



Incremental Learning and Forgetting in Stochastic Turn-Taking Models

Kornel Laskowski, Jens Edlund and Mattias Heldner

KTH Speech Music and Hearing, Stockholm, Sweden

kornel@cs.cmu.edu, edlund@speech.kth.se, mattias@speech.kth.se

Abstract

We present a computational framework for stochastically modeling dyad interaction chronograms. The framework’s most novel feature is the capacity for incremental learning and forgetting. To showcase its flexibility, we design experiments answering four concrete questions about the systematics of spoken interaction. The results show that: (1) individuals are clearly affected by one another; (2) there is individual variation in interaction strategy; (3) strategies wander in time rather than converge; and (4) individuals exhibit similarity with their interlocutors. We expect the proposed framework to be capable of answering many such questions with little additional effort.

Index Terms: interaction, chronogram modeling, turn-taking, incremental learning.

4. *Is there interlocutor similarity within dyads?* If so, then adapting a general model, towards the interlocutor’s past decisions as they are taken, will improve predictions.

We present experiments which answer all four questions in the affirmative. They show that conditioning predictions on the dyadic history is consistently advantageous, independently of history duration. Models estimated using one’s own conversation-specific decisions outperform general models surprisingly soon after the start of a conversation; interpolation with a general model leads to even better cumulative performance. Forgetting one’s old decisions also appears to improve prediction, suggesting that interaction strategies¹. Finally, while a participant’s conversation-specific model poorly predicts their interlocutor’s incipient decisions, its interpolation with a general model leads to better predictions, sooner.

We believe that the described framework may have significant impact on our understanding of interaction dynamics, not only with respect to the deployment of speech, but also to laughter [5], gaze [6], and blinking [7]; the study of group activities other than conversation is equally likely to benefit. The relevance of the proposed techniques is particularly immediate for those human behaviors which can be coded as binary.

1. Introduction

“Interaction” has proven to be a slippery term, in conversational conduct as elsewhere. Despite a plethora of qualitative and even quantitative descriptions, analyses which are *prescriptive* are lacking. We can tell *if* two or more entities are interacting, but we have trouble explaining *how* any one of them should behave “instant by instant” in order to achieve interaction, and to render it overt to interlocutors and observers alike. We ourselves can do it, but the machines we build to interact with cannot.

It is our belief that a source of this seeming impasse is uncertainty about *what* to measure, to then potentially implement in autonomous entities. The space of candidate phenomena is large, and search through it is likely to take decades. Unfortunately, results such as the mean duration of overlap will still not directly shed light on how tactical human behavior changes *because* of overlap, or how that of a synthetic entity should. What is missing is a *quantitative grammar* of interaction.

We approach this problem by inferring stochastic models of a party’s actions, conditioned on an interaction’s recent history [1, 2, 3, 4], from arbitrarily large collections of human-human conduct. Model predictions are easy to evaluate, once the current instant passes. The incremental framework we develop allows us to ask concrete questions about the nature of interactional systematics, with respect to a participant’s incipient binary decision to speak:

1. *Is one individual affected by the other?* If so, then including the interlocutor’s past decisions in a general model’s conditioning history will improve predictions.
2. *Is there individual variation?* If so, then adapting a general model towards an individual’s past decisions as they are taken, will improve predictions.
3. *Are the systematics we seek time-dependent?* If so, then forgetting one’s least recent decisions will improve predictions.

2. Data

To demonstrate techniques, we use a set of 35 dialogues from the Spontal corpus [8]. Each dialogue consists of 30 minutes of free and unscripted dyadic conversation. We split the set into TRAINSET, DEVSET and EVALSET, of 23, 6 and 6 dialogues, respectively. No participant occurs in more than one subset.

Given the automatic speech/non-speech segmentation (cf. [9], Figure 11.17a, with 200-ms minimum duration constraints), each dialogue in the corpus is viewed as a speech interaction chronogram [10, 11]. We discretize the activity of both parties in a frame-synchronous manner, using non-overlapping frames 100 ms in duration, to yield the discrete chronogram $\mathbf{Q} \in \{\square, \blacksquare\}^{K \times T}$. \square and \blacksquare are the absence and presence of speech activity, $K \equiv 2$ is the number of parties, and T is the number of frames. The t th column of \mathbf{Q} , \mathbf{q}_t , is the vector concatenation of the states of both parties.

3. Representation

\mathbf{Q} is assumed to be the output of a Markov process. Our task is to develop a model Θ which provides for the likelihood of \mathbf{Q} ,

$$P(\mathbf{Q} | \Theta) = \prod_{t=1}^T P(\mathbf{q}_t | \dots, \mathbf{q}_{t-1}; \Theta), \quad (1)$$

¹In this work, for the lack of a more suitable term, we refer to the probabilities of speaking, in all possible contexts, as an participant’s strategy.

where the ellipsis represents \mathbf{q}_{t-2} and earlier emissions, reflecting the order of the Markov process. Θ is commonly known as a turn-taking model [13, 14], although in principle it is an interaction model which captures many phenomena, of which one-speaker-at-a-time is but one. We assume, as we have elsewhere [4], that the behavior of both participants is *conditionally independent* (CI), given their joint behavior in the immediate past; this renders each term on the right-hand side of Equation 1 equal to

$$\begin{aligned} P(\mathbf{q}_t | \dots, \mathbf{q}_{t-1}; \Theta) & \\ &= P(\mathbf{q}_t[1] | \dots, \mathbf{q}_{t-1}[1], \mathbf{q}_{t-1}[2]; \Theta) \\ &\quad \times P(\mathbf{q}_t[2] | \dots, \mathbf{q}_{t-1}[2], \mathbf{q}_{t-1}[1]; \Theta). \end{aligned} \quad (2)$$

Square brackets index participants.

As an example, we consider the chronogram snippet

$$\mathbf{Q} = \left[\dots \begin{array}{cccccccc} \blacksquare & \blacksquare & \blacksquare & \blacksquare & \square & \square & \square & \blacksquare \\ \square & \square & \square & \blacksquare & \blacksquare & \square & \square & \square \end{array} \dots \right] \quad (3)$$

The expression for the conditional probability of observing the 5th of the 9 frames depicted, given the preceding 2 frames, is

$$\begin{aligned} P \left(\begin{array}{c} \square \\ \blacksquare \end{array} \middle| \begin{array}{cc} \blacksquare & \blacksquare \\ \square & \square \end{array}; \Theta \right) & \\ &= P(\square | \square, \blacksquare, \blacksquare, \blacksquare; \Theta) \times P(\blacksquare | \blacksquare, \square, \blacksquare, \blacksquare; \Theta) \end{aligned} \quad (4)$$

This results in a 5-gram, over a vocabulary of 2 symbols². An alternative to treating participants as conditionally independent is to assume that they are *unconditionally independent* (UI), and to remove the states describing each participant’s interlocutor’s past behavior in both factors, yielding $P(\square | \blacksquare, \blacksquare; \Theta) \times P(\blacksquare | \square, \blacksquare; \Theta)$. The resulting model is a 3-gram.

4. n -Gram Modeling

n -gram techniques, as used in language models (LMs), are largely suitable for modeling sequences of the discrete speech activity states here. In the ensuing discussion, we replace the symbols specific to chronograms (e.g. \mathbf{q} , \square , \blacksquare , etc.) [9] with those in the LM literature (e.g. [12]), namely w for words and w_{i-n+1}^{i-1} for their $(n-1)$ -gram word histories, to permit comparison. We assume that the 2-participant chronogram has been marshalled into a 1-dimensional sequence as in Equation 4.

4.1. Maximum Likelihood

The maximum likelihood (ML) estimate involves the count sum over alternative n -gram completions,

$$C(w_{i-n+1}^{i-1} \bullet) = \sum_{w_i} c(w_{i-n+1}^i), \quad (5)$$

where $w_i \in \{\square, \blacksquare\}$. The ML estimate is then given by

$$p_{ML}(w_i | w_{i-n+1}^{i-1}) = \frac{c(w_{i-n+1}^i)}{C(w_{i-n+1}^{i-1} \bullet)}. \quad (6)$$

Figure 1 shows the normalized negative log-likelihood (NLL), as defined in [14], on DEVSET given a model Θ inferred using TRAINSET. It is seen that ML models, of the CI and UI varieties, can be trained with up to 9 and 4 frames of context (i.e., a 10-gram and a 9-gram), respectively. There are clearly n -grams in DEVSET, which are longer, that do not occur in TRAINSET.

²Somewhat arbitrarily, we choose to sequence the conditioning units by frame, with target speaker state first.

4.2. Order Interpolation

A common solution to this problem is *smoothing*; for word sequences [12], many n -gram techniques are in use. Unfortunately, the parameters of most of them assume a categorical distribution over a Zipf’s law-conforming vocabulary of thousands of words. Our experience is that many of these techniques fare poorly on Bernoulli outcomes over $\{\square, \blacksquare\}$.

Instead, we propose to recursively interpolate n -gram models with their $(n-1)$ -gram next-lower-order model,

$$\begin{aligned} p_{int}(w_i | w_{i-n+1}^{i-1}) &= \lambda(w_{i-n+1}^{i-1}) p_{ML}(w_i | w_{i-n+1}^{i-1}) \\ &\quad + (1 - \lambda(w_{i-n+1}^{i-1})) p_{int}(w_i | w_{i-n+2}^{i-1}), \end{aligned} \quad (7)$$

a form of Jelinek-Mercer interpolation [15]. We set the value of the *history-specific* interpolation parameter λ to that used in additive smoothing [16],

$$\lambda(w_{i-n+1}^{i-1}) = \frac{C(w_{i-n+1}^{i-1} \bullet)}{C(w_{i-n+1}^{i-1} \bullet) + \rho}. \quad (8)$$

In this work, ρ is a global relevance parameter; we optimize it by minimizing NLL on DEVSET.

Figure 1 shows that for large ρ , the NLL curve for the CI model, which takes the interlocutor’s past into account, can be uncurled to remain at a fixed offset from that of the UI model (which ignores interlocutors and does not overfit). Although NLLs are currently difficult to interpret in absolute terms, we allow the following qualitative interpretation: (1) accounting for the interlocutor, with one 100-ms frame of history, leads to performance which approximately matches that of the UI model with four frames; and (2) with two frames, the CI model outperforms the UI model for all explored history durations.

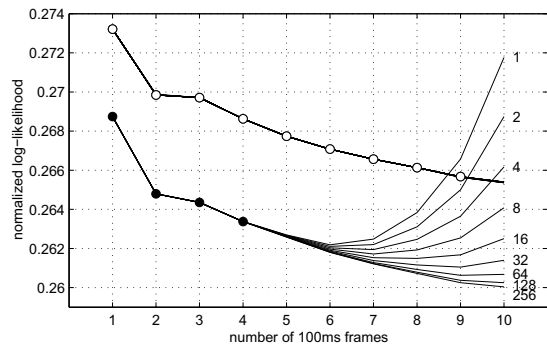


Figure 1: Normalized negative log-likelihoods (NLLs; along the y -axis) for DEVSET, as a function of the number of frames in the conditioning history (along the x -axes). Model parameters estimated using TRAINSET. Unfilled and filled circles refer to ML UI and CI models, respectively. Lines (shown for visualization) without markers connect CI models smoothed as in Equations 7 & 8, with ρ as annotated at the right of the diagram.

In the remainder of this work (except §7), we retain the TRAINSET-trained and smoothed CI model curve of Figure 1 as the *general* model for comparison in figures (referring to it as Θ_G), and its instance with 10 frames of historical context as the *universal background model* (UBM). We also define a new scale, the *relative* normalized negative log-likelihood (RNLL), whose origin is the NLL of the UBM (0.2600) and whose unit

is given by the average difference between the CI and UI curves (0.0052) in Figure 1. This represents the cost of ignoring interlocutors in the general model.

5. Learning from One’s Past Decisions

5.1. Incremental Re-estimation

To contrast with the performance of Θ_G , we infer a separate model Θ_t for each party in each dialogue in DEVSET. The models do not see the future; to score speech activity at instant t , Θ_t relies only on earlier instants of the dialogue, up to instant $t - 1$. Model parameters thus evolve in time; when an n -gram w_{i-n+1}^i is observed at instant t , the model from instant $t - 1$ is used to score it, and then that n -gram’s count is incremented,

$$c_t(w_{i-n+1}^i) = c_{t-1}(w_{i-n+1}^i) + 1. \quad (9)$$

The counts for other n -grams remain unchanged. Equations 5, 6, 8 and 10, yielding time-dependent estimates C_t , $p_{t;ML}$, λ_t and $p_{t;int}$, respectively, are then applied (ρ is held constant). This results in an incremental, dialogue- and participant- dependent “OWN” model Θ_t .

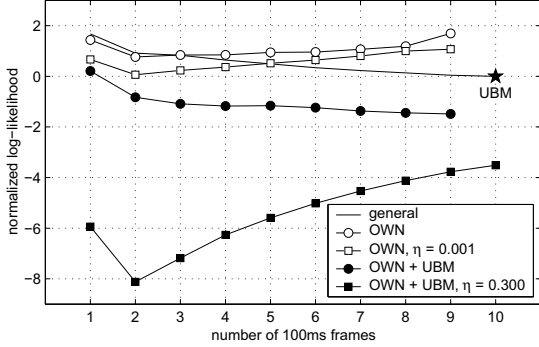


Figure 2: DEVSET RNLLs (along the y -axis), as a function of the number of frames in the conditioning history (along the x -axis), for incremental CI models learned (with and without forgetting) from the target participant’s (“OWN”) past decisions.

Figure 2 shows the results with unfilled circles. When models look only 200 ms back in time, Θ_t yields better DEVSET predictions than does Θ_G . This is a surprising result, since the amount of data in TRAINSET is 39–48 times larger than that in a single side of any DEVSET dialogue. Furthermore, the incremental models implicitly start out with the zero-gram ($1/2$) model for the first frame. For longer histories, as can be expected, Θ_G outperforms Θ_t .

5.2. Incremental Adaptation

We can also interpolate the incremental model Θ_t , inferred as above, with the UBM from Section 4. We do this according to

$$p_{comb}(w_i|w_{i-n+1}^{i-1}) = \lambda(w_{i-n+1}^{i-1})p_{int}(w_i|w_{i-n+1}^{i-1}) + (1 - \lambda(w_{i-n+1}^{i-1}))p_{ubm}(w_i|w_{i-n+1}^{i-1}), \quad (10)$$

where p_{int} is the order-interpolated Θ_t (using relevance parameter ρ) and p_{ubm} is the order-interpolated UBM. For simplicity, λ in Equation 10 is the same as that used for order interpolation

of Θ_t . The filled circles of Figure 2 show that model combination helps enormously: its improvement over the UBM is slightly larger than the cost of not conditioning on interlocutor behavior in the general model.

5.3. Unlearning

Finally in this section, we explore forgetting, by exponentially decaying old counts in favor of more recent ones. At each instant, we score the observed behavior at instant t using Θ_{t-1} , and then apply a “forgetting” factor η to all counts in Θ_{t-1} ,

$$c'_{t-1}(w) = (1 - \eta) \cdot c_{t-1}(w), \quad \forall w. \quad (11)$$

Only then do we increment the count of the single n -gram observed at instant t (cf. Equation 9) to yield a new model Θ_t , and proceed to score the speech activity behavior at instant $t + 1$.

The results, shown in Figure 2, indicate that forgetting helps. For the incrementally re-estimated model, a modest $\eta = 0.001$ makes Θ_t competitive with the UBM with only 200 ms of history. For the model adapted from the UBM, an tremendous improvement (8 times larger than the cost of not conditioning on interlocutor behavior) is achieved with a more aggressive $\eta = 0.300$.

6. Learning from Partner’s Past Decisions

We repeat the experiments of §5.1, re-estimating one model Q_t for each participant in each dyad of DEVSET, but this time we score the subsequently observed decisions of each participant using their interlocutor’s model. We do this for every instant t . The results are shown in Figure 3 with unfilled circles, as “OTH”; they indicate that models based only on the interlocutor’s past decisions are much worse than the general model Θ_G .

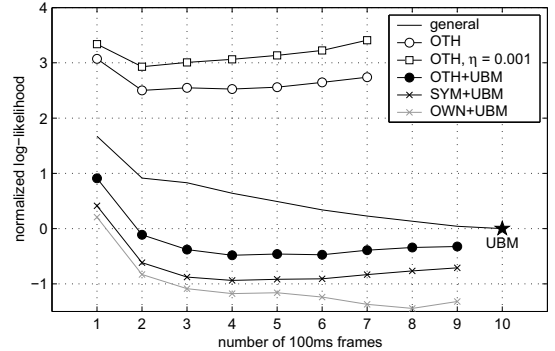


Figure 3: DEVSET RNLLs (along the y -axis), as a function of the number of frames in the conditioning history (along the x -axis), for incremental CI models learned (with and without forgetting) from the target participant’s interlocutor’s (“OTH”) past decisions. “SYM” is a model which includes both “OWN” and “OTH” counts.

Similarly, we repeat the experiments of §5.2, but interpolating the interlocutor’s “OTH” (rather than each participant’s “OWN”) incremental model with the UBM for scoring. The filled circles in Figure 3 shows that these models improve over the UBM, demonstrating within-dyad turn-taking similarity [17, 18]. A symmetric “SYM” model, containing counts from both parties, does still better when adapted from the UBM — but not as well as the combination of the UBM with the target

participant’s “OWN” model only. The interlocutor’s decision model appears to be a weak reflection of the target participant’s; when the latter is not available it helps [17, 18].

In contrast to the finding in §5.3, forgetting the interlocutor’s behavior (the unfilled squares versus unfilled circles in Figure 3) appears to degrade performance.

7. Generalization

The lowest RNLLs achieved by the proposed methods, on DEVSET, are shown in Table 1. Also shown, are the same parameter settings, are the RNLLs for the unseen EVALSET dialogues.

Model	τ	n	ρ	η	DEV	EVAL
UI Models (Figure 1)						
UBM	10	11	256	—	+1.0222	+3.7870
CI Models (Figures 1, 2 & 3)						
UBM	10	21	256	—	0	+2.8922
OWN	2	5	8	0.001	+0.0595	+2.0516
+ UBM	2	5	1	0.300	−8.1209	−4.8206
OTH	2	5	64	0.000	+2.5002	+7.9201
+ UBM	4	9	256	0.000	−0.4822	+2.6600

Table 1: RNLLs for DEVSET (DEV) and EVALSET (EVAL), at DEVSET-determined parameter settings. Symbols as in the text; τ is the number of 100-ms frames in the conditioning history.

Although EVALSET appears to be consistently harder to predict than DEVSET, even for the simpler UI models which ignore interlocutor history. The trends for both sets are similar in terms of rank: incremental models interpolated with the UBM outperform the UBM alone, by a large margin for OWN; interpolating with OWN is always better than with OTH. The main difference is that for EVALSET, an incrementally re-estimated OWN model, without interpolation with UBM, outperforms the 10-frame-history UBM with its optimal 2 frames of context.

8. Conclusions

We have presented a framework for the stochastic modeling of interaction, inclusive of turn-taking, in spontaneous two-party conversation. Its most novel feature, with respect to our past work, is incremental learning and unlearning.

The utility of the framework has been validated by answering four basic question about what it means to interact, as seen from one party’s point of view in dyadic conversation. First, we extended an earlier finding, to large n -gram orders, showing that augmenting the conditioning context with one’s interlocutor’s past behavior improves predictions of one’s incipient behavior; individuals are therefore clearly affected by one another. Second, it was shown that short-history incremental models, beginning with no knowledge whatsoever, quickly begin to outperform same-history-duration time-independent models trained on a large number of other dialogues; this proves that there is individual variation. Interpolation of incremental models with time-independent models helps tremendously. Third, incrementally forgetting one’s oldest decisions improves predictions, which suggests that interactional systematics are time-dependent since that behavioral strategies drift rather than converge as more data becomes available. Finally, an interlocutor’s incremental model improves on the performance of a time-independent model, but only when one’s own incremental model is unavailable. Participants’ strategies thereby appear

to offer noisy renditions of the strategies of their interlocutors, proving interlocutor similarity within dyads.

We expect the framework, due to the ease with which such findings can now be obtained, to significantly impact future work on descriptive and prescriptive analyses of interaction.

9. Acknowledgments

The work was supported in part by the Riksbankens Jubileumsfond (RJ) project *Samtalets Prosodi*. Experiments were performed at Carnegie Mellon University’s interACT lab.

10. References

- [1] Brady, P., “A model for generating on-off speech patterns in two-way conversation”, *Bell Systems Tech J* **48**(9):2445–2472, 1969.
- [2] Jaffe, J. and Feldstein, S., *Rhythms of Dialogue*, Academic Press, 1970.
- [3] Raux, A. and Eskenazi, M., “Finite state turn taking model for spoken dialog systems”, *Proc HLT*, Boulder CO, USA, 629–637, 2009.
- [4] Laskowski, K., Edlund, J. and Heldner, M., “A single-port non-parametric model of turn-taking in multi-party conversation”, *Proc ICASSP*, Praha, Czech Republic, 2011 (to appear).
- [5] Glenn, P., *Laughter in Interaction*, Cambridge University Press, 2003.
- [6] Kendon, A., “Some functions of gaze-direction in social interaction”, *Acta Psychologica* **26**:22–63, 1967. doi:10.1016/0001-6918(67)90005-4
- [7] Cummins, F., “Gaze and blinking in dyadic conversation: A study in coordinated behavior among individuals”, *Language & Cognitive Processes* (submitted), 2011.
- [8] Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S. and House, D., “Spontal: A Swedish spontaneous dialogue corpus of audio, video and motion capture”. *Proc LREC*, La Valletta, Malta, 2992–2995, 2010.
- [9] Laskowski, K., “Predicting, detecting and explaining the occurrence of vocal activity in multi-party conversation”, PhD Thesis CMU-LTI-11-001, Carnegie Mellon University, 2011.
- [10] Chapple, E., “The Interaction Chronograph: Its evolution and present application”, *Personnel* **25**(4):295–307, 1949.
- [11] Dabbs Jr., J. and Ruback, R., “Dimensions of group process: Amount and structure of vocal interaction”, *Advances in Experimental Social Psychology* **20**:123–169, 1987. doi:10.1016/S0065-2601(08)60413-X
- [12] Chen, S. and Goodman, J., “An empirical study of smoothing techniques for language modeling”, Tech Report TR-10-98, Harvard University, 1998.
- [13] Wilson, T., Wiemann, J. and Zimmerman, D., “Models of turn-taking in conversational interaction”, *J Language and Social Psychology* **3**(3):159–183, 1984. doi:10.1177/0261927X8400300301
- [14] Laskowski, K., “Modeling norms of turn-taking in multi-party conversation”, *Proc ACL*, Uppsala, Sweden, 999–1008, 2010.
- [15] Jelinek, F. and Mercer, R., “Interpolated estimation of Markov source parameters from sparse data”, *Proc Workshop on Pattern Recognition in Practice*, Amsterdam, The Netherlands, 1980.
- [16] Lidstone, G., “Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities”, *Trans Faculty of Actuaries* **8**(182–192):80, 1920.
- [17] Edlund, J., Heldner, M. and Hirschberg, J., “Pause and gap length in face-to-face interaction”, *Proc INTERSPEECH*, Brighton, UK, 2779–2782, 2009.
- [18] Branigan, H., Pickering, M., Pearson, J. and McLean, J., “Linguistic alignment between people and computers”, *J Pragmatics* **42**(9):2355–2368, 2010. doi:10.1016/j.pragma.2009.12.012