. *Proc. ICASSP*, Taipei, Taiwan, pp. 4745–4748.
- Joseph Jaffe and Stanley Feldstein. 1970. *Rhythms of Dialogue*. Academic Press, New York NY, USA.
- Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI Meeting Corpus. *Proc. ICASSP*, Hong Kong, China, pp. 364–367.
- Frederick Jelinek. 1999. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge MA, USA.
- Hanae Koiso, Yasui Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. *Language and Speech*, 41(3-4):295–321.
- Kornel Laskowski and Tanja Schultz. 2006. Unsupervised learning of overlapped speech model parameters for multichannel speech activity detection in meetings. *Proc. ICASSP*, Toulouse, France, pp. 993–996.
- Kornel Laskowski and Susanne Burger. 2007. Analysis of the occurrence of laughter in meetings. *Proc. INTERSPEECH*, Antwerpen, Belgium, pp. 1258–1261.
- Kornel Laskowski and Tanja Schultz. 2007. Modeling vocal interaction for segmentation in meeting recognition. *Machine Learning for Multimodal Interaction*, A. Popescu-Belis, S. Renals, and H. Bourlard, eds., Lecture Notes in Computer Science, 4892:259–270, Springer Berlin/Heidelberg, Germany.
- Stephen C. Levinson. 1983. *Pragmatics*. Cambridge University Press.
- National Institute of Standards and Technology. 2002. Rich Transcription Evaluation Project, www.itl.nist.gov/iad/mig/tests/rt/ (last accessed 15 February 2010 1217hrs GMT).
- A. C. Norwine and O. J. Murphy. 1938. Characteristic time intervals in telephonic conversation. *Bell System Technical Journal*, 17:281–291.
- Lawrence Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286.
- Antoine Raux. 2008. Flexible turn-taking for spoken dialogue systems. PhD Thesis, Carnegie Mellon University.
- Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest semantics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.
- Emanuel A. Schegloff. 2007. *Sequence Organization in Interaction*. Cambridge University Press, Cambridge, UK.
- Mark Seligman, Junko Hosaka, and Harald Singer. 1997. “Pause units” and analysis of spontaneous Japanese dialogues: Preliminary studies. *Dialogue Processing in Spoken Language Systems* E. Maier, M. Mast, and S. LuperFoy, eds., Lecture Notes in Computer Science, 1236:100–112. Springer Berlin/Heidelberg, Germany.
- Elizabeth Shriberg, Andreas Stolcke, and Don Baron. 2001. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. *Proc. EUROSPEECH*, Genève, Switzerland, pp. 1359–1362.
- Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. *Proc. SIGDIAL*, Boston MA, USA, pp. 97–100.
- David Traum and Peeter Heeman. 1997. Utterance units in spoken dialogue. *Dialogue Processing in Spoken Language Systems* E. Maier, M. Mast, and S. LuperFoy, eds., Lecture Notes in Computer Science, 1236:125–140. Springer Berlin/Heidelberg, Germany.
- Britta Wrede and Elizabeth Shriberg. 2003. Spotting “hot spots” in meetings: Human judgments and prosodic cues. *Proc. EUROSPEECH*, Aalborg, Denmark, pp. 2805–2808.
- Victor H. Yngve. 1970. On getting a word in edgewise. *Papers from the Sixth Regional Meeting Chicago Linguistic Society*, pp. 567–578. Chicago Linguistic Society, Chicago IL, USA.